

Methods in Pharmacology
and Toxicology

Springer Protocols

Kunal Roy *Editor*

Ecotoxicological QSARs

EXTRAS ONLINE

 Humana Press

METHODS IN PHARMACOLOGY AND TOXICOLOGY

Series Editor

Y. James Kang

Department of Pharmacology and
Toxicology, University of Louisville
Louisville, KY, USA

For further volumes:
<http://www.springer.com/series/7653>

Methods in Pharmacology and Toxicology publishes cutting-edge techniques, including methods, protocols, and other hands-on guidance and context, in all areas of pharmacological and toxicological research. Each book in the series offers time-tested laboratory protocols and expert navigation necessary to aid toxicologists and pharmaceutical scientists in laboratory testing and beyond. With an emphasis on details and practicality, *Methods in Pharmacology and Toxicology* focuses on topics with wide-ranging implications on human health in order to provide investigators with highly useful compendiums of key strategies and approaches to successful research in their respective areas of study and practice.

Ecotoxicological QSARs

Edited by

Kunal Roy

*Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology,
Jadavpur University, Kolkata, India*



Editor

Kunal Roy
Drug Theoretics and Cheminformatics
Laboratory, Department of
Pharmaceutical Technology
Jadavpur University
Kolkata, India

ISSN 1557-2153

ISSN 1940-6053 (electronic)

Methods in Pharmacology and Toxicology

ISBN 978-1-0716-0149-5

ISBN 978-1-0716-0150-1 (eBook)

<https://doi.org/10.1007/978-1-0716-0150-1>

© Springer Science+Business Media, LLC, part of Springer Nature 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Dedication

For Aatreyi, Arpit, and Chaitali

Preface

In the world of the twenty-first century, we are surrounded by a large number of chemicals in various forms, starting from those in industrial, commercial, and agricultural uses to those for domestic and personal uses. These chemicals have an impact on different living organisms in the environment, including humans. Studies have confirmed that these effects are mostly hazardous in the case of unintended exposure at a significant level of concentration [1]. There is evidence that species from different biomes of the biosphere, like aquatic, terrestrial, and aerial in different trophic levels, are affected due to uncontrolled production, usage, and disposal of chemicals used in our day-to-day life, apart from those used in bulk quantities in chemical industries. The increasing environmental pollution caused by old and new chemicals has led to an ever-growing concern about the potential effects that they may have on the environment and also their direct or indirect effects on human health. According to the Organization for Economic Co-operation and Development (OECD), contaminants of emerging concern (CECs) are “a vast array of contaminants that have only recently appeared in water, or that are of recent concern because they have been detected at concentrations significantly higher than expected, or their risk to human and environmental health may not be fully understood” [2]. CECs include pharmaceuticals, industrial and household chemicals, personal care products, pesticides, plasticizers, flame retardants, surfactants, manufactured nanomaterials, microplastics, their transformation products, etc. [2]. The most common concern with most of the CECs is their endocrine-disrupting effects. Apart from this, CECs may show genotoxicity, cytotoxicity, carcinogenicity, antibiotic resistance, etc., depending on the use category [3]. The effects of CECs on human and ecosystem health are largely unknown, and relatively little is known about the ways they travel, transform, or degrade in the environment. The number of CECs is continuously evolving as new chemical compounds are produced, and improvements in chemical analysis increase our understanding of the effects of current and past contaminants on human and environmental health. In spite of the hazardous effects of perilous chemicals, pharmaceuticals, cosmetics, biocides, pesticides, dyes, solvents, and other pollutants on the ecosystem, relatively few of these have been subjected to sufficient experimental evaluation for their hazardous environmental properties. The experimental determination of environmental parameters (e.g., soil sorption, bioconcentration, biodegradation and biotransformation, toxic effects) of commercial chemicals is a costly and time-consuming process. Also, there exist an incredibly large number of environmental endpoints for which it is difficult or even impossible to gather experimental data for a large number of chemicals due to available limited resources. This results in data gaps to a significant extent, which draw the attention of different regulatory bodies like OECD (Organization for Economic Co-operation and Development), ECVAM (European Centre for the Validation of Alternative Methods), ECHA (European Chemical Agency), FDA (Food and Drug Administration), EPA (Environmental Protection Agency), etc. These aspects are also relevant to different regulations, for example, REACH (Registration, Evaluation and Authorization and Restriction of Chemicals) and CLP (Classification, Labelling and Packaging) regulations in the EU and Toxic Substances Control Act (TSCA) in the US [4–9]. Specific regulations exist in the EU for specific families of products, such as fertilizers (EU Fertiliser Regulation, to be adopted), biocides (EU Biocidal Products Regulation, 2012), explosives (EU Regulations on

Explosives, 2013), drug precursors (EU Regulation on Drug Precursors, 2004), cosmetics (EU Regulations on Cosmetics Products, 2009), etc. The regulatory bodies recognize the use of *in silico* models as a supplement or even as a replacement for experimental testing. It is necessary to develop quantitative models that will accurately and readily predict the environmental behavior of large sets of chemicals. A chemistry approach to predictive toxicology relies on quantitative structure–activity relationship (QSAR) modeling to predict biological activity from chemical structure [10]. Such approaches have proven capabilities when applied to well-defined toxicity end points or regions of chemical space. These also support the “3Rs” (replacement, refinement, and reduction of animals in research) principles of Russell and Burch [11]. Most importantly, these models might help in designing “greener” alternatives replacing the original toxic chemicals. It is also possible to study the effects of possible degradation products and metabolites using QSAR approaches.

QSARs are essentially statistical models derived from the correlation of the response being modeled with quantitative chemical structure information presented in the form of descriptors. There are a plethora of descriptors currently available due to the availability of several descriptor computing software tools. In ecotoxicological QSARs, usually 0D–3D descriptors are applied, although the usage of higher dimensional QSAR is still possible in appropriate cases. Curation of the data set is very important before the development of any QSAR model. In view of the large number of descriptors available, it is important to use appropriate feature selection and learning algorithms to derive models that are statistically meaningful and not overfitted. Recent literature has also described the application of different machine learning methods in modeling different ecologically important endpoints. Ecotoxicological QSAR models for regulatory purposes should be developed based on OECD five-point principles [12]. Such models should be validated using stringent methodologies, including external validation ensuring the reliability of predictions for unseen compounds. These models might be either regression based, giving quantitatively precise predictions of the response, or classification based, giving qualitative gradation of endpoints that might be helpful in the initial screening process. One might be choosing classical QSAR methodologies for more interpretable and simple models with a clear mechanistic interpretation, while others might be more inclined to using various machine learning algorithms such as deep neural nets, support vector machine, random forest, etc., focusing more on the quality of predictions instead of interpretability. Both of these two types of QSARs are equally important and useful, and the choice between them depends on the problem to be solved. The concept of applicability domain is of paramount importance for the prediction of a new compound from a previously developed QSAR model. Apart from QSARs, read-across based on a similarity principle has also recently emerged as a useful tool in ecotoxicological modeling.

This current volume of *Ecotoxicological QSARs* presents the background of the application of QSARs in the predictive toxicology field in a regulatory context and covers the protocols for descriptor computation, data curation, feature selection, machine learning algorithms, validation of models, applicability domain assessment, confidence estimation for predictions, and so on. The book also presents diverse case studies for ecotoxicological QSARs applied to different chemical classes, including industrial chemicals, solvents, pollutants, pharmaceuticals, personal care products, biocides, agrochemicals, nanomaterials, etc. Compilations of different databases relevant to ecotoxicological QSARs are presented. There are a total of 32 chapters in the 4 parts of this book.

The first part presents an introduction to ecotoxicological risk assessment and modeling. The first chapter in this part, authored by *García-Fernández*, gives an overview of different

EU laws and regulations in the context of ecotoxicological risk assessment. The second chapter, authored by *Aher, Khan, and Roy*, briefly introduces the concept of quantitative structure–activity relationships (QSARs) as useful tools in predictive ecotoxicology. The third chapter, authored by *Nantasesamat*, provides general guidelines and best practices for constructing reproducible QSAR models. The fourth chapter of Part I, contributed by *García-Fernández, Espín, Gómez-Ramírez, Martínez-López, and Navas*, discusses the advantages and disadvantages on the use of potential sentinel species for biomonitoring purposes.

The second part of the book deals with the methods and protocols of ecotoxicological QSARs. The first chapter of this part has been contributed by *Ambure and Cordeiro*. This chapter focuses on several data curation tools that are used before QSAR model development, paying special attention to those that can be used to semi-automate the curation process. The next chapter, contributed by *Gini and Zanolí*, presents different machine learning and deep learning methods that are used in ecotoxicological QSAR modeling. Chapter 7, contributed by *Barros, Sousa, Scotti, and Scotti*, reviews different machine learning and classical QSAR methods applied in computational ecotoxicology. *Concu and Cordeiro* have contributed Chapter 8 on the relevance of the feature selection algorithms while developing nonlinear QSARs. *Moura and Cordeiro* have coauthored Chapter 9, which discusses the methodologies and fundamentals of classical and perturbation-based QSAR models within the environmental risk assessment framework. *Rasulev* has authored Chapter 10, in which the recent advances in the development and application of 3D-QSAR and protein–ligand docking approaches in the studies of nanostructured materials, such as fullerenes and carbon nanotubes, have been outlined. *Tondo, Montaruli, Mangiatordi, and Nicolotti* have presented computational methods and open-source computational tools for the evaluation of the ecotoxicological effects of pharmaceutical impurities. In Chapter 12, *Svensson and Norinder* discuss a type of confidence predictor called conformal prediction, which can be used to generate predictions with a guaranteed error rate. *Tugcu, Önlü, Aydin, and Saçan* discuss the application of read-across in regulatory toxicology in Chapter 13. *Pedrazzani and colleagues* present in the last chapter of this part a methodological protocol for the experimental assessment of environmental footprints.

The third part deals with literature reviews of ecotoxicological QSARs and case studies. The first chapter of this part, authored by *Ebbrell, Cronin, Ellison, Firman, and Madden*, develops predictive approaches for acute and chronic toxicity in fish, Daphnia, and algae utilizing baseline toxicity models. In the second chapter of this part, *Khan, Sanderson, and Roy* review information related to the impact and occurrence of personal care products and biocides, as well as their persistence, environmental fate, risk assessment, and risk management, with a special emphasis given on in silico tools such as QSAR, which can be employed in predicting the ecotoxicity of personal care products and biocides mainly to aquatic species. Chapter 17, contributed by *Gómez-Ganau, Marzo, Gozalbes, and Benfenati*, presents some computational models for the estimation of ecotoxicity of biocides in micro-organisms and fish developed in the context of the EU LIFE+ project titled COMBASE. In Chapter 18, *Funar-Timofei and Ilia* review QSAR/QSPR reports in the estimation of dye ecotoxicity. *Khan, Kar, and Roy* have illustrated the basic concepts of mixture toxicity assessment and reviewed QSAR reports on the ecotoxicity of mixtures in Chapter 19. In the following chapter, *Ojha, Mandal, and Roy* have reviewed QSPR modeling of adsorption of pollutants by carbon nanotubes. Chapter 21, contributed by *Bora, Crisan, Borota, Funar-Timofei, and Ilia*, presents successful QSAR models for the ecotoxicological data of

organophosphorus and neonicotinoid pesticides. *Nendza, Ahlers, and Schwartz* have presented a case study, in the next chapter, on QSAR and read-across for thiochemicals. In Chapter 23, *Chang, Chang, Wu, Chang, and Liu* have reviewed in silico ecotoxicological modeling of pesticide metabolites and mixtures. *Furuhamma* has presented in Chapter 24 a case study of read-across and QSAR for green algae growth inhibition toxicity data. Chapter 25, contributed by *Hamadache, Benkortbi, Amrane, and Hanini*, presents a literature review of ecotoxicological QSAR modeling of pesticides, ionic liquids, pharmaceuticals, and other pollutants. *Speck-Planche* presents in Chapter 26 a case study of the development of a multiscale QSAR model that is able to assess the ecotoxicity of the pesticides by considering different measures of ecotoxic effects, many bioindicator species, several different assay guidelines, and the multiple times during which the bioindicator species have been exposed to pesticides. *Jana, Pal, Sural, and Chattaraj* have presented in Chapter 27 a case study of the development of quantitative structure-toxicity models using hydrophobicity and electrophilicity. The last chapter of this part, contributed by *Sanderson and colleagues*, presents a survey of environmental toxicity (Q)SARs for polymers as an emerging class of materials in regulatory frameworks, with a focus on challenges and possibilities regarding cationic polymers.

The last part of the book deals with software tools and databases. The first chapter of the part, contributed by *Ghosh, Kar, and Leszczynski*, deals with ecotoxicity databases for QSAR modeling. The next chapter, contributed by *Benfenati and Lombardo*, discusses VEGAHub for ecotoxicological QSAR modeling. In the third chapter of this part, *Varou, Tsoumanis, Afantitis, and Melagraki* present and discuss the Enalos Cloud Platform and model development and validation using three web services hosted in the Enalos Cloud. The last chapter of the book, authored by *Mauri*, presents alvaDescriptor, a software to calculate and analyze molecular descriptors and fingerprints for QSAR modeling.

This collection of 32 chapters presents the current status and recent developments in ecotoxicological QSAR modeling, especially in the context of different EU regulations. This book will certainly update readers in the field with current practices and introduce to them new developments and hence should be very useful for researchers in academia, industries, and regulatory bodies.

Kolkata, West Bengal, India

Kunal Roy

References

1. van Leeuwen CJ, Vermeire TG (2007) Risk assessment of chemicals: an introduction. Springer, Dordrecht
2. OECD workshop on Managing Contaminants of Emerging Concern in Surface Waters: Scientific developments and cost-effective policy responses, 5 February 2018. <https://www.oecd.org/water/Summary%20Note%20-%20OECD%20Workshop%20on%20CECs.pdf>
3. Raghav M, Eden S, Mitchell K, Witte B (2013) Contaminants of emerging concern in water. Water Resources Research Center, The University of Arizona. https://wrrc.arizona.edu/sites/wrrc.arizona.edu/files/Arroyo2013LR_0.pdf
4. European Union (EU) (2001) White paper: strategy for a future chemicals policy. Commission of the European Committees, Brussels, Belgium, COM, pp 1–32

5. European Commission, Directive 2006/121/EC of the European Parliament and of the Council of 18 December 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No. 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and establishing a European Chemicals Agency. Official Journal of the European Union, L 396/850 of 30.12.2006, Office for Official Publications of the European Communities (OPOCE), Luxembourg, 2006
6. Toxic Substances Control Act of 1976, US Environmental Protection Agency
7. Kar S, Roy K (2010) Predictive toxicology using QSAR: a perspective. *J Indian Chem Soc* 87:1455–1515
8. Kar S, Roy K (2012) Risk assessment for ecotoxicity of pharmaceuticals – an emerging issue. *Expert Opin Drug Saf* 11:235–274
9. Roy K, Kar S (2016) In silico models for ecotoxicity of pharmaceuticals. In: Benfenati E (ed) *Silico methods for predicting drug toxicity*. Springer, New York, pp 237–304
10. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, New York
11. Russell WMS, Burch RL (1958) *The principles of humane experimental technique*. Johns Hopkins University, Baltimore
12. <http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm>

Contents

Preface	vii
Contributors	xvii

PART I INTRODUCTION

- 1 Ecotoxicological Risk Assessment in the Context of Different EU Regulations 3
Antonio Juan García-Fernández
- 2 A Brief Introduction to Quantitative Structure-Activity Relationships as Useful Tools in Predictive Ecotoxicology 27
Rahul Balasaheb Aher, Kabiruddin Khan, and Kunal Roy
- 3 Best Practices for Constructing Reproducible QSAR Models 55
Chanin Nantasenamat
- 4 Wildlife Sentinels for Human and Environmental Health Hazards in Ecotoxicological Risk Assessment 77
Antonio Juan García-Fernández, Silvia Espín, Pilar Gómez-Ramírez, Emma Martínez-López, and Isabel Navas

PART II METHODS AND PROTOCOLS

- 5 Importance of Data Curation in QSAR Studies Especially While Modeling Large-Size Datasets 97
Pravin Ambure and M. Natália Dias Soeiro Cordeiro
- 6 Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling 111
Giuseppina Gini and Francesco Zanoli
- 7 Use of Machine Learning and Classical QSAR Methods in Computational Ecotoxicology 151
Renata P. C. Barros, Natália F. Sousa, Luciana Scotti, and Marcus T. Scotti
- 8 On the Relevance of Feature Selection Algorithms While Developing Non-linear QSARs 177
Riccardo Concu and M. Natália Dias Soeiro Cordeiro
- 9 Got to Write a Classic: Classical and Perturbation-Based QSAR Methods, Machine Learning, and the Monitoring of Nanoparticle Ecotoxicity 195
Ana S. Moura and M. Natália D. S. Cordeiro
- 10 Ecotoxicological QSAR Modeling of Nanomaterials: Methods in 3D-QSARs and Combined Docking Studies for Carbon Nanostructures 215
Bakhtiyor Rasulev

11	Early Prediction of Ecotoxicological Side Effects of Pharmaceutical Impurities Based on Open-Source Non-testing Approaches	235
	<i>Anna Rita Tondo, Michele Montaruli, Giuseppe Felice Mangiavordi, and Orazio Nicolotti</i>	
12	Conformal Prediction for Ecotoxicology and Implications for Regulatory Decision-Making	271
	<i>Fredrik Svensson and Ulf Norinder</i>	
13	Read-Across for Regulatory Ecotoxicology.....	289
	<i>Gulcin Tugcu, Serli Önlü, Ahmet Aydin, and Melek Türker Saçan</i>	
14	Methodological Protocol for Assessing the Environmental Footprint by Means of Ecotoxicological Tools: Wastewater Treatment Plants as an Example Case.....	305
	<i>Roberta Pedrazzani, Pietro Baroni, Donatella Feretti, Giovanna Mazzoleni, Nathalie Steimberg, Chiara Urani, Gaia Viola, Ilaria Zerbini, Emanuele Ziliani, and Giorgio Bertanza</i>	

PART III CASE STUDIES AND LITERATURE REPORTS

15	Development of Baseline Quantitative Structure-Activity Relationships (QSARs) for the Effects of Active Pharmaceutical Ingredients (APIs) to Aquatic Species	331
	<i>David J. Ebbrell, Mark T. D. Cronin, Claire M. Ellison James W. Firman, and Judith C. Madden</i>	
16	Ecotoxicological QSARs of Personal Care Products and Biocides	357
	<i>Kabiruddin Khan, Hans Sanderson, and Kunal Roy</i>	
17	Computational Approaches to Evaluate Ecotoxicity of Biocides: Cases from the Project COMBASE	387
	<i>Sergi Gómez-Ganau, Marco Marzo, Rafael Gozalbes, and Emilio Benfenati</i>	
18	QSAR Modeling of Dye Ecotoxicity	405
	<i>Simona Funar-Timofei and Gheorghe Ilia</i>	
19	Ecotoxicological QSARs of Mixtures	437
	<i>Pathan Mohsin Khan, Supratik Kar, and Kunal Roy</i>	
20	QSPR Modeling of Adsorption of Pollutants by Carbon Nanotubes (CNTs)	477
	<i>Probir Kumar Ojha, Dipika Mandal, and Kunal Roy</i>	
21	Ecotoxicological QSAR Modeling of Organophosphorus and Neonicotinoid Pesticides.....	513
	<i>Alina Bora, Luminita Crisan, Ana Borota, Simona Funar-Timofei, and Gheorghe Ilia</i>	
22	QSARs and Read-Across for Thiochemicals: A Case Study of Using Alternative Information for REACH Registrations	545
	<i>Monika Nendza, Jan Ahlers, and Dirk Schwartz</i>	

23	In Silico Ecotoxicological Modeling of Pesticide Metabolites and Mixtures	561
	<i>Chia Ming Chang, Chiung-Wen Chang, Fang-Wei Wu Len Chang, and Tien-Cheng Liu</i>	
24	Combination of Read-Across and QSAR for Ecotoxicity Prediction: A Case Study of Green Algae Growth Inhibition Toxicity Data	591
	<i>Ayako Furuhama</i>	
25	QSAR Approaches and Ecotoxicological Risk Assessment	615
	<i>Mabrouk Hamadache, Othmane Benkorthbi, Abdeltif Amrane, and Salah Hanini</i>	
26	Multi-scale QSAR Approach for Simultaneous Modeling of Ecotoxic Effects of Pesticides	639
	<i>Alejandro Speck-Planche</i>	
27	Quantitative Structure-Toxicity Relationship Models Based on Hydrophobicity and Electrophilicity	661
	<i>Gourhari Jana, Ranita Pal, Shamik Sural, and Pratim Kumar Chattaraj</i>	
28	Environmental Toxicity (Q)SARs for Polymers as an Emerging Class of Materials in Regulatory Frameworks, with a Focus on Challenges and Possibilities Regarding Cationic Polymers	681
	<i>Hans Sanderson, Kabiruddin Khan, Anna M. Brun Hansen, Kristin Connors, Monica W. Lam, Kunal Roy, and Scott Belanger</i>	
PART IV TOOLS, DATABASES, AND WEB SERVERS		
29	Ecotoxicity Databases for QSAR Modeling	709
	<i>Shinjita Ghosh, Supratik Kar, and Jerzy Leszczynski</i>	
30	VEGAHUB for Ecotoxicological QSAR Modeling	759
	<i>Emilio Benfenati and Anna Lombardo</i>	
31	Enalos Cloud Platform: Nanoinformatics and Cheminformatics Tools	789
	<i>Dimitra-Danai Varsou, Andreas Tsoumanis, Antreas Afantitis, and Georgia Melagraki</i>	
32	alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints	801
	<i>Andrea Mauri</i>	
	<i>Index</i>	821

Contributors

- SERLİ ÖNLÜ • *Bogazici University, Institute of Environmental Sciences, Istanbul, Turkey; Corporate Product Safety/Henkel AG & Co. KGaA, Düsseldorf, Germany*
- ANTREAS AFANTITIS • *Nanoinformatics Department, Novamechanics Ltd, Nicosia, Cyprus*
- RAHUL BALASAHEB AHER • *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India*
- JAN AHLERS • *Consultant, Berlin, Germany*
- PRAVIN AMBURE • *LAQV@REQUIMTE, Department of Chemistry and Biochemistry, University of Porto, Porto, Portugal*
- ABDELTIF AMRANE • *Univ Rennes, Ecole Nationale Supérieure de Chimie de Rennes, CNRS, ISCR - UMR 6226, F-35000, Rennes, France*
- AHMET AYDIN • *Yeditepe University, Faculty of Pharmacy, Department of Toxicology, Istanbul, Turkey*
- PIETRO BARONI • *MISTRAL c/o DSCS – University of Brescia, Brescia, Italy; DII – Department of Information Engineering, University of Brescia, Brescia, Italy*
- RENATA P. C. BARROS • *Laboratory of Cheminformatics, Postgraduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, PB, Brazil*
- SCOTT BELANGER • *The Procter and Gamble Company, Mason, OH, USA*
- EMILIO BENFENATI • *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Department of Environmental Health, Milan, Italy*
- OTHMANE BENKORTBI • *Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Fares de Médéa, Medea, Algeria*
- GIORGIO BERTANZA • *DICATAM – Department of Civil Engineering, Architecture, Land, Environment and Mathematics, University of Brescia, Brescia, Italy*
- ALINA BORA • *“Coriolan Dragulescu” Institute of Chemistry, Timisoara, Romania*
- ANA BOROTA • *“Coriolan Dragulescu” Institute of Chemistry, Timisoara, Romania*
- ANNA M. BRUN HANSEN • *Aarhus University, Department of Environmental Science, Roskilde, Denmark*
- CHIA MING CHANG • *Environmental Molecular and Electromagnetic Physics (EMEP) Laboratory, Department of Soil and Environmental Sciences, National Chung Hsing University, Taichung, Taiwan*
- CHIUNG-WEN CHANG • *Food and Drug Administration, Ministry of Health and Welfare, Taipei, Taiwan*
- LEN CHANG • *Environmental Molecular and Electromagnetic Physics (EMEP) Laboratory, Department of Soil and Environmental Sciences, National Chung Hsing University, Taichung, Taiwan*
- PRATIM KUMAR CHATTARAJ • *Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology Kharagpur, Kharagpur, India; Department of Chemistry, Indian Institute of Technology Bombay, Mumbai, India*
- RICCARDO CONCU • *Department of Chemistry and Biochemistry, Faculty of Science, University of Porto, Porto, Portugal*
- KRISTIN CONNORS • *The Procter and Gamble Company, Mason, OH, USA*
- M. NATÁLIA DIAS SOEIRO CORDEIRO • *LAQV@REQUIMTE, Department of Chemistry and Biochemistry, University of Porto, Porto, Portugal; Department of Chemistry and*

- Biochemistry, Faculty of Science, University of Porto, Porto, Portugal; LAQV-REQUIMTE, Department of Chemistry and Biochemistry, Faculdade de Ciências, Universidade do Porto, Porto, Portugal
- LUMINITA CRISAN • “Coriolan Dragulescu” Institute of Chemistry, Timisoara, Romania
- MARK T. D. CRONIN • School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK
- DAVID J. EBBRELL • School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK
- CLAIRE M. ELLISON • School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK
- SILVIA ESPÍN • Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Campus de Espinardo, Murcia, Spain
- DONATELLA FERETTI • MISTRAL c/o DSCS – University of Brescia, Brescia, Italy; Department of Medical and Surgical Specialities, Radiological Sciences and Public Health, University of Brescia, Brescia, Italy
- JAMES W. FIRMAN • School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK
- SIMONA FUNAR-TIMOFEI • “Coriolan Dragulescu” Institute of Chemistry, Timisoara, Romania
- AYAKO FURUHAMA • Center for Health and Environmental Risk Research, National Institute for Environmental Studies (NIES), Tsukuba, Ibaraki, Japan
- SERGI GÓMEZ-GANAU • ProtoQSAR SL (www.protoqsar.com), Centro Europeo de Empresas Innovadoras (CEEI), Parque Tecnológico de Valencia, Valencia, Spain
- PILAR GÓMEZ-RAMÍREZ • Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Campus de Espinardo, Murcia, Spain
- ANTONIO JUAN GARCÍA-FERNÁNDEZ • Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Murcia, Spain; Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Campus de Espinardo, Murcia, Spain
- SHINJITA GHOSH • School of Public Health, Jackson State University, Jackson, MS, USA; Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, USA
- GIUSEPPINA GINI • DEIB, Politecnico di Milano, Milan, MI, Italy
- RAFAEL GOZALBES • ProtoQSAR SL (www.protoqsar.com), Centro Europeo de Empresas Innovadoras (CEEI), Parque Tecnológico de Valencia, Valencia, Spain; MolDrug AI Systems SL, Valencia, Spain
- MABROUK HAMADACHE • Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Fares de Médéa, Medea, Algeria
- SALAH HANINI • Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Fares de Médéa, Medea, Algeria
- GHEORGHE ILIA • “Coriolan Dragulescu” Institute of Chemistry, Timisoara, Romania; West University Timisoara, Faculty of Chemistry–Biology and Geography, Timisoara, Romania
- GOURHARI JANA • Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology Kharagpur, Kharagpur, India

- SUPRATIK KAR • *Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, USA*
- KABIRUDDIN KHAN • *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India*
- PATHAN MOHSIN KHAN • *Department of Pharmacoinformatics, National Institute of Pharmaceutical Educational and Research (NIPER), Kolkata, India*
- MONICA W. LAM • *The Procter and Gamble Company, Cincinnati, OH, USA*
- JERZY LESZCZYNSKI • *Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, USA*
- TIEN-CHENG LIU • *Bureau of Animal and Plant Health Inspection and Quarantine (BAPHIQ), Council of Agriculture, Executive Yuan, Taipei, Taiwan*
- ANNA LOMBARDO • *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Department of Environmental Health, Milan, Italy*
- JUDITH C. MADDEN • *School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK*
- DIPAKA MANDAL • *Department of Pharmaceutical Technology, University of North Bengal, Darjeeling, West Bengal, India*
- GIUSEPPE FELICE MANGIATORDI • *Istituto di Cristallografia, Consiglio Nazionale delle Ricerche, Bari, Italy*
- EMMA MARTÍNEZ-LÓPEZ • *Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Campus de Espinardo, Murcia, Spain*
- MARCO MARZO • *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy*
- ANDREA MAURI • *Alvascience srl, Lecco, Italy*
- GIOVANNA MAZZOLENI • *MISTRAL c/o DSCS – University of Brescia, Brescia, Italy; DSCS – Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy*
- GEORGIA MELAGRAKI • *Nanoinformatics Department, Novamechanics Ltd, Nicosia, Cyprus*
- MICHELE MONTARULI • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari “Aldo Moro”, Bari, Italy*
- ANA S. MOURA • *LAQV-REQUIMTE, Department of Chemistry and Biochemistry, Faculdade de Ciências, Universidade do Porto, Porto, Portugal*
- CHANIN NANTASENAMAT • *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand*
- ISABEL NAVAS • *Toxicology and Risk Assessment Group, Department of Health Sciences, Biomedical Research Institute of Murcia (IMIB-Arrixaca), Faculty of Veterinary, University of Murcia, Campus de Espinardo, Murcia, Spain*
- MONIKA NENDZA • *Analytical Laboratory, Luhnstedt, Germany*
- ORAZIO NICOLOTTI • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari “Aldo Moro”, Bari, Italy*
- ULF NORINDER • *Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden*
- PROBIR KUMAR OJHA • *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India*
- RANITA PAL • *Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology Kharagpur, Kharagpur, India*
- ROBERTA PEDRAZZANI • *DIMI – Department of Mechanical and Industrial Engineering, University of Brescia, Brescia, Italy; MISTRAL c/o DSCS – University of Brescia, Brescia, Italy*

BAKHTIYOR RASULEV • *Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND, USA*

KUNAL ROY • *Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India*

MELEK TÜRKER SAÇAN • *Bogazici University, Institute of Environmental Sciences, Istanbul, Turkey*

HANS SANDERSON • *Aarhus University, Department of Environmental Science, Section for Toxicology and Chemistry, Roskilde, Denmark; Aarhus University, Department of Environmental Science, Roskilde, Denmark*

DIRK SCHWARTZ • *Bruno Bock Thiochemicals, Marschacht, Germany*

LUCIANA SCOTTI • *Laboratory of Chemoinformatics, Postgraduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, PB, Brazil*

MARCUS T. SCOTTI • *Laboratory of Chemoinformatics, Postgraduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, PB, Brazil*

NATÁLIA F. SOUSA • *Laboratory of Chemoinformatics, Postgraduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, PB, Brazil*

ALEJANDRO SPECK-PLANCHE • *Department of Chemistry, Institute of Pharmacy, I.M. Sechenov First Moscow State Medical University, Moscow, Russian Federation*

NATHALIE STEIMBERG • *MISTRAL c/o DSCS – University of Brescia, Brescia, Italy; DSCS – Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy*

SHAMIK SURAL • *Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India*

FREDRIK SVENSSON • *Alzheimer's Research UK UCL Drug Discovery Institute, University College London, London, UK; The Francis Crick Institute, London, UK*

ANNA RITA TONDO • *Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy*

ANDREAS TSOUMANIS • *Nanoinformatics Department, Novamechanics Ltd, Nicosia, Cyprus*

GULCIN TUGCU • *Yeditepe University, Faculty of Pharmacy, Department of Toxicology, Istanbul, Turkey*

CHIARA URANI • *MISTRAL c/o DSCS – University of Brescia, Brescia, Italy; DISAT – Department of Earth and Environmental Sciences, University of Milan – Bicocca, Milan, Italy*

DIMITRA-DANAI VARSOU • *Nanoinformatics Department, Novamechanics Ltd, Nicosia, Cyprus*

GAIA VIOLA • *Department of Medical and Surgical Specialities, Radiological Sciences and Public Health, University of Brescia, Brescia, Italy*

FANG-WEI WU • *Environmental Molecular and Electromagnetic Physics (EMEP) Laboratory, Department of Soil and Environmental Sciences, National Chung Hsing University, Taichung, Taiwan*

FRANCESCO ZANOLI • *DEIB, Politecnico di Milano, Milan, MI, Italy*

ILARIA ZERBINI • *Department of Medical and Surgical Specialities, Radiological Sciences and Public Health, University of Brescia, Brescia, Italy*

EMANUELE ZILIANI • *DICAr – Department of Civil Engineering & Architecture, University of Pavia, Pavia, Italy*

About the Editor



KUNAL ROY is a professor in the Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India. He has been a recipient of Commonwealth Academic Staff Fellowship (University of Manchester, 2007) and Marie Curie International Incoming Fellowship (University of Manchester, 2013). The fields of his research interest are quantitative structure–activity relationship (QSAR) and chemometric modeling with application in drug design and ecotoxicological modeling. Dr. Roy has published about 300 research articles in refereed journals (current SCOPUS *h* index 41; SCOPUS Author ID 56962764800). He has also coauthored two QSAR related books (Academic Press and Springer), edited five QSAR books (Springer, Academic Press, and IGI Global) and published 12 book chapters. Dr. Roy is a coeditor-in-chief of *Molecular Diversity* (Springer Nature) and editor-in-chief of *International Journal of Quantitative Structure-Property Relationships* (IGI Global). He also serves in different capacities in the Editorial Boards of several International Journals.

Part I

Introduction



Chapter 1

Ecotoxicological Risk Assessment in the Context of Different EU Regulations

Antonio Juan García-Fernández

Abstract

For an appropriate environmental risk assessment, it is necessary to perform a set of ecotoxicological tests in the different environmental compartments. The number and type of ecotoxicological assays that must be performed to introduce a substance, mixture, or product into the market will depend on the properties and characteristics of the chemical itself, its persistence, bioaccumulation, toxicity, and ecotoxicity. In addition, the intended use also determines the type and number of tests to be performed. During the last decades, the European Union has approved many regulations and directives on chemicals, such as those under REACH Regulation, Biocidal Products, Plant Protection Products, Human and Veterinary medicines, Nanoforms, etc. All of them are subjected to rigorous legislative and regulatory frameworks to ensure human health and the environment and, in some circumstances also, animal health.

In addition to these laws focused on the chemicals, there are other crosscutting laws, like the Water Framework Directive (2000/60/EC), Directive 2010/63/EU on the protection of animals used for scientific purposes, or Directive 2004/10/EC on good laboratory practice, which are continuously mentioned in the directives and regulations on chemicals.

Finally, the European Union is currently working to implement a strategy for a nontoxic environment, paying special attention to promoting innovation and the development of sustainable substitutes including nonchemical solutions.

Key words Ecotoxicology, EU laws, EU environment policy, REACH Regulation, CLP Regulation, Biocidal Products Regulation, Plant Protection Products Regulation, Pharmaceuticals, Nanoforms, Water Framework Directive

1 Introduction to the European Union

The European Union (EU) is composed of 28 Member States (sorted by acceptance year into the EU): Belgium, France, Germany, Italy, Luxembourg, and the Netherlands in 1958; Denmark, Ireland, and the United Kingdom in 1973; Greece in 1981; Portugal and Spain in 1986; Austria, Finland, and Sweden in 1995; Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia in 2004; Bulgaria and Romania in 2007; and the last, Croatia in 2013. At the moment, five

other countries are in the process of integrating EU legislation into national law, Albania, Montenegro, North Macedonia, Serbia, and Turkey; and two countries (Kosovo and Bosnia and Herzegovina) are being considered as potential candidates but do not yet fulfil the requirements for EU membership [1].

Four EU institutions have responsibilities in the proposal, development, and final adoption of any legislation under the Treaty of Rome although the EU is considered to be a single institutional body. These four EU institutions are the European Parliament, the European Council, the Council of the EU, and the European Commission [2]. The first EU institution, the European Parliament, is the EU's law-making body, and its members are representatives (directly elected, by the EU voters every 5 years) with legislative, budgetary, and supervisory responsibilities. The second one, the European Council, representing the highest level of political cooperation between EU countries, sets the EU's broad priorities and the general political direction. The third one, the Council of the EU, is the institution where the respective governments defend their own country's national interests, adopting EU laws and coordinating EU policies. In this Council, the government ministers from each EU country discuss, amend, and adopt the laws. This institution is considered, together with the European Parliament, the main decision-making body of the EU. The last one, the European Commission, is the EU's politically independent executive arm, and it is responsible for drawing up proposals for new European legislation, and it also implements the decisions of the European Parliament and the Council of the EU [2]. It is composed of 28 Commissioners (one from each EU country). The European Commission is organized into 53 departments known as Directorates-General (DGs), each responsible for a specific policy area, one of them being the DG for Environment.

1.1 European Union Law

In the EU, there are two types of law: primary and secondary. The primary law includes all EU treaties which constitute the legal basis of all the Union's binding acts, and therefore they are the starting point for all EU laws. On the other hand, the rest of the legislative acts that come from the principles and objectives of the treaties are considered as "secondary law." Regulations, directives, decisions, recommendations, and opinions constitute the secondary law [3].

The EU treaties set out EU objectives, rules for EU institutions, how decisions are made, and the relationship between the EU and its member countries. Every action taken by the EU is founded on treaties [4] which are binding agreements between all EU member countries, so that is why they have to be amended every time a new country is accepted as a member state. In addition, they are amended when it is necessary to make the EU more transparent and efficient and especially when new areas of cooperation must be introduced and regulated. In 2016, every one of the

Table 1
Main treaties and other specific treaties in force in the European Union

<i>Treaty in force and year</i>	<i>Consolidated version</i>	<i>Previous</i>
Treaty on the European Union (TEU) (2016)	OJ C 202, 7.6.2016 http://data.europa.eu/eli/treaty/teu_2016/oj	2010, 2008, 2006, 2002, 1997, 1992
Treaty on the Functioning of the European Union (TFEU) (2016)	OJ C 202, 7.6.2016 http://data.europa.eu/eli/treaty/tfeu_2016/oj	2012, 2010, 2008
Treaty establishing the European Atomic Energy Community (2016)	OJ C 203, 7.6.2016 http://data.europa.eu/eli/treaty/euratom_2016/oj	2012, 2010, 1957
Charter of Fundamental Rights of the EU (2016)	OJ C 202, 7.6.2016, p. 391-407 http://data.europa.eu/eli/treaty/char_2016/oj	2012, 2010, 2007
Treaty establishing the European Community (2006)	OJ C 321E, 29.12.2006	2002, 1997, 1992, 1957
Treaty establishing the European Economic Community (1957)	http://data.europa.eu/eli/treaty/teec/sign	Not in English. Only in French, German, Italian, and Dutch
<i>Specific treaties and year</i>	<i>Published in</i>	
Treaty of Lisbon (2007)	OJ C 306, 17.12.2007, 1-271 http://data.europa.eu/eli/treaty/lis/sign	
Treaty establishing a Constitution for Europe (2004)	OJ C 310, 16.12.2004 http://data.europa.eu/eli/treaty/tcons_2004/oj	
Treaty of Nice (2001)	OJ C 80, 10.3.2001, 1-87 http://data.europa.eu/eli/treaty/nice/sign	
Schengen Convention (1985)	OJ L 239, 22.9.2000	
Treaty of Amsterdam (1997)	OJ C 340, 10.11.1997, 173-306 http://data.europa.eu/eli/treaty/tec_1997/oj	
Agreement of the European Economic Area (1992)	OJ L 1, 3.1.1994	
Treaty of Greenland (1984)	OJ L 29, 1.2.1985, 1-7 http://data.europa.eu/eli/treaty/tgreenl/sign	
Merger Treaty (1965)	OJ 152, 13.7.1967, 2-17 http://data.europa.eu/eli/treaty/fusion/sign	

treaties currently in force was published in the following consolidated versions [5]: the Treaty on European Union (TEU), the Treaty on the Functioning of the European Union (TFEU), the Treaty establishing the European Atomic Energy Community, and the Charter of Fundamental Rights of the EU (Table 1). TEU and TFEU consolidated versions were published together with the

annexes and protocols thereto, as they result from the amendments introduced by the Treaty of Lisbon, which was signed on 13 December 2007 in Lisbon and which entered into force on 1 December 2009.

1.2 Legislative and Nonlegislative Acts in the EU

The EU is based on the rule of law, and the European Commission is the only institution empowered to initiate legislation. Legislative acts in EU are adopted following one of the legislative procedures set out in the EU treaties: ordinary or special [3]. The EU's standard decision-making procedures follow what is known as *ordinary legislative procedure* (formerly known as *codcision*) which gives the same weight to the European Parliament and the Council of the European Union on some areas like economic governance, consumer production, energy, or environment, among others; that is why both institutions, the Parliament and the Council, adopt jointly the majority of the European laws [6]. The other type of procedure, *special legislative procedure*, is followed only in certain cases. The most usual case is when the EU Council is the sole legislator and the EU Parliament is consulted in regard to a legislative proposal or is required to give its consent. Sometimes, but rarely, the Parliament alone adopts legal acts but always after consulting the Council. Following these legislative procedures, the EU adopts the main binding legislative acts: Regulations, Directives, and Decisions; and two minor acts: Delegated and Implementing acts (Table 2).

Nonlegislative acts (Recommendation and Opinion) do not follow the standardized legislative procedures, so they are adopted by the EU institutions according to their specific rules, and they are not binding acts. The first one, *Recommendation*, allows the institutions to make their views known and to suggest a line of action without imposing any legal obligation on those to whom it is addressed. The second one, *Opinion*, allows the institutions to make a statement, without imposing any legal obligation on those to whom it is addressed [3].

1.3 EU Policies on Environment

One of the EU priorities is to ensure that chemicals are safe for human health and the environment and simultaneously keep EU industry competitive internationally. That is why *Chemicals* is the name of one of the policies closely linked to EU Environment Area. Apart from *Chemicals*, other environment-related policies are considered in the EU: *Circular economy*, *Marine and coastal environment*, *Clean air*, *Noise pollution*, *Soil quality*, *Urban environment*, *Waste and recycling*, *Water resources*, and *Endocrine disruptors*. In the EU webpage, one can read on *Environment Policy*: “EU environmental policies and legislation aim to enable EU citizens to live well, within the planet’s ecological limits. These are centred on an innovative, circular economy, where biodiversity is protected, valued and restored and environment-related health risks are

Table 2
Types of legislative acts in the European Union

<i>Binding legislative acts</i>	
Legislative acts	Adopted by a legislative procedure (ordinary or special legislative procedure), Article 289 Treaty on the Functioning of the European Union
Regulation	Legal act that has general application. It is binding in its entirety and directly applicable in all Member States. EU countries as soon as they enter into force, without needing to be transposed into national law
Directive	Legislative binding act that sets out a goal that all EU countries must achieve. It leaves, however, to the national authorities the choice of form and methods. Transposition into national law must take place by the deadline set in the directive but generally is within 2 years
Decision	Legislative binding and directly applicable act in its entirety on those to whom it is addressed (e.g., one, several, or all Member States or an individual company)
<i>Other binding legislative acts</i>	
Delegated acts	They are acts legally binding that enable the Commission to supplement or amend nonessential parts of EU legislative acts. If Parliament and Council have no objections, it enters into force
Implementing acts	They are legally binding acts that enable the Commission (under supervision of committees consisting of EU countries' representatives) to set conditions that ensure that EU laws are applied uniformly

minimised—enhancing our society's resilience, and decoupling growth from resource use” [7].

The EU considers that chemicals participate in an essential manner in our daily lives, but sometimes, under certain circumstances and conditions, they can provoke severely deleterious effects on our health and/or on the environment. The EU is aware that there has been an increase in health problems due, partially, to the use of chemicals. Similarly, the environment is suffering from the consequences of the increased chemical presence not only where they are generated but also in the most remote places, recognizing that chemicals are everywhere [8]. In this sense, the EU has a comprehensive chemical legislation, spearheaded by REACH Regulation [9] (Registration, Evaluation, Authorisation and Restriction of Chemicals) and CLP Regulation [10] (Classification, Labeling and Packaging of chemical substances and mixtures). These legislative documents are complemented with legislation on specific groups of chemicals, such as pesticides, biocides, pharmaceuticals, etc. In addition, nowadays, endocrine disruption is a concern in the European Union, so the European Commission is focusing its interest on those chemicals able to interfere with the hormone system [8].

2 EU Laws Related to Ecotoxicology Risk Assessment

2.1 REACH

Regulation (EC) 1907/2006

Regulation (EC) 1907/2006 of the European Parliament and the Council, commonly known as REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals), was published in December 2006, coming into force on 1 June 2007 [9]. The adoption of this Regulation was probably one of the most relevant legislative acts on chemicals in the world, with a priority on ensuring a high level of protection of human health and the environment. This regulation also considers it a priority to avoid or reduce the number of tests on live animals for hazard assessment of chemicals. These aims should be compatible with the free movement of substances and products containing them, as well as enhancing competitiveness and innovation. REACH Regulation is based on the principle that the responsibility for the identification and management of risks from substances and products lies with manufacturers, importers, and downstream users. According to this, industry should act with the highest level of responsibility to ensure the safety of human health as well as the environment, and always in accordance with the precautionary principle. Another important aim of this new system is to ensure that substances of high concern must be replaced by other less dangerous substances or technologies where viable alternatives are available.

REACH Regulation foresees that the assessment of the chemical safety of a substance should include not only hazard assessments on human health and the environment but also assessments of persistent, bioaccumulative, toxic (PBT), and very persistent and very bioaccumulative (vPvB) substances. Focusing attention on environmental hazards, its objective is not only to determine the classification and labeling of a substance but also to set out the threshold level of the substance below which effects on the environment are not expected to occur. This level is known as Predicted No-Effect Concentration (PNEC). As a first step, the assessment must evaluate all available information on potential effects on the aquatic (including sediment), terrestrial, and atmospheric compartments, microbiological activity of sewage treatment systems, and also potential effects via food chain accumulation. The information on these effects will be included in the Chemical Safety Report (Table 3); and its evaluation will comprise the hazard identification and the establishment of the quantitative dose-response or concentration-effect relationships. All information used to assess the effects on the environment or on a particular environment shall be briefly presented together with any other relevant information. With respect to identification of the PNEC, it must be established for each environmental sphere.

The chemical safety assessment must be reported for all substances subject to registration according to REACH in annual

Table 3
Chemical safety report under REACH Regulation

Chemical safety report (REACH Regulation)
<i>Part A</i>
Summary of risk management measures
Declaration that risk management measures are implemented
Declaration that risk management measures are communicated
<i>Part B</i>
Identity of the substance and physical and chemical properties
Manufacture and uses
Classification and labeling
Environmental fate properties
Human health hazard assessment
Human health hazard assessment of physicochemical properties
<i>Environmental hazard assessment</i>
Aquatic compartment (including sediment)
Terrestrial compartment
Atmospheric compartment
Microbiological activity on sewage treatment systems
Persistent, bioaccumulative, and toxic (PBT) and vPvB assessment
Exposure assessment (scenario 1, scenario 2, . . . , overall exposure)
Risk characterization (scenario 1, scenario 2, . . . , overall exposure)

quantities of 10 tonnes or more per registrant. If, as a result of these assessments cited above, the registrant concludes that the substance can be considered as persistent, bioaccumulative, and toxic (PBT) or very persistent and very bioaccumulative (vPvB) or it must be classified as dangerous according to Directive 67/548/EEC (repealed by Regulation EC 1272/2008 [10]), the chemical safety assessment must be implemented with additional information on exposure estimation, exposure assessment in the possible exposure scenarios, and the risk characterization in all possible scenarios, addressing all identified uses by the registrant.

The volume of manufacture or importation of a substance is related to the potential for exposure to human beings, other living beings, and environment to the substances; that is why the information on substances that is required by REACH Regulation depends on the volumes of importation or manufacture of each substance. In this sense, Article 12 details the type of information

to be submitted depending on tonnage. Concretely, it is specified that the technical dossier shall include all physicochemical, toxicological, and ecotoxicological information that could be relevant and available to the registrant of the product or substance. REACH Regulation sets out four ranges of tonnages (1–10, 10–100, 100–1000, and more than 1000 tonnes) of substances manufactured or imported to require the corresponding information (see Annexes VII–X, REACH Regulation) (Table 4). All ecotoxicological and toxicological tests and analyses required by REACH must be carried out in compliance with the principles of good laboratory practice.

In REACH, Annex XI describes the general rules for adaptation of the standard testing regime set out in the previous Annexes VII to X (information required according to tonnage). This annex includes some circumstances that could introduce adaptations to the standard testing regime (Table 5).

REACH dedicates an article to “Review” its effectiveness in achieving these aims, based on the experience in its application. Some of the reviews in Article 138 are as follows:

- To evaluate the necessity to extend the obligation to do a chemical safety assessment with the pertinent chemical safety report to substances not covered by this obligation at the time REACH Regulation entered into force (i.e., substances manufactured or imported in quantities below 10 tonnes). The deadline is 1 June 2019 for all substances, with the exception of the substances considered carcinogenic, mutagenic, or toxic for reproduction substances, whose review shall be carried out by 1 June 2014. The review takes into consideration all relevant factors including those with regard to costs of drawing up the chemical safety reports, distribution of costs, and benefits for human health and the environment.
- On the experience acquired with the application of REACH, the requirements for registration of substances manufactured or imported in quantities between 1 and 10 tons per manufacturer or importer could be reviewed. The Commission may modify these requirements taking into consideration the latest developments, especially in alternative methods and procedures and (quantitative) structure-activity relationships (Q)SAR.
- To promote the use of alternative methods of non-animal testing and the application of the 3R principles in animal testing (replacement, reduction, and refinement), the requirements to evaluate reproductive toxicity (Subheading 8.7) of Annex VIII (more than 1 tonne) will be reviewed before 1 June 2019.

With the entry into force of REACH Regulation, Directive 91/155/ECC was automatically repealed, but other laws were also repealed, with effects 1 or more years later, such as Directives

Table 4

Standard ecotoxicological information required for substances manufactured or imported in accordance to the tonnage (information extracted from REACH Annex VII to X)

REACH TONNES	REACH TONNES	REACH TONNES	REACH TONNES
9.1. Aquatic toxicity			
9.1.1. Short-term toxicity testing on invertebrates (preferred species Daphnia)			
9.1.2 Growth inhibition study aquatic plants (algae preferred)			
9.1.3. Short-term toxicity testing on fish: The registrant may consider long-term toxicity testing instead of short-term			
9.1.4. Activated sludge respiration inhibition testing			
9.1.5. Long-term toxicity testing on invertebrates			
9.1.6. Long-term toxicity testing on fish			
9.1.6.1. Fish early-life stage toxicity test			
9.1.6.2. Fish short-term toxicity test on embryo and sac-fry stages			
9.1.6.3. Fish, juvenile growth test			
9.2. Degradation			
9.2.1. Biotic			
9.2.1.1. Ready biodegradability			
9.2.1.2. Simulation testing on ultimate degradation in surface water			
9.2.1.3. Soil simulation testing			
9.2.1.4. Sediment simulation testing			
9.2.2. Abiotic			
9.2.2.1. Hydrolysis as a function of pH			
9.2.3. Identification of degradation products			
9.3. Fate and behaviour in the environment			
9.3.1. Adsorption/desorption screening			
9.3.2. Bioaccumulation in aquatic species, preferably Fish			
9.3.3. Further information on adsorption/desorption depending on the results of the study required in Annex VIII			
9.3.4. Further information on the environmental fate and behaviour of the substance and/or degradation products			
9.4. Effects on terrestrial organisms			
9.4.1. Short-term toxicity to invertebrates			
9.4.2. Effects on soil micro-organisms			
9.4.3. Short-term toxicity to plants			
9.4.4. Long-term toxicity testing on invertebrates, unless already provided as part of Annex IX requirements.			
9.4.6. Long-term toxicity testing on plants, unless already provided as part of Annex IX requirements.			
9.5.1. Long-term toxicity to sediment organisms			
9.6.1. Long-term or reproductive toxicity to birds			

Table 5

General rules for adaptation of the standard testing regime set out in annexes VII to X (information extracted from REACH Annex XI)

1. Testing does not appear scientifically necessary
1.1. Use of existing data
<ul style="list-style-type: none"> • On physico-chemical properties from experiments not carried out according to GLP or the test methods referred to in Article 13. • On human health and environmental properties from experiments not carried out according to GLP or methods referred to in Article 13. • Historical human data (i.e. epidemiological studies and others)
1.2. Weight of evidence. Where sufficient weight of evidence for the presence or absence of a particular dangerous property is available:
<ul style="list-style-type: none"> • Further testing on vertebrate for that property shall be omitted • Further testing not involving vertebrate animals may be omitted.
1.3. Qualitative or Quantitative structure-activity relationship. (Q)SAR
results may be used instead of testing when:
<ul style="list-style-type: none"> • Results are derived from a (Q)SAR model scientific validated • Substance falls within the applicability domain of the (Q)SAR model • Results are adequate for classification, labelling and/or RA • Adequate and reliable documentation of the applied method
1.4. Results obtained from in vitro methods
<ul style="list-style-type: none"> • In vitro method is scientific validated by a suitable validation study • Results are adequate for classification and labelling and/or RA • Adequate and reliable documentation of the applied method
1.5. Grouping of substances and read-across approach. Substances whose physicochemical, toxicological and ecotoxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity may be considered as a group of substances. The similarities may be based on:
<ul style="list-style-type: none"> • A common functional group • The common precursors and/or the likelihood of common breakdown products via physical and biological processes, which result in structurally similar chemicals • A constant pattern in the changing of the potency of the properties across the category
2. Testing is technically not possible
<ul style="list-style-type: none"> - Very volatile compound - Highly reactive substance - Unstable substance
3. Substance-tailored exposure-driven testing
3.1. Testing in accordance with sections 8.6 (Repeated dose toxicity) and 8.7 (Reproductive toxicity) of Annex VIII (10–100 tonnes), Annex IX (100–1,000 tonnes), and Annex X (> 1,000 tonnes) may be omitted, based on the exposure scenario(s) developed in the Chemical Safety Report.
3.2. In all cases, adequate justification and documentation shall be provided.

93/105/EC and 2000/21/EC and Regulations EEC/793/93 and EC/1488/94 (effect 1 June 2008), Directive 93/769/EEC (effect 1 August 2008), and Directive 76/769/EEC (effect 1 June 2009).

Shortly after the entry into force of the REACH Regulation, Regulation (EC) 440/2008 was approved, which includes a list of 23 test methods for ecotoxicity assessment to be applied for the purposes of REACH Regulation [11]. Ten of the tests included in this list are performed using live animals, both vertebrates (fish) and invertebrates, and are the following: three acute toxicity tests for fish, *Daphnia* sp., and algae, respectively, two more tests using fish for bioconcentration and for growth of juveniles, a short-term toxicity test on fish embryos, a toxicity test using earthworms, two acute tests on honeybees, and a reproduction test using *Daphnia magna*. The remaining 13 tests do not use animals: a test for determination of ready biodegradability, three tests to evaluate degradation (two for biochemical and chemical oxygen demands, and a test for abiotic degradation measuring hydrolysis as a function of pH), four tests to evaluate biodegradation (Zahn-Wellens test, Modified SCAS test and two tests using activated sludges), a method for evaluation of adsorption/desorption, two tests using soil microorganisms (nitrogen and carbon transformation tests), and finally, two tests for aerobic and anaerobic transformation in soil and in aquatic sediment systems.

2.2 Nanoforms in Amended REACH Regulation (EU) 2018/1881

Recently the Commission Regulation (EU) 2018/1881 has amended REACH Regulation regarding Annexes I, III, VI, VII, VIII, IX, X, XI, and XII to address nanoforms of substances [12]. On the basis of the Commission Recommendation of 18 October 2011 on the definition of nanomaterial, a nanoform *is a form of a natural or manufactured substance containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50% or more of the particles in the number size distribution, one or more external dimensions is in the size range 1–100 nm, including also by derogation fullerenes, graphene flakes and single wall carbon nanotubes with one or more external dimensions below 1 nm.*

From a risk assessment point of view, the Commission considers that the exposure pattern and toxicological and ecotoxicological profiles of the nanoforms, as well as their behavior in the environment, may be influenced by particle size, shape, and surface properties. That is why they require a specific risk assessment and therefore an appropriate specific risk management. The Commission also considers that the existing qualitative or quantitative structure-activity relationship (QSAR) does not yet allow us to prioritize specific substances, and thus the information on insolubility should be used alternatively to evaluate the toxicological and ecotoxicological issues.

Regarding the aspects mentioned above, Annexes VII, VIII, and IX of REACH Regulation [9] were modified by the inclusion of the following text in the corresponding introductory text of each annex: “Without prejudice to the information submitted for other forms, any relevant physicochemical, toxicological and ecotoxicological information shall include characterisation of the nanoform tested and test conditions. A justification shall be provided where QSARs are used or evidence is obtained by means other than testing, as well as a description of the range of the characteristics/ properties of the nanoforms to which the evidence can be applied.”

On the other hand, nanoforms may be grouped in sets of nanoforms when they have similar physicochemical properties and toxicological and ecotoxicological characteristics or they follow similar patterns as a result of having similar chemical structures (REACH Regulation Annex I modified). In any case, a nanoform must only correspond to a single group of nanoforms. This Regulation also mentions that certain physicochemical properties like water solubility or partition coefficient in octanol/water serve as a data input to QSAR and other predictive models. In any case, any adaptation must be appropriately justified from a scientific point of view.

2.3 Biocidal Products Regulation (EU) 528/2012

Regulation EU 528/2012 (Biocidal Products Regulation—BPR) defines *biocide product* as *any substance or mixture in the form in which it is supplied to the user, consisting of, containing or generating one or more active substances, with the intention of destroying, deterring, rendering harmless, preventing the action of, or otherwise exerting a controlling effect on, any harmful organisms by any means other than mere physical or mechanical action*. The Regulation also includes in this definition “any substance or mixture, generated from substances or mixtures which do not themselves fall under the first indent, to be used” *with the same purposes cited above* [13].

The Biocides Regulation has been in force since 1 September 2013, repealing Directive 98/8/EC [14]. This Regulation has the purpose to improve the functioning of the internal market of biocidal products, guaranteeing a high level of protection of both human and animal health and the environment. For that, all provisions concerning the protection of human and animal health and the environment are based on the precautionary principle, taking special care of the protection of vulnerable groups.

In this Regulation, the active substance (substance that has an action on or against harmful organisms) and the biocidal products are treated separately. In regard to the active substances, the main scope of this Regulation is the establishment at Union level of a list of substances that may be contained in biocidal products. With respect to biocidal products, this Regulation lays down rules for authorization for the market and use.

In accordance to the BPR Regulation, manufacturers and importers must supply data on the substances, such as chemical identity, annual volume manufactured or imported, concentration of active substance in the biocidal product, and, when needed, any information on potential risks of significant and relevant impurities and non-active substances present in the product. In addition, they must supply information on its residues of toxicological or environmental significance. Regarding the assays submitted for approval of an active substance as biocide, these must be done according to the methods included in the Regulation (EC) 440/2008 [11] laying down test methods pursuant to REACH Regulation. If validated methods are not available, appropriated and internationally recognized scientific methods could be used, but their appropriateness must be adequately justified. In the case of nanomaterials, justifications on scientific appropriateness of the methods must be also submitted. In the case of mixtures, when valid data on each of the components are available and synergistic effects are not expected, the mixture can be classified according to REACH [9] and CLP [10] Regulations.

The general rules for the adaptation of the data requirements (Table 5) according to REACH Regulation are also applied to biocides (Annex IV Regulation EU 528/2012) [13].

2.4 Plant Protection Products Regulation (EC) 1107/2009

In October 2009, the European Parliament and the Council approved Regulation (EC) 1107/2009 concerning the placing of plant protection products (PPPs) on the market [15]. The purposes of this Regulation are to state rules for the authorization of plant protection products (PPP) and their commercial forms and to improve the functioning of their internal market together with an improvement of the agricultural production; and all of this should go hand in hand with the highest level of protection of both human and animal health and the environment.

This Regulation shall apply to both plant protection products and active substances. The definition of PPP according to this Regulation is *any product supplied to the user, consisting of or containing active substances, safeners or synergists, and intended for one of the following uses: protecting plants, influencing the life processes of plants, preserving plant products, destroying undesired plants, checking or preventing undesired growth of plants.*

For the approval of active substances, safeners, and synergists, a specific procedure and criteria must be followed. In this sense, from an ecotoxicological point of view, they will only be approved if potential risks may be considered as acceptable under PPP Regulation. The uncertainties of the available data, severity of the harmful effects, and the group of organisms that is expected to suffer adverse effects by the intended use must be taken into consideration in the assessment. The substance will be approved if, depending on the basis of the assessment of internationally recognized test

guidelines, there is no evidence of endocrine disruption on nontarget organisms.

Also, substances may be approved if assessment of effects on honeybee larvae and honeybee behavior demonstrates that their appropriate use is not able to provoke a hazardous exposure for honeybees or no adverse effects on colony survival and development are expected.

In this Regulation, the presence of residues of the PPPs, due to applications consistent with good plant protection practice, shall not have any adverse effects on human health (including vulnerable groups), animal health, and the environment, considering known cumulative and synergistic effects.

Certain active substances contained in PPPs may be considered as low risk (Article 22) to human and animal health and the environment. According to this Regulation, carcinogenic and mutagenic active substances shall not be considered of low risk. It is the same with substances considered toxic to reproduction, very toxic, or sensitizing. In addition, it shall also not be considered of low risk if its half-life in soil is more than 2 months, its bioconcentration factor (BCF) is higher than 100, and it is suspected to be an endocrine disruptor, neurotoxic or immunotoxic substance.

2.5 Pharmaceuticals in the Environment: EU Strategic Approach

More than 3000 active pharmaceutical compounds are being currently used in Europe, and the sales of human medicines have been increasing over the past three decades. Recently, in March 2019, the European Commission has communicated to the European Parliament, the Council, and the European Economic and Social Committee its position on pharmaceuticals in the environment, entitled *European Union Strategic Approach to Pharmaceuticals in the Environment* [16]. In this communication, the EU recognizes that residues may enter into the environment during the production, use, and disposal of pharmaceuticals, which explains their detection in surface water and groundwater systems, soil, wildlife tissues, and even drinking water. The concentrations of pharmaceuticals in different parts of the ecosystems will depend on certain properties of the pharmaceutical itself, like chemical and metabolic stability; but they will also depend on the nature and proximity of the sources of release into the environment. Pharmaceuticals reach the environment in different ways, such as effluents from wastewater treatment plants, animal manure, aquaculture, sewage sludge as fertilizer, the pharmaceutical industry, veterinary treatments, etc. A growing number and variety of pharmaceuticals are being detected in the environment, such as antibiotics, anti-inflammatories, contraceptives, antidepressants, painkillers, hormones, or antiparasitics. The European Commission recognizes that this is an emerging problem, and well-documented evidence exists on the risks to the environment, and, particularly, of great concern is the increase in antimicrobial resistance in humans.

It should be noted that the term pharmaceutical is applied to any medical product, whether it be for human or veterinary use, where the substances of main concern are the active pharmaceutical ingredients, although their degradation products and relevant metabolites must be also taken into account. Also, excipients and the packaging material must be included. Although the EU recognizes both the presence and effects of these substances on the environment, it also recognizes that there is as yet no clear link between pharmaceuticals in the environment and direct impacts on human health. Of special concern are antimicrobials.

The Commission has selected the following four main objectives of the strategic approach: (1) identify actions to be taken or further investigated to address the potential risks from pharmaceutical residues in the environment, with special attention to antimicrobial resistance; (2) encourage innovation and promote the circular economy through recycling of water, sewage sludge, and manure; (3) identify other knowledge gaps proposing possible solutions; and (4) take special care so that proposals to address the risks do not impede access to safe and effective pharmacological treatments.

The European Commission Communication on the *European Union Strategic Approach to Pharmaceuticals in the Environment* set out the following six areas for action [16]:

1. Increase awareness and promote prudent use of pharmaceuticals (including antimicrobials), including cooperation, among others, with the World Health Organization.
2. Support the development of pharmaceuticals intrinsically less harmful for the environment, and promote greener manufacturing.
3. Improve environmental risk assessment and its review, mainly collaborating with the European Medicines Agency (EMA) and Member States and considering the findings of recent REACH evaluations.
4. Reduce wastage and improve the management of waste, in collaboration also with the EMA and Member States, improving urban wastewater treatment, improving the Codes of Good Agricultural Practices, considering the next evaluation of the Industrial Emissions Directive, etc.
5. Expand environmental monitoring to know more about the concentrations of pharmaceuticals in the environment in order to improve the ERAs and to give better coverage to certain pharmaceuticals in all parts of the environment, for example, cytotoxics, X-ray contrast media, antimicrobial-resistant microorganisms, etc.
6. Fill other knowledge gaps, such as further research on ecotoxicity and environmental fate of those pharmaceuticals not yet

subject to ERA, linkage between antimicrobial in the environment and development and spread of antimicrobial resistance, possible effects on humans exposed chronically to low levels of pharmaceuticals via the environment, and cost-effective methods for reducing the presence of pharmaceuticals in slurry, manure, and sewage sludge.

2.5.1 Environmental Risk Assessment of Medicinal Products

As with other chemical substances, EU legislation on medicinal products for human and veterinary uses has been elaborated on the premise that they will be safe not only for human and animal health, respectively, but also for the environment. In the particular case of veterinary medical products, the *assessment of ecotoxicity* required by the Directive 2001/82/EC [17] has changed to an *environmental risk assessment (ERA)* in the new Regulation 2019/6 [18] (repealing Directive 2001/82/EC) for all new applications that require marketing authorization. Also, in this last Regulation, the ERA for veterinary medicinal products will be different depending on whether the product contains or consists of genetically modified organisms. Finally, this Regulation sets out a provision for a review of rules for ERA by 28 January 2022.

In the case of medicinal products for human use, the assessment of potential risks to the environment [19] is a stepwise, phased procedure, consisting of two phases (Table 6). This guideline is applied to all human medicinal products excepting those containing genetically modified organisms (GMOs) [20].

Phase I. Estimation of exposure. In this phase substances should be screened for persistence, bioaccumulation, and toxicity in accordance to the Test Guidelines of the European Chemical Bureau [21] followed by calculation of Predicted Environmental Concentration (PEC), which is only carried out in the aquatic compartment.

Table 6
The phased approach in the environmental risk assessment (EMEA 2006)

Stage in regulatory evaluation	Stage in risk assessment	Objective	Method	Test/data requirement
Phase I	Pre-screening	Estimation of exposure	Action limit	Consumption data, logKow
Phase II Tier A	Screening	Initial prediction of risk	Risk assessment	Base set aquatic toxicology and fate
Phase II Tier B	Extended	Substance and compartment-specific refinement and risk assessment	Risk assessment	Extended data set on emission, fate, and effects

Phase II. It is conducted by evaluating PEC/PNEC ratio and is divided into two parts, Tiers A and B. In Tier A (initial environmental fate and effect analysis), information on physicochemical properties of the substance, as well as the fate in the environment, is obtained. In Phase II Tier A, seven types of study may be carried out, all of them in the aquatic compartment. The recommended protocols are adsorption/desorption test, ready biodegradability test, aerobic and anaerobic transformation in aquatic sediment systems, growth inhibition test on algae, *Daphnia* sp. reproduction test, fish early-life stage toxicity test, and activated sludge respiration inhibition test [22]. In addition, a groundwater assessment will be performed. When needed the substance will be assessed following procedures of the Phase II Tier B (extended environmental fate and effects analysis). In Tier B, water sediment effects, specific effects on aquatic microorganisms, and another five tests in the terrestrial compartment will be performed. These five tests are aerobic and anaerobic transformation in soil, nitrogen transformation test on soil microorganisms, growth test on terrestrial plants, acute toxicity test on earthworm, and reproduction test on *Collembola* [22].

2.6 Classification, Labeling, and Packaging Regulation (EC) 1272/2008

Since June 2015, Classification, Labeling and Packaging (CLP) Regulation (EC) 1272/2008 [10] is the only law in force in the European Union on classification and labeling of substances and mixtures, being directly applicable to industrial sectors. Its main objective is to ensure a high protection level for human health and the environment, as well as free movement of substances, mixtures, and products. Regulation amends REACH (EC) 1907/2006 [9] and both the Dangerous Substances (67/548/EEC) and Preparations (1999/45/EC) Directives. This Directive determines how substances and their mixtures must be classified and labeled according to their human and environmental hazards. Manufacturers, importers, or downstream users must assess all available information related to hazardous properties of a substance or mixture; and when no data are available, toxicological and ecotoxicological assays must be carried out according to the OECD principles of good laboratory practice (GLP) and the REACH Regulation or any recognized methods validated. It must be noted that testing on humans and nonhuman primates is prohibited, but in certain cases assays using laboratory animals have to be performed; in which case, all procedures must follow the legal requirements for the protection of animal experimentation (Directive 2010/63/EU) [23]. This Directive includes an explicit reference to the 3Rs principle: replace, reduce, and refine. The replacement is to use alternative methods replacing the use of animals with non-animal procedures or protocols. The aim of reduction is to obtain the needed information with lower number of animals killed. Finally, with refinement measures, it is possible to minimize the distress,

suffering, or pain in animals used in experimentation, improving their welfare. This type of alternative method includes chemical properties, QSAR models and predictions, *in vitro* tests, and application of new technologies such as proteomics or genomics. In this sense, one of the ECHA priorities is to promote alternatives to animal testing to assess risks to the environment and human health.

2.7 EU Water Framework Directive 2000/60/EC

The Water Framework Directive (WFD) 2000/60/EC [24] is, together with REACH Regulation [9], one of the most relevant legislative acts of the European environmental policy for ecotoxicologists. In accordance with WFD, the Commission shall submit a list of priority pollutants selected from among those presenting a significant risk to or via the aquatic environment as part of a strategy against water contamination. This strategy should include measures of control such as the progressive reduction and/or cessation of discharges, emissions, and losses. In 2001, the Commission adopted the Decision 2455/2001/EC [25] establishing the first list of priority substances, and 7 years later, the Environmental Quality Standards Directive (EQSD) 2008/105/EC set the quality standards as required in the WFD [26]. The priority list includes 33 substances or groups of substances showing a significant risk; 11 of them are considered as priority hazardous substances. This list includes plant protection products, polycyclic aromatic hydrocarbons (PAH), existing chemicals, biocides, metals, polybrominated diphenyl ethers (PBDE), etc. Regarding groundwater, the European Parliament and the Council adopted measures to control pollution of groundwater, promoting also a specific strategy (Directive 2006/118/EC) [27].

On the other hand, Directive 2013/39/EU [28], amending Directive 2008/105/EC [26], considers that it is necessary to collect high-quality monitoring data of substances together with toxicological and ecotoxicological effects related to them to carry out risk assessments able to justify appropriately their inclusion as new priority substances. The Directive in force also considers that there is a lack of routine monitoring data for many emerging contaminants at the Union level, which could pose risks derived from their potential toxicological and ecotoxicological effects. Accordingly, Article 8b (*Watch list*) considers the preparation of a new list of substances for which monitoring data at the Union level is needed to support the priority list. This list will contain no more than ten substances or groups of substances indicating complementary information about appropriate analytical methods, matrices to be monitored, etc. Concretely, for pharmaceutical substances, Article 8c requires the EC to implement a strategic approach to contamination of water by the presence of pharmaceutical compounds, including the environmental impacts of pharmaceuticals to be considered in the procedure for placing medical products on the market. Finally, this article also requires the Commission to propose

measures to reduce emissions, discharges, and losses of hazardous medical products in aquatic ecosystems, balancing costs and effectiveness and always under the premise of the safeguarding of public health.

2.8 Crosscutting Legislation in Ecotoxicology Tests

In the sections dealing with REACH [9] and CLP [10] Regulations, it has been mentioned the need to prioritize the selection of methods and tests not using animals, thereby avoiding needless animal sacrifices (replacement of animal tests). Together with the replacement measures, the other 2R measures (reduction and refinement) must be implemented. In addition to REACH [9] and CLP [10] Regulations, all European regulations and directives abovementioned, and any other European legislative act that requires information on risks to human health and the environment, special attention is paid to the application of 3Rs principles for animals used for experimentation. In accordance to these regulations, companies are responsible for providing information on the hazards, risks, and safe use of chemical substances that they manufacture or import. It is for these reasons that all toxicological and ecotoxicological information required in each one of these regulations or directives, independently of the type of active substance, mixture, product, or article, have to be performed complying with the requirements of protection of laboratory animals set out in the Directive 2010/63/EU of the European Parliament and the Council of 22 September 2010 on the protection of animals used for scientific purposes [23].

According to Directive 2010/63/EU [23], the Member States must ensure that a procedure is not carried out if there is a method recognized under Union legislation, not entailing the use of live animals. In addition, this Directive sets out four key criteria to choose between procedures with live animals: (a) use the minimum number of animals throughout the procedure; (b) involve animals with the lowest capacity to experience pain, suffering, distress, or lasting harm; and (c) cause the least pain, suffering, distress, or lasting harm to the animals. In any case the chosen method must be the most likely to provide satisfactory results. The Directive also recommends replacing death as the end point of a procedure by early and humane end points.

The European legislative acts (Regulations and Directives) also pay special attention to the application of good laboratory practice (GLP) in all ecotoxicological and toxicological assays, which are set out in Directive 2004/10/EC of the European Parliament and of the Council [29]. In spite of this, all Regulations on chemicals include the possibility to comply with other international standards recognized by the Commission.

3 European Strategy for a Nontoxic Environment

In 2013, the European Parliament and the Council adopted the 7th Environment Action Programme (7th EAP) to 2020 [30]—*Living well, within the limits of our Planet*, mandating the European Commission to implement and develop by 2018 a “Union strategy for a non-toxic environment that is conducive to innovation and the development of sustainable substitutes including non-chemical solutions.” In mid-2017, the final report of this mandate was published [31]. In it, the main gaps and deficits in the current status in policies and legislation were identified, proposing the need for a global framework additional to REACH Regulation [9] for protection of human health and the environment (i.e., minimizing the exposure) from harm provoked by hazardous chemicals. Some of these knowledge gaps and deficits are copied literally in Table 7.

Following a detailed analysis of the weaknesses or deficiencies identified, the report concluded that the efforts should be concentrated at improving the knowledge on chemicals; promoting

Table 7

Gaps and deficits identified in policies and legislation in the EU (DG Environment 2017)

Remaining gaps in knowledge on health and environment hazardous properties of chemical substances
Slow progress in identification of substances of very high concern (SVHC) and in substitution of hazardous chemicals in industrial processes and products
Lack of information concerning chemicals in articles, including imported articles, and the resulting exposure
Insufficient attention to hazardous chemicals in material flows important for a circular economy
Deficits in the framework for protection of children and other vulnerable groups, e.g., from chemicals in products such as textiles, electronics, and other consumer products
The still insufficient management of a number of aspects related to exposure and toxicity (sometimes termed “emerging issues”), such as combination effects; cumulative, low-dose, and long-term exposure; endocrine disruptors; neurotoxicity; protection of children and vulnerable groups; and chemicals in articles including in waste, material recycling, and the circular economy
Insufficient knowledge of the occurrence of chemical substances in the environment and technosphere, as well as the societal costs of the resulting exposure
Insufficient means to address risks posed by chemicals on the basis of persistence alone
Lack of monitoring of environmental compartments concerning possible buildups of chemical contamination and health and environmental risks thereof, in particular with respect to sources of water intended for human consumption
Need for better incentives for development of new, nontoxic substances as well as nonchemical solutions
Need for more comprehensive compilation of monitoring data at EU level and establishment of an early warning system

measures of substitution, especially of very persistent compounds; promoting innovation and development of nontoxic chemicals and materials in products and articles; and reducing chemical exposures as well as promoting circular economy. In addition, special attention should be paid to more efficiently protect vulnerable groups, mainly children, along with the development of early warning systems for detecting chemical threats to human health and the environment. The report also proposed, in risk assessment and in risk management, changing from the current chemical by chemical to groupings of chemical approaches.

The study of the DG for Environment [31], finally, proposes a type of hierarchy in chemical policy and management that may be explained with an inverted pyramid with six steps, starting in the upper level, with the principle of avoiding the production and use of substances of concern, followed by elimination of all unessential uses of substances of concern, including very persistent substances, thus minimizing the chemical exposure. The following steps seek to design non-/less-toxic chemicals and products, and, finally, the last step is for remediation measures to mitigate risks from legacy chemicals. This last phase includes technologies for decontamination of recycled materials, recovery of substances from wastes, and destruction of the hazardous substances.

4 Overview

The free movement of all types of chemical substances and products containing them, as well as enhancing competitiveness and innovation in EU, should be compatible with the highest standards of protection on the environment, animal health, and public health. For that, all provisions concerning them are based on three main sets of principles: precautionary, 3Rs in animal testing, and good laboratory practice principles.

Looking further ahead, the EU is working on a strategy toward a nontoxic environment that is conducive to innovation and the development of sustainable substitutes including nonchemical solutions.

References

1. European Union (2019). https://europa.eu/european-union/about-eu/countries_en. Accessed 10 Jun 2019
2. European Union (2019). https://europa.eu/european-union/about-eu/institutions-bodies_en. Accessed 10 Jun 2019
3. European Commission (2019). https://ec.europa.eu/info/law/law-making-process/types-eu-law_en. Accessed 10 Jun 2019
4. European Union Law (2019). <https://eur-lex.europa.eu/homepage.html?locale=en>. Accessed 10 Jun 2019
5. European Union Law (2019). <https://eur-lex.europa.eu/collection/eu-law/treaties/treaties-force.html>. Accessed 10 Jun 2019
6. European Parliament (2019). <http://www.epgenpro.europarl.europa.eu/static/ordinary-legislative-procedure/en/ordinary>

- <legislative-procedure/handbook-on-the-ordinary-legislative-procedure.html>. Accessed 10 Jun 2019
7. European Commission (2019). <https://ec.europa.eu/info/policies/environment/>. Accessed 10 Jun 2019
 8. European Commission (2019). http://ec.europa.eu/environment/chemicals/index_en.htm. Accessed 10 Jun 2019
 9. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council REGULATION (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. OJ L 3961, 1-849
 10. Regulation (EC) No 1272/2008 of The European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006. Official Journal of the European Union, 31/12/2008, L353/1-1355
 11. Regulation (EC) No 440/2008 of 30 May 2008 laying down test methods pursuant to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). Official Journal of the European Communities, 31/05/2008, L 142/1-734
 12. Regulation (EU) 2018/1881 of the Commission of 3 December 2018 amending Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards Annexes I, III, VI, VII, VIII, IX, X, XI, and XII to address nanoforms of Substances. Official Journal of the European Union, 4/12/2018, L308/1-20
 13. Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products. Official Journal of the European Union. L 167/1-123
 14. Directive 98/8/EC of The European Parliament and of the Council of 16 February 1998 concerning the placing of biocidal products on the market. Official Journal of the European Communities, 24.4.98, L/123-63
 15. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. Official Journal of the European Communities, 24/11/2009, L 309/1-50
 16. Communication from the Commission to the European Parliament, The Council and the European Economic and Social Committee. European Union strategic approach to pharmaceuticals in the environment. Brussels, 11.3.2019 COM(2019) 128 final, 12 pp
 17. Directive 2001/82/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to veterinary medicinal products. Official Journal of the European Communities, 28/11/2001, L 311/1-66
 18. Regulation (EU) 2019/6 of the European Parliament and of the Council of 11 December 2018 on veterinary medicinal products and repealing Directive 2001/82/EC. Official Journal of the European Union. 7/1/2019, L-4/43-163
 19. European Medicines Agency (2006) Guideline on the environmental risk assessment of medicinal products for human use. *Pre-authorisation evaluation of medicines for human use*. London, 01 June 2006. EMEA/CHMP/SWP/4447/00 corr 2. <https://www.ema.europa.eu/>. Accessed 10 Jun 2019
 20. European Medicines Agency (2004) Environmental risk assessments for medicinal products containing, or consisting of, genetically modified organisms (GMOs) (Module 1.6.2). *Evaluation of medicines for human use*. London, 20 January 2005. EMEA/CHMP/BWP/135148/2004. <https://www.ema.europa.eu/>. Accessed 10 Jun 2019
 21. European Chemicals Bureau (2008) Technical guidance document in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. <https://publications.europa.eu/en>
 22. OECD Guidelines for the Testing of Chemicals, Section 2 (Effects on biotic systems) and Section 3 (Environmental fate and behaviour). <http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm>. Accessed 10 Jun 2019

23. Directive 2010/63/EU of the European Parliament and the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European Union, 20/10/2010, L-276/33-79
24. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities, 22/12/2000, L 327/1-73
25. Decision No 2455/2001/EC of the European Parliament and of the Council of 20 November 2001 establishing the list of priority substances in the field of water policy and amending Directive 2000/60/EC. Official Journal of the European Communities. 15/12/2001, L-331/1-5
26. Directive 2008/105/EC of the European Parliament and of the Council of 16 December 2008 on environmental quality standards in the field of water policy, amending and subsequently repealing Council Directives 82/176/EEC, 83/513/EEC, 84/156/EEC, 84/491/EEC, 86/280/EEC and amending Directive 2000/60/EC of the European Parliament and of the Council. Official Journal of the European Communities, 24/12/2008, L 348/84-97
27. Directive 2006/118/EC of the European Parliament and of the Council of 12 December 2006 on the protection of groundwater against pollution and deterioration. Official Journal of the European Union, 27/12/2006, L-372/19-31
28. Directive 2013/39/EU of the European Parliament and of the Council of 12 August 2013, amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy. Official Journal of the European Communities, 24/8/2013, L 226/1-17
29. Directive 2004/10/EC of the European Parliament and of the Council of 11 February 2004 on the harmonisation of laws, regulations and administrative provisions relating to the application of the principles of good laboratory practice and the verification of their application for tests on chemical Substances. Official Journal of the European Union, 20/02/2004, L 50/44-59
30. Decision No 1386/2013/EU of the European Parliament and of the Council of 20 November 2013 on a General Union Environment Action Programme to 2020 ‘Living well, within the limits of our planet’. Official Journal of the European Union, 28/12/13, L-354/171-200
31. Directorate General for Environment (2017) Study for the strategy for a non-toxic environment of the 7th Environment Action Programme. Final Report. Written by Milieu Ltd, Ökopol, Risk & Policy Analysts (RPA) and RIVM. <http://ec.europa.eu/environment/chemicals/non-toxic/pdf/NTE%20main%20report%20final.pdf>



Chapter 2

A Brief Introduction to Quantitative Structure-Activity Relationships as Useful Tools in Predictive Ecotoxicology

Rahul Balasaheb Aher, Kabiruddin Khan, and Kunal Roy

Abstract

This introductory chapter highlights the applications of quantitative structure-activity relationships (QSARs) in the assessment of ecotoxicological risk posed by the chemicals used in our day-to-day life and in the industries. A wide variety of chemicals (industrial substances/toxicants/pollutants) are emitted into the environment from various sources. These chemicals may be pharmaceuticals, personal care products, nanomaterials, plasticizers, flame retardants, endocrine disruptors, pesticides, persistent organic pollutants (POPs), etc. The continuous emissions of chemicals into the environment and the resultant pollution effects and potential exposure of living organisms and humans to these noxious substances may pose a risk to the ecosystem and human health. The experimental determination of toxicities of these chemicals involving different aquatic organisms and laboratory animals is a lengthy, time-consuming, and costly process. In this scenario, QSAR is quite useful for the prediction of toxicities of these chemicals prior to their use on a large scale. QSAR models could also be used further to predict the toxicity of any designed chemicals and would thus be helpful for green chemical design.

Key words QSAR, Pollutants, Toxicants, Ecotoxicity, Aquatic organisms, Contaminants of emerging concern (COEC), Data gap, Read-across

1 Introduction

The industrial chemicals have become an essential part of human life due to their applications in different facets. The major domains where these chemicals play a crucial role include health care, veterinary medicine, agriculture, research, and day-to-day utilities. Exponential rises in the demand of industrial chemicals have served as a major source for contamination of the surroundings (air, water, and land). The slow release of these chemical entities into the environment is attributed to their persistence, bioaccumulation, and toxicity (PBT) behaviors [1]. However, the major concern arises when these chemicals prove to have potential to behave like chronic accumulator slowly leading to deformities of body organs or internal body functions in the living systems mainly due to hormonal

imbalances. Substances like engineered nanoparticles are proved to have carcinogenic effect in several organisms [2]. The progressive accumulation of synthetic chemicals mainly in water bodies is attributed to overexploitation, irrational use, and improper pre-treatment by sewage treatment plants prior to release in river bodies. Although the attentions paid to study ecotoxicity profile of these contaminants have increased several folds in the recent decades, we cannot manage experimental toxicity determinations of all chemicals against a huge number of endpoints. However, one can rely upon intelligent in silico tools which can be utilized in data gap filling of large number of chemicals using a small number of experimental data. The computational tools employed in ecotoxicity of environmental contaminants include quantitative structure-activity relationship (QSAR), toxicophore modeling, and related approaches [3]. This introductory book chapter focuses on QSAR and its applications in ecotoxicological studies of environmental pollutants. The QSAR approach has several merits as this can be used in predicting responses of unknown and untested chemicals utilizing a limited data, thus proving to be time- and cost-effective. It also serves as a tool or medium to protect animal's lives in the laboratory thus being acceptable from the ethical point of view [4]. Due to these encouraging features, QSAR is recommended for risk assessment of chemicals by various regulatory agencies like the US Environmental Protection Agency (US EPA); Agency for Toxic Substances and Disease Registry (ATSDR); European Centre for the Validation of Alternative Methods (ECVAM) of the European Union; European Union Commission's Scientific Committee on Toxicity, Ecotoxicity, and Environment (CSTEE); etc. [5].

2 Definition and Constituents of QSAR

2.1 What Is QSAR?

QSAR modeling is a statistical approach correlating the structural information of chemicals (including pharmaceuticals, cosmetics, agrochemicals, ionic liquids, nanomaterials, etc.) with endpoints/response values (activity/property/toxicity) using chemometrical techniques. The chemical information corresponds to the chemical domain space, which is derived in terms of descriptors (independent variables) using different software tools. The endpoint is the dependent variable obtained from an experiment, which is modeled using the following general equation:

$$\begin{aligned} \text{Endpoint (Activity/toxicity/property)} \\ = f(\text{chemical structure}) \end{aligned} \quad (1)$$

Equation 1 can be used to determine the biological activity, toxicity, property, etc. in a quantitative manner. Depending upon

the type of endpoint modeled, the modeling technique may be named as quantitative structure-activity relationship (QSAR), quantitative structure-toxicity relationship (QSTR), quantitative structure-property relationship (QSPR), etc. Sometimes, the activity or toxicity terms themselves may be used as additional descriptors as in case of quantitative structure-activity-activity relationship (QSAAR), quantitative structure toxicity-toxicity relationship (QSTTR), etc. Here, the current chapter has been written by considering only the QSAR/QSTR aspects of modeling.

The selection of data (chemical/biological) is one of the primary steps for the success of any cheminformatic study, including QSAR. For the development of QSAR models, two different types of data information are required, i.e., biological data (endpoints) and chemical data in terms of molecular descriptors. The feature selection method is then used to select appropriate number of meaningful and informative descriptors before applying any modeling algorithm for model development. The flowchart for the development of QSAR/QSTR model is given in Fig. 1.

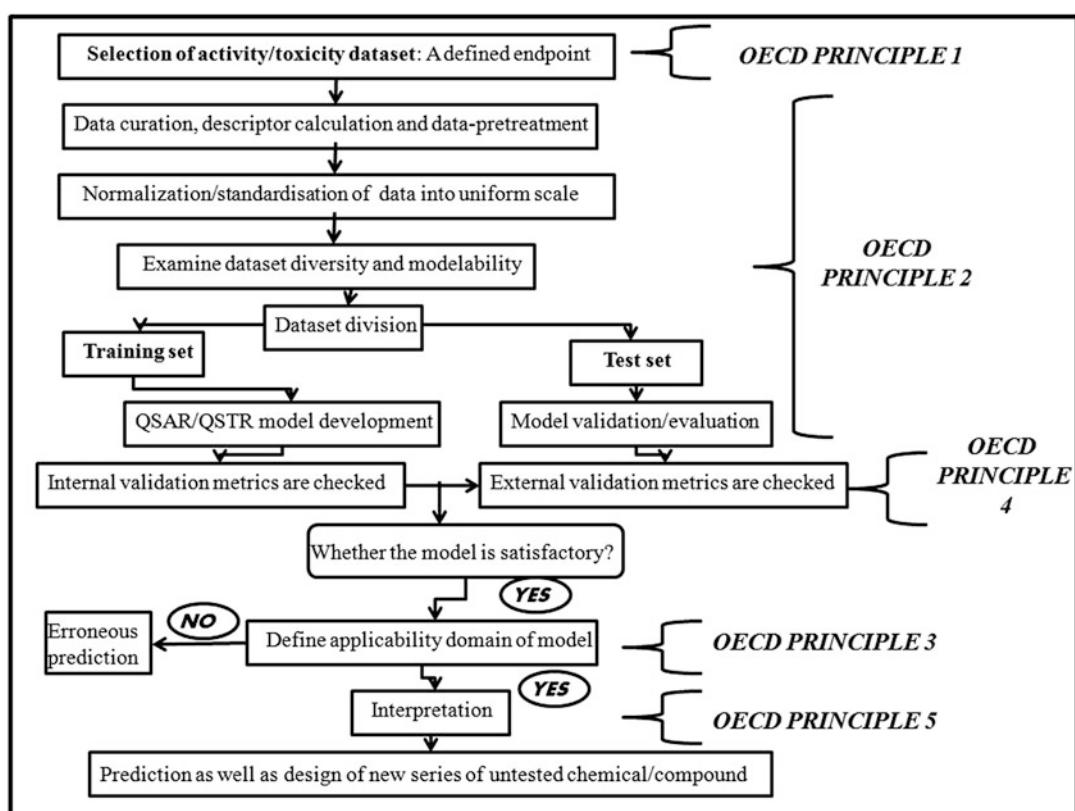


Fig. 1 Flowchart for development of a QSAR/QSTR model

2.1.1 Biological Data

The biological data are experimental endpoints, which may be activity/toxicity values (IC_{50} , EC_{50} , LD_{50} , etc.). The activity/toxicity values are obtained from the dose-response curve. The concentration values such as IC_{50} , EC_{50} , and LD_{50} are inhibitory, effective, and lethal concentration which inhibit/effect/kill 50% of the test population. The concentration values are expressed in a molar unit and then converted to negative logarithmic scale, so that a higher value in the positive scale represents higher activity and vice versa. The following points are to be considered while selecting any biological data for modeling:

1. As per Organisation for Economic Co-operation and Development (OECD) guideline no. 1 [6], the QSAR model should have a defined endpoint. There should be transparency in the endpoint being predicted by the model. The compounds having well-defined activity should only be taken for the model development. Another important point is that the compounds used for the model development must have same mode/mechanism of action.
2. The range of the endpoint values should be at least logarithmic scale of 3–4 units. The conventional metrics of validation (R^2 , Q^2 , r^2_{pred}) are response range-based parameters [7]; hence, care should be taken that the data points are uniform and there should not be any gaps.
3. All the compounds should be tested using same assay protocol under same experimental conditions.
4. The dataset is to be properly curated prior to the use. One should also be careful for considering the data points with activity cliff information [8].
5. One should be careful while selecting/working with small datasets. The issue of working with small data points is really a big problem in QSAR modeling. The scarcity of data points is due to unavailability of the experimental observation such activities/toxicities or properties of nanomaterials [9].

2.1.2 Chemical Data/Descriptors

The descriptors give information relevant to the chemical space, which is considered during model development to find the correlation between activity values and molecular descriptors. The following points are to be considered while calculating or selecting the descriptors:

1. All the chemical structures should be thoroughly checked and curated to remove any possible error before calculation of descriptors
2. 3D descriptors should be taken into consideration while considering the enantiomeric form of compounds. For 3D

descriptors, conformational aspects and energy minimization should be considered.

3. Compounds in racemic and salt forms should not be used for descriptor calculation. Any salt form is to be converted to the corresponding acidic or basic form prior to the calculation.
4. The descriptor pool is to be thinned or reduced by considering the criteria of variance or correlation coefficient cutoff values to remove constant and intercorrelated descriptors.

2.1.3 Molecular Descriptors: Different Types

According to Todeschini and Consonni [10], a chemical descriptor is defined as “the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.” A key step in classical QSAR/QSTR modeling is the encoding of a chemical compound into a vector of numerical descriptors. Different descriptors used for the QSAR study can be classified into 2D and 3D classes as shown in Table 1.

Table 1
Description of 2D and 3D descriptors for the QSAR study

Dimension of descriptors	Class of the descriptors	Representative examples
2D	Topological	Balaban index, kappa shape index, molecular connectivity index, subgraph count, Chi indices, Wiener, Zagreb, electrotopological
	Structural	Molecular weight, number of rotatable bonds, H-bond acceptor, H-bond donor, chiral centers
	Physicochemical parameters	LogP, ALogP, ALogP98, AlogP_atypes, MolRef
	Extended topochemical atom (ETA) indices	First- and second-generation ETA indices
	Constitutional indices	Number of atoms, number of non-H atoms, number of bonds, number of aromatic bonds, sum of atomic van der Waals volumes (scaled on carbon atom), etc.
	Functional indices	Number of terminal primary C(sp ₃), number of total secondary C(sp ₃), number of ring secondary C(sp ₃), number of unsubstituted benzene C(sp ₂), number of isocyanates (aliphatic), etc.
3D	Electronic	Dipole, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), superdelocalizability
	Spatial	Radius of gyration, Jurs descriptors, area, density, volume, PMI _{mag}

2.1.4 Division of the Dataset

In order to ascertain the performance of a predictive model, the whole dataset is to be divided into a training set and a test set based on the chemical similarity. The training set is employed for developing the model (i.e., the equation), while the test set (not used during model development) is used to judge the external predictivity of the model. Since the training set is employed for building the model, usually a higher number of compounds are allotted to it as compared to the test set. The total dataset is divided such that the test set compounds lie within the chemical space of the training set, i.e., the training set becomes representative of the test set. The methods of the dataset division may involve (a) Euclidean distance (diversity-based) [11], (b) Kennard-Stone [12], (c) k -means clustering [13], (d) sorted response [14], etc.

2.1.5 Feature Selection

Descriptor selection is one of the most important steps while selecting the meaningful descriptors from a large pool of descriptor set. One cannot use the entire descriptor pool for modeling, since it is computationally expensive and time-consuming. For regression-based modeling, methods like stepwise selection, genetic method, factor analysis, etc. are to be used for selecting an appropriate number of descriptors [15].

2.1.6 Modeling Algorithms and Chemometric Tools Used in QSAR

As per the OECD guideline no. 2 [6], the model should be developed by using an unambiguous algorithm, i.e., there should be transparency in the description of the modeling algorithm. This includes the formalisms implemented during data pretreatment, dataset division, selection of features, and model development. The commonly employed linear modeling algorithms suggested by the OECD include univariate linear regression (ULR), multiple linear regression (MLR) [16], ordinary least squares (OLS), partial least squares (PLS), principal component analysis (PCA) [17], principal component regression (PCR), etc. The OECD guideline also suggests performing a priori feature selection using mechanistic basis or an evolutionary technique, e.g., genetic algorithm (GA), as well as techniques such as principal component analysis (PCA) or factor analysis (FA), etc. [15].

Different model building tools used in QSAR can simply be grouped into three broad categories, namely, regression-based approach, classification-based approach, and machine learning [18]. The regression-based approach is applied when quantitative (continuous) dependent (response) and independent (descriptors) variable values are available (as seen in case of multiple linear regression). The classification-based approach is for graded response data where response is available in a Boolean form such as active/inactive and positive/negative (as seen in linear discriminant analysis, logistic regression, and cluster analysis). The machine learning approach does not follow explicit programmed instructions and thus constructs and develops its own learning based on

data provided (as seen in case of artificial neural network, Bayesian neural network, decision tree, and random forest protocol).

2.1.7 Checking Domain of Applicability of Developed Models

A QSAR model developed using a set of chemicals possesses a specific theoretical space, and it is considered to provide reliable predictive result within that domain [19]. Thus, the determination of applicability domain of a model using the training set molecules is necessary to check whether the prediction of test set molecules is trustworthy or not (as per OECD guideline no. 3) [6]. The applicability domain of a model depends on three major attributes, (a) structural information, (b) physicochemical features, and (c) response space. Because of the possible involvement of multiple mechanistic basis in various regulatory endpoints, QSAR models can be developed on specific chemical classes acting via same mechanism of action. Sometimes, a single general QSAR model might be unable to distinguish chemical classes and thereby might not provide definite estimate with respect to a specific chemical class. In order to achieve global applicability, the OECD guideline suggests the development of (a) multiple predictive models for the same endpoint on different domain of applicability combined to give a global estimation or (b) use of statistical method giving global modeling attribute across multiple mechanism of actions with respect to a same endpoint. By the suitable use of the domain of application of available QSAR models, one can identify the data gaps by comparing the domain of chemicals with respect to each defined regulatory endpoint.

2.1.8 Validation of QSAR Models

Once the model is developed, it is necessary to check whether the model is statistically significant or not [20]. The fourth principle of the proposed OECD guideline [6] emphasizes statistical validation of models in terms of goodness-of-fit, robustness, and predictivity. The predictive quality of a developed QSAR model can be statistically characterized by computing several quality parameters as well as validation metrics. The aim of different parameters is to judge the accuracy of prediction, i.e., determination of closeness between the experimental and model derived predicted value. The model fitness can be determined by computing metrics such as coefficient of determination or correlation coefficient using the training set, i.e., the set used for model development. This parameter portrays the extent of achieved correlation between the experimental and predicted response value, while robustness or stability can be determined by introducing a perturbation into the model, e.g., by deletion of samples from the training set and redeveloping the model. The external predictivity refers to the predictive quality determined using test set chemicals which were not employed during the model development. The internal validation metrics included correlation coefficient (R^2), adjusted R^2 (R_{adj}^2), leave-one-out (Q^2), etc., whereas the external validation metrics included

R^2_{pred} , etc. Additional metrics such as $\overline{r_m^2}$, Δr_m^2 for the training, test and overall sets [21], and mean absolute error (MAE)-based criteria can also be used to check quality and performance of the developed models [7].

2.1.9 Mechanistic Interpretation of the QSAR Model

The fifth OECD principle is associated with a “mechanistic interpretation,” wherever such an interpretation can be made. Clearly, it is not always feasible to provide a mechanistic interpretation of a given QSAR model, and thus the principle suggests the modeler to report if any such information is available facilitating the future research on that endpoint. The specific information on the mechanism of action of chemicals toward a process can guide the design and development of desired analogues [22].

2.1.10 A Few Important Issues in QSAR

There are several prerequisites for the preparation of experimental data ready for QSAR/QSTR analysis [18]. Usually the concentrations or doses required for a fixed response such as EC₅₀, ED₅₀, IC₅₀, or LD₅₀ values are used as the response for activity- or toxicity-based QSAR analyses. The concentration values should be expressed in a molar unit and in a negative logarithmic scale so that a higher value represents higher activity or toxicity. There should be a good degree of freedom to ascertain statistical soundness of a QSAR model. Therefore, the number of observations based on which a model is developed should be considerably high with respect to the number of descriptors (constraints) used in the modeling. Although this aspect is less important for more robust techniques, the use of sufficient number of training compounds cannot be ignored even in case of machine learning techniques. QSAR researchers frequently experience the problem of modeling small datasets, as for several endpoints, sufficient number of experimental observations might be unavailable. Multiple linear regression (MLR) is a commonly used method for activity- and toxicity-based classical-type QSARs, while it presents several problems like intercorrelation among descriptors, bias in descriptor selection due to a fixed composition of the training set, inability to handle many descriptors in the model, etc. This problem may be overcome by using a more robust modeling technique like partial least squares (PLS) [23], which converts the original set of descriptors into a lower number of latent variables (LVs) which are functions of the original descriptors. The dataset with a small number of data points needs a special attention during modeling. A double cross-validation technique [24, 25] may be of help in such cases. In this approach, the validation is done in two loops: in the inner loop, the training set is further divided into “n” calibration and validation sets resulting in diverse compositions, which are further utilized for model building and model selection, while the test set in the external loop is exclusively used for the model assessment. In

another approach, consensus predictions have been applied in several studies as more reliable than individual model derived predictions, as the former takes into account contribution of maximum possible combination of important descriptors. This approach can also afford greater chemical space coverage. Recently, an intelligent consensus modeling method has been reported considering that a single QSAR model may not be equally good for predictions for all query compounds [26]. It is also important to evaluate the reliability of predictions [27, 28] for untested compounds, which may not be dependent solely on applicability domain.

Those readers who wish to learn in detail about the basics of QSAR are encouraged to refer a relevant published literature [18] as the current introductory chapter emphasizes more on application of QSAR in ecotoxicity predictions.

2.2 Ecotoxicity Predictions

Ecotoxicity deals with the ability of chemical, biological, or physical stressors to have an adverse effect on the environment and the organisms living in it, such as fish, wildlife, insects, plants, and microorganisms. Such stressors might occur in the natural environment at densities, concentrations, or levels high enough to disrupt the natural biochemistry, physiology, behavior, and interactions of the living organisms that comprise the ecosystem (<https://en.wikipedia.org/wiki/Ecotoxicity>). The various forms of environmental toxicity include aquatic toxicity, developmental toxicity, carcinogenicity and genotoxicity, and toxicity to human and environment due to drugs, cosmetics, biological products, hazardous chemicals [29], food, and agrochemicals [30–32]. All these forms of toxicities will prove fatal to humans and every other form of life if not controlled at the beginning stage.

The toxicity assessment of environmental pollutants or contaminants of emerging concern (CECs) is a daunting task that involves multiple testing conditions and endpoints. The lack of experimental data, and gaps in the existing data points, has created an opportunity to the computational chemists to fill the missing data points by performing the modeling studies.

3 Why QSAR in Ecotoxicity Predictions: Can It Really Reduce Animal Experimentation?

The aroused concern of regulatory/monitoring bodies in implementation of QSAR in ecotoxicity of various environmental pollutants is already discussed above. In addition, the idea of developing alternative testing strategies (ATS) also known as intelligent testing strategies (ITS) or risk assessment strategies (RAS) has become a central point of extensive discussion in the last 5 years through many scientific research projects. The common environmental

pollutants include diethyl phthalate, bisphenol A (BPA), pharmaceuticals (climbazole), pesticides, cleaning products, laundry detergents, fabric softeners, fragrance chemicals, oven cleaners, disinfectants, phosphates, plasticizers, oil spills, etc. The accumulations of these chemicals in the environmental bodies are enough to disrupt the natural flora (plant life) and fauna (animals) of the environment.

QSARs are the potential tools for predicting the properties or toxicity of chemicals including their physicochemical attributes, health effects, and ecotoxicity. QSAR models thus are used to categorize chemicals in terms of their potentially hazardous nature. The prediction of toxicity by QSAR does not require lengthy and costly experiments involving the use of plants or animals. Hence, the QSAR models can be utilized for the assessment of new and existing chemicals in conformity with regulatory requirements within the scope of Organisation for Economic Co-operation and Development (OECD).

Researchers can use the developed QSAR models comprising the information of known chemical substances/toxicants/pollutants (training set compounds) to predict/classify the activity/property/toxicity of new substance. The new substances may be pesticides, solvents, pharmaceuticals, industrial chemicals, or class of persistent organic pollutants (POPs). POPs are organic compounds that are resistant to environmental degradation through chemical, biological, and photolytic processes. Examples of some of the POPs, which are present on the Stockholm Convention list, are given in Table 2 (https://en.wikipedia.org/wiki/Persistent_organic_pollutant). Because of their persistence, POPs bioaccumulate and have potential adverse impacts on human health and the environment. The prior information of these types of compounds that are predicted as hazardous earlier prior to their use will enable the researchers either to skip the use of that new substance or find some eco-friendly alternatives, which are safe and less hazardous. Compounds predicted as POPs by the QSAR models will be immediately discarded, which will prevent the bioaccumulation, persistence, and toxicity of these chemicals to the humans and aquatic life. This is a cost-effective and efficient way to eliminate potentially toxic substances and focus on those that appear not to be harmful. This way we can avoid spending time on animal testing of the substances which have been predicted to be harmful. Hence, the QSAR models will remain a good alternative to animal testing today and in future as well (<http://scienzenordic.com/can-we-avoid-animal-testing-entirely>).

Moreover, the use of expert system for the toxicity screenings of chemicals and pharmaceuticals further reduces the expenditure and avoid the sacrifice of a large number of animals. The expert systems are enriched with the broader information of structural and activity regions in comparison with the local QSAR models. Thus,

Table 2
Name of POPs present on the Stockholm convention list

S. no.	POPs	S. no.	POPs
1	Aldrin	12	Polychlorinated dibenzofurans
2	Chlordane	13	Chlordecone
3	Dieldrin	14	α -Hexachlorocyclohexane (α -HCH) and β -hexachlorocyclohexane (β -HCH)
4	Endrin	15	Hexabromodiphenyl ether (hexaBDE) and heptabromodiphenyl ether (heptaBDE)
5	Heptachlor	16	Lindane
6	Hexachlorobenzene (HCB)	17	Pentachlorobenzene (PeCB)
7	Mirex	18	Tetrabromodiphenyl ether
8	Toxaphene	19	Perfluorooctanesulfonic acid (PFOS)
9	Polychlorinated biphenyls (PCBs)	20	Endosulfans
10	Dichlorodiphenyltrichloroethane (DDT)	21	Hexabromocyclododecane
11	Dioxins		

they have advantages over the use of traditional QSAR models in toxicity prediction.

4 Ecotoxicological Data Sources, Expert Systems, and Freely Available QSAR Tools

The experimental data for the ecotoxicological modeling could be collected from different scientific journals or web-based databases depending upon the toxicity endpoints. Some of the publicly available databases which contain the information of toxicity data are given in Table 3.

Expert systems are knowledge-based computer prediction systems available to predict the endpoints related to chemicals and pharmaceutical toxicity. The expert system also provides structural alerts to identify fragments mediating different toxicities. There is a continuous need to enrich the existing expert systems with the plethora information of local models and to develop new expert systems in future for the ease of toxicity screenings of chemicals and pharmaceuticals in less time. The different freely available and commercial expert systems to predict the endpoints of chemicals and pharmaceutical toxicity are given in Table 4.

A software tool is one of the major components of the study required for performing different tasks involved in the QSAR modeling. Different steps where the software is required are

Table 3
Freely available databases containing information on the human, animal, and environmental toxicity

Types of toxicity	S. no.	Name of databases	Websites
Pesticide-induced toxicity	1	EXTOXNET	http://extoxnet.orst.edu/ghindex.html
	2	NPIC	http://npic.orst.edu/
	3	PAN pesticide	http://www.pesticideinfo.org/
	4	ToxRefDB	https://catalog.data.gov/dataset/toxcast-toxrefdb
Aquatic toxicity	5	ECOTOX	http://cfpub.epa.gov/ecotox/
	6	ESIS	https://old.datahub.io/dataset/esis
	7	TEXTRATOX	https://vetmed.tennessee.edu/Pages/utcvm_home.aspx
	8	TOXNET	https://toxnet.nlm.nih.gov/
	9	USGS	https://www2.usgs.gov/science/cite-view.php?cite=1336
	10	OECD HPV database	https://hpvchemicals.oecd.org/ui/Default.aspx
	11	N-class database, KemI	https://www.kemi.se/en/prio-start/before-starting/contents-of-the-database/substance-names-and-synonyms
	12	Riskline, KemI	http://www.inchem.org/pages/kemi.html
Carcinogenesis and genotoxicity	13	Cal/EPA	https://oehha.ca.gov/chemicals
	14	CCRIS	https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS
	15	CPDB	https://toxnet.nlm.nih.gov/cpdb/
	16	GENE-TOX	https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX
	17	IARC Monograph	http://monographs.iarc.fr/
	18	ISSCAN	http://alttox.org/resource-center/databases/
	19	LAZAR	http://www.in-silico.de/
	20	Oncology Tools	https://www.accessdata.fda.gov/scripts/cder/onctools/animalquery.cfm
	21	rita	https://reni.item.fraunhofer.de/reni/public/rita/
Developmental toxicity	22	BDSM	https://kundoc.com/pdf-data-input-module-for-birth-defects-systems-manager-.html
	23	DevTox	https://www.devtox.org/index_en.php

(continued)

Table 3
(continued)

Types of toxicity	Name of S. no. databases	Websites
Industrial chemical-induced toxicity to human and environment	24 ACToR 25 CEBS	https://actor.epa.gov/actor/home.xhtml https://tools.niehs.nih.gov/cebs3/views/index.cfm?action=main.dataReview&bin_id=2781
	26 Danish (Q)SAR Database	http://qsar.food.dtu.dk/
	27 DSSTox	https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database
	28 HERA	https://www.heraproject.com/RiskAssessment.cfm
	29 Household Products Database	https://hpdb.nlm.nih.gov/
	30 IRIS	https://www.epa.gov/iris
	31 ITER	https://toxnet.nlm.nih.gov/newtoxnet/iter.htm
	32 JECDB	http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp
	33 JRC QSAR Database	https://qsardb.jrc.ec.europa.eu/qmrf/
	34 MRL	https://www.atsdr.cdc.gov/mrls/index.html
	35 NTP	https://ntp.niehs.nih.gov/
	36 RAIS	https://rais.ornl.gov/
	37 SCOGS	https://www.fda.gov/food/ingredientspackaginglabeling/gras/scogs/default.htm
	38 STITCH	http://stitch.embl.de/
	39 Toxtree	http://toxtree.sourceforge.net/
Drug, biological products, food, and agrochemical-induced toxicity	40 AERS 41 CEDI/ADI Database	https://healthdata.gov/dataset/adverse-event-reporting-system-aers https://www.fda.gov/food/ingredientspackaginglabeling/fcs/cedi/default.htm
	42 CERES	https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCCT&dirEntryId=231472
	43 DITOP 44 Drugs@FDA	https://omictools.com/ditop-tool https://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
	45 EAFUS	https://www.fda.gov/food/ingredientspackaginglabeling/foodadditivesingredients/ucm115326.htm
	46 FDA Poisonous Plant Database	https://www.accessdata.fda.gov/scripts/plantox/index.cfm
	47 MRTD	http://mediformatica.com/index.php?option=com_content&view=article&id=540&Itemid=9

Table 4
Freely available and commercial expert systems to predict endpoints related toxicity predictions

S. no.	Expert system	Manufacturer and website
1	<i>ASTER</i> (ASsessment Tools for the Evaluation of Risk)	https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NHEERL&dirEntryID=74887
2	<i>CAESAR</i> (Computer Assisted Evaluation of industrial chemical Substances According to Regulations)	http://www.caesar-project.eu/
3	<i>DEREK</i> (Deductive Estimation of Risk from Existing Knowledge)	Lhasa Ltd. https://www.lhasalimited.org/?cat=2&sub_cat=64
4	<i>ECOSAR</i> (Ecological Structure Activity Relationships)	https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model
5	<i>HazardExpert Pro</i>	CompuDrug Inc. https://www.compudrug.com/
6	<i>MCASE/ MC4PC</i>	http://www.multicase.com/products/
7	<i>OASIS&TIMES</i>	http://oasis-lmc.org/products/software/times.aspx
8	<i>OECD (Q)SAR</i> Application Toolbox	http://www.oecd.org/env/ehs/risk-assessment/oecdquantitativestructure-activityrelationshipsprojectqsars.htm
9	<i>ONCOLOGIC</i>	https://www.epa.gov/tsca-screening-tools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals
10	<i>OSIRIS</i> property explorer	Organic Chemistry Portal http://www.organicchemistry.org/prog/peo/tox.html
11	<i>SARET</i> (Structure-Activity Relationships for Environmental Toxicology)	https://www.tandfonline.com/doi/full/10.1080/10590500802135578?src=reccsys
12	<i>TERA</i> (Tools for Environmental Risk Assessment)	TERAbase is a part of expert system created by Prof. S.M. Novikov and co-authors from the A.N. Sysin Research Institute of Human Ecology and Environmental Health of Russian Academy of Medical Sciences
13	<i>TerraQSTR- FHM</i>	http://www.terrabase-inc.com
14	<i>TIMES-SS</i> Times MEtabolism Simulator platform	Marketed by LMC University “As Zlatarov,” Bourgas, Bulgaria
15	<i>TOPKAT</i> (TOxicity Prediction by C(K)omputer A ssisted Technology)	Accelrys Inc. http://www.3dsbiovia.com/products/topkat/

Table 5
Freely available software tools for QSAR modelling

S. no.	Software	Developer/institute	Platform	Website
1	ACD/ChemSketch	ACD labs, Toronto, Ontario, Canada	Windows	https://www.acdlabs.com/resources/freeware/chemsketch/
2	CORAL-QSAR/QSPR	Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy	Windows	http://www.insilico.eu/coral/
3	DTC-lab tools	DTC-lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India	Windows	http://teqip.jdvu.ac.in/QSAR_Tools/
4	QSARINS	University of Insubria, Department of Theoretical and Applied Sciences, Italy	Windows	http://www.qsar.it/
5	PaDEL-Descriptor	Department of Pharmacy, Pharmaceutical Data Exploration Laboratory, National University of Singapore	Windows	http://www.yapcwsoft.com/dd/padeldescriptor/
6	PBT profiler	United States Environmental Protection Agency	Windows	http://www.pbtprofiler.net/

normalization/standardization of data, data pretreatment and curation, dataset division, model development and validation, determining applicability domain, etc. A freely available software tool enables any researchers to run, copy, and distribute the tool anywhere in the scientific community. We have enlisted some of the freely available QSAR software tools in Table 5.

5 Applications of QSAR in Ecotoxicological QSAR Studies

The application of QSAR modeling is increasing in the areas of pharmaceuticals (drug design, predictive toxicology) [33], chemicals (ionic liquids [34, 35], agrochemicals [31], persistent organic pollutants (POPs), reaction optimization, etc.), nanotechnology [36–38], material sciences, cosmetics [39], and food sciences and also in regulatory field. Here, we have discussed in brief selectively the application of QSAR modeling in toxicity prediction of ionic liquids (ILs), nanomaterials, and contaminants of emerging concern (CECs) [40]. Ionic liquids are being considered as green replacements (“environmental friendly”) for industrial volatile organic compounds. But still the solubility of ILs in water and a number of literature documenting toxicity of ILs to aquatic organisms highlight a real cause for concern. More and more

nanomaterials are expected to be used in consumer products. This is expected to lead to an increased human exposure to nanomaterials in their daily lives. Therefore, the effect of nanomaterials present in human environment is an area of increasing scientific interest. Chemicals of emerging concern (also called “contaminants of emerging concern” or “CECs”) can include nanoparticles, pharmaceuticals, personal care products, estrogen-like compounds, flame retardants, detergents, and some industrial chemicals with potential significant impact on human health and aquatic life. These three classes of chemicals are widely being used on a larger scale in chemical industries, and continuous discharge of these chemicals into the environment poses a major serious health risk to humans, animals, and other forms of life.

5.1 QSAR Models for Toxicity of Ionic Liquids

The ionic liquids are gaining attention as future “green solvents” within scientific and industrial community. They are designed to be inflammable, nonvolatile, with lower vapor pressure, nontoxic, and nonexplosive media with a high thermal stability. Due to lower vapor pressure, they are not expected to accumulate in the atmospheric environment. But, due to their high aqueous solubility, they may contribute to the aquatic pollution or toxicity.

Despite having these desirable physicochemical properties, ionic liquids must have been tested for their ecotoxicity before commercialization. As more combination of ILs could be developed, the experimental toxicity prediction of every liquid is laborious, costly, and time-consuming. This necessitates the development of predictive models as an alternative to replace costly and laborious manual toxicity measurements.

Numerous studies have reported the toxicity of diverse IL subfamilies against different organisms including green algae, *Vibrio fischeri* (*V. fischeri*), gram-positive and gram-negative bacteria, and fish. These studies suggest that the ionic liquids cannot be fully considered as the greener class of chemicals, and it is important to predict its toxicities to the environment prior to their production and usage on the industrial scale [41]. We have discussed below some of the recent applications of QSAR models in ecotoxicity prediction of the ionic liquids.

A diverse set of 269 ILs containing 9 cationic cores and 44 types of anions were used to develop 3D-QSTR models. The models were developed to make a correlation between the structural information of the ionic liquids (ILs) and their cytotoxicity toward leukemia rat cell line IPC-81 using partial least squares (PLS) and support vector regression (SVR) methods. Genetic algorithm (GA) was used to select the best and interpretative subset of variables for the predictive model building. These models can reduce the amount of cellular testing necessary by predicting the toxicological functions of the chemical structures [42].

A QSAR model in accordance with the OECD guidelines was developed using a larger dataset of 305 ionic liquids and their

ecotoxicity values based on the luminescence inhibition of *V. fischeri* bacterium species. *V. fischeri* is a gram-negative, rod-shaped bacterium and considered as an important member in the marine ecosystem. It can be easily applied as a test organism for assessment of toxicity of chemicals in the aquatic systems. The models were developed using topological and quantum chemical descriptors. The developed QSAR models using PLS method were validated experimentally, by designing a low predicted toxicity ILs, subsequently synthesized and experimentally tested their toxicity against *V. fischeri*. These models could be used to design and prepare new ionic liquids with reduced toxicity profile [43].

Das et al. (2005) developed the interspecies quantitative structure-toxicity-toxicity relationship (QSTTR) models of ionic liquids which allow the extrapolation of data when the toxicity data toward one organism are absent. They utilized the toxicity data of ionic liquids toward three aquatic organisms, viz., a bacterium (*V. fischeri*), a cladoceran (*Daphnia magna*), and a green alga (*S. vacuolatus*). These models could be used to fill the data gaps and aid future studies on assessment of hazard of ILs [44].

Imidazolium-based ionic liquids containing different functionalized and unsaturated side chains were evaluated for cytotoxicity toward the channel catfish ovary (CCO) cell line. The experimental cytotoxicity data of 14 different imidazolium ionic liquids in CCO cells, with EC₅₀ values, were used to develop quantitative structure-toxicity relationship (QSTR) models using regression- and classification-based approaches. It was observed that the toxicity of ILs toward CCO was chiefly related to the shape and hydrophobicity parameters of cations. These models could be utilized to predict the environment risk assessment of new ILs before production and application at the industrial scale [45].

Ghanem et al. 2016 developed linear and nonlinear QSAR models using a diverse set of 110 ILs comprising a combination of 49 cations and 29 anions along with their ecotoxicity data against bioluminescent bacterium *V. fischeri*. The model was developed using σ -profile descriptor and multiple regression method. The selected descriptor set from the linear model was then used in high multilayer perceptron (MLP) technique to develop the non-linear model. These models can be used as the primary step for screening and designing inherently safer ILs [41].

5.2 QSAR Models for Nanomaterial Toxicity (Nano-QSAR)

The use of nanoparticles such as metal oxide NPs, C60 NPs (fullerene), etc. has grown exponentially during this decade due to their extraordinary properties, covering a wide range of products in the optoelectronics, pharmaceutical, medical, cosmetics and sunscreens, solar batteries, space technology, environmental engineering, self-cleaning windows, textile industries, and so on. However, the risk to human health (cytotoxic, mutagenic or carcinogenic effects) and environment for most of them is still not well established.

Alternative routes for risk assessment based on in silico methods avoid the highly expensive and time-consuming toxic evaluation of nanoparticles (NPs) in the laboratory. Lack of sufficient data and low adequacy of experimental protocols hinder comprehensive risk assessment of nanoparticles (NPs). QSAR is one of the chemometric methods which correlates the physicochemical properties of nanomaterials with their cytotoxic, mutagenic, or carcinogenic potential, and the developed mathematical models could be very efficiently utilized to predict their toxicity in an efficient manner. We have discussed below some of the recent applications of QSAR models in toxicity prediction of the NPs.

“Quantitative conditions-property/activity relationships” (QCPR/QCARs) models of fullerene C60 NPs were developed using an experimental data of two endpoints. The endpoint-1 is the bacterial reverse mutation test (Ames) that was conducted using *Salmonella typhimurium* strains TA100; and the endpoint-2 is mutagenic effect of fullerene for *E. coli* strain WP2. The regression models were developed by means of optimal descriptors calculated with the Monte Carlo method by using CORAL software. These models could be utilized to predict the mutagenic potential of fullerene nanoparticles under different conditions [46].

Multi-target quantitative structure-toxicity relationship (mt-QSTR) models were developed for diverse metal oxide NPs using a multiple (four different) toxicity endpoints based on the experimental cytotoxicity data in *E. coli* (under dark-induced and photoinduced conditions) and in human keratinocyte (HaCaT) cell line following the OECD guidelines. The models were constructed for an individual toxicity prediction, and an mt-QSTR model was developed for simultaneous prediction of the multiple toxicity endpoints. These models would largely help to reduce the cost and computational efforts in generating information on the toxicities of new NPs for their safety evaluation [47].

Kar et al. (2014) developed the QSTR models using simple periodic table-based descriptors (cost-effective) for prediction of cytotoxicity of metal oxide NPs to bacteria *E. coli*. The cytotoxicity data of 17 metal oxides to bacteria *E. coli* have been taken to develop and validate the QSTR models. These simple descriptors included metal electronegativity (χ), the charge of the metal cation corresponding to a given oxide (χ_{ox}), atomic number, and valence electron number of the metal. The models were developed using stepwise MLR and PLS methods, respectively. The simple descriptors highlighted in this study and the developed models would be utilized for future prediction of cytotoxicity of metal NPs with probable mechanistic interpretation [9].

Nano-QSAR models for metal oxide NPs were developed using novel descriptors to predict the cytotoxicity of various NPs [48]. The nano-specific theoretical descriptor was proposed by integrating codes of certain physicochemical features into

SMILES-based optimal descriptors to characterize the nanostructure information of NPs. The new descriptors were applied to model metal oxide NP cytotoxicity to both *E. coli* bacteria and HaCaT cells. These developed models would reliably predict the cytotoxicity of novel NPs solely from the newly developed descriptors and provide guidance for prioritizing the design and manufacture of safer nanomaterials with desired properties.

5.3 QSAR Models for Toxicity of Contaminants of Emerging Concern (CECs)

The United States Geological Survey (USGS) defined the CECs as “any synthetic or naturally occurring chemical or any microorganism that is not commonly monitored in the environment but has the potential to enter the environment and cause known or suspected adverse ecological and/or human health effects [49].” The contaminants of emerging concern (CECs), including pharmaceuticals and personal care products (PPCPs) [50], are increasingly being detected at low levels in surface water, and there is concern that these compounds may have an impact on aquatic life. Conventional waste water and recycled water treatment are only partially effective in their removal or for their degradation, so they are discharged into the environment with treated waste water effluent, recycled water, and waste water plant sludge [49].

PPCPs are the unique group of emerging environmental contaminants, which include numerous chemical classes with very unique physiochemical properties and biological activities. Pharmaceuticals are used to treat a variety of human and animal diseases depending on their pharmacological action. Personal care products are commonly used in applications to improve the quality of daily life and include cosmetics, shampoos, soaps, deodorants, sunscreens, and toothpastes. There are many CECs and PPCPs that act as so-called endocrine disruptors (EDCs) [51]. EDCs are compounds that alter the normal functions of hormones resulting in a variety of health effects. EDCs can alter hormone levels leading to reproductive effects in aquatic organisms, and evaluating these effects may require testing methodologies not typically available along with endpoints not previously evaluated using current guidelines.

The accumulation of PPCPs and EDCs in different compartments of environments even in microlevel is raising a serious concern to the human, aquatic, and other forms of life. It is necessary to evaluate the potential impact of PPCPs and EDCs on the aquatic life and have an approach for determining protective levels for aquatic organisms (<https://www.epa.gov/wqc/contaminants-emerging-concern-including-pharmaceuticals-and-personal-care-products>). We have discussed here some of the recent application of QSAR models in ecotoxicity prediction of the PPCPs and EDCs.

Khan et al. (2017) [39] developed the QSTR models for toxicity of cosmetic ingredients on three different ecotoxicologically relevant organisms, namely, *Pseudokirchneriella subcapitata*,

D. magna, and *Pimephales promelas*, following the OECD guidelines. The dataset was collected from ECOTOX database and consisted of variety of PCPs including soaps, creams, shampoo, body lotion, sunscreen agent, and fragrances. The models were developed using partial least squares method. Predictions obtained by the derived QSTR models and ECOSAR tools were used to compare and rank the PCPs based on their average scaled aquatic toxicity values. These models could be utilized for the design and synthesis of safer cosmetics.

Three different QSTR models were developed by Gramatica et al. (2016) [52] using a dataset of 534 PCPs showing ecotoxicity against three aquatic tropic levels of organisms, i.e., algae (*P. subcapitata*), *Daphnia* (*D. magna*), and fish (*P. promelas*). These models were developed by the GA-OLS method and were applied to prioritize the most toxic compounds among about 500 PCP ingredients without experimental data. The predicted values obtained from these models are more similar to the available experimental values, if compared with those obtained by the commonly used software ECOSAR. These models could be utilized for the prediction of the acute aquatic toxicity of organic ingredients of personal care products (PCPs) and to design the new PCP that could be a possible “safer alternative” of a recognized hazardous chemical.

A quantitative activity-activity relationship (QAAR) models were developed by Sangion et al. (2016) [53], using the ecotoxicological data of standard organisms internationally accepted in the standard guidelines for the testing of chemicals (*D. magna*, *P. promelas*, and *Oncorhynchus mykiss*). The interspecies models were developed using the simple linear regression and multiple linear regression methods. These models are helpful tools for the prioritization of the most hazardous compounds. Also the proposed invertebrate-fish interspecies models can reduce the more complex experimental tests on upper trophic organisms and save animal lives. The models could also be applied to help the REACH requirement of a reduction of animal testing by gathering and extrapolating information from tested to untested animals, as well as from tested to untested chemicals through the integration of testing and in silico approaches.

Khan et al. (2019) [5] developed QSTR models for pharmaceuticals using the ecotoxicological data of four different aquatic species, namely, *P. subcapitata*, *D. magna*, *O. mykiss*, and *P. promelas* [54]. Genetic algorithm (GA) was used for feature selection followed by partial least squares regression technique according to the OECD guidelines. A double cross-validation methodology was employed for selecting the best models. The obtained robust consensus models were utilized to predict the toxicity of a large dataset of approximately 9300 drug-like molecules in order to prioritize the existing drug-like substances in

accordance to their acute predicted aquatic toxicities. These models could be utilized to predict the acute toxicity of pharmaceuticals before their usage on the industrial scale.

QSTR models for CECs were developed [55] using the toxicity data of 75 CEC compounds to *D. japonica*, 47 compounds to *D. magna*, and 19 compounds against *P. promelas*. The considered CECs included ionic and nonionic surfactants, UV filters, hormones and endocrine disrupting agents (EDCs), preservatives, pharmaceuticals, and organophosphates. Besides the QSTR models, QSTTR models were also developed for the toxicity prediction of CEC compounds through interspecies relationship for 47 CECs between *D. japonica* and *D. magna* (*daphnia*) and for 19 compounds between *D. japonica* and *P. promelas* (fish). These models can be used in toxicity evaluation, screening and prioritization, and development of risk management measures in a scientific and regulatory frame.

Khan et al. (2019) [56] developed QSTR and i-QSTTR models using toxicity data of 144 endocrine disruptor chemicals (EDCs) toward 14 different species falling in 4 different trophic levels. The models were developed using genetic algorithm followed by PLS regression method. These models can be employed in library screening, in regulatory decisions, and in the design of safer alternatives in order to reduce environmental hazards caused by EDCs.

6 Read-Across (RA) as a Tool to Predict Missing Ecotoxicological Data

In recent years, with emergence of ITS, predicting missing ecotoxicological data with read-across (RA) has become an attractive and pragmatic alternative. The RA works on the principle based on the structural similarity, i.e., following an assumption that similar structures should exhibit similar physicochemical, toxicological, and ecotoxicological properties [2]. Conceptually, RA may act like local QSAR models. The read-across model is derived from molecules with similar nature in terms of structure or functions. In RA, similarly grouped chemicals with a defined endpoint are used to predict the same endpoint for other chemicals. The following four schemes have been proposed in read-across data gap filling, (1) one-to-one, one-to-many, many-to-one, and many-to-many. As per the official OECD guidelines [6], read-across (quantitative) can be conducted with either of the following concepts:

- Performing read-across using the endpoint value of a similar chemical entity
- Applying a scale (mathematical) to the trend in experimental data with two/more similar chemicals with respect to target molecule

- Processing response values from two/more source chemicals
- Taking the most conservative value among the source data if sufficient data is available

Here, we have discussed in brief the application of RA in toxicity prediction or data gap filling.

6.1 Application of Read-Across in Ecotoxicological Data Gap Filling

The applications of RA in ecotoxicological data gap filling have been restricted mainly to nanomaterials/nanoparticles so far. Gajewicz et al. [2] proposed a novel quantitative read-across (Nano-QRA) approach using a simple and effective algorithm to give reliable predictions of the missing data against *E. coli* and human keratinocyte (HaCaT) cell line for untested metal oxide nanoparticles. Nel et al. (2013) [57] and Cockburn et al. (2012) [58] demonstrated the use of RA in ranking of nanomaterials based on their level of acute concentration in rodents. Sellers et al. (2015) [59] employed RA for data gap filling of silver and titanium dioxide nanoparticles mainly in fish. The data gap filling of in vitro genotoxicity using RA approach was demonstrated by Lamon et al. (2018) [60], whereas for bacteria, algae, protozoa, and human keratinocyte cell, it was demonstrated by Sizachenko et al. (2018) [61]. George and colleague [62] ranked metal oxide NPs based on their hazard-causing potential in zebra fish using read-across technique.

7 Overview and Conclusion

Computational modeling of ecotoxicity prediction of diverse organic chemicals, environmental pollutants, and contaminants of emerging concern is one of the crucial aspects of ecotoxicological risk assessment. These chemical substances/toxicants/pollutants may be ionic liquids, nanomaterials, personal care products, pharmaceuticals, persistent organic pollutants, endocrine disruptors, etc. The experimental determination of toxicities of all these chemicals involves multiple testing protocols and costly laboratory experiments requiring sacrifice of a large number of animals. Therefore, standard indicator organisms accepted in the guidelines for the testing of chemicals are used. Different organisms of varied trophic levels differ in their susceptibility to specific chemicals, most likely due to their differences in accessibility, metabolic rate, excretion rate, genetic factors, dietary factors, and stress level of the organism. According to the literature reports, ionic liquid toxicity against *V. fischeri*, *D. magna*, and *S. vacuolatus*; nanomaterial toxicity against *S. typhimurium* and *E. coli*; cosmetic and pharmaceutical toxicities against *P. subcapitata*, *D. magna*, and *P. promelas*; and contaminants of emerging concern toxicities against *D. japonica*, *D. magna*, and *P. promelas* have been experimentally studied to determine the toxicity of those chemicals to the ecosystem (mainly aquatic).

QSAR modeling could be used for the development of statistically significant, robust, and reliable models of diverse chemicals and environmental pollutants, and these models would be used efficiently to predict the toxicity of new chemicals which are expected to become pollutants or toxicants after its use at the present or in the future. The chemical domain of these models could be made wider by incorporating the chemical information from huge toxicity databases. The wider domain of the models may enable the researcher to predict the toxicity of almost any class of chemicals reliably. These models do not require lengthy, timely, and costly experiments which involve the use of plants or animals. Hence, the QSAR models can be utilized for the assessment of new and existing chemicals in conformity with regulatory requirements within the scope of OECD. Finally, with the introduction of read-across technique, one can evaluate the potential negative impact of nanomaterials (and also other classes of toxicants) to the human health and the surrounding without the necessity of performing expensive and time-consuming experiments. The applications of QSAR and read-across in predictive toxicology are depicted in Fig. 2.

The present chapter has reviewed the applications of QSAR in predictive ecotoxicology as a tool to replace costly experimental procedures. It also gives a bird's-eye view of the basic methodology implemented in QSAR such as data collection, descriptor calculation, modeling algorithms, model validation, and interpretation. Representative examples of QSAR studies in ecotoxicity prediction for different chemical classes such as pharmaceuticals, cosmetics, CECs, and NPs have also been discussed. The chapter also lists the important databases containing ecotoxicity data for various

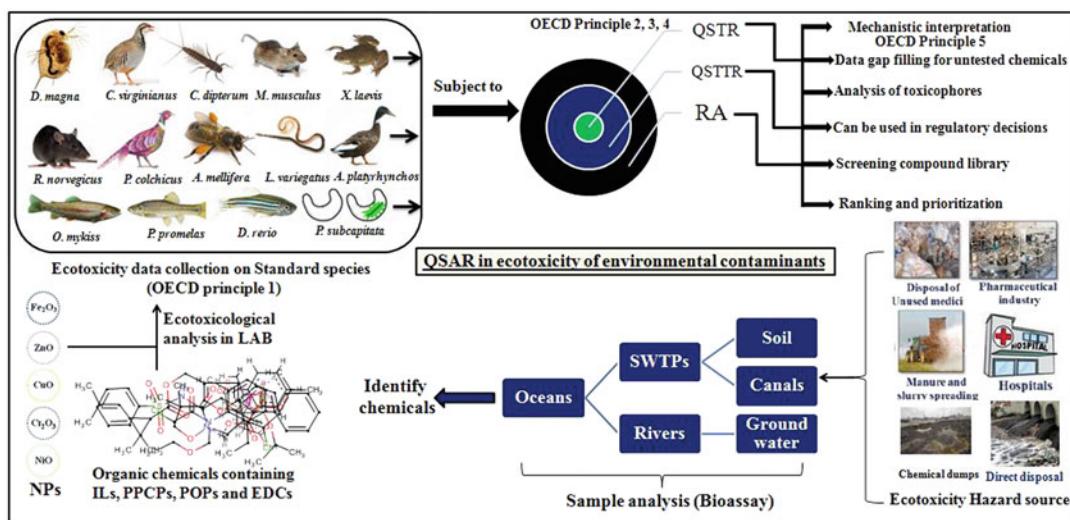


Fig. 2 Application of QSAR/QSTTR/RA in ecotoxicity of environmental contaminants

ecologically important endpoints alongwith available software tools used for toxicity predictions. We have also listed important regulatory bodies which encourage the use of QSAR in early detection of hazardous materials. Finally, the application of read-across as tool to fill the data gap in species where experimental data availability is limited has also been mentioned. The chapter highlights the need for developing ample number of QSAR models in order to computationally derive missing ecotoxicity data for different chemical classes against various endpoints. In summary, QSARs have the ability to make fast and reliable predictions which are faster than experimental methods, thus justifying their relevance in early risk assessment of chemicals.

Acknowledgment

KK thanks the Indian Council of Medical Research, New Delhi, for financial support in the form of a senior research fellowship.

References

1. De P, Roy K (2018) Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR QSAR Environ Res* 29:319–337
2. Gajewicz A, Jagiello K, Cronin M, Leszczynski J, Puzyn T (2017) Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environ Sci Nano* 4:346–358
3. Roy K (2019) *In silico drug design: repurposing techniques and methodologies*. Academic Press, New York
4. Dearden JC (2017) The history and development of quantitative structure-activity relationships (QSARs). In: *Oncology: breakthroughs in research and practice*. IGI Global, Hershey, pp 67–117
5. Khan K, Benfenati E, Roy K (2019) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotox Environ Safe* 168:287–297
6. OECD (2014) Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models, OECD series on testing and assessment, no. 69. OECD Publishing, Paris. Available at <https://doi.org/10.1787/9789264085442-en>
7. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *CHEMOM INTELL LAB SYST* 152(229):18–33
8. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J Chem Inf Model* 52:1138–1145
9. Kar S, Gajewicz A, Puzyn T, Roy K, Leszczynski J (2014) Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach. *Ecotoxicol Environ Saf* 107:162–169
10. Todeschini R, Consonni V (2000) Methods and principles in medicinal chemistry. In: Kubinyi H, Timmerman H (Series eds) *Handbook of molecular descriptors*. Wiley-VCH, Weinheim
11. Golmohammadi H, Dashtbozorgi Z, Acree WE Jr (2012) Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci* 47:421–429
12. Kennard RW, Stone LA (1969) Computer aided design of experiments. *Technometrics* 11:137–148
13. Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc C-Appl* 28:100–108
14. Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. *J Indian Chem Soc* 95:1497–1502

15. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). *Expert Opin Drug Dis* 13:1075–1089
16. De P, Aher RB, Roy K (2018) Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector Aedes aegypti: application of ETA indices. *RSC Adv* 8:4662–4670
17. De P, Kar S, Roy K, Leszczynski J (2018) Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environ Sci-Nano* 5:2742–2760
18. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, Amsterdam
19. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemometr Intell Lab Syst* 145:22–29
20. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 14:450–474
21. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52:396–408
22. Roy K (2019) Multi-target drug design using chem-bioinformatic approaches. Springer, New York
23. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130
24. Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* 6:47
25. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
26. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J Chemom* 32:2992
27. Roy K, Ambure P, Kar S (2018) “Prediction reliability indicator”: a new tool to judge the quality of predictions from QSAR models for new query compounds. In: 24 May 2018 in MOL2NET 2018. International conference on multidisciplinary sciences, MDPI AG, Basel
28. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* 3:11392–11406
29. Khan K, Khan PM, Lavado G, Valsecchi C, Pasqualini J, Baderna D, Marzo M, Lombardo A, Roy K, Benfenati E (2019) QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors. *Chemosphere*. <https://doi.org/10.1016/j.chemosphere.2019.04.204>
30. Kar S, Roy K, Leszczynski J (2017) On applications of QSARs in food and agricultural sciences: history and critical review of recent developments. In: *Advances in QSAR modeling*. Springer, Cham, pp 203–302
31. Khan PM, Roy K, Benfenati E (2019) Chemometric modeling of Daphnia magna toxicity of agrochemicals. *Chemosphere* 224:470–479
32. Roy K (2017) Advances in QSAR modeling. In: *Applications in pharmaceutical, chemical, food, agricultural and environmental sciences*. Springer, Cham, p 555
33. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2017) Ecotoxicological assessment of pharmaceuticals using computational toxicology approaches: QSTR and interspecies QTTR modeling. In: *Proceedings of MOL2-NET 2017*, international conference on multidisciplinary sciences, 3rd edn. MDPI AG, Basel, p 1
34. Das S, Ojha PK, Roy K (2017) Multilayered variable selection in QSPR: a case study of modeling melting point of bromide ionic liquids. *Int J Quant Struct-Prop Relat (IJQSPR)* 2:106–124
35. Das S, Ojha PK, Roy K (2017) Development of a temperature dependent 2D-QSPR model for viscosity of diverse functional ionic liquids. *J Mol Liq* 240:454–467
36. Ojha PK, Kar S, Roy K, Leszczynski J (2019) Toward comprehension of multiple human cells uptake of engineered nano metal oxides: quantitative inter cell line uptake specificity (QICLUS) modeling. *Nanotoxicology* 31:14–34
37. Ghosh S, Ojha PK, Roy K (2019) Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. *Chemosphere* 228:545–555
38. Roy J, Ojha PK, Roy K (2019) Risk assessment of heterogeneous TiO₂-based engineered nanoparticles (NPs): a QSTR approach using simple periodic table based descriptors. *Nanotoxicology* 13:701–716
39. Khan K, Roy K (2017) Ecotoxicological modelling of cosmetics for aquatic organisms: a

- QSTR approach. *SAR QSAR Environ Res* 28:567–594
40. Hossain KA, Roy K (2018) Chemometric modeling of toxicity of contaminants of emerging concern to *Dugesia japonica* and its interspecies correlation with daphnia and fish: QSTR and i-QSTTR approaches. In: MOL2-NET 2018, international conference on multidisciplinary sciences, 4th edn. <https://sciforum.net/paper/view/conference/5266>
 41. Ghanem OB, Mutalib MIA, Leveque J-M, El-Harbawi M (2017) Development of QSAR model to predict the ecotoxicity of *Vibrio fischeri* using COSMO-RS descriptors. *Chemosphere* 170:242–250
 42. Farahani SR, Sohrabi MR, Ghasemi JB (2018) A detailed structural study of cytotoxicity effect of ionic liquids on the leukemia rat cell line IPC-81 by three dimensional quantitative structure toxicity relationship. *Ecotoxicol Environ Saf* 158:256–265
 43. Das RN, Sintra TE, Coutinho JAP, Ventura SPM, Roy K, Popelier PLA (2016) Development of predictive QSAR models for *Vibrio fischeri* toxicity of ionic liquids and their true external and experimental validation tests. *Toxicol Res* 5:1388–1399
 44. Das RN, Roy K, Popelier PLA (2015) Interspecies quantitative structure-toxicity-toxicity (QSTTR) relationship modeling of ionic liquids. Toxicity of ionic liquids to *V. fischeri*, *D. magna* and *S. vacuolatus*. *Ecotoxicol Environ Saf* 122:497–520
 45. Bubalo MC, Radošević K, Srček VG, Das RN, Popelier P, Roy K (2015) Cytotoxicity towards CCO cells of imidazolium ionic liquids with functionalized side chains: preliminary QSTR modeling using regression and classification based approaches. *Ecotoxicol Environ Saf* 112:22–28
 46. Toropova AP, Toropov AA, Veselinović AM, Veselinović JB, Benfenati E, Leszczynska D, Leszczynski J (2016) Nano-QSAR: model of mutagenicity of fullerene as a mathematical function of different conditions. *Ecotoxicol Environ Saf* 124:32–36
 47. Basant N, Gupta S (2017) Multi-target QSTR modeling for simultaneous prediction of multiple toxicity endpoints of nano-metal oxides. *Nanotoxicology* 11:339–350
 48. Pan Y, Li T, Cheng J, Telesca D, Zink JI, Jiang J (2016) Nano-QSAR modeling for predicting the cytotoxicity of metal oxide nanoparticles using novel descriptors. *RSC Adv* 6:25766–25775
 49. Raghav M, Eden S, Mitchell K, Witte B (2013) Contaminants of emerging concern in water. Water Resources Research Center College of Agriculture and Life Sciences, Arizona
 50. Kar S, Roy K, Leszczynski J (2018) Impact of pharmaceuticals on the environment: risk assessment using QSAR modeling approach. In: Computational toxicology. Springer, New York, pp 395–443
 51. Kar S, Sepālveda MS, Roy K, Leszczynski J (2017) Endocrine-disrupting activity of per-and polyfluoroalkyl substances: exploring combined approaches of ligand and structure based modeling. *Chemosphere* 184:514–523
 52. Gramatica P, Cassani S, Sangion A (2016) Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives design. *Green Chem* 18:4393–4406
 53. Sangion A, Gramatica P (2016) Ecotoxicity interspecies QAAR models from *Daphnia* toxicity of pharmaceuticals and personal care products. *SAR QSAR Environ Res* 27:781–798
 54. Kar S, Das RN, Roy K, Leszczynski J (2016) Can toxicity for different species be correlated?: the concept and emerging applications of interspecies quantitative structure-toxicity relationship (i-QSTR) modeling, (IJQSPR) 1:23–51
 55. Hossain KA, Roy K (2018) Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in *Dugesia japonica* and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. *Ecotoxicol Environ Saf* 166:92–101
 56. Khan K, Roy K, Benfenati E (2019) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. *J Hazard Mater* 369:707–718
 57. Nel A, Xia T, Meng H, Wang X, Lin S, Ji Z, Zhang H (2012) Nanomaterial toxicity testing in the 21st century: use of a predictive toxicological approach and high-throughput screening. *Acc Chem Res* 46:607–621
 58. Cockburn A, Bradford R, Buck N, Constable A, Edwards G, Haber B, Hepburn P, Howlett J, Kampers F, Klein C (2012) Approaches to the safety assessment of engineered nanomaterials (ENM) in food. *Food Chem Toxicol* 50:2224–2242
 59. Sellers K, Deleebeck NM, Messiean M, Jackson M, Bleeker EAJ, Sijm D, Van Broekhuizen F (2015) Grouping nanomaterials: a strategy towards grouping and read-across, Rijksinstituut voor Volksgezondheid en Milieu RIVM
 60. Lamon L, Asturiol D, Richarz A, Joossens E, Graepel R, Aschberger K, Worth A (2018) Grouping of nanomaterials to read-across hazard endpoints: from data collection to assessment of the grouping hypothesis by application

- of chemoinformatic techniques. Part Fibre Toxicol 15:37
61. Sizachenko N, Mikolajczyk A, Karolina J, Puzyn T, Leszczynski J, Rasulev B (2018) How the toxicity of nanomaterials towards different species could be simultaneously evaluated: a novel multi-nano-read-across approach. Nanoscale 10:582–591
62. George S, Tian X, Robert R, Yan Z, Zhaoxia J, Sijie L, Xiang W (2011) Use of a high-throughput screening approach coupled with *in vivo* zebrafish embryo screening to develop hazard ranking for engineered nanomaterials. ACS Nano 5:1805–1817



Chapter 3

Best Practices for Constructing Reproducible QSAR Models

Chanin Nantasenamat

Abstract

Quantitative structure-activity/property relationship (QSAR/QSPR) has been instrumental in unraveling the origins of the mechanism of action for biological activity of interest by means of mathematical formulation as a function of the physicochemical description of chemical structures. Of the growing number of QSAR models being published in the literature, it is estimated that the majority of these models are not reproducible given the heterogeneity of the components of the QSAR model setup (e.g., descriptor, learning algorithm, learning parameters, open-source and commercial software, different software versions, etc.) and the limited availability of the underlying raw data and analysis source codes used to construct these models. This inherently poses a challenge for newcomers and practitioners in the field to reproduce or make use of the published QSAR models. However, this is expected to change in light of the growing momentum for open data and data sharing that are being encouraged by funders, publishers, and journals as well as driven by the next generation of researchers who embrace open science for pushing science forward. This chapter examines these issues and provides general guidelines and best practices for constructing reproducible QSAR models.

Key words Quantitative structure-activity relationship, Quantitative structure-property relationship, Structure-activity relationship, QSAR, QSPR, SAR, Research reproducibility, Reproducibility, Reproducible, Jupyter, Python

1 Quantitative Structure-Activity Relationship

Quantitative structure-activity relationship (QSAR) is an exciting field that harnesses past biological activity data to drive further experimentations by enabling the prediction/design of biological activity of new compounds, deducing the important molecular features giving rise to good or poor biological activity, prioritizing compounds from a large chemical library, etc. [1, 2]. QSAR has successfully been demonstrated to be useful for modeling a wide range of biological and chemical endpoints as summarized in Table 1.

In almost 60 years since the coining of the QSAR term by Hansch [4], the field has evolved from classical QSAR models (i.e., consisting of a few compounds and described by simple

Table 1
Summary of target biological and chemical endpoints investigated by QSAR models

Endpoints	Examples
Physical and chemical properties	Boiling point, melting point, octanol-water partition coefficient, water solubility, etc.
Environmental fate	Biodegradation, bioconcentration, adsorption/desorption in soil, etc.
Ecotoxicity	Acute toxicity to fish, short-term toxicity to Daphnia, toxicity to plants, etc.
Human health	Acute inhalation toxicity, skin irritation, mutagenicity, etc.
Toxicokinetics	Blood-brain barrier penetration, skin penetration, metabolism, etc.
Drug discovery	Enzyme inhibition, enzyme activation, pharmacokinetics, etc.

Aside from the drug discovery class, other endpoints are categorized according to the convention described by Piir et al. [3]

descriptors) to complex machine learning-based QSAR models (i.e., encompassing several hundred to thousands of descriptors as modeled by nonlinear learning algorithms) [5]. The field of QSAR has witnessed its ups and downs [6] and has become pillars for drug discovery [7] and regulatory purposes [8].

2 Laboratory Notebooks: Past and Present

Historically, the documentation of experimental results had traditionally been kept within the confinement of paper-based notebooks whereby the scientific benefit of which is to allow subsequent reproduction of the documented experiment, while its legal use is to serve as a proof of inventorship [9].

The electronic laboratory notebooks have been introduced as a digital alternative to the paper-based version but with augmented capabilities such as search capability, integration with instrumentation [10], as well as collaborative writing and archiving of results figures and tables. Scientists are increasingly adopting the use of electronic laboratory notebooks in their research laboratories owing to the inherent need to organize the growing volume of biological data [11] with the benefit of being able to access these documents via the Internet at any place and time.

With the rising awareness on research reproducibility, scientists are increasingly sharing these notebooks publicly so as to support the open science initiative and in doing so fosters the sharing of associated raw data and analysis code that would have otherwise remained within the confinements of laboratory computers or individual researcher's personal computer (i.e., known as *dark data*).

3 Data Sharing

Publishers of research journals are encouraging or requiring that researchers share the scripts and codes used to analyze the data as a condition of publication. Failure to do so (i.e., owing to privacy or safety) may require a written statement justifying the reason. A notable example is the appointment of *reproducibility editors* for overseeing the code and data sets submitted by authors to the Applications and Case Studies (ACS) section of the Journal of the American Statistical Association (JASA). In an editorial article by the Editor-in-Chief of the Journal of Chemical Information and Modeling, William L. Jorgensen formulated a set of guidelines for submitting QSAR work to the journal. A key issue pertaining to data sharing is highlighted in one of the recommendation as follows: *All data and molecular structures used to carry out a QSAR/QSPR study are to be reported in the paper and/or in its Supporting Information, or be readily available, without infringements or restrictions.* Furthermore, publishers (Springer Nature [12]) and journals (PeerJ, PLoS One [13], etc.) have established similar requirements. Of particular note, Vasilevsky et al. [14] performed an analysis of the pervasiveness and quality of data sharing policies in the biomedical literature and found that 11.9% of journals explicitly stated that data sharing was required as a condition of publication.

The advantage of imposing data sharing as a condition for publishing is that potential unintended errors (e.g., missing data, mislabeling, corrupted files, etc.) may be identified prior to publication, which would consequently solve any future problems that may hamper the reproducibility of subsequent works [15]. On the other side of the coin, possible reasons that authors may be reluctant to share the proprietary data is that it would reveal the confidentiality of compounds. In addressing this issue, Gedeck et al. [16] described an approach for facilitating data sharing and the development of collaborative QSAR models while not revealing the structural information. Polanski et al. [17] reviewed the contributing factors for robust QSAR models, and of particular note is their proposition that QSAR is highly data dependent and that the underlying data may inherently produce noise that may arise from many factors such as the molecular conformation, computed descriptors, algorithms used, etc.

4 Data, Chemical Structure, Conformation, and Descriptors

As we have seen, the data availability is an important prerequisite for model reproducibility. Aside from this is a series of additional hurdles and challenges that may affect the reproducibility of the

Table 2
Summary of different dimensions of molecular descriptors

Dimensions	Description
Zero-dimensional (0D)	Molecules are directly described by the chemical formula pertaining to atom counts, molecular weight, sum/average of molecular property, etc.
One-dimensional (1D)	Molecules are characterized by substructural features that consider the presence/absence of molecular fragments or functional groups
Two-dimensional (2D)	Molecules are described by the presence and type of chemical bonds that are used to connect atoms together
Three-dimensional (3D)	Molecules are perceived as a geometrical object in space that are characterized by the nature and connectivity of atoms together with their spatial representation
Four-dimensional (4D)	Representation of the molecule-receptor interaction by means of molecular interaction fields that is generated from grid-based mapping of probes in relation to thousands of evenly spaced grid points
Higher dimensions	These high dimensional models may be characterized by different induced-fit and solvation models

Adapted from Grisoni et al. [20]

QSAR model. Inherently, QSAR models are reliant on the underlying chemical structures that may produce a myriad of possible descriptors that may range anywhere from simple descriptors to various other dimensions ranging from zero-dimensional to six-dimensional descriptors [18–20] as summarized in Table 2.

The concept of structure-activity cliffs [21, 22] demonstrated that even a minor change in the chemical structure (i.e., addition or deletion of a methyl group or even the stereoisomeric placement of functional groups) may be a deciding factor whether the compound can or cannot bind to the intended target protein) can give rise to significant changes to the observed activity. Such induced-fit of ligands to their target proteins may be affected by the structure-activity cliff concept, but a question arises as to the importance of conformation on other sets of compounds. The bioactive conformations of compounds are known to be principal drivers of their resulting biological activity, and thus several QSAR studies have addressed this area.

A notable example is the work of Guimarães et al. [23] in which they performed an investigation comparing 2D and 3D QSAR models for a set of halogenated anesthetics. Surprisingly, their results indicated that the 2D model provided comparable performance to that of the 3D model, thereby suggesting that the 2D descriptors were also robust in their particular investigation. Thus, one can conclude that the influence of the molecular conformation on the resulting QSAR model is system dependent and must therefore be subjected to careful investigation on a case-by-case basis.

Another interesting work by Pissurlenkar et al. [24] tackled the traditional paradigm of QSAR that is the concept of *one chemical, one structure, one parameter value* as proposed by the authors. Their development of the so-called ensemble QSAR (*e*QSAR) model takes into account descriptors generated from a set of low-energy conformers instead of the traditional approach of using only one low-energy conformer. The study for the first time establishes the possibility of incorporating conformation flexibility into QSAR models and thus opens up a new area for further exploration of this important paradigm. Recently, Wicker and Cooper [25] proposed a new molecular descriptor nConf₂₀ based on chemical connectivity for capturing the conformational space of a molecule. To facilitate usage by the scientific community, the authors also provided the Python code and the accompanying data set (i.e., containing both the calculate molecular descriptors and the class label that can be used for QSAR model building) in the Supporting Information of their article.

5 QSAR Model Building Process

The general procedures for constructing QSAR models are summarized in chronological order in Table 5 and Fig. 1. A more in-depth treatment on recommendations and best practices for QSAR model development is described by excellent review articles by Dearden et al. [26] and Tropsha et al. [27, 28]. The concepts presented in Table 5 and the aforementioned articles on best practices of QSAR model development help to ensure that robust and

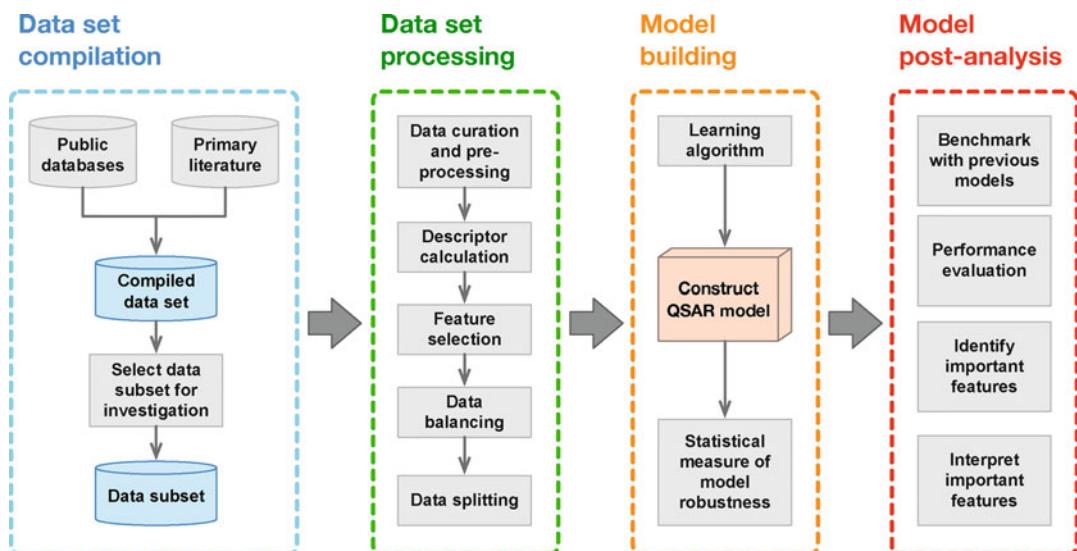


Fig. 1 Schematic representation of the QSAR modeling workflow

accurate models are built. In addition to this, there are emerging efforts in the QSAR literature that is targeted at the following issues:

1. Determine the confidence level for predictions obtained from QSAR models through the use of conformal predictions.
2. Assess the modelability of data sets so as to elucidate the feasibility of obtaining robust models [29].
3. Constructing interpretable QSAR models that can be of practical use for biologists and medicinal chemists [18].
4. Ensuring the reproducibility of QSAR models such that other research groups can make use of or extend published models.

In efforts to encourage the development of high-quality QSAR models, the Organization for Economic Cooperation and Development (OECD) had formulated a simple set of rules as outlined in Table 3. Criteria 2 of the OECD principles stressed that robust QSAR models should have *unambiguous algorithm*. At first glance, one would assume that details on the components used in the formulation of the QSAR model that are described in the Materials and Methods section of research articles would be enough to allow reproducibility of the model. As such information are descriptive in nature and as detailed as it may be, one can assume that there may potentially be some elements of ambiguity that may consequently lead to slightly different outcomes (if not different results) from that of the original model. Roy et al. [30] had pointed out in their investigation that QSAR are highly dynamic models that can easily be perturbed upon changes in the underlying algorithm for descriptor calculation, software version, or software availability using the Dragon software. Moreover, a summary of factors influencing the reproducibility of QSAR models based on our lab's own experience

Table 3
Summary of OECD principles for QSAR model building

No. OECD principles	Description
1 Defined endpoint	To ensure clarity in the endpoint being predicted as they may be derived from different experimental methods or conditions
2 Unambiguous algorithm	To ensure that underlying details of the model is transparent so as to facilitate model reproducibility
3 Defined applicability domain	To define the biological/chemical landscape in which the model can reliably make predictions
4 Measures of model performance	To evaluate the internal and external predictive ability of the model
5 Mechanistic interpretation	To ensure that the model can be interpreted such that the underlying mechanism of action of compounds is revealed

Table 4
Key factors influencing the reproducibility of QSAR models

No. Factors	Description
1 Data set	To achieve reproducibility of a QSAR model, the original data set should be available. At a minimum, this entails the provision of the chemical representation and bioactivity values. Other useful information may include references to the original data source
2 Chemical representation	Availability of chemical representation such as IUPAC name, SMILES notation or other forms of identifier number
3 Descriptors	The provision of computed descriptors would help to solve any potential issues pertaining to accessibility to commercial software or software updates that may alter descriptor calculation results
4 Model's parameters/ details	Name and version of software used for multivariate analysis; learning parameters used in the formulation of the model; classical QSAR models readily provide this from the MLR equations
5 Predicted endpoint values	Availability of the experimental and calculated endpoint values enable readers to compare their own reproduction of the model with that of the original model's results
6 Data splits	Availability of precise details as to which compound belongs to which data splits (e.g., internal, external, calibration, or validation sets) would facilitate comparison with the user's reproduction of models. Details on data split ratios (80/20 split or 70/30 split) or whether undersampling or oversampling were used

is provided in Table 4 while the summary of procedures for QSAR model building is provided in Table 5.

Piir et al. [3] performed a systematic review of the QSAR literature consisting of 1533 articles pertaining to 79 biological and chemical endpoints. Their results indicated that 42.5% of articles may be potentially reproducible (i.e., and thus complies with the five OECD principles) given that interested readers invest the necessary effort in retracing the protocol step-by-step using the same software and version. Furthermore, it was suggested that of the machine learning algorithm used in the QSAR literature, multiple linear regression seemed to be afforded the most reproducibility owing to its simplicity (i.e., inclusion of MLR equations in the research article). In spite of this, it was found that only 51% were technically complete, while the other majority were lacking significant details for reproducibility. Moreover, the authors also provided recommendations and best practices for QSAR reporting.

Early efforts by Spjuth et al. [34] had laid important foundations for interoperable QSAR data sets via the use of a QSAR markup language (QSAR-ML) in which the authors established the markup language to house meta data information that defines pertinent information about the QSAR data set consisting of

Table 5
Summary of procedures for QSAR model building

No.	Procedures	Description
1	Data compilation	The very essence of QSAR models lies in the compilation of the data set. Potential data sources include the primary literature and curated/semi-curated bioactivity databases
2	Select data subset for investigation	An extension of the previous step is selecting a subset of from the original data set for further investigation data
3	Data curation and pre-processing	This is probably the most time-consuming phase of the entire model building process as it entails cleaning the data (e.g., dealing with missing data), normalizing variables (e.g., logarithmic transformation), removal of salts/metals/duplicates, normalizing chemical structures (e.g., selecting appropriate tautomers), etc.
4	Descriptor calculation	An important component of the model building process is deciding how to represent the molecular features and physicochemical properties of the compounds of interest. A wide range of open-source and commercial software are available. Decisions will have to be made as to use a few interpretable features or to use a large set of features that may or may not be interpretable
5	Feature selection	Once descriptors are generated, the initial set of descriptors are normally subjected to removal of low variance variables followed by removal of collinear (redundant) variables. Again, there exists a large collection of algorithms for reducing the features (e.g., stepwise linear regression, genetic algorithm, particle swarms optimization, etc.)
6	Data balancing	A common problem for the development of classification models is that the classes of active and inactive compounds are often imbalanced where either classes may be significantly smaller or larger than the other. Such imbalanced data set is not suitable for model building, and the classes will have to be balanced via either undersampling or oversampling as well as via more sophisticated approaches such as the SMOTE algorithm
7	Data splitting	Partitioning the data set into various subsets (e.g., training, calibration, external validation, and cross-validation sets) is a common practice for validating the model robustness whether it is capable of reliable prediction on unseen data samples or for optimizing and tuning the model parameters
8	Learning algorithm	The highlight of the QSAR model building process is making use of the aforementioned curated data for multivariate analysis so as to correlate computed descriptors with the endpoint values of interest. Learning algorithm can be either supervised or unsupervised (i.e., making use of or not making use of the endpoint variable in the learning process), and the resulting model can be interpretable or not interpretable (black-box models) [18]
9	Statistical measures of model robustness	Model robustness and its reliability are traditionally assessed via various metrics such as R^2 , Q^2 , RMSE, and Y-scrambling. In recent years, conformal predictions [31–33] and other metrics have also been introduced

chemical structures, descriptors, endpoint, and meta data (e.g., authors, license, source reference, etc.). QSAR-ML is implemented via a set of plug-ins in the Bioclipse software via simple to use graphical tools [35]. Although useful, QSAR-ML considers only the pre-modeling phases, which encompasses procedures 1–4, while the modeling phases spanning procedures 5–9 were not covered.

Further efforts in driving the reproducibility of QSAR models forward were set forth by the works of Ruusmann et al. [36, 37] in which they introduced the QSAR DataBank repository (QsarDB). The QsarDB data format is conceptually similar to that of QSAR-ML but extends it to also include model information. Particularly, Predictive Model Markup Language (PMML) is an open standard for encoding information pertaining to the machine learning model, thereby allowing model sharing. The flexibility of PMML permits it to act as an intermediary in encoding the essence of the model from among the different machine learning softwares and tools that are available (i.e., which is comparable to an interpreter who can speak many languages). In their work, the authors propose the use of the R language for carrying out the model building procedures in the R programming environment followed by using the author’s own R package *rQsarDB* [38] for data conversion from CSV format to the QsarDB format as well as modifying the contents of existing QsarDB archive directories from within the R environment.

In regard to the pre-modeling phase, data compilation and curation can be considered to be the most time-consuming procedures in the QSAR model building process. Aside from the issue of time consumption, the quality of the resulting model is dependent on the quality of the curated data. Thus, the important concept in computer science of *garbage in, garbage out* has become ever more important in the context of QSAR model building as attested by the important articles from Fourches et al. [39–41].

In later sections of this chapter, we describe the use of the Jupyter notebook for performing all of the aforementioned procedures encompassing the pre-processing, construction, validation, and evaluation of the robustness of the QSAR model. Such coverage naturally facilitates reproducible construction of QSAR models as the precise protocol, learning function, learning parameters, and performance metrics are housed within the Jupyter notebook file. It is increasingly becoming common practice for researchers to share their Jupyter notebook along with accompanying data sets on public repositories such as GitHub or Bitbucket as well as the QsarDB.

6 Interactive Notebooks

Electronic notebooks merely refer to the archiving of explanatory text of what was done and how, while the associated data and analysis code may or may not be provided with the notebooks. There are now *interactive notebooks* that make it possible for the code used to perform the data analysis to be shown alongside the explanatory text and visualizations (e.g., images, plots, etc.). As a result, this affords easy comprehension of the experimental results and the underlying code while also facilitating reproducible research.

A widely adopted interactive notebook that is used in the scientific community is known as the Jupyter notebook (i.e., previously known as iPython notebook). The original iPython notebook was created in 2001 by Fernando Perez and had since evolved to the more general and powerful Jupyter notebook (<http://www.jupyter.org/>) with support for more than 40 programming languages (e.g., Python, R, Javascript, Latex, etc.).

For the sake of data sharing, it is common practice to store the Jupyter notebooks (i.e., used hereafter to also refer to the iPython notebook) on GitHub (i.e., or other web repository such as Bitbucket). Such notebook files can then be rendered as static HTML via the nbviewer (<http://nbviewer.jupyter.org/>). Moreover, GitHub also makes it possible for Jupyter notebook files to render directly on its repositories. Owing to the static nature of the rendered notebook, the resulting HTML is consequently not interactive and therefore not amenable to modifications. A first step toward solving this limitation is made by the Freeman laboratory at Janelia Research Campus in their development of *binder* (<http://mybinder.org/>), a web service that converts Jupyter notebook files hosted on GitHub to executable and interactive notebooks. Recently, there is a web service known as the *Code Ocean* that not only allows the sharing of the raw data and associated analysis codes but also enables users the capability of running the analysis codes (i.e., supports several open-source languages such as R and Python as well as commercial languages such as MATLAB and Stata).

7 Tutorials on Using Jupyter Notebook for QSAR Modeling

7.1 Tutorial 1: Installing Miniconda

Before we begin, let's familiarize ourselves with Conda, which is a package manager that we will be using to manage the installation of packages in supported languages such as R and Python. The reason for using this package manager is that it will simplify the installation of packages by automatically taking care of installing the prerequisites (dependencies) that are needed to run the package of interest. Some packages that may be a challenge to install if performed

manually (i.e., requiring the compilation of C++ code via the use of additional libraries namely Boost) such as the *rdkit* can be easily installed via a one-line command (shown below).

Conda comes in two versions: (1) Anaconda and (2) Miniconda. In this tutorial, we will be using the Miniconda version owing to its requirement of less computer resources. To get started, we will need to install Miniconda by following the steps below:

1. In a web browser, go to <https://docs.conda.io/en/latest/miniconda.html>
2. Download the appropriate installer that matches your operating system (Windows, Mac OS X or Linux) and bit version (32-bit or 64-bit).
3. Once Miniconda is installed, try running the *conda* command in a terminal window (i.e., also known as the command prompt window). To a blank prompt that looks like the following:

```
$
```

Run the *conda* command as follows (press the **Enter** button after typing the command):

```
$ conda
```

If installation went successfully, the following output should be displayed:

```
$ conda
usage: conda [-h] [-V] command ...
conda is a tool for managing and deploying applications, environments and packages.
```

Options:

```
positional arguments:
  command
    clean      Remove unused packages and caches.
    config     Modify configuration values in .condarc. This is modeled
               after the git config command. Writes to the user .condarc
               file (/Users/chanin/.condarc) by default.
    create     Create a new conda environment from a list of specified
               packages.
    help       Displays a list of available conda commands and their help
               strings.
    info       Display information about current conda install.
    install    Installs a list of packages into a specified conda
               environment.
    list       List linked packages in a conda environment.
```

```

package      Low-level conda package utility. (EXPERIMENTAL)
remove       Remove a list of packages from a specified conda environment.
uninstall    Alias for conda remove. See conda remove --help.
search       Search for packages and display associated information. The
            input is a MatchSpec, a query language for conda packages.
            See examples below.
update       Updates conda packages to the latest compatible version. This
            command accepts a list of package names and updates them to
            the latest versions that are compatible with all other
            packages in the environment. Conda attempts to install the
            newest versions of the requested packages. To accomplish
            this, it may update some packages that are already installed,
            or install additional packages. To prevent existing packages
            from updating, use the --no-update-deps option. This may
            force conda to install older versions of the requested
            packages, and it does not prevent additional dependency
            packages from being installed. If you wish to skip dependency
            checking altogether, use the '--force' option. This may
            result in an environment with incompatible packages, so this
            option must be used with great caution.
upgrade     Alias for conda update. See conda update --help.

```

optional arguments:

```

-h, --help      Show this help message and exit.
-V, --version   Show the conda version number and exit.

```

conda commands available from other packages:

```
env
```

7.2 Tutorial 2: Installing Packages in Conda

As stated previously, conda supports the management of packages of languages such as R and Python. Thus, in this tutorial we will present examples for installing packages in both languages.

7.2.1 Installing Python Packages

Installing Python packages is very simple, which can be performed by typing the following command:

```
$ conda install package-name
```

where package-name refers to the Python package name. For instance, if we would like to install the jupyter package, then enter the following:

```
$ conda install jupyter
```

Then, it will ask to confirm that we would like to proceed with the installation, which we will enter y as the answer and press the Enter button.

```
Proceed ([y]/n) ?
```

Instead of manual confirmation for every package that we would like to install, the confirmation process can be automated by invoking the `-y` function as follows:

```
$ conda install jupyter -y
```

Now this time, the installation process will proceed without manual confirmation, which becomes very convenient if installing more than one package.

To check if the package has been successfully installed, use the `list` function as follows:

```
$ conda list
```

```
$ conda list
# packages in environment at /Users/chanin/miniconda2:
#
# Name           Version        Build       Channel
appnope          0.1.0          py27_0    
asn1crypto       0.24.0         py27_0    
backports         1.0           py27_0    
backports functools_lru_cache 1.5           py27_1    
backports_abc    0.5           py27_0    
beautifulsoup4   4.7.1          py27_1    
blas              1.0           mkl        
bleach            1.5.0          py27_0    
boost             1.56.0         py27_3    rdkit
bzip2             1.0.6          3          
ca-certificates  2019.1.23    0          
cairo             1.14.8         0          
certifi           2019.3.9    py27_0    
cffi              1.9.1          py27_0    
chardet          3.0.4          py27_1    
chembl-webresource-client 0.8.51        pypi      
chembl_webresource_client 0.9.31        chembl    
click              6.7           pypi_0    pypi
conda              4.6.14         py27_0    
conda-env         2.6.0          1          
configparser      3.5.0          py27_0    
cryptography     2.6.1          py27ha12b0ac_0
cycler             0.10.0         py27_0    
dash               0.17.5         pypi_0    pypi
dash-core-components 0.5.0          pypi_0    pypi
dash-html-components 0.6.1          pypi_0    pypi
dash-ly            0.17.3         pypi_0    pypi
```

dash-renderer	0.7.3	pypi_0	pypi
decorator	4.0.11	py27_0	
easydict	1.6	pypi_0	pypi
entrypoints	0.2.2	py27_1	
enum34	1.1.6	py27_0	
flask	0.12.2	pypi_0	pypi
flask-compress	1.4.0	pypi_0	pypi
flask-seasurf	0.2.2	pypi_0	pypi
fontconfig	2.12.1	3	
freetype	2.5.5	2	
functools32	3.2.3.2	py27_0	
futures	3.2.0	py27_0	
get_terminal_size	1.0.0	py27_0	
gevent	1.1.2	pypi_0	pypi
gevent-openssl	1.2	py27_0	chembl
glew	1.13.0	0	mw
greenlet	0.4.12	pypi_0	pypi
grequests	0.2.0	pypi_0	pypi
html5lib	0.999	py27_0	
icu	54.1	0	
idna	2.8	py27_0	
ipaddress	1.0.18	py27_0	
ipykernel	4.5.2	py27_0	
ipython	5.3.0	py27_0	
ipython_genutils	0.2.0	py27_0	
ipywidgets	6.0.0	py27_0	
itsdangerous	0.24	pypi_0	pypi
jinja2	2.9.5	py27_0	
jsonschema	2.5.1	py27_0	
jupyter	1.0.0	py27_7	
jupyter_client	5.0.0	py27_0	
jupyter_console	5.1.0	py27_0	
jupyter_core	4.3.0	py27_0	
libiconv	1.14	0	
libpng	1.6.27	0	
libxml2	2.9.4	0	
linecache2	1.0.0	py27_0	
lxml	3.8.0	pypi_0	pypi
markupsafe	0.23	py27_2	
matplotlib	2.0.0	np111py27_0	
mistune	0.7.4	py27_0	
mkl	2017.0.1	0	
nbconvert	5.1.1	py27_0	
nbformat	4.3.0	py27_0	
nose	1.3.7	py27_1	
notebook	4.4.1	py27_0	
numpy	1.11.3	py27_0	

openbabel	2.4.1	py27_3	openbabel
openssl	1.1.1b	h1de35cc_1	
pandas	0.19.2	np111py27_1	
pandocfilters	1.4.1	py27_0	
path.py	10.1	py27_0	
pathlib2	2.2.0	py27_0	
pexpect	4.2.1	py27_0	
pickleshare	0.7.4	py27_0	
pip	18.1	pypi_0	pypi
pixman	0.34.0	0	
pkgconfig	1.2.2	pypi_0	pypi
plip	1.3.4	pypi_0	pypi
plotly	2.0.10	pypi_0	pypi
pmw	2.0.1	py27_0	mw
prompt_toolkit	1.0.13	py27_0	
ptyprocess	0.5.1	py27_0	
pyasn1	0.1.9	py27_0	
pycosat	0.6.3	py27h1de35cc_0	
pycparser	2.17	py27_0	
pygments	2.2.0	py27_0	
pymol	1.8.0.0.r4144	py27_0	mw
pyopenssl	16.2.0	py27_0	
pyparsing	2.1.4	py27_0	
pyqt	5.6.0	py27_2	
pysocks	1.6.8	py27_0	
python	2.7.6	0	
python-dateutil	2.6.0	py27_0	
pytz	2017.2	py27_0	
pyzmq	16.0.2	py27_0	
qt	5.6.2	0	
qtconsole	4.3.0	py27_0	
rdkit	2016.09.4	np111py27_1	rdkit
readline	6.2	2	
requests	2.5.3	pypi_0	pypi
requests-cache	0.4.13	pypi_0	pypi
ruamel_yaml	0.11.14	py27_1	
scandir	1.5	py27_0	
scikit-learn	0.18.1	np111py27_1	
scipy	0.19.0	np111py27_0	
seaborn	0.8.1	pypi_0	pypi
setuptools	41.0.1	py27_0	
simplegeneric	0.8.1	py27_1	
singledispatch	3.4.0.3	py27_0	
sip	4.18	py27_0	
six	1.10.0	py27_0	
soupsieve	1.8	py27_0	
sqlite	3.13.0	0	

ssl_match_hostname	3.4.0.2	py27_1
subprocess32	3.2.7	py27_0
terminado	0.6	py27_0
testpath	0.3	py27_0
tk	8.5.18	0
tornado	4.4.2	py27_0
traceback2	1.4.0	py27_0
traitlets	4.3.2	py27_0
unittest2	1.1.0	py27_0
unittest2six	0.0.0	py27_0 chembl
urllib3	1.20	pypi_0 pypi
wcwidth	0.1.7	py27_0
werkzeug	0.12.2	pypi_0 pypi
wheel	0.29.0	py27_0
widgetsnbextension	2.0.0	py27_0
yaml	0.1.6	0
zlib	1.2.8	3

A quick inspection of the first column indicated that the *jupyter* package was indeed installed.

7.2.2 Installing Multiple Python Packages at Once

Furthermore, we can also install multiple packages at once as follows:

```
$ conda install jupyter scipy numpy matplotlib scikit-learn
```

7.2.3 Installing Python Packages from Channels

If the package that we want to install is not available in the default conda repository, then we may need to install these packages from channels (i.e., third-party package repositories other than that provided by Anaconda.org containing the package of interest).

For example, let's say that we would like to install the *rdkit* package, then we will call the following commands:

```
$ conda install -c rdkit rdkit
```

where *-c rdkit* represents the *rdkit* channel for which we subsequently call upon the *rdkit* package for installation.

Now that the packages are in place, we are ready to proceed to the next step in running the Jupyter notebooks.

7.3 Tutorial 3: Running the Jupyter Notebook

To launch the Jupyter notebook, open up a terminal window and enter the following commands:

```
$ jupyter notebook
```

If everything went well, the terminal should display the following message, and a new Internet browser window will automatically

appear (as shown in Fig. 2) where the Internet browser will be directed to the URL `http://localhost:8888/tree/`.

```
[I 11:16:10.203 NotebookApp] Serving notebooks from local  
directory: /Users/chanin  
[I 11:16:10.203 NotebookApp] 0 active kernels  
[I 11:16:10.203 NotebookApp] The Jupyter Notebook is running
```

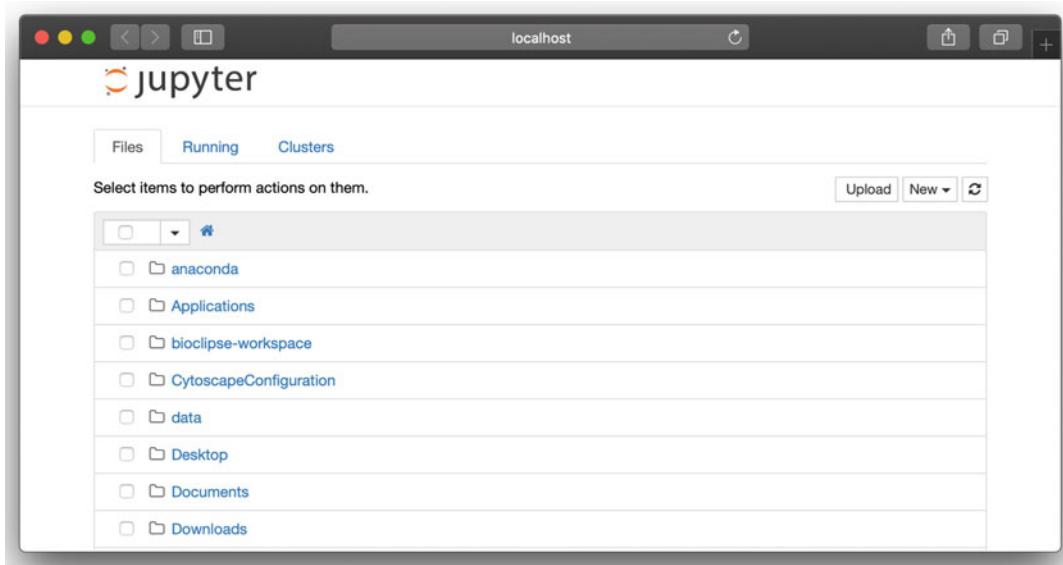


Fig. 2 Screenshot of the default page of Jupyter showing the hard disk content

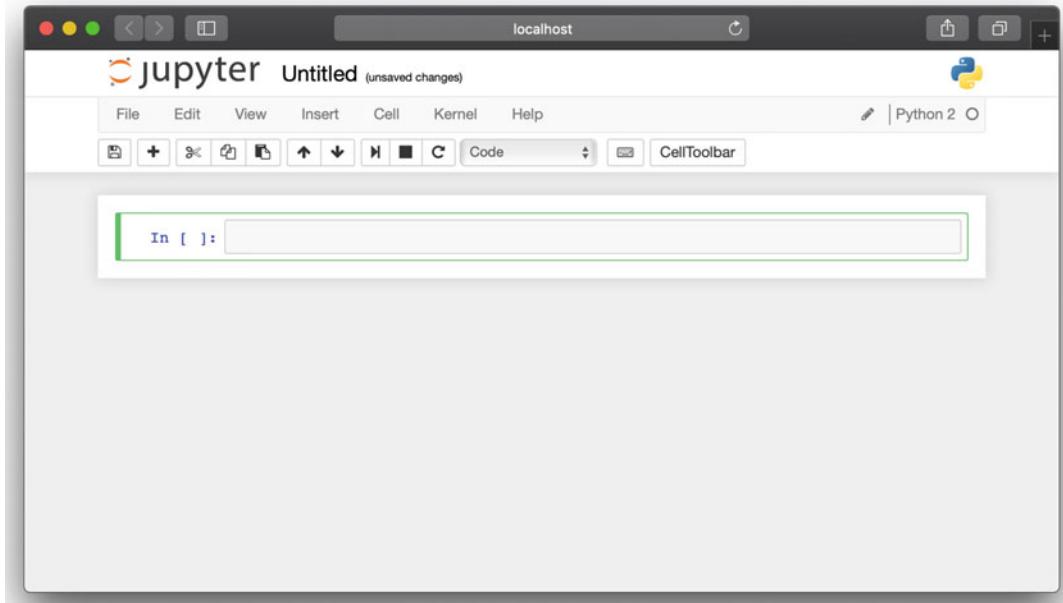


Fig. 3 Screenshot of a newly created Jupyter notebook

jupyter 001_ReadingMolecules1 (autosaved)

File Edit View Insert Cell Kernel Help | Python 2 | CellToolbar

Reading and writing molecules 1

This is a short overview of creating molecules from and writing molecules to various file formats. It is intended to be a complement to, not replacement for, the contents of the [main RDKit documentation](#)

©TAGS: #basics #molecule_input

```
In [1]: from rdkit import Chem
from rdkit.Chem.Draw import IPythonConsole
from rdkit.Chem import Draw
# uncomment this if you try the tutorial and end up with low-quality images
# IPythonConsole.ipython_useSVG=True
```

```
In [2]: import time
print(time.ctime()) # doctest: IGNORE
```

Sun Oct 9 07:11:37 2016

Working with SMILES

If you have a SMILES string, the easiest thing to use is MolFromSmiles:

```
In [3]: m = Chem.MolFromSmiles('COc1ccc(c1)[nH]c(n2)[S@H](=O)C=CCN(c1cc(C)OC)c1')
```

```
Out[3]:
```

Note that the coordinates used for the drawing are not present in the molecule, the RDKit generates them when the molecule is drawn.

Reading Mol file data

```
In [4]: molblock = """phenol
Mrv1682210081607082D
 7 7 0 0 0 0 999 V2000
 -0.6473 1.0929 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 -1.3618 0.6804 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 -1.3618 -0.1447 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 -0.6473 -0.5572 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.0671 -0.1447 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.0671 0.6804 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.7816 1.0929 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 2 1 0 0 0 0
 2 3 2 0 0 0 0
 3 4 1 0 0 0 0
 4 5 2 0 0 0 0
 5 6 1 0 0 0 0
 1 6 2 0 0 0 0
 6 7 1 0 0 0 0
M END
"""
m = Chem.MolFromMolBlock(molblock)
```

```
Out[4]:
```

Here the molecule has coordinates that were read in from the Mol block. We can see this because the molecule has a conformer:

```
In [5]: m.GetNumConformers()
```

```
Out[5]: 1
```

The conformer that is present is 2D (we can see that from the coordinates above):

```
In [6]: m.GetConformer(0).Is3D()
```

```
Out[6]: False
```

Fig. 4 Screenshot of a Jupyter notebook from an rdkit tutorial by Greg Landrum [42] demonstrating how to read and write molecules

```
at: http://localhost:8888/
[I 11:16:10.203 NotebookApp] Use Control-C to stop this
server and shut down all kernels (twice to skip confirmation).
```

A newly created Jupyter notebook file will display an empty cell box (Fig. 3) that can house either the code (i.e., both the input code and their corresponding output) or their descriptive text (in Markdown language). The combined use of code and descriptive text in a Jupyter notebook is the hallmark of this platform as it facilitates easy sharing and comprehension of the code's input and output results in an intuitive and rapid manner. An example of a Jupyter notebook showing the step-by-step procedures of how to read and write molecules using the rdkit package in Python is shown in Fig. 4.

8 Conclusion

The field of QSAR has grown rapidly and has become a pillar of drug discovery and for regulatory purposes owing to its robustness in effectively predicting endpoints of interests as well as providing pertinent insights for model interpretation. In spite of its usefulness, the literature is still predominated by QSAR models that may not be reproducible. As such, this limiting factor hinders future usage of QSAR models especially in situations where the molecular descriptors may not be computed due to updates or changes to the software or simply due to their unavailability. Similar situations may apply if in the future, significant updates to operating systems may render incompatibility issues with the descriptor or multivariate software. In light of these challenges, interactive notebooks together with exported environment file (i.e., containing information on the modules and specific versions used at the time of code runtime) make it possible to share the exact replica of the computing environment from the researcher's own computer to their reader's computer. Furthermore, the emergence of container technologies such as Docker and Singularity (not discussed in this chapter) paves further road in creating a suitable environment for facilitating research reproducibility. It is anticipated that the next generation of data-driven biologists would embrace such technologies as a gold standard or best practice for performing computer-based research.

Acknowledgement

This work is supported by the Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund.

References

1. Nantasesamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2009) A practical overview of quantitative structure-activity relationship. *EXCLI J* 8(7):74–88
2. Nantasesamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010) Advances in computational methods to predict the biological activity of compounds. *Exp Opin Drug Discov* 5 (7):633–654
3. Piir G, Kahn I, Garcia-Sosa AT, Sild S, Ahte P, Maran U (2018) Best practices for QSAR model reporting: physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. *Environ Health Perspect* 126(12):126001
4. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194(4824):178–180
5. Fujita T, Winkler DA (2016) Understanding the roles of the “Two QSARs”. *J Chem Inf Model* 56(2):269–274
6. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M *et al* (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57 (12):4977–5010
7. Sprous DG, Palmer RK, Swanson JT, Lawless M (2010) QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr Top Med Chem* 10 (6):619–637
8. Fjodorova N, Novich M, Vrachko M, Smirnov V, Kharchevnikova N, Zholdakova Z *et al* (2008) Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 26 (2):201–236
9. Garabedian TE (1997) Laboratory record keeping. *Nat Biotechnol* 15(8):799–800
10. Rubacha M, Rattan AK, Hosselet SC (2011) A review of electronic laboratory notebooks available in the market today. *J Lab Autom* 16 (1):90–98
11. Mascarelli A (2014) Research tools: jump off the page. *Nature* 507(7493):523–525
12. Macmillan Publishers Limited (2016) Announcement: where are the data? *Nature* 537(7619):138
13. Celi LA, Citi L, Ghassemi M, Pollard TJ (2019) The PLOS ONE collection on machine learning in health and biomedicine: towards open code and open data. *PLoS ONE* 14(1): e0210232
14. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE (2017) Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ* 5:e3208
15. Greenwald NF, Bandopadhayay P, Beroukhim R (2017) Open data: spot data glitches before publication. *Nature* 550(7676):333
16. Gedeck P, Skolnik S, Rodde S (2017) Developing collaborative QSAR models without sharing structures. *J Chem Inf Model* 57 (8):1847–1858
17. Polanski J, Bak A, Gieleciak R, Magdziarz T (2006) Modeling robust QSAR. *J Chem Inf Model* 46(6):2310–2318
18. Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N *et al* (2017) Towards the revival of interpretable QSAR models. In: Roy K (ed) *Advances in QSAR modeling: applications in pharmaceutical, chemical, food, agricultural and environmental sciences*. Springer International Publishing, Cham, pp 3–55. Available from: https://doi.org/10.1007/978-3-319-56850-8_1
19. Guha R, Willighagen E (2012) A survey of quantitative descriptions of molecular structure. *Curr Top Med Chem* 12(18):1946–1956
20. Grisoni F, Consonni V, Todeschini R (2018) Impact of molecular descriptors on computational models. In: Brown JB (ed) *Computational chemogenomics*. Humana Press, New York, pp 171–209
21. Guha R, Van Drie JH (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model* 48 (3):646–658
22. Sisay MT, Peltason L, Bajorath J (2009) Structural interpretation of activity cliffs revealed by systematic analysis of structure–activity relationships in analog series. *J Chem Inf Model* 49(10):2179–2189
23. Guimarães MC, Duarte MH, Silla JM, Freitas MP (2016) Is conformation a fundamental descriptor in QSAR? A case for halogenated anesthetics. *Beilstein J Org Chem* 12:760–768
24. Pissurlenkar RR, Khedkar VM, Iyer RP, Coutinho EC (2011) Ensemble QSAR: a QSAR method based on conformational ensembles and metric descriptors. *J Comput Chem* 32 (10):2204–2218
25. Wicker JG, Cooper RI (2016) Beyond rotatable bond counts: capturing 3D

- conformational flexibility in a single descriptor. *J Chem Inf Model* 56(12):2347–2352
26. Dearden J, Cronin M, Kaiser K (2009) How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ Res* 20 (3–4):241–266
27. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29(6–7):476–488
28. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* 22(1):69–77
29. Golbraikh A, Muratov E, Fourches D, Tropsha A (2014) Data set modelability by QSAR. *J Chem Inf Model* 54(1):1–4
30. Roy PP, Kovarich S, Gramatica P (2011) QSAR model reproducibility and applicability: a case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. *J Comput Chem* 32(11):2386–2396
31. Svensson F, Aniceto N, Norinder U, Cortes-Ciriano I, Spjuth O, Carlsson L *et al* (2018) Conformal regression for quantitative structure-activity relationship modeling—quantifying prediction uncertainty. *J Chem Inf Model* 58(5):1132–1140
32. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR (2019) Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J Cheminform* 11(1):4
33. Lampa S, Alvarsson J, Arvidsson Mc Shane S, Berg A, Ahlberg E, Spjuth O (2018) Predicting off-target binding profiles with confidence using conformal prediction. *Front Pharmacol* 9:1256
34. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JE (2010) Towards interoperable and reproducible QSAR analyses: exchange of datasets. *J Cheminform* 2:5
35. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J *et al* (2007) Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinfo* 8:59
36. Ruusmann V, Sild S, Maran U (2014) QSAR DataBase – an approach for the digital organization and archiving of QSAR model information. *J Cheminform* 6:25
37. Ruusmann V, Sild S, Maran U (2015) QSAR DataBase repository: open and linked qualitative and quantitative structure-activity relationship models. *J Cheminform* 7:32
38. Ruusmann V, Sild S, Maran U (2012) r-qsardb R package. <https://code.google.com/archive/p/r-qsardb/>
39. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 50(7):1189–1204
40. Fourches D, Muratov E, Tropsha A (2016) Trust, but verify II: a practical guide to chemogenomics data curation. *J Chem Inf Model* 56 (7):1243–1252
41. Fourches D, Muratov E, Tropsha A (2015) Curation of chemogenomics data. *Nat Chem Biol* 11(8):535
42. Landrum G (2016) Reading and writing molecules 1. https://raw.githubusercontent.com/greglandrum/rdkit-tutorials/master/notebooks/001_ReadingMolecules1.ipynb



Chapter 4

Wildlife Sentinels for Human and Environmental Health Hazards in Ecotoxicological Risk Assessment

Antonio Juan García-Fernández, Silvia Espín, Pilar Gómez-Ramírez,
Emma Martínez-López, and Isabel Navas

Abstract

Can animals reflect human and environmental health risks? This is a frequently asked question in the research community. Sentinel species are organisms that can provide early warning signs of potential risks to humans, so that preventive measures can be taken in time to avoid serious adverse health consequences. In spite of the well-known cases of use of sentinel species, animals are continuously offering information that in most cases is underestimated or incorrectly interpreted. Many species may be successfully used as sentinels or monitors of environmental and health hazards; however, there is no ideal species for all types of scenarios and conditions. For this reason, the advantages and disadvantages on the use of potential sentinel species and the main characteristics they should gather to be effective sentinels are discussed. Although a wide range of wildlife species are considered good candidates for biomonitoring purposes, bird species are especially suitable as biomonitoring of environmental quality and to estimate human health risks.

During the last two decades, the effects induced by endocrine-disrupting chemicals (EDCs) on hormone action have been widely tested. Since the mid-twentieth century, it is well-known that humans and wildlife species are simultaneously exposed to multiple chemicals from multiple sources with potential ability to disrupt the endocrine system by different pathways and/or interfere with hormone actions. Moreover, additive effects related to this chemical cocktail exposure could be expected, increasing the potential risks to animal and human health. In addition, carcinogenic, immunotoxic, neurotoxic, behavioral, and other chronic effects are observed in wildlife, which are closely linked to human diseases.

Key words Wildlife, Sentinel animals, Ecotoxicological risk, Free-living animals, Human health, Environmental health, Biomonitoring, Endocrine disruption, Cancer, Immunotoxicity, Neurotoxicity, Behavioral effects

1 Animals Can Reflect Human and Environmental Health Risk: Background

Sentinel species are organisms that can provide early warning signs of potential risks to humans, so that preventive measures can be taken in time to avoid serious adverse health consequences.

Industrial activities, chemical use, and pollution have been linked to different contaminant-related diseases, but also to unspecific alterations on both human and animal health. Some historical

events, especially those related to chronic exposure at sublethal doses, were not recognized until human beings became affected. Likely, the best-known case of animal usage as sentinels is the use of canaries in coal mines. In the early twentieth century, the Bureau of Mines (Department of the Interior, Washington CD), issued a circular regarding the use of small birds in coal mines to detect the presence of carbon monoxide (CO) in the air, due their higher sensitivity to CO toxic effects compared to humans, cautioning that the breathing equipment should be used wherever the birds showed signs of distress [1]. Throughout the twentieth century, several alerts and disasters have shown that an appropriate observation of animal species and populations could have minimized deleterious effects on humans and the environment. In the UK, at the end of the nineteenth and in the middle of the twentieth century, cows died due to smog, a cause of death in humans, especially in persons with chronic respiratory problems and elderly people. During the second half of the twentieth century, the relationship between exposure to organochlorine pesticides (i.e., DDT, DDE, etc.) and eggshell thinning in endangered raptor species due to endocrine disruption was widely reported, and later, endocrine alterations were also evidenced in humans. Particular consideration merits the neurological alterations observed incidentally in cats in Minamata (South of Japan), while a neurological human epidemic was being investigated in the population of the bay. The symptoms appeared in cats and other wild species before than in humans, but they were not appropriately interpreted regarding the potential similar effects in humans. The final diagnosis was intoxication due to the consumption of fish containing high levels of methylmercury, a compound discharged by a chemical company into the bay. This would probably be the best-known case of underappreciated use of sentinel animals. In the same sense, Fox [2] suggested that the prevalence, occurrence, and severity of health effects in wildlife populations were being underestimated, because individuals observed were only the survivors.

To be effective sentinel species, animals should present some characteristics [3, 4]:

1. They should be sensitive to the pollutant of concern at levels relevant for the risk assessment.
2. They should provide measurable and interpretable health effects.
3. They must show the measured changes or effects before they are produced in humans.
4. Their ecology and biology should be well known to adequately interpret contaminant levels.
5. They should be able to be monitored and captured and tolerant to handling and sampling in a cost-effective way.

6. They should ideally share the habitat with other animal species and humans, so they are exposed to the same pollutants and other stressors.

Using sentinel wildlife species as indicators of human and environmental health hazards provides several advantages [5]:

1. Sentinel animals normally present physiological mechanisms similar to humans, sharing exposure/effect biomarkers, which allows comparisons and interpretations of how responses in animals reflect human health risks (e.g., acetylcholinesterase inhibition due to organophosphorus pesticide exposure).
2. Effects on a wide range of systems (i.e., reproductive, behavioral, carcinogenic, immune effects, etc.) can be readily observed in animal populations in response to pollutant exposure.
3. Some sentinel species are sensitive to environmental changes in their habitat and can act as an early warning system of potential risks before the effects are developed or found in human or in other animal species.
4. Exposure conditions can be similar under some circumstances (e.g., people living in the same habitat than some wildlife species and feeding on similar food sources).
5. Practical and ethical issues make it easier to collect samples from some sentinel animal species than from protected species and humans.

However, there are also some disadvantages or barriers on the use of sentinel species for evaluating human and environmental health [5]:

1. Further information on the ecology and biology of some species is needed.
2. Sampling methods and analytical techniques used in wildlife sentinel species should be standardized, as well as the reports and information provided.
3. A database compiling and organizing data on effects in different sentinel species would be needed to facilitate the access to the information and coordinate studies.
4. Exposure route for some sentinel species may be not relevant for other animal species or humans.
5. A good understanding of the mechanism of toxicity for the chemical of concern in both sentinel species and humans is needed.
6. Extrapolation of wildlife data to humans may pose some difficulties depending on the sentinel species, the exposure route, the biomarker observed, and the dose-response relationship.

7. Appropriate control sites are essential to know the background levels of contaminants or effect biomarkers in sentinel organisms.
8. Frequent and appropriate communication and coordination between researchers and regulatory agencies should be improved to incorporate nontraditional data on wildlife sentinels into the decision-making process.
9. The development of new approaches, standardization, validation, and coordination will imply additional costs.

When human data is available, it is usually preferable to data coming from sentinel species or laboratory animals to evaluate risks from pollutants. However, human data are rarely available due to low sample size, confounding variables, or other reasons.

2 Biomonitoring Studies in Wildlife

A wide range of wildlife species could be used for biomonitoring purposes, and many of them have been proposed to be particularly valuable. Since there is a vast knowledge of many bird species, they are especially suitable as biomonitorors of environmental quality and to estimate human health risks [6]. Other wildlife species have been also successfully used as sentinels in specific geographical areas and for some effects (e.g., marine mammals, polar bears) [7]. In order to carry out a biomonitoring scheme using wildlife species, it is crucial to know the specific objectives to be achieved within each scheme. Collecting data about spatial and temporal trends of the environmental contaminant exposure and their potential effects is one of the most relevant objectives [8]. These trends are of special concern regarding the exposure to endocrine disrupting chemicals (EDCs), since, in most of the cases, the effects on the endocrine system are critical in human and wildlife health risk assessment. Therefore, biomonitoring studies are essential to compile information on contaminant exposure trends in different scenarios and for different risk groups. Wildlife biomonitoring studies provide valuable data that can be used in the risk assessment process for the species and populations under study, but also for other species of interest with similar feeding habits and for humans inhabiting the same area. In addition, biomonitoring studies allow the detection of spatiotemporal changes in contaminant exposure [8, 9], which can be used to track if legislative measures prohibiting or restricting the use of specific compounds or their emission are being successfully applied [10–13].

2.1 Endocrine Disruption: Reproductive and Development Effects

In 2002, the WHO/IPCS defined “endocrine disruptor compound” (EDC) as an exogenous substance or mixture that possesses properties that might be expected to lead to endocrine disruption in an intact organism, or its progeny, or (sub)populations [14]. More recently, the Endocrine Society statement defined endocrine disruptor as “an exogenous chemical, or mixture of chemicals, that can interfere with any aspect of hormone action” [15], highlighting the differences between endocrine function and hormone action. For its part, the European Commission (EC) defined endocrine disruptors as “chemicals, which under certain conditions can impact on the hormonal system of humans and animals”. However, according to Zoeller et al. [15], the definition assumed by the WHO/IPCS will probably not change. The majority of chemical compounds considered as endocrine disruptors are man-made chemicals used in a huge variety of processes, goods, and materials.

In the middle of the twentieth century, mimetic effects on hormone action were related to chemical substances present in foodstuffs for livestock. During the 1970s, certain chemicals were found to be associated to cancer and reproductive effects in both humans and wildlife species [16]. In regard to wildlife, the best-known examples were the population decline of bald eagles (*Haliaeetus leucocephalus*), double-crested cormorants (*Phalacrocorax auratus*), and herring gulls (*Larus argentatus*) in the Great Lakes, in the 1950s and 1960s (see [2]). In these three species, the reproductive failures were associated with the exposure to dichlorodiphenyl aliphatic compounds, concretely with the pesticides DDT and DDE. The exposure to these organochlorine pesticides was also associated with effects such as eggshell thinning above 20%, embryo toxicity, and hatching failures. However, reproductive effects were not the only adverse effects observed; an extensive adult mortality was also described due to the neurotoxic effects induced by these pesticides. Together with the reproductive impairments, congenital malformations had also been detected in the last 30 years in fish-eating birds nesting on the Great Lakes.

In July 1991, Theo Colborn (1927–2014) organized at the Wingspread Conference Center in Racine (Wisconsin, USA), a conference whose main purpose was to integrate and evaluate findings about the problem of endocrine disruptors in the environment. Colborn brought together experts in fields as diverse as anthropology, comparative endocrinology, immunology, medicine, law, psychiatry, ecology, histopathology, reproductive physiology, toxicology, wildlife management, tumor biology, zoology, mammalogy, and psychoneuroendocrinology to share their knowledge on endocrine disruption from their respective research disciplines. A consensus statement was reached by participants entitled “Chemically-induced alterations in sexual development: The Wildlife/Human Connection”. A few years later, in December 1996, in Weybridge (UK), the EC sponsored the first international meeting

(*The Weybridge Conference*) entitled “European workshop on the impact of endocrine disruptors on human health and wildlife.” The conclusions were reflected in the well-known “Weybridge Report”. At the beginning of the first decade of the twentieth century, Fox [2] addressed the issue in his interesting review on wildlife as sentinels of human health effects in the Great Lakes, paying special focus on the alterations on endocrine function. Fifteen years after the Weybridge Conference, in 2011, the EC reviewed the advancement of scientific knowledge on EDCs and recorded their conclusions in “The impacts of endocrine disrupters on wildlife, people and their environments: Weybridge+15 (1996–2011) Report.” This report established the priority future research lines: studies on EDCs in humans and wildlife, mechanisms of action of EDCs, exposure (analytical methods) and measurement of effects, development of in vivo and in vitro test methods, and establishment of monitoring programs. In this sense, clearly, wildlife should be used as sentinels for human risks.

In the last two decades, effects induced by EDCs on hormone action have been widely tested, and nowadays, close to 800 substances have been proven to cause endocrine alterations. In spite of this, only a small fraction of the chemicals with potential endocrine activity are currently known, which introduces significant uncertainties about the real extent of the problem in both human and wildlife populations [17].

Endocrine disruption is a special type of toxicity which, according to Bergman et al. [17], must be taken into account in both ecological and toxicological risk assessments. It should not be forgotten that mechanisms of action on hormone receptors respond to mimetic or antagonistic effects, while, in occasions, they act directly on proteins controlling hormone delivery to cell or tissue targets.

Humans and wildlife species are simultaneously exposed to multiple chemicals with potential ability to both disrupt the endocrine system by different pathways and/or interfere with hormone actions. In any case, additive effects could be expected, increasing the potential risks to their health.

“Human and wildlife health depends on the ability to reproduce and develop normally which is not possible without a healthy endocrine system” is the first of the key concerns included in the report published by the United Nations Environment Programme (UNEP/WHO) entitled “State of the Science of Endocrine Disrupting Chemicals-2012” [17]. In the last decade, the role of the EDC exposure to induce, promote, or increase the prevalence of alterations on the endocrine function during critical periods of childhood development and growth has been underestimated [17]. The scientific studies in experimental animals, together with an increasing number of studies on both wildlife biomonitoring and human epidemiology, are allowing to achieve a better scientific

understanding of the relationship between exposure to EDCs and their effects on wildlife and human health.

The first studies on EDCs in wildlife were mainly focused on female reproductive effects induced by the exposure to polychlorinated biphenyls (PCBs), DDT, dioxins, phthalates, etc., but also on males (diethylstilbestrol (DES), bisphenol A-BPA) and on the sex ratio described in wild fish and shellfish. In the early 1990s, a large number of scientific papers on wild marine and terrestrial mammals, fish, and birds connected reproductive impairments with exposure to PCB congeners and other persistent compounds. A lot of information on endocrine effects and hormone action has been related to disorders in the thyroid function (PCBs, polybrominated diphenyl ethers-PBDEs) and in the neurodevelopmental system (lead, mercury, PCBs, etc.), inducing behavioral and cognitive alterations during embryonic and postnatal periods. Moreover, other endocrine disorders have been described, including hormone-related cancers in both female and male reproductive organs (PCBs, dioxins, some pesticides, cadmium, arsenic, etc.), immune dysfunction (polycyclic aromatic hydrocarbons (PAH), PCBs, BPA, etc.), or functional disorders in both adrenal glands and hypothalamic–pituitary–adrenal axis (DDTs, PCBs). Other less studied effects include alterations in bones affecting mineral density and increasing bone fractures (persistent organic pollutants (POPs) in general) or metabolic disorders such as obesity or diabetes (BPA, PCBs, dioxins, arsenic, phthalates), which have been of concern mainly in human health and rarely in wildlife. On the contrary, since the 1950s, many studies of wildlife species have been focused on population declines [17]. The hierarchical organization of ecotoxicology [18] allows us a better understanding of the consequences of wildlife populations declines on communities and ecosystems. However, the origin of these declines usually finds their explanation at the lowest levels of this hierarchical organization, where the chemicals interact with biomolecules or subcellular and cellular fractions, starting a cascade of molecular, biochemical, structural, or systemic effects upon which the survival of the individuals and populations depends. Frequently, wildlife population declines are the direct consequence of reproductive failures, and in most cases, they have been related to the exposure to EDCs. In the same way, thousands of studies have been focused on the potential effects of EDCs on human health, and most of them paid special attention on the association between reproductive failures and the exposure to EDCs [19]. However, they also inform that there are many other studies in which such associations were not observed, so uncertainties about the effects of background levels of EDCs on human reproduction must be taken in account, especially in ecotoxicological risk assessments.

2.2 Research Challenges on EDCs in Human and Wildlife

In spite of the increasing knowledge on EDCs and their effects on human and wildlife health, nowadays there are still many gaps and questions to resolve. Shug et al. [16] reviewed the top-priority areas in the future research on EDCs. The response of organisms to chemical mixtures exposure is one of the main questions requiring answer. All living beings are exposed simultaneously to several compounds or group of compounds, with or without known endocrine effect when they act individually, but little is known about the potential effects due to the interaction among them and about the synergistic or antagonistic response of the organisms exposed. This issue is closely related to the similar or different mechanisms of action of the EDCs present in the environment. In this connection, Shug et al. [16] suggested that additional understanding on the properties allowing a chemical to mimic hormone action and their mechanisms of action is needed, especially, in regard to latent effects. In addition, reports of nonmonotonic (e.g., U-shaped) and low-dose effects and no threshold effects for EDCs continue to be a challenge in chemical risk assessment [19, 20]. Other concerns with relevant gaps of knowledge are the windows of susceptibility. In accordance with Grandjean et al. [21], under certain circumstances, organisms can be more sensitive to the toxicity of chemicals. In this line, although much information is available about the increased susceptibility to EDC effects during critical periods as gestation, infancy, or puberty, there remains much to learn about other potential susceptibility windows and how the exposure to EDCs during these periods can induce health effects on human and wildlife. Associated to it, much remains unknown about development effects and how environmental factors can influence phenotypes. In this sense, minimal attention has been paid to cell-to-cell interactions, in spite of their crucial role to determining the phenotypic changes observed during fetal exposure to EDCs [16]. Other gaps needed to be covered are related to the potential health effects of EDCs that have not been studied in detail yet, for example, the relationships between EDCs and the effects in cardiovascular system, nervous system, obesity, metabolic syndrome, diabetes, bone development, etc. [22]. Also, there are substantial gaps of knowledge related with exposure scenarios and biomonitoring of EDCs. In this sense, much more needs to be known about the persistence of EDCs in organisms' bodies (human and wildlife) and especially on the exposure in different geographic locations and across socioeconomic and ethnic status, etc. To achieve these aims, it is necessary to improve the methods for identifying EDCs. Schug et al. [16] suggested to improve the detection of chemicals and metabolites in the urine, cord blood, and other tissues of animals with known EDC-associated abnormalities, such as reduced anogenital distance. However, although it is necessary to improve our knowledge on the presence of EDCs in living organisms, it is as important—if not more—to find predictive

biomarkers able to relate the exposure to EDCs with health effects in a shorter period of time than for epidemiological studies [23]. Last but not the least, translation of animal research to human is, from the beginning of the research on EDCs, one of the most outstanding challenges. Comparable kinetics and mechanisms of action have been described in human and wildlife, but it is necessary to better understand how to translate the effects observed in animal models and especially in wildlife to human exposure to EDCs. Therefore, in accordance with Shug et al. [16], new studies are needed to improve the translation of results obtained with laboratory animals and wildlife to benefit humans.

2.3 Carcinogenic Effects

Animal populations inhabiting different environments and exposed to different stressors may be used as natural “experiments” to study the impact of pollutants on cancer in wildlife and identify potential risks to humans. However, little attention has been given to carcinogenic effects on wild animals.

Some xenobiotics have a direct role in cancer development, inducing somatic mutations and disrupting oncogenes (e.g., *RAS*) or tumor suppressor genes (e.g., *TP53*). In addition, certain compounds may induce oxidative stress and oxidative damage, inflammation processes, telomere shortening, and epigenetic effects as DNA methylation that may influence cancer development. Along with this, contaminants such as organochlorines and dioxins are also immunosuppressive and may increase the susceptibility of animals to oncogenic pathogens. Therefore, it is possible that pollutants operate not only as mutagens but by several mechanisms ending up in cancer development.

Different cancer types have been described in wildlife species associated to contaminant exposure. Some examples are gastrointestinal adenocarcinomas in beluga whale (*Delphinapterus leucas*) related to PAH exposure, lymphoma in bottlenose dolphin (*Tursiops truncatus*) associated with PCBs, hepatic and biliary neoplasia in white suckers (*Catostomus commersonii*) related to EDCs, and neoplasm (predominantly a poorly differentiated carcinoma of urogenital origin) in California sea lions (*Zalophus californianus*) associated with PCBs and other organochlorines. In this sense, exposure to PAHs, organochlorines, and, in general, to EDCs has been associated with the development of similar tumors in humans [24].

Although evidences of cancer have been reported in mammals and several fish species, they are not frequently found in wild birds, with neoplasms only observed in 9 of ca. 18,000 free-living avian species examined at the US National Wildlife Health Center from 1975 to 1981, probably related to genetic or viral etiology [25]. Animals generally have a shorter life span in the wild due to food or nutrient restrictions, diseases, predation, adverse weather conditions, and other parameters of stress affecting survival. In

general, the risk of cancer development increases with age. Thus, all the potential factors affecting life expectancies in free-living animals could lead to lower cancer incidence in wild animals compared to their captivity counterparts. In addition, cancer rates could be expected to be higher in some species of marine mammals and fish that, because of their longevity, are exposed to and affected by xenobiotics for longer time periods.

The case of the beluga whales in St. Lawrence Estuary, Quebec (Canada), is particularly relevant considering the unusually high occurrence of malignant neoplasms [26]. From 1983 to 2012, 39 malignant neoplasms were diagnosed in 35 belugas, causing the death of 31 mature adults (>19 years old) [26]. The median age of cancer diagnosis was 48 years, and the most frequently observed neoplasm was adenocarcinoma of the gastrointestinal mucosa (11 cases), followed by mammary carcinoma (8 cases). Different issues related to changes in local aluminum production in the area, the consequent decrease of PAH concentrations in the environment, and the decrease of cancer occurrence in this population in the following years support a possible link between exposure to industrial PAHs and cancer in belugas from this study [26]. In addition, most of the cancers had a gastrointestinal origin, which has been linked with belugas habit to feed on invertebrates living in sediments and accumulating PAHs. This association is also supported by the hepatic neoplastic lesions found in fish exposed to PAHs in the same area [27, 28]. Interestingly, associations between occupational exposure to PAHs and cancer (lung, bladder, laryngeal, and stomach) in workers of the aluminum plants located in the beluga habitat have been documented [29–31]. Therefore, these findings linking contaminant exposure to cancer in wild animal species and humans are of particular interest.

Carcinogenic effects due to pollutant exposure in wildlife are understudied and rarely described, and further research in a long-term basis and with large sample sizes is needed to better understand the dose and type of xenobiotics at which wildlife species are exposed and the associated cancer incidence across free-living species. In spite of the link reported between contaminant exposure and cancer in some wild animals, it is difficult to prove the etiologic role of pollutants in carcinogenesis in wildlife studies, and in many cases the cancer development may be multifactorial. Since some carcinogen compounds are known to cause specific gene mutation patterns, the study of these mutations in tumors of wild animal species and humans in future studies may support the role of specific xenobiotics in carcinogenesis.

2.4 Behavioral and Neurotoxic Effects

Neurotoxicity-related diseases are a significant percentage of those affecting human health. Environmental contaminants can influence in behaviors, learning, and other cognitive abilities using various mechanisms. Persistent organic pollutants (POPs, including some

pesticides), heavy metals, and pharmaceuticals classified as neurotoxicants are in the environment together with thousands of compounds with unknown neurotoxic potential to different species and life stages. It is thought that 30% of all products which are commercialized can have neurotoxic potential. Some of these chemicals may affect early ages of human and wildlife, increasing the incidence of neurodevelopmental disorders and affecting the life quality of future population. It is important to take into account that the exposure in early life stages to neurotoxic compounds may impact in adult phenotype, so exposure during pre- or postnatal period has been linked with impairment of human and animal behavior followed with neurodegenerative diseases in adults [32, 33]. Nevertheless, knowledge about the neurotoxic potential of pollutants in ecosystems is scarce. According to Legradi et al. [34], there is a clear lack of developmental neurotoxicity assessment studies. Additionally, a 2009 report indicated that little more than 100 compounds had been checked for potential human developmental neurotoxicity [35].

Although most of this scarce information concerning neurotoxic effects of environmental pollutants comes from studies in laboratory animals or in vitro techniques, the use of wild animals in neurotoxicity studies can reflect the real exposure, which occurs in an environment of genetic diversity, with different stressors and with all interactions that occur in the reality. Wildlife, such as bald eagles, pigeons, mink, and polar bears, can be used as sentinel species due to their susceptibility to bioaccumulate neurotoxicants from the local environment [2]. Thus, for decades, the Arctic Monitoring and Assessment Programme and related programs have monitored the health of Arctic wildlife and humans, spatial and temporal trends, and human exposure using ringed seals and polar bears as key monitoring species [36]. Studies on brain tissue of polar bears corroborated or presumed that several POPs detected in them are neurotoxicants in humans and experimental animals (see [36]).

The behavioral effects of pollutants are particularly difficult to observe in field studies. However, a retrospective analysis carried out in the late 1990s already gave evidence that a big proportion of dead birds found in the field would have suffered sublethal neurotoxic effects by cyclodienes. These compounds could have produced behavioral disturbances, which could have caused the decline of these species [37]. In epidemiological studies, the exposure to organochlorine compounds (OCs, widely used for agricultural in the last century) in pregnant mothers produced a deterioration of neurodevelopment and postnatal neuropsychological faults, such as diminished motor functions, bare cognitive development, lack of attention and alteration of activity, and autism, and also an augmented risk of serious chronic illnesses. Besides, an increasing evidence indicated that OCs exposure is

connected to increased risk of several neurodegenerative disorders including dementia, Alzheimer's disease and others [38]. In studies carried out in the Great Lakes, St. Lawrence Basin, behavior and neurobehavioral development changes were similar among minks, herring gulls, and other wildlife species and infants born from mothers feeding on fish from the area. This similarity has been found in the distribution pattern of pesticides in blood samples of raptors and humans and egg and human milk samples [12, 39, 40]. On the other hand, many studies have shown that sublethal doses of organophosphorus compounds can cause behavioral effects in birds, and these effects have been often related to cholinesterase inhibition.

Besides, some studies have evaluated the relationships among toxicant exposure, mainly to heavy metals and pesticides, and alterations in several neurochemical biomarkers such as monoamine oxidase, cholinesterase, muscarinic acetylcholine receptor, and dopamine-2 in both wildlife and human studies at a population level. These changes in biomarkers can be used in the early stages of neurotoxicity in all species (human and wildlife). In this sense, Stamler et al. [41] show that the neurotoxicity biomarkers aforementioned can be detected in wild mink and human samples, in good storage conditions, and can be used as early biomarkers. These changes in neurochemical biomarkers that precede other behavioral changes are fundamental for the survival of species.

2.5 Immune Effects

The relation between xenobiotics and detrimental effects on the immune system has been known even since ancient Egypt, when workers exposed to asbestos suffered lung diseases [42]. However, from a historical point of view, the field of "immunotoxicology" is considered quite recent as the term was firstly coined in the 1970s [42] and laboratory studies that relate toxic compounds to immune effects were not formally established until the late 1980s [43]. The reason to include these experimental assays for risk assessment was the consideration of effects in the immune system as a mode of action and not a side effect of the substances, mainly because the concentrations that cause immune disruption are much lower than those related to other toxicological endpoints such as mortality [42].

Toxic compounds can alter immune system at different levels in complex ways, which complicates the assessment of immunotoxicity [44].

1. Immunotoxicity should be distinguished between direct (interaction between the chemical and the immune cells) and indirect (due to a systemic stress response to the chemical).
2. The effects of xenobiotics may not be evidenced in the resting immune system, but the response to pathogens may be compromised.

3. The response of the immune system may be either increased (causing autoimmune diseases or allergies) or decreased (increasing the risk for infections and diseases), so the interpretation of the toxicant-induced immunomodulation is complicated.
4. The potential targets and effects at immune system are diverse, from molecular to cellular and organ level (i.e., immune cell proliferation, differentiation and survival, functioning of the immune organs and cells). In addition, the immune system network is highly sophisticated, with varied signal transduction pathways, multiple cellular components, and a diversity of mediators and receptors for communication and activation.
5. Alterations may be either transient or persistent, as they may be evidenced both in the mature immune system as well as in the developing immune system.

In addition, the assessment of risks for immunotoxic contaminants in wildlife is complicated due to the lack of dose-response data [45]. However, the impacts of environmental contaminants in the immune system of these species are currently acknowledged and of increasing concern. Some of the most relevant examples include global decline of populations, such as amphibian species, due to parasite infections that seem to be favored by toxicant-induced immunosuppression [46]. In the case of marine mammals, OCs such as PCBs have been related to wildlife diseases such as the distemper virus outbreak in harbor porpoises [47]. In a recently published overview, Desforges et al. [45] combined field and laboratory data to establish effect threshold levels for immune suppression in polar bears and several pinniped and cetacean species. The assessment of contaminant exposure in these species can be considered a useful tool for risk assessment as they are considered to be the most exposed to high levels of pollutants of all wildlife [45]. Birds have also been often affected by immunotoxic compounds. In fact, one of the first studies in immunotoxicology was about the relation of increased susceptibility to hepatitis virus in young mallards (*Anas platyrhynchos*) exposed to PCBs [48]. Further studies in different bird species have evidenced immune alterations due to different contaminant types, mainly metals and OCs [49].

Abundant literature describes the relation of immunotoxicity with perfluorinated compounds (PFCs), another relevant group of contaminants quite frequent nowadays in human and wildlife. Some of these effects include decreased spleen and thymus weights and cellularity, reduced specific antibody production, reduced survival after influenza infection, and altered cytokine production. It should be remarked that these effects have been evidenced in experimental animals at doses within or just above the range for highly exposed humans and wildlife [50].

Other compounds that are involved in the alteration of the immune system are the 4-alkylphenol ethoxylates (APEOs), a group of surfactants included in many cleaning formulations and used as industrial process aids. Monitoring studies have found APEO metabolites in many environmentally relevant matrices including human tissues. Generally, related immunotoxic effects (i.e., increase in IgE and antigen-specific IgG and aggravation of atopic dermatitis-like skin lesions and asthma in mice) were found at higher concentrations than those found in the environment, although recent studies have shown that 4-*tert*-octylphenol (OP) is able to cause an immune response in human macrophage-like THP-1 cell [51].

As mentioned above, the immune system may be affected indirectly, and here is where the endocrine system may play an important role. The endocrine system facilitates interorgan communication by steroid and protein hormones. Organs of the immune system are also regulated by hormones. Therefore, any disruption affecting these hormones or their receptors may consequently alter the immune response of the individuals. Briefly, *in vivo* corticosteroids, androgens, progesterone, and adrenocorticotropic hormone (ACTH) suppress the specific immune responses, whereas prolactin, growth hormone, insulin, and thyroid hormones enhance it [52].

2.6 Other Chronic Effects

Other adverse effects than those described previously (cancer, immunodepression or immunotoxicity, neurotoxicity, behavior alterations, or endocrine disruption, including reproductive and developmental effects) have never been widely described. In most cases, other chronic effects often go unnoticed. However, it is sometimes possible to find in animals effects rarely studied associated to chemical exposure in humans. This is the case of a study on dogs living in Mexico City, which presented histologic images of neuronal inflammation and an increased amount of messenger ribonucleic acid (RNAm) from two inflammatory genes in the brains. Similar findings could be related to lower scores on psychometric tests in children living in a similar air-polluted environment [53].

In susceptible species, certain pollutants are able to inhibit specific enzymes involved in the hemosynthesis, provoking the accumulation of porphyrins, such as uroporphyrin and other highly carboxylated porphyrins [2]. Metabolic disorders have been described as consequence of the chronic dietary exposure to pollutants. In this regard physiological synthetic or degradative processes could suffer severe alterations as consequence of this type of exposure. Several piscivorous and carnivorous bird species showed glucose intolerance due to lack of glucokinase, which is responsible for hepatic clearance of glucose. Similarly, adult individuals of herring gulls inhabiting in the Great Lakes showed a mild hyperglycemia

(not related to body condition or stress) compared to the birds from the control areas [2]. This author also described a wasting process with loss of muscle mass in chicks of terns from the Great Lakes associated to exposure to PCBs. Wasting, characterized by loss of body mass, had been related to the exposure to TCDD [54].

3 Overview

It is well known that humans and wildlife species are simultaneously exposed to low doses of chemical mixtures potentially able to interfere with hormone activities, disrupt the endocrine system, impair the immune system, or induce cancer or other chronic effects.

Although wildlife species are continuously offering information about environmental and health risks, their use as sentinels has been frequently underestimated or incorrectly interpreted. Comparable kinetics and mechanisms of action have been described in human and wildlife, but a better understanding on how to extrapolate the effects observed in animal models and wildlife to humans is still needed.

The scientific community must still assess a number of factors: How are organisms, species, or populations responding to low-dose exposure to chemical mixtures? How do different development periods, environments, or social factors affect chemical-related effects? Are we trained to obtain appropriately all findings that experimental, wildlife biomonitoring and human epidemiology studies offer us in each of the possible scenarios of chemical exposure? And finally, are we properly skilled to interpret them adequately under an integrative wildlife-human approach?

References

- Burrell GA, Seibert FM (1916) Gases found in coal mines. *Miners' Circular* 14. Bureau of Mines, Department of the Interior, Washington, DC
- Fox GA (2001) Wildlife as sentinels of human health effects in the Great Lakes-St. Lawrence Basin. *Environ Health Perspect* 109(suppl 6):853–886
- Halliday JE, Meredith AL, Knobel DL, Shaw DJ, Bronsvoort BM, Cleaveland S (2007) A framework for evaluating animals as sentinels for infectious disease surveillance. *J R Soc Interface* 4:973–984
- Stephen C, Ribble C (2001) Death, disease and deformity; using outbreaks in animals as sentinels for emerging environmental health risks. *Glob Change Hum Health* 2:108–117
- van der Schalie WH, Gardner HS, Bantle JA, De Rosa CT, Finch RA, Reif JS, Reuter RH, Backe LC, Burger J, Folmar LC, Stokes WS (1999) Animals as sentinels of human health hazards of environmental chemicals. *Environ Health Perspect* 107:309–315
- García-Fernández AJ (2014) Avian ecotoxicology. In: Wexler P (ed) *Encyclopedia of toxicology*, 3rd edn, vol 2. Amsterdam Academic Press, Elsevier, pp 289–294. ISBN: 9780123864543
- Reif JS (2011) Animal sentinels for environmental and public health. *Publ Health Rep* 126:50–57
- García-Fernández AJ, Calvo JF, Martínez-López E, María-Mojica P, Martínez JE (2008) Ecotoxicology of Raptors in Spain: a review of

- persistent environmental contaminants. *Ambio* 37:432–439
9. Espín S, García-Fernández AJ, Herzke D, Shore RF, van Hattum B, Martínez-López E, Coeurdassier M, Eulaers I, Fritsch C, Gómez-Ramírez P, Jaspers VL, Krone O, Duke G, Helander B, Mateo R, Movalli P, Sonne C, van den Brink NW (2016) Tracking pan-continental trends in environmental contamination using sentinel raptors—what types of samples should we use? *Ecotoxicology* 25 (4):777–801
 10. García-Fernández AJ, Romero D, Martínez-López E, Navas I, Pulido M, María-Mojica P (2005) Environmental lead exposure in the European kestrel (*Falco tinnunculus*) from southeastern Spain: the influence of leaded gasoline regulations. *Bull Environ Contam Toxicol* 74:314–319
 11. Gómez-Ramírez P, Shore RF, van den Brink NW, van Hattum B, Bustnes JO, Duke G, Fritsch C, García-Fernández AJ, Helander BO, Jaspers V, Krone O, Martínez-López E, Mateo R, Movalli P, Sonne C (2014) An overview of existing raptor contaminant monitoring activities in Europe. *Environ Int* 28 (2):300–306
 12. Martínez-López E, María-Mojica P, Gómez-Ramírez P, Calvo JF, Martínez JE, García-Fernández AJ (2012) DDT residues in breeding population of booted eagle (*Aquila pennata*) associated with agricultural land practices. In: Jokanovic M (ed) The impact of pesticides. AcademyPublish.org, Cheyenne, pp 321–338. ISBN: 978-0-9835850-9-1
 13. Valverde I, Espín S, Navas I, María-Mojica P, Gil JM, García-Fernández AJ (2019) Lead exposure in common shelduck (*Tadorna tadorna*): tracking the success of the Pb shot ban for hunting in Spanish wetlands. *Regul Toxicol Pharmacol* 106:147–151
 14. Damstra T, Barlow S, Bergman A, Kavlock RJ, van der Kraak G (eds) (2002) Global assessment of the state-of-the-science of endocrine disruptors. World Health Organization, Geneva
 15. Zoeller RT, Bergman Å, Becher G, Bjerregaard P, Bornman R, Brandt I, Iguchi T, Jobling S, Kidd KA, Kortenkamp A, Skakkebaek NE, Toppari J, Vandenberg LN (2014) A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environ Health* 13:118
 16. Shug TT, Johnson AF, Birnbaum LS, Colborn T, Guillette LJ Jr, Crews DP, Collins T, Soto AM, vom Saal FS, McLachlan JA, Sonnenschein C, Heindel JJ (2016) Minireview: endocrine disruptors: past lessons and future directions. *Mol Endocrinol* 30:833–847
 17. Bergman A, Heindel JJ, Jobling S, Kidd KA, Zoeller RT (2013) State of the science of endocrine disrupting chemicals 2012. WHO/UNEP, p 289
 18. Newman MC (2015) Fundamentals of ecotoxicology. The science of pollution, 4th edn. CRC Press, Taylor & Francis Group, Boca Raton
 19. Hotchkiss AK, Rider CV, Blystone CR, Wilson VS, Hartig PC, Ankley GT, Foster PM, Gray CL, Gray LE (2008) Fifteen years after “Wingspread”—Environmental endocrine disrupters and human and wildlife health: where we are today and where we need to go. *Toxicol Sci* 105(2):235–259
 20. Kortenkamp A (2014) Low dose mixture effects of endocrine disrupters and their implications for regulatory thresholds in chemical risk assessment. *Curr Opin Pharmacol* 19:105–111
 21. Grandjean P, Barouki R, Bellinger DC, Casteleyn L, Chadwick LH, Cordier S, Etzel RA, Gray KA, Ha EH, Junien C, Karagas M, Kawamoto T, Paige LB, Perera FP, Prins GS, Puga A, Rosenfeld CS, Sherr DH, Sly PD, Suk W, Sun Q, Toppari J, van den Hazel P, Walker CL, Heindel JJ (2015) Life-long implications of developmental exposure to environmental stressors: new perspectives. *Endocrinology* 156:3408–3415
 22. Miller MD, Crofton KM, Rice DC, Zoeller RT (2009) Thyroid-disrupting chemicals: interpreting upstream biomarkers of adverse outcomes. *Environ Health Perspect* 117 (7):1033–1041
 23. Guillette LJ Jr, Iguchi T (2012) Ecology. Life in a contaminated world. *Science* 337 (6102):1614–1615
 24. Pesavento PA, Agnew D, Keel MK, Woolard KD (2018) Cancer in wildlife: patterns of emergence. *Nat Rev Cancer* 18:646
 25. Siegfried LM (1983) Neoplasms identified in free-flying birds. *Avian Dis* 27:86–99
 26. Lair S, Measures LN, Martineau D (2016) Pathologic findings and trends in mortality in the Beluga (*Delphinapterus leucas*) population of the St Lawrence Estuary, Quebec, Canada, from 1983 to 2012. *Vet Pathol* 53:22–36
 27. Couillard CM, Hodson PV, Castonguay M (1997) Correlations between pathological changes and chemical contamination in American eels, *Anguilla rostrata*, from the St. Lawrence River. *Can J Fish Aquat Sci* 54:1916–1927

28. Mikaelian I, de Lafontaine Y, Menard C, Tellier P, Harshbarger J, Martineau D (1998) Neoplastic and nonneoplastic hepatic changes in lake whitefish (*Coregonus clupeaformis*) from the St. Lawrence River, Quebec, Canada. *Environ Health Perspect* 106:179–183
29. Armstrong B, Tremblay C, Baris D, Thériault G (1994) Lung cancer mortality and polynuclear aromatic hydrocarbons: a case-cohort study of aluminum production workers in Arvida, Quebec, Canada. *Am J Epidemiol* 139:250–262
30. Gibbs GW, Labrèche F, Busque MA, Duguay P (2014) Mortality and cancer incidence in aluminum smelter workers: a 5-year update. *J Occup Environ Med* 56:739–764
31. Tremblay C, Armstrong B, Thériault G, Brodeur J (1995) Estimation of risk of developing bladder cancer among workers exposed to coal tar pitch volatiles in the primary aluminum industry. *Am J Ind Med* 27:335–348
32. Thirtamara Rajamani K, Doherty-Lyons S, Bolden C, Willis D, Hoffman C, Zelikoff J et al (2013) Prenatal and early-life exposure to high-level diesel exhaust particles leads to increased locomotor activity and repetitive behaviors in mice: diesel exhaust particles and autism. *Autism Res* 6:248–257
33. Raciti M, Ceccatelli S (2018) Epigenetic mechanisms in developmental neurotoxicity. *Neurotoxicol Teratol* 66:94–101
34. Legradi J, Di Paolo C, Kraak MHS, van der Geest HG, Schymanski EL, Williams AJ, Dingemans MML, Massei R, Brack W, Cousin X et al (2018) An ecotoxicological view on neurotoxicity assessment. *Environ Sci Europe* 30 (1):46
35. Crofton KM, Mundy WR, Shafer TJ (2012) Developmental neurotoxicity testing: a path forward. *Cong Anomal* 52:140–146
36. Sonne C, Letcher RJ, Janssen BM, Desforges JP, Eulaers I, Andersen-Ranberg E, Gustavson K, Styrihave B, Dietz R (2017) A veterinary perspective on One Health in the Arctic. *Acta Vet Scand* 59:84
37. Sibly RM, Newton I, Walker CH (2000) Effects of dieldrin on population growth rates of UK sparrowhawks. *J Appl Ecol* 37:540–546
38. Saeedi Saravi S, Dehpour AR (2016) Potential role of organochlorine pesticides in the pathogenesis of neurodevelopmental, neurodegenerative, and neurobehavioral disorders: a review. *Life Sci* 145:255–264
39. Martínez-López E, Romero D, María-Mojica P, Martínez JE, Calvo JF, García-Fernández AJ (2009) Changes in blood pesticide levels in Booted eagle (*Hieraetus pennatus*) associated with agricultural land practices. *Eco-toxicol Environ Saf* 72:45–50
40. Navas I (2017) Contaminantes ambientales persistentes (Metales Pesados y Plaguicidas Organoclorados) en Rapaces del Sur de España. Doctoral thesis, University of Murcia
41. Stamler CJ, Basu N, Man CH (2005) Biochemical markers of neurotoxicity in wildlife and human populations: considerations for method development. *J Toxicol Environ Health A* 68(16):1413–1429
42. Luster MI (2014) A historical perspective of immunotoxicology. *J Immunotoxicol* 11:197–202
43. Dean JH, House RV, Luster MI (2001) Immunotoxicology: effects of, and response to, drugs and chemicals. In: Hayes AW (ed) *Principles and methods of toxicology*, 4th edn. Taylor & Francis, London, pp 1415–1450
44. Rehberger K, Werner I, Hitzfeld B, Segner H, Baumann L (2017) 20 Years of fish immunotoxicology – what we know and where we are. *Crit Rev Toxicol* 47(6):509–535
45. Desforges JP, Sonne C, Levin M, Siebert U, De Guise S, Dietz R (2016) Immunotoxic effects of environmental pollutants in marine mammals. *Environ Int* 86:126–139
46. Rohr JR, Schotthoefer AM, Raffel TR, Carrick HJ, Halstead N, Hoverman JT, Johnson CM, Johnson LB, Lieske C, Piwoni MD, Schoff PK, Beasley VR (2008) Agrochemicals increase trematode infections in a declining amphibian species. *Nature* 455:1235–1239
47. Hall AJ, Hugunin K, Deaville R, Law RJ, Allchin CR, Jepson PD (2006) The risk of infection from polychlorinated biphenyl exposure in the harbor porpoise (*Phocoena phocoena*): a case-control approach. *Environ Health Perspect* 114:704–711
48. Friend M, Trainer DO (1970) Polychlorinated biphenyl: interaction with duck hepatitis virus. *Science* 170:1314–1316
49. Grasman KA (2002) Assessing immunological function in toxicological studies of Avian wildlife. *Integr Comp Biol* 42:34–42
50. Corsini E, Luebke RW, Germolec DR, DeWitt JC (2014) Perfluorinated compounds: emerging POPs with potential immunotoxicity. *Toxicol Lett* 230:263–270
51. Acir IH, Guenther K (2018) Endocrine-disrupting metabolites of alkylphenol ethoxylates – a critical review of analytical methods, environmental occurrences, toxicity, and regulation. *Sci Total Environ* 635:1530–1546

52. Besedovsky HO, Del Rey A (1996) Immuno-neuro-endocrine interactions: facts and hypotheses. *Endocr Rev* 17:64–102
53. Calderón-Garcidueñas L, Mora-Tiscareño A, Ontiveros E, Gómez- Garza G, Barragán-Mejía G, Broadway J et al (2008) Air pollution, cognitive deficits and brain abnormalities: a pilot study with children and dogs. *Brain Cogn* 68:117–127
54. Peterson RE, Seefeld MD, Christian BJ, Potter CL, Kelling CK, Keesey RE (1984) The wasting syndrome in 2,3,7,8-tetrachlorodibenzo-p-dioxin toxicity: basic features and their interpretation. In: The Banbury report – biological mechanisms of dioxin action. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp 291–308

Part II

Methods and Protocols



Chapter 5

Importance of Data Curation in QSAR Studies Especially While Modeling Large-Size Datasets

Pravin Ambure and M. Natália Dias Soeiro Cordeiro

Abstract

A huge amount of chemical and biological data that is available in several online databases can now be easily retrieved and studied by many researchers (including QSAR modelers) to extract meaningful information. Everyone is naturally aware, however, of the errors in chemical structures and biological data that are possibly present in the retrieved data from these online databases. Implications of those might be severe, particularly for QSAR modelers since developing models using such erroneous data will certainly lead to false or non-predictive models. Proper curation of the retrieved chemical and biological data is therefore crucial and mandatory prior to any QSAR modeling. For large datasets, manual data curation becomes highly impossible, nevertheless. This chapter reviews and discusses the several data curation tools normally applied for such endeavors, paying special attention to those that can be used to semiautomate the curation process, like resorting to a workflow by employing the freely available KNIME software.

Key words Data curation, Online databases, Structural errors, Duplicate analysis, Activity cliffs, Curation tools, QSAR

1 Introduction

Recent advances in computational power, storage capacity, and efficient algorithms/tools have resulted in easier handling of extraordinary amounts of data. Nowadays, a huge amount of data is being collected, analyzed, stored, retrieved, and/or utilized daily. Researchers from across the world are now giving efforts on extracting useful information from the dozens of available online databases (public or commercial) to gain new or improved insights in their own field of interest. Similarly, quantitative structure-activity relationship (QSAR) specialists or modelers are also taking the advantage of the enormous relevant data that is available for modeling biological activity (QSAR), physicochemical properties (QSPR), toxicity (QSTR), and so forth. The models then built are highly desired in several research fields, such as drug design, toxicity assessment of chemicals, probing material properties, etc. Though

one might consider that most of the data present in the online databases are accurate, it is also true that a substantial fraction of these data – reaching up to 10% – may be not accurate [1]. This amount of error is significant enough to seriously affect the reliability of QSAR models developed upon them. However, QSAR modeling should not be limited by the quality of the available data, as it is still possible to perform a proper curation of data to avoid the development of misleading/non-predictive models or put forward false hypotheses [2]. This chapter aims firstly at understanding the need and the importance of data curation. Then, the key steps involved in data curation of large-size datasets that are employed for QSAR modeling will be discussed. Finally, the recent advances in the development of freely available automated or semiautomated tools for performing curation of large datasets used in QSAR, QSPR, QSTR, etc. studies will be reviewed and discussed.

2 Importance of Data Curation in QSAR Studies

One can comprehend the importance of data curation only after realizing the existence of errors in the available online databases. Such errors can be observed in the chemical structures or biological activity/toxicity data or in any relevant information that is provided by the online databases. Also, users of these available online databases should be aware that they might find differences in the chemical structure representation of compounds whenever retrieving those from different databases [3]. It is also important to understand how such errors get incorporated into the databases since even database' service providers never want to provide data with errors. Let us then look at the most common ways how these errors might end up in databases, namely:

1. *Scientific literature*: For a chemical database, one of the primary sources for collecting the chemical structural information are scientific publications, where unfortunately most of the chemical structures are available in the image format. Thus, for depositing these chemical structures in the database, the structures are redrawn and then stored in a proper structure readable format, or often an image-to-structure converting software is employed. Inevitably, there is a high possibility of introducing structural errors due to human errors or sometimes due to software limitations. Another means of propagating structural errors from the literature might be that the published chemical structures are frequently drawn in relevance to the context of the publication. For instance, in a paper reporting molecular docking study, an acidic or a basic molecule might be drawn as a negatively or positively charged molecule as this is the relevant form for binding to the protein. In another instance, a review

article presenting the biological activity information for a series of relevant compounds may accidentally show the parent structure of a molecule even though the activity is shown by its salt form [3].

2. *Chemical file format system:* All chemical databases use a text-based chemical file format system, like molfile in V2000 format, so that chemoinformatic software tools can easily read, interpret, and store single or multiple chemical structures in a plain text format. However, the mostly used molfile V2000 also has some shortcomings [3]. For instance, it cannot represent compounds that have two stereocenters and are a mixture of two enantiomers but do not contain any of the diastereoisomers, just like drug milnacipran – *a mixture of the 1R,2S and 1S,2R enantiomers*. Another issue with the V2000 molfile version is that there is no way for properly representing dative or coordinate bonds. Such shortcoming may lead to an incomplete and thus incorrect structural representation of chemicals stored in the databases.
3. *Units:* A major cause of errors pertaining to biological activity or toxicity values is due to miscalculations occurring during the units' conversions or due to typos/errors seen in the published literature from where the data is collected [2]. For example, the incorrect substitution of “ μM ” (micromolar) with “ mM ” (millimolar) is often detected.
4. *Normalization of chemical structures:* Each database service provider has its own set of rules for standardizing or normalizing the chemical structures, for instance, the way nitro groups and sulfoxides are normalized or how the tautomers are canonicalized and whether the compounds are “merged” at a parent level, etc. This also results in different chemical representations of the same compound appearing in different databases [3].
5. *Interlinked databases:* As some databases collect the information from other databases and are in some way interlinked, there remains a huge possibility of transferring the errors that are present in the parent database. For example, the PubChem database [4] collects data from around 668 sources including ChEMBL [5], BindingDB [6], and ChEBI [7]. Further, ChEMBL contains data from PubChem, ZINC [8] comprises data from ChEMBL, and ChemSpider [9] is a chemical structure database that holds information of 71 million chemicals gathered from about 259 other data sources.

There can be many more possible ways of inclusion of unintentional errors in the online databases. Several investigators or users of such databases have already noticed that there are indeed numerous errors present, and they have pointed out that fact in various publications [1, 3, 10–14, 15] highlighting the severity and

amount of errors present in the online databases including some well-known databases. Therefore, it is highly risky to employ the chemical structures or biological data from the online databases at their face value since, as already mentioned, the models set up based on those can be completely flawed. Another issue is that there are no absolute standard guidelines on how to assemble and integrate data from different primary or secondary sources, as it appears that each database service provider follows its own unverified approach (es) for gathering and storing the data. There is thus an urgent need for the database service providers along with the scientific community to work together and ascertain guidelines for best practices to reduce the number of avoidable errors.

Moreover, several studies have also been carried out to evaluate the impact of data curation on the development of QSAR models. For instance, Young et al. [10] developed QSAR models using chemical structures from a database: (1) with an error rate of 3.4% (not curated) and (2) all correct structures (curated). The authors found that even slight errors in the chemical structures, such as misplacing a Cl atom or swapping hydroxy and methoxy functional groups on a multiple rings' structure, resulted in significant differences in the prediction quality for those chemicals. In another recent study, Mansouri et al. [16] developed also QSAR models using $\log P$ data before and after curation. The validation statistics obtained clearly showed better performance of the model trained with the curated data over that trained with the original non-curated dataset. Comparison of the derived QSAR models (non-curated vs. curated) was shown in terms of r^2 (0.59 vs. 0.70) and $RMSEP$ (1.13 vs. 0.96) statistics on the test sets as well as on the basis of the number of test set chemicals that were found outside of the applicability domains (98 vs. 46).

3 Key Steps Involved in Data Curation for QSAR/QSTR/QSPR Studies

Prior to initiating a QSAR study, one should always be concerned with the quality of at least two categories of data, namely, (1) the chemical structures and (2) the biological data, which are employed for QSAR modeling. The workflow demonstrating the steps involved in the chemical and biological curation is shown in Figs. 1 and 2, respectively. Here we will discuss the basic curation steps that should be always performed for datasets used in QSAR modeling. Note that wherever possible, we will also refer to a special tool for such purpose, that is, the free nodes available in KNIME workflow platform [17], that can be used to implement the respective curation step. An illustration of a KNIME workflow used for data curation is shown in Fig. 3.

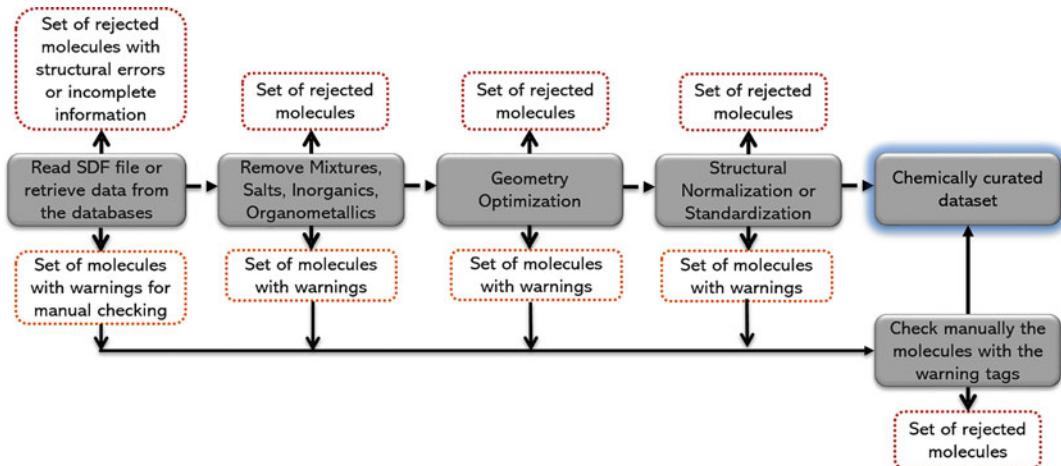


Fig. 1 A workflow illustrating the key steps involved in the chemical data curation

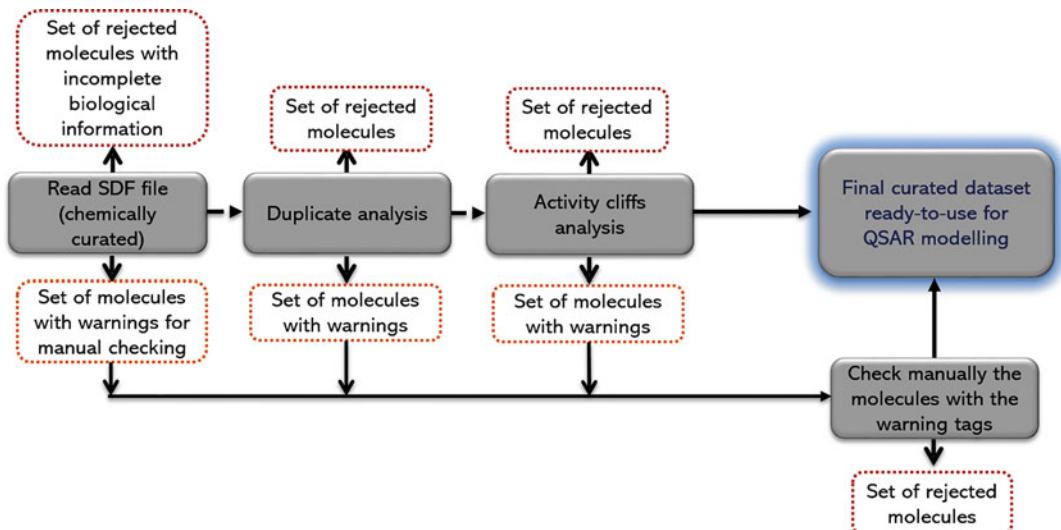


Fig. 2 A workflow illustrating the key steps involved in the biological data curation subsequent to chemical curation

1. *Data retrieval:* Data retrieval becomes the first step in data curation, particularly when the data is extracted from a database. Usually, when one downloads a set of chemical structures from any database, this is done in the form of a structural data format (SDF) file. Such SDF file mainly contains some basic structural information such as the molecule name, count of atoms and bonds, 2D or 3D coordinates of each atom present in the molecule, the connectivity, bond types, stereotypes, etc. for each chemical present in it. Along with the requisite structural information, it may also include various properties like unique chemical ID allotted by the databases, molecular

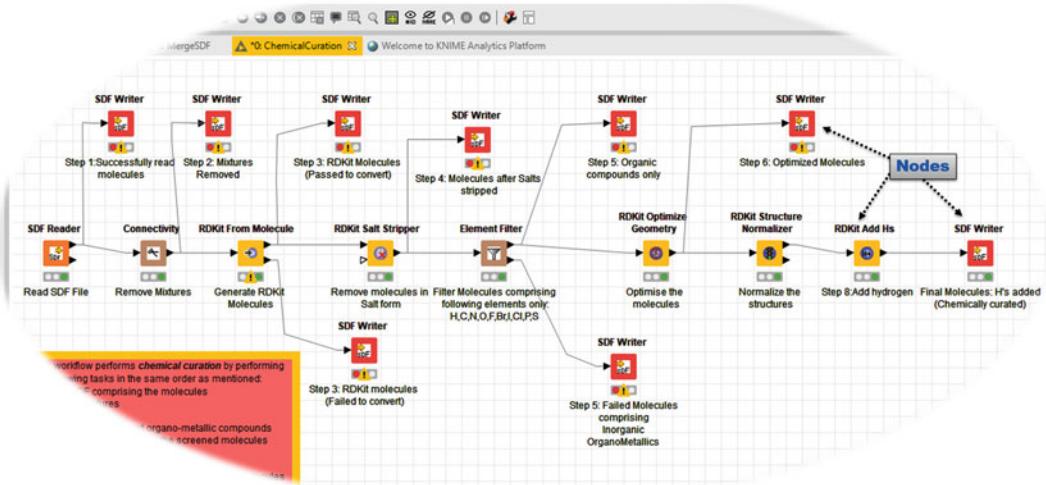


Fig. 3 An illustration of a KNIME workflow employed for performing the chemical data curation

weight, the biological property ($IC_{50}/EC_{50}/K_i$), charges, SMILES notation, etc. For the curation of chemical structures, the structural information provided in the SDF file is enough to identify and remove the erroneous chemical structures, but for the curation of the biological data, one needs both the structural information and the biological data under study. Thus, the molecules with incomplete structural or biological data information such as any missing atom name or coordinates or requisite biological data should be discarded. This can be easily done in the KNIME workflow platform using the “SDF Reader” node which helps in reading the input SDF file apart from removing the chemical structures with incomplete structural information. Finally, reading and storing all the mandatory information correctly is the first step of data curation. Also, whenever possible, it is always better to retrieve the data from multiple databases and compare the accuracy of structural and biological information therein.

2. *Chemical curation:* The next step is to perform the identification and correction of the structural errors for a set of chemicals. Note that the following chemical curation steps are particularly recommended for QSAR studies.
 - (a) *Inorganics or organometallics:* One should be aware that several chemoinformatic software cannot handle the inorganic or even organometallic chemicals, for instance, most of the software tools used for calculating molecular descriptors can only handle organic chemicals. Thus, the

inorganics or organometallic compounds in the datasets may end up with incorrect descriptor values. Thus, it is recommended either to remove all the inorganic and organometallic chemicals before the descriptors are calculated or to use some dedicated software such as CORAL [18, 19] that may handle such compounds. To remove inorganics and organometallics, one may employ the “Element Filter” node available in KNIME platform.

- (b) *Mixtures:* It is advised to remove mixtures prior to the descriptor calculation since the treatment of a mixture is not an easy task unless the active component is known [20]. To remove the mixtures, one may use the “Connectivity” node available in KNIME. However, in studies focusing on QSAR modeling of mixtures, one can resort to software such as ISIDA [21] to calculate suitable descriptors for those.
- (c) *Removal of salts:* Like inorganics or organometallic compounds, salts are usually mishandled by the descriptor-calculating software, which might result in errors in the values of the descriptors. If essential, salts can be identified and removed using the “RDKit Salt Stripper” node that is available in KNIME. However, the recent version of the well-known commercial software Dragon 7.0 allows the calculation of descriptors for a disconnected structure such as salts, like ionic liquids.
- (d) *Normalization or standardization of chemical structures:* Normalization or standardization means to transform the chemical structures into customized, canonical representations, which is important as there is a possibility of representing the same functional group (such as nitro groups) using different structural patterns. The different representations of the same chemical structure may create issues such as being unable to identify duplicates based on the values of the molecular descriptors, because the calculated descriptor values for these distinct representations of the same functional group could be significantly different. One may use the “RDKit Structure Normalizer” node available in KNIME or the “Standardizer” facility available in “JChem” of the ChemAxon software [22] (free for academic organizations) for standardizing the chemical structures. Further, the “Structure Checker” tool also provided by ChemAxon (<https://chemaxon.com/>) can be used also to identify and correct several structural problems such as invalid bond lengths, overlapping bonds or atoms, molecule charges, incorrect chiral flags, invalid valences, etc.

3. *Biological curation:* Biological curation is performed to confirm the accuracy of the extracted experimental data from the online database. If possible, the biological data for the compounds present in the dataset should always be retrieved from multiple databases, and the experimental activity/toxicity values should be compared at this stage. The experimental values are more reliable when found same in most of the databases. In the next step, two important analyses, namely, duplicate analysis and activity cliff analysis, should be performed to finally create a ready-to-use dataset for QSAR modeling. Note that both these analyses require the structural as well as the biological information, and it not only assists one to remove the redundant or problematic compounds but also helps to examine further possible errors particularly in the experimental data.

(a) *Duplicate analysis:* Identification of duplicates is a crucial and challenging task, especially for large datasets, where manual visual inspection is not possible. Particularly, the unique identification number that is provided by the databases should not be used for duplicates' identification, as there can be errors or redundancies present in the unique itself identification of numbers. The duplicates can be efficiently identified by carrying out similarity searches using distance metrics such as the Tanimoto index, Euclidean distance, etc. These distance metrics are usually computed using fingerprints (*a requisite for the Tanimoto index*) and/or molecular descriptor values. However, simple identification of duplicates based on the structural similarity index is not the desirable final achievement, since their associated experimental data (i.e., the biological activity or toxicity values under study) should also be properly analyzed before randomly deleting the extra identical compound(s). Thus, for a given pair of duplicate structures, if their experimental properties are identical, then one of the compounds can be selected randomly and then deleted. But, if their experimental properties are numerically different, then one should always consider the following scenarios.

- If the experimental values are nearly similar, then one of the compounds (duplicates) can be kept with the arithmetic average of the experimental values.
- If the experimental values are significantly different, then it means that either or both the experimental values are incorrect. In that case, such duplicates should be always removed, unless one can confirm the correct experimental value from the primary sources. Another option is to keep these compounds in a query set (i.e., a

set *not used* in the QSAR model development in any way) so that one might check the predicted value for these compounds using the developed model, which might help in rectifying the error.

- One more case might be observed [23] when both the experimental values are found significantly different but are in fact correct. The reason is that one of the previous curation steps has modified the original compounds to create such duplicates. For instance, the two identified duplicate compounds might correspond to two different salts of the same compound, and the error is introduced when the counterion was removed during the step concerning the neutralization of salts. The experimental properties can be significantly different if they are directly influenced by the counterion. In those cases, the safer suggestion is to remove such salts for QSAR modeling [23].

The duplicate analysis can be performed using KNIME employing any structural similarity index such as the “3D D-Similarity” node (molecular shape similarity measure) or using software like the HTS Navigator [24] or the OCHEM [25]. However, as far as our knowledge goes, there is no software tool for performing the detailed duplicate experimental value joint analysis as discussed above, and thus the analysis part must be performed manually.

- (b) *Activity cliff analysis:* Activity cliffs are the regions where large changes in activity are observed for relatively small changes in the chemical structure [26]. Thus, compounds that may show activity cliffs are having high structural similarity but a large difference in their biological property values. Such compounds are difficult to understand or interpret using QSAR modeling, which is based on the chemical similarity principle. To identify and verify activity cliffs especially in the large datasets, one can perform a matched molecular pair (MMP) analysis [26]. MMP is simply defined as a pair of molecules that differ in only a minor single point change. With this respect, single point changes in the molecule pairs are termed a molecular transformation. For finding activity cliffs, the transformation is considered significant, if it leads to a drastic change in the biological property value with minor single point change. The MMP analysis can be easily performed in KNIME using the “Automated Matched Pairs” node.

Though it is not always advisable to remove activity cliff compounds due to their importance [26], at least some of them are explainable using specific descriptors or 3D descriptors that can highlight the minor difference in the structure

which is responsible for a large difference in the activity [27]. However, the modeler should finally decide whether to keep or discard such compounds.

4 Recent Advances in Data Curation Tools

Manual data curation is comparatively easy when one is dealing with small datasets (i.e., less than 50 compounds), but it becomes more and more difficult as the dataset size progresses. For very large datasets involving thousands of compounds, manual data curation becomes almost impossible. Thus, it is highly essential to develop efficient tools to automate the data curation process with least manual intervention. Recently, many reported studies have been focused on developing such tools mainly through the freely available KNIME workflow to perform the data curation. Here, we will discuss some of the representative studies along with the website links (if available) where one may download the dedicated data curation tools.

In 2010, a detailed workflow revealing the important steps required to curate a chemical dataset for QSAR modeling was published by Fourches et al. [20]. Though previous publications have raised the curation issue and suggested some solutions, the authors provided a detailed and systematic workflow to perform chemical curation. Later on, the same research group published another work [27] describing both the chemical and biological curation process. Both such articles are highly informative for QSAR modelers, and at least some of the data curation tools proposed by the authors included KNIME workflows.

In another study, Mansouri et al. [16] have developed a semi-automated KNIME workflow solely to curate and correct errors in the structure and identity of chemicals. The workflow was then tested using the physicochemical properties and environmental fate datasets available in the PHYSPROP database. In that study, the workflow first collects structure-identity pairs using four chemical identifiers, including chemical name, CASRNs, SMILES, and MolBlock, which are then employed to identify and rectify problems such as errors and mismatches in chemical structure formats, identifiers, duplicates, and several structural validation issues including hypervalency and incorrect stereochemistry. The curated datasets and the developed KNIME workflow in this study are available for download at ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/.

Kim et al. [28] also developed a KNIME workflow to curate and prepare high-throughput screening data for QSAR modeling purposes. This workflow loads chemical structures in SMILES or SDF format, keeps only organic molecules, and then standardizes the structural representation using InChI tautomerism and

additional SMARTS transformation. The output files comprise the structurally curated structures, rejected structures, and the structures with warnings. The workflow can be downloaded as a zip file at <https://github.com/zhu-lab/>.

Kausar and Falcao [29] developed a fully automated QSAR modeling framework using KNIME workflow, which includes almost all requisite modeling tasks starting from data curation to QSAR model building and validation. The data curation part of the workflow facilitates retrieval of data directly from databases such as ChEMBL, removal of irrelevant data by selecting only the bioactivity type of interest, filtering out missing data, identifying duplicates, and dealing with several forms of the same molecule (including salt groups). The developed framework was tested on datasets with 30 different problems. This workflow is available at <https://github.com/Saminakausar/Automated-framework-for-QSAR-model-building>.

Ambure et al. [30] have developed semiautomated KNIME workflows to perform both chemical and biological curation. Here, the workflow dedicated to chemical curation loads the data in SDF format, keeps organic molecules (only), removes mixtures and salts, optimizes the geometry, and finally carries out the normalization of the chemical structures to the screened set of chemicals. The workflow dedicated to biological curation performs duplicate analysis using molecular shape similarity measure and activity cliff analysis using MMPs. Ambure et al. [30] discussed in detail how these analyses were performed with some task needing manual involvement. Five datasets curated using these workflows were then used to develop multiple QSAR models and further employed to identify multi-target directed ligands against Alzheimer's disease. These workflows are free to download at <https://sites.google.com/site/dtclabdc/>.

Finally, Gadaleta et al. [23] have designed and implemented a semiautomated workflow integrating structural data retrieval from several web-based databases, automated comparison of these data, chemical structure cleaning, and selection and standardization of data into a consistent, ready-to-use format that can be employed for modeling. This workflow integrates almost all the vital tasks for data curation, and the output files comprise not only the information about the structures that are retained or rejected but some helpful additional information such as reliability (high or medium), possible warnings, removed counterion information for salts, structures that need manual checking, etc. The respective KNIME workflow is freely available at https://github.com/DGadaleta88/data_curation_workflow.

5 Prospects

From the discussion, it is clear that data curation has already become and will remain as an important part of QSAR modeling. Thus, all the datasets extracted from the online databases and/or published literature must be curated prior to using it for QSAR modeling. Identifying the true relationship between the structural features and the response (activity/toxicity) under study is only expected if the dataset is curated, while the non-curated dataset might always result in a false relationship. Though many service providers of databases like ChEMBL [15] have already taken initiatives to provide curated data, however, it is still sensible for users to always confirm the accuracy of extracted data. The freely available and shared data curation tools especially KNIME workflows will surely help to further improve the automatic data curation efforts in the near future.

Acknowledgments

This work was supported by UID/QUI/50006/2019 with funding from FCT/MCTES through national funds.

References

- Williams AJ, Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. *Drug Discov Today* 16:747–750
- Waldman M, Fraczkiewicz R, Clark RD (2015) Tales from the war on error: the art and science of curating QSAR data. *J Comput Aided Mol Des* 29(9):897–910
- Hersey A, Chambers J, Bellis L, Bento AP, Gaulton A, Overington JP (2015) Chemical databases: curation or integration by user-defined equivalence? *Drug Discov Today Technol* 14:17–24
- Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA (2015) PubChem substance and compound databases. *Nucleic Acids Res* 44(D1): D1202–D1213
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40(D1): D1100–D1107
- Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2006) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 35 (Suppl_1):D198–D201
- Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2007) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 36(Suppl_1):D344–D350
- Irwin JJ, Shoichet BK (2005) ZINC— a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45 (1):177–182
- Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87(11):1123–1124
- Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? *QSAR Comb Sci* 27 (11–12):1337–1345
- Williams AJ, Ekins S, Tkachenko V (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov Today* 17(13–14):685–701
- Kramer C, Kalliokoski T, Gedeck P, Vulpetti A (2012) The experimental uncertainty of

- heterogeneous public Ki data. *J Med Chem* 55(11):5165–5173
13. Tiikkainen P, Bellis L, Light Y, Franke L (2013) Estimating error rates in bioactivity databases. *J Chem Inf Model* 53(10):2499–2505
 14. Kallikoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC₅₀ data—a statistical analysis. *PLoS One* 8(4):e61007
 15. Papadatos G, Gaulton A, Hersey A, Overington JP (2015) Activity, assay and target data curation and quality in the ChEMBL database. *J Comput Aided Mol Des* 29(9):885–896
 16. Mansouri K, Grulke C, Richard A, Judson R, Williams A (2016) An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. *SAR QSAR Environ Res* 27(11):911–937
 17. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME—the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explorations Newsletter* 11(1):26–31
 18. Toropova A, Toropov A, Benfenati E, Gini G (2011) QSAR modelling toxicity toward rats of inorganic substances by means of CORAL. *Open Chem* 9(1):75–85
 19. Toropova A, Toropov A, Benfenati E, Gini G (2011) Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: an unexpected good prediction based on a model that seems untrustworthy. *Chemom Intell Lab Syst* 105(2):215–219
 20. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 50(7):1189–1204
 21. Oprisiu I, Varlamova E, Muratov E, Artemenko A, Marcou G, Polishchuk P, Kuz'min V, Varnek A (2012) QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. *Mol Inform* 31(6–7):491–502
 22. Csizmadia F (2000) JChem: Java applets and modules supporting chemical database handling from web browsers. *J Chem Inf Comput Sci* 40(2):323–324
 23. Gadaleta D, Lombardo A, Toma C, Benfenati E (2018) A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *J Chem* 10(1):60
 24. Fourches D, Sassano MF, Roth BL, Tropsha A (2013) HTS navigator: freely accessible cheminformatics software for analyzing high-throughput screening data. *Bioinformatics* 30(4):588–589
 25. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25(6):533–554
 26. Kramer C, Fuchs JE, Whitebread S, Gedeck P, Liedl KR (2014) Matched molecular pair analysis: significance and the impact of experimental uncertainty. *J Med Chem* 57(9):3786–3802
 27. Fourches D, Muratov E, Tropsha A (2016) Trust, but verify II: a practical guide to chemogenomics data curation. *J Chem Inf Model* 56(7):1243–1252
 28. Kim MT, Wang W, Sedykh A, Zhu H (2016) Curating and preparing high-throughput screening data for quantitative structure-activity relationship modeling. In: High-throughput screening assays in toxicology. Springer, Humana Press, New York, NY, pp 161–172
 29. Kausar S, Falcao AO (2018) An automated framework for QSAR model building. *J Chem* 10(1):1
 30. Ambure P, Bhat J, Puzyn T, Roy K (2019) Identifying natural compounds as multi-target-directed ligands against Alzheimer's disease: an in silico approach. *J Biomol Struct Dyn* 37:1282–1306



Chapter 6

Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling

Giuseppina Gini and Francesco Zanoli

Abstract

Today the registered chemical structures are about 28 millions, while experimental toxicity data are available for a few hundred thousands of them. Defining properties and effects for all the available chemicals is a huge task due to the cost of the experimentation and to legislative restrictions. Therefore, prediction is the only available solution, but it poses many challenges in terms of accuracy and interpretability. Predictive toxicology systems use statistics as well as methods based on machine learning (ML). While ML has been widely used in the pharmaceutical domain, its use in ecotoxicology is more limited. After reviewing the experiences in quantitative structure-activity relationships (QSARs) for modeling CMR (carcinogenic, mutagenic, reproductive) toxicity and PBT (persistent, bioaccumulative, and toxic) chemicals, we look at the advancements of technology in ML. Recently, the investigation of the neural basis for many cognitive functions has provided the tools to create new systems that can think, solve problems, find patterns, and recognize images and texts; these new methods are named deep learning (DL). We modified the most successful DL architecture, implemented Toxception as a tool to generate QSAR models, and tested it in a real case, on a dataset of about 20,000 molecules tested for mutagenicity with the Ames test. The results obtained challenge the current state of the art. In addition, Toxception does not use any chemistry knowledge besides the 2D structures derived from SMILES. We conclude examining advantages, open challenges, and drawbacks of building QSARs with DL.

Key words Machine learning, Neural networks, Deep learning, Mutagenicity, Ames test

1 Introduction

In 2009, the paper “The Toxicity Data Landscape for Chemicals” [1] reported 28 million chemicals discovered. Only three millions were tested on animals and humans, and about one million had some toxic assay summary. To fill the data gap, successive studies on other compounds are using predictions or simulations. It is understandable that with those numbers, finding an accurate and optimized model is a big challenge. Moreover, the data heterogeneity is very high, and this reduces, even more, the reachable precision. It must also be considered that knowledge about the functioning of

the organisms is still insufficient with respect to the complexity of the systems.

Chemoinformatics has appeared as an alternative way to physical models (quantum chemistry or molecular dynamics simulation). Chemoinformatics started by considering the chemical structures as graphs. One of the simplest representations of a molecular graph is the adjacency matrix that supports the computation of many topological descriptors. Other chemical descriptors that account for electrical and physical properties are also extracted from chemical structures. 1D and 2D descriptors are the most used in chemoinformatics. 3D descriptors require 3D coordinate representations and are sensitive to structural variations, since they usually are built from the most common optimized 3D structure and cannot account for the other less common structures of the same chemical. Four-dimensional chemical descriptors are necessary to simultaneously consider multiple structural conformations. Chemical fingerprints are vectors of large dimension that represent in each position 0 or 1 if a substructure of a given list is present or not.

Chemical similarity is another fundamental technique in chemoinformatics. Its objective is to group the compounds with structures and bioactivity similar to each other, according to the principle that similar compounds have similar activity. This assumption is not always true, and the concept of activity cliffs has been introduced to explain when minor modifications of functional groups cause a dramatic change in the activity.

QSARs are typically used to develop models that are specific to a single defined endpoint, for which usually an *in vivo* or *in vitro* test is available. As implicit in the name, such models use only the chemical structure of the molecule to create an association with the biological endpoint. Moreover, in the development of new drugs, such models can be used to flag compounds that are likely to cause adverse effects.

We have to mention also the methods used to create SAR models, which are based on the individuation of toxicity-conferring molecular fragments, that are small parts of a molecule that can be associated with toxicity effect following statistical methods. The main problem with such a method is the consistency of the structural alerts (SA) provided. Frequently the fragments used are too small to be unambiguously linked to the chemicals. Moreover, the set of rules to search inside the input dataset must be decided, and these rules are usually linked to the specific database, therefore, not transferable to other dataset [2].

The use of QSAR and SAR models within regulatory bodies, however, is in an initial phase and still under active development [3]. The safety assessment of the huge quantities of chemicals in daily use has been of concern at least in the last decades, together with the release of specific registration regulations. For environmental QSARs the main problems are the availability of good

quality toxicity data, ideally obtained from certified laboratories, and about valid toxicological endpoints. A number of regulatory authorities are considering the utility of QSAR in toxicity prediction for tasks as prioritization, classification and labeling, and chemical assessment. While QSARs are of common use in the pharmaceutical industry, as they are apt to screen large amount of data, they encounter acceptability problems in the environmental risk assessment, where studies are often interested in a few substances and in understanding how they interact with the physical and biological agents in the environment. Moreover, for the environmental protection, there is more interest in the chronic toxicities than in the acute toxicities, which are a big concern in drug design.

Roughly there are two main streams for making models: data modeling stream and algorithmic modeling. Data modeling is the stream commonly developed by statisticians: from the data analysis, they postulate the kind of relation between data and response and use mathematical tools to derive the model. Initial QSAR studies used simple linear or multi-linear regressions with a small number of descriptors (features in the modeling terminology). Those models were apt to model small series of similar compounds.

Algorithmic modeling has been developed more recently, starting in the mid-1980s when powerful new algorithms for fitting data became available. They are generally named machine learning (ML) methods. To create QSARs from a large series of compounds in a wider chemical space, those more effective methods are needed [4]. Algorithmic methods include decision trees, production rules, neural networks (NN), genetic algorithms, support vector machine (SVM), random forest (RF), and naïve Bayes just to name the main families. This field is in rapid evolution and is boosted by the introduction of massively parallel hardware. ML uses pattern recognition to find the mathematical relationship between experimental observations and biological or chemical properties. ML techniques are usually more efficient and more feasible than physical models and can scale up to big data. Many good reviews have recently appeared on the topic, as [5], which reviews the main ML methods adopted in QSAR, and [6], who also discusses ML methods for feature selection.

A literature analysis of QSAR in the years 2009–2015 appeared recently [7]. It observed that the number of QSAR papers using standard regression tools was decreasing, while more papers used ML methods, especially RF and naïve Bayes. The reasons may be the transformation of QSAR studies into routine work done using the available tools or some saturation of models for available data. In conclusion, we may expect that the progress in ML methods can help QSAR to enter in the productivity cycle in drug design. The same could be true for ecotoxicology.

ML methods can be broadly divided in supervised and unsupervised. Most of the QSAR applications use supervised learning, where data used for learning are labeled (i.e., they contain both the chemical structure and the property under investigation). The property values can be real values, as in the case of dose-response, or integer labels, as in the case of active/not active.

The most adopted supervised ML methods are so far neural networks (NN), support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and naïve Bayes (NB). The most common unsupervised techniques that are aimed at discovering the unknown pattern from unlabeled data are hierarchical and non-hierarchical clustering and k-means clustering.

Many-layer NNs have been recently introduced to learn highly complex functions. Those networks contain a large number of hidden layers; adding more layers requires a full new way of building the net, as conceptualized by deep neural networks (DNN). DNNs are now being used to learn directly useful features from data, without the need of computing features. Convolutional neural networks (CNNs) as well as recurrent neural networks (RNN) have been successful in recognizing images and text, respectively, and are now being used in chemoinformatics, especially in drug discovery [8]. Today the needed computer power is easily available from systems as GPU-accelerated computers, for instance powered by NVIDIA.¹

In the rest of this chapter, we will shortly introduce the environmental properties successfully investigated with ML methods. To go further in the new directions of ML, we will look at the basic deep NN architecture that has gained attention due to its capability to understand images, and we will show how to use it to create a QSAR model.

Usually ML methods are applied to QSAR in a pattern matching style, so using a selected set of descriptors. Descriptors are indeed some specific or local views of the molecular structure. Why not to use the structure itself, i.e., the chemical graph, and nothing else? The reason for building again another QSAR model is that we want to explore how the basic hypothesis of QSAR, i.e., “similar molecules have similar properties” can be challenged by just using the structure and nothing else, in particular no measures of similarity. We will apply those concepts in the construction of Toxception, a new deep learning model of the Ames mutagenicity test. An important point of Toxception with respect to the previous use of NNs is that it is able to auto generate the relevant descriptors.

Of course modeling is only one aspect of QSAR, since making the model robust, understandable, and acceptable is another big task, as we will see in the final discussion.

¹ <https://www.nvidia.com/en-us/data-center/dgx-1/>

2 Machine Learning in QSAR for Environmental Protection

Initial QSAR models used statistical methods and parametric models to model a property on a family of compounds with similar chemical characteristics. With new studies on different chemical classes, some nonparametric techniques, as the ones taken from ML, were considered.

ML has the task of learning a function that minimizes an error. Minimization means optimization: learning can be seen as an optimization problem. The mathematical formula for optimization is simple one, but its computation can be really hard since we cannot accept trying all the possible combinations of parameters setting. The goal for ML is so to optimize the performance of a model given an objective function and the training data.

In particular scientists have used traditional AI methods such as decision trees and NN and then adopted new emerging pattern recognition methods as SVM and RF, which allow for classifying anything based on its features. All those methods have a common characteristic: they make an optimization. Moreover, in the creation process, they require the interaction with a human expert in order to select the features (chemical descriptors) on which to base the whole process. In fact, all of the methods just cited are extremely powerful if they are applied to a consistent set of features.

Other methods derived from data mining are instead able to autonomously mine data and find patterns. Those methods are at the basis of systems that extract substructures from the chemical structures; after the substructures are created, an optimization method is called to make a correlation among some of them and the target. This line of research originated from the work of [9], which developed a way to learn rules of rodent carcinogenicity; it used an inductive learning program to work on the data about short-term assays of mutagenic and cancerogenic toxicity together with maximum tolerated dose. This work improved the prediction of nongenotoxic chemicals. Extracting rules from data is the way of operating SARpy, a tool publicly available in VEGAhub,² which works both to find fragments correlated to toxic and nontoxic substances; the rules are then used to make a SAR model. CORAL, another tool in VEGAhub, uses an optimization method to find the best combination of small pieces of the molecules in the training dataset, so providing a QSAR model and an indication about the fragments more strongly related to the toxic effect.

Another common method used in ML is ensembling. Ensembling means combining different models together. An ensemble is an algorithm trained with the results of a number of models, which is then used to make predictions. Empirically, ensembles tend to

² <https://www.vegahub.eu/download/>

give better results when there is significant diversity among the models used to build them. This conclusion is further highlighted by the fact that models of close similarity tend to have similar prediction errors that cannot be corrected. Model diversity can be achieved using a number of strategies, for example, using different datasets to train individual classifiers or using different training parameters for different classifiers. Alternatively, entirely different types of classifiers, such as linear regression, decision trees, and SVM, can be combined to enhance model diversity.

The initial examples of ensembles were based on bagging and boosting techniques. Using the bagging technique, a number of training data subsets are randomly drawn, with replacement, from the training data. Each subset is used to train a different classifier of the same type; the classifiers are then combined using majority vote. A variant of bagging is RF, which merges simple decision trees constructed with different parameters. The boosting technique uses a pool of classifiers that are sequentially trained on subsets of data, each time including data misclassified by the previous classifiers. The classifiers are then combined using majority vote.

Another way of improving models is to create hybrid systems, i.e., systems that use different knowledge representation methods. Compared with the approaches previously described, where pre-existing models are integrated, in this case, the integration of different models is planned upfront. In this context, NN with fuzzy systems and NN with symbolic rules have been the most used. The rationale of this choice is that any technique may have limitations; however, these can be overcome through integration with complementary methods. For instance, considering the pros and cons of the most common systems, we can obtain better explanation ability by using rules, or tolerance to noise by using NN, or learning ability by using RF.

The pharmaceutical industry is making a large use of computational methods, including QSAR, in various stages of the development of candidate new drugs. From large numbers (even millions) of potential candidates, it is necessary to extract the best ones, i.e., the most active for the disease under study and the less toxic. The use of QSAR in other industrial areas is less advanced, due to lack of available data and to different needs. As reported in [10], regulators take decisions on the basis of scientific evidence. This evidence, initially given by experimental tests on the chemicals, is more and more substituted by evidence given from in-silico toxicological studies. Problems about introducing QSAR methods for the environmental protection have been presented in [11]. Since then the Environmental Protection Agency (EPA) in US played a pioneering role in developing and applying QSAR [3].

Today many regulations define rules for registering and using chemical substances assessing, at least, two properties: CMR (carcinogenic, mutagenic and reprotoxic) and PBT (persistent,

bioaccumulative and toxic) characterizations. CMR products are supposed to be of high concern for humans. PBT chemicals are substances that are not easily degraded, accumulate in different organisms, and exhibit an acute or chronic toxicity. vPvB substances are defined as substances that are very persistent and very bioaccumulative. They are supposed to have long-term adverse impacts on the environment.

The role played by ML methods in developing models useful for PBT and CMR assessment is relevant. We just review some of those models that applied ML methods.

2.1 CMR Assessment

CMRs chemicals are chronically toxic and have very serious impacts on health. The CMR assessment is necessary for industrial chemicals as well as for other uses. For instance, the European Cosmetics regulation defines that substances classified as CMR are banned in cosmetic products. Mutagenicity, cancerogenicity, and reproductive toxicity are usually taken as binary values, i.e., active or non-active. The potency of toxicity (expressed as a dose) is however considered at least for cancerogenicity in the well-known dataset built by L. Gold and coworkers, and available from the Carcinogenic potency project.³

2.1.1 ML Methods in QSAR for Cancerogenicity

ML methods effectively used to predict carcinogenicity include NN and RF. The earliest models were the result of the Predictive toxicology challenge, which ran in 2000–2001 [12], and was aimed at predicting the cancerogenicity of a set of molecules.

In 1999, [13] described a successful method based on back-propagation NN to predict the carcinogenicity potency of aromatic compounds. Data were taken from the before mentioned Gold dataset. The NN system was integrated with a rule based system to obtain the IARC classification [14]. This was an early example of applying the methods of ensembling in QSAR.

Ensemble models have been proposed also in [15], which developed a set of neural QSAR models for the prediction of the carcinogenicity TD₅₀ index. It used a self-organizing feature map algorithm to select subsets of molecular descriptors, then trained an ensemble of predictive fuzzy ARTMAP networks. Results show that the diversity introduced by the predictors trained using different subsets of descriptors produces better generalization than single models.

A quite large population of chemicals has been used in [16], with data about rat toxicity extracted from the Gold dataset and used to train a counterpropagation NN. The best model uses eight MDL descriptors. For the test set, it obtained accuracy equal to 73%, sensitivity 75%, and specificity 69%. This result is quite interesting and balanced, so it has been further studied and interpreted,

³ <https://toxnet.nlm.nih.gov/cpdb/>

considering that its results are also compatible with the structural alerts so far identified.

Extracting the SAs from the dataset and using them to create a classifier have been proposed in [17]. The method of extraction is provided in SARpy [2], and its advantage is that the QSAR interpretation is straightforward, since the relevant substructures are immediately available. The model uses a large number of SAs, as the idea of SARpy is to privilege large SAs instead of finding their maximum common substructures.

In a recent study [18], the carcinogenicity of polycyclic aromatic hydrocarbons (PAHs) is modeled by using RF. The used dataset contains 91 PAHs, and several molecular descriptors were computed. Different models were developed using partial least squares (PLS), ANN, and RF. The best model, with highest classification accuracy and modeling time, was the RF model. This finding is compatible with the fact that RF is very apt to work on small datasets.

A novel use of QSAR linked to text mining has been proposed in [19]. In this study QSAR data are combined with literature profiles of carcinogenic modes of action automatically generated by a text-mining tool. Using these two methods, individually and combined, the authors evaluated 96 rat carcinogens of the hematopoietic system, liver, lung, and skin. They found that skin and lung rat carcinogens were mainly mutagenic, while the carcinogens affecting the hematopoietic system and the liver were often non-mutagens. The automatic analysis of texts showed how different endpoints as mutagenicity, immunosuppression, and hormonal receptor-mediated effects were found in connection with some of the carcinogens, so allowing identifying more detailed information on biological mechanisms and the relation with chemical structures.

2.1.2 ML Methods in QSAR for Mutagenicity

Mutagenicity has been often approached through SAR systems, based on a number of substructure search. Since many mutagenicity data for the Ames test are available, most of the models predict the results of the Ames test.

The CAESAR mutagenicity model [20] within the VEGAhub is a hybrid model, composed of a SVM model and a rule system. In order to increase the sensitivity, the rule model is applied to substances that have been predicted to be negative (non-mutagenic) by SVM; the rules are a subset of the rules available in ToxTree.⁴ If the output of the second step predicts mutagenicity, the software stops. In contrast, if the output of the second step predicts non-mutagenicity, a third model is applied based on another set of rules to obtain additional assessment.

⁴ ToxTree: <http://toxtree.sourceforge.net/>

The SARpy mutagenicity model [2] in VEGA instead uses a data mining approach to automatically extract SAs from a dataset containing the SMILES of the molecules and their Ames test results. An optimal number of alerts is then kept to create a SAR program. Those rules, integrated with many more SAs extracted from specific datasets, are displayed to the expert in a graphical interface where it is possible to examine the similar compounds [21]. The use of SMILES as the only description of the molecular structure proved to be useful also in optimization techniques, as in [22], which found better results with SMILES than molecular graphs in modeling mutagenicity of aromatic amines.

KNN is another popular strategy in mutagenicity. Lazar [23] is a generic tool developed to create QSAR from any dataset using information on the neighbor compounds. It has been applied to various endpoints, including mutagenicity.

RF also has been applied to mutagenicity, by [24]. NN have been used in a number of models for non-congeneric compounds. Among them let us cite the work [25], which used 2D and 3D descriptors to develop a NN model for genotoxicity.

2.1.3 ML Methods in QSAR for Reproductive Toxicity

Reproductive toxicity indicates an endpoint much more complex in the definition, which has been addressed quite recently. Reproductive toxicity indicates the adverse effects induced by chemicals on fertility and developmental toxicity in the offspring.

Few experimental data are available, and the construction of QSAR just at the beginning. As reported in [26] there are many problems in applying QSAR to reproductive toxicology since there is a variety of endpoints associated to reproductive toxicology and a lack of data to make models. Models about ADME relating to reproductive toxicity and to endocrine disruption are more developed than global models, using a large family of chemical compounds, for reproductive endpoints. Initial models were based on expert systems technology, incorporating expert knowledge, and are sometimes part of commercial systems.

Among the first freely available models is the TEST-VEGA model [27]. It applies RF and has been chosen after developing other models with different methods; it is today available both in VEGA and in TEST.⁵

2.2 PBT Assessment

The term PBT, introduced in Japan's Chemical Substances Control Law in 1973, was later adopted by regulations in several countries including European Union (REACH regulation), the USA, and Canada. Under REACH, the PBT/vPvB assessment is an important part of the chemical safety assessment that must be conducted for the registration dossiers, alongside with the CMR assessment.

⁵ <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>

2.2.1 ML Methods in QSAR for Persistence

The authors of [28] discussed methods based on artificial intelligence for the prediction of chemical biodegradability; in particular, they considered regression, expert systems, and ML models. They proposed an inductive logic model, which is expressed as a set of rules, that can be compared to the concept of SAs already seen in other SAR systems.

Another reported model for ready biodegradability [29] is based on SARpy to extract rules for the prediction of ready biodegradable, non-ready biodegradable, and possibly biodegradable.

2.2.2 ML Methods in QSAR for Bioaccumulation

Bioconcentration factor (BCF) describes the behavior of a chemical in terms of its likelihood of concentrating in organisms in the environment.

Miller and coworkers [30] have published a review of methods used in bioconcentration factor. They extensively compared linear and ML models for the prediction of bioconcentration in fish *Cyprinus carpio*. An optimized multilayer perceptron with 14 descriptors was selected for further testing on invertebrates, with good results.

The BCF CAESAR model, available in VEGA, is a hybrid system integrating two models built with SVM [31].

Making use of models that directly use SA is a quite popular method that addressed the need of explaining the results of the QSAR. It is much more explicative that the interpretation of the molecular descriptors that are in general selected from huge numbers. Using SARpy, the SAs for BCF have been extracted by [32]. Indeed the use of SAs as rules has been proposed in a stand-alone system [33], which shows in a graphical interface the similar molecules and the relevant rules to help the expert in the toxicity assessment of BCF.

2.2.3 ML Methods in QSAR for Toxicity

Environmental toxicity according to the different legislation requires the analysis of acute and/or chronic toxicity for water and terrestrial animals and plants. A very large number of endpoints can be relevant to assess toxicity. Of great concern is aquatic toxicity, for which fish, alga, and daphnia are the most common organisms addressed. For terrestrial toxicity rats, bees, and birds are the most common targets used. Among the many developed models, we can make a distinction between models for generic industrial chemicals and models for pesticides.

Some models for pesticides using ML methods have been developed in the DEMETRA project and are described in [34]. Endpoints include acute toxicity toward fish, birds, and bees.

Fish toxicity is one of the most studied endpoints for QSAR. Many ML methods have been developed for industrial chemicals due to the large dataset of test data available from EPA on the fathead minnow fish. A model making use of this dataset for training, and using trout toxicity data as an external test set, is presented

in [35]. It uses SARpy to extract the SAs that characterize a classification in three toxicity classes. Using the same dataset, other models used ML methods. One of them [36] is an NN ensemble, and another [37] used adaptive fuzzy partitioning.

A model about Daphnia toxicity [38] uses Monte Carlo optimization as offered in the before mentioned CORAL system.

Of great interest for terrestrial toxicology are endpoints in rat and birds.

Median lethal death, LD₅₀, is an indicator of acute oral toxicity (AOT). There is a recent interest in ML methods to predict LD₅₀, a value of high interest in both pharmaceutical and industrial chemicals assessment. In recent years, [39] developed three models for AOT using a convolutional NN that outperformed previously reported models. Moreover the authors performed automatic feature learning, to map activation values into fragment space and derive AOT-related chemical substructures.

For regulatory purposes, [40] developed QSARs using deep learning as part of the Predictive Models for Acute Oral Systemic Toxicity project hosted by the ICCVAM Acute Toxicity Workgroup. The networks on fingerprint descriptors [41] demonstrated a way to combine multiple models with a neuro-fuzzy system to greatly improve the prediction and the interpretability of results for avian toxicity LD₅₀. Mazzatorta and coworkers [42] have explored SVM methods to build QSAR for oral bird toxicity using genetic algorithms for reducing the number of descriptors. They conducted a study on 116 pesticides for avian oral toxicity. The analysis of the descriptors indicates the prominent role of the interaction of pesticides with macromolecules and/or proteins in the mechanism of action. On a similar dataset, [43] developed an ensemble QSAR combining together more QSARs; they used also an extension of the ROC curve method to show the initial and combined models.

3 The New Methods of Deep Learning

In the following sections, we show the steps to construct a QSAR model using the recently developed methods of deep NN (DNN); we adapt an architecture used for image recognition and called Inception [44] and take inspiration from the Chemception network [45] developed in the chemical domain. It is necessary to explain the principles of those neural architectures before presenting our model.

3.1 From Neural Networks to Deep Learning

NN are a biologically inspired programming paradigm, firstly described by [46], which enables a computer to learn from data. They can also be described as a mathematical function that maps a given input to the desired output. NN is not an algorithm but can be considered a framework to process complex input data. The

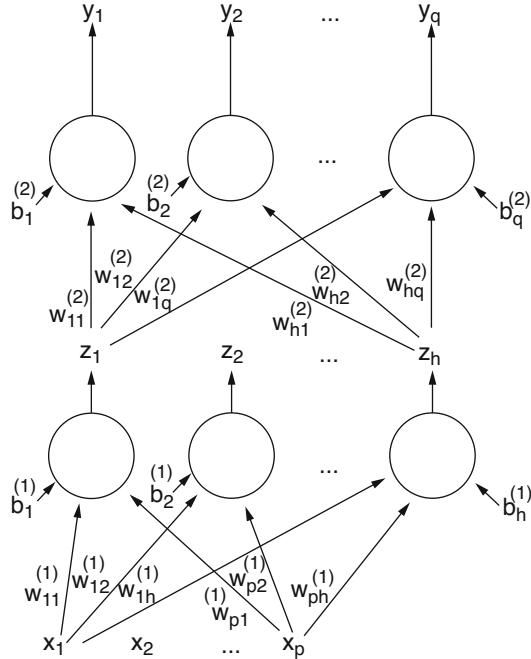


Fig. 1 A neural network with weights, bias, and two fully connected hidden layers. The input x goes into the first inner layer; the output of each hidden layer is sent to all the second layer neurons

same network can be applied to many different tasks depending on the dataset it has been trained with. A NN is a collection of connected units, called nodes, that are divided in three categories, input layers, hidden layers, and output layers, as shown in Fig. 1. Each node is connected to the successive with a link. Each link has a weight w_i that is a real number that is multiplied for the output of each node, and it is passed to the next adjacent one. A bias can be added to the sum of these weights to facilitate the network's training.

The neurons are fully connected in a hierarchical order. The breakthrough for NN can be considered the introduction of the backpropagation algorithm [47], a method to train multilayer networks in a feasible and efficient way. This algorithm allows calculating the gradient of the loss function with respect to the weights in the NN. The weights updates can be done via stochastic gradient descent using Eq. 1:

$$w_{ij}(t+1) = w_{ij}(t) + \eta (\delta C / \delta w_{ij}) + \xi(t) \quad (1)$$

where η is the learning rate, C is the loss function, and $\xi(t)$ is a stochastic term.

The choice of the loss and the activation functions depends on the problem analyzed. The most used are mean square error (MSE) and cross entropy (xentropy).

The activation function is used to get the output from a neuron. The most used activation function is the sigmoid function that exists between 0 and 1 and is especially useful for models that predict the probability as an output. Moreover the function is differentiable (it is possible to find the slope at any two points) and monotonic. The softmax function is instead used for multiclass classification.

The sigmoid function can cause a neural network to get stuck at the training time. In fact it squishes a large input space into a small input space between 0 and 1; a large change in the input causes a small change in the output. Hence, the derivative becomes small, making it difficult for the backpropagation algorithm to converge; this is called the vanishing gradient problem. In 2011, the ReLU (rectified linear unit) activation function was proposed to solve it [48].

ReLU is a half-rectified function (Eq. 2) that returns its argument x whenever it is greater than 0 and returns 0 otherwise. Its first derivative is 1 for $x > 0$, so its value is never too small:

$$f(x) = \max(0, x) \quad (2)$$

This function allowed the birth of deep learning, a new machine learning method that gave rise to DNN. In the 1990s of the last century, data scientists and computational toxicologists used NN in different ways, especially to predict toxicity of substances and to make feature selection [49]. Indeed in computational toxicology, the main interesting aspect about NNs is that, opposite to the parametric methods, they do not require any user expertise or any a priori knowledge to analyze and discover patterns. Their main drawback is, however, the difficulty in creating the architecture; the number of layers and of neurons for each layer is determined in a heuristic way and with trials. Moreover, early NNs were prone to over fitting, so their training required care.

The resurging interest on NNs is due to the new amounts of data. Since the 1990s, new technologies have been developed and widely applied to produce large amounts of chemical and biological data; in particular high-throughput screening (HTS) and high-content screening (HCS) provide information on the biological activity of thousands of compounds.

DNNs have seen tremendous development in the last decade due to their application to image analysis and object recognition. The main idea is to train the computer by showing different examples of the same object, in order to make the machine understand and recognize the object based on the images features. This problem is called Large-Scale Visual Recognition Challenge (LSVRC); every year, the ImageNet contest,⁶ i.e., a challenge on the big

⁶ <http://image-net.org/challenges/LSVRC/>

dataset ImageNet, declares the winner network that proved to have the best accuracy and the smallest error. Error reduction has been impressive, going from 16% in 2012 to 3.5% in 2015; this error is lower than the human error, which is about 5%.

Recognizing objects from images is the same as recognizing a molecule as toxic or not. In chemistry, DNN have gained popularity after recent achievements that include DNN-based models winning the Merck Kaggle challenge⁷ for activity prediction in 2012 and the NIH Tox21 challenge for toxicity prediction in 2014. What is common in those networks is that they are deep, i.e., they have many hidden layers, and that the neurons in the net are not fully connected.

The most powerful model to solve LSVRC is the convolutional neural network (CNN). The most active actors of this field are Google, Microsoft, and Facebook, which today define the current state of the art. CNNs [50] typically adopt a standard structure with stacked convolutional layers, followed by one or more fully connected layers. CNNs use many of the same ideas as the NNs, such as backpropagation, gradient descent, regularization, nonlinear activation functions, and so on.

3.2 Convolutionary NN

NNs have an important theoretical advantage over other methods: the universality theorem states that a NN with at least one hidden layer can approximate any continuous or discontinuous function. Moreover, adding layers in the net increases the number of parameters to fit and so allows to better approximate complex functions.

Why adding fully connected layers in NN without considering the properties of the data representation? The architecture of the network should take into account the spatial structure of images. For instance, why treat pixels that are far apart and close together? So, instead of starting with a generic NN multilayer architecture, we take advantage of the spatial structure. For images it is accepted that the neighborhood rather than the pixel carries the geometrical interpretation; using this principle, *Convolutional NN (CNN)* adds trainable filters and neighboring local pooling operations in an alternate sequence.

DNNs contain simple nonlinear processing units, each transforming the representation at one level (starting from the input image) into a representation at a higher level. In practice, the network works as a representation learning method, learning from low to high level features, without the need of computing and selecting the relevant features (chemical descriptors in our case).

The first CNN was developed in 1995 to classify handwritten digits [51]. The basic CNN, as in Fig. 2, is composed of

⁷<https://www.kaggle.com/c/MerckActivity>

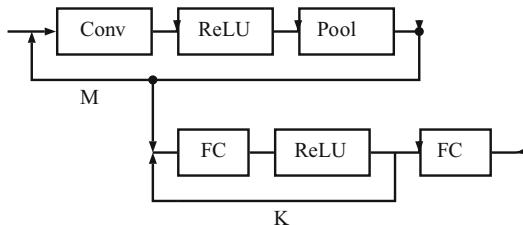


Fig. 2 A basic CNN network

convolutional layers (Conv), rectified linear units (ReLU), pooling layer (Pool), and fully connected layers (FC). The operations are performed for M and K number of times as indicated.

3.2.1 Convolutional Layer

We connect the pixels of the input image to a layer of hidden neurons. But we won't connect every input pixel to every hidden neuron: we only make connections in small, localized regions of the input image. Each neuron in the first hidden layer is connected to a small region of the input neurons, for example, a 5×5 region, corresponding to 25 input pixels. Each connection learns a weight, and there is a common bias for the region. We slide this 5×5 window across the entire input image, each time acting on a different neuron in the hidden layer. The sliding can be done one pixel at a time or using a different number of pixels; this number is called stride length. We use the same weights and bias for each of the hidden neurons, so reducing the number of parameters of a convolutional network, which is at least one order of magnitude less than the number of parameters of the fully connected layer. A Conv layer is composed of several feature maps (with different weight vectors), so that multiple features can be extracted at each location.

Each Conv layer is followed by the pooling layer, which performs local averaging and subsampling, so reducing the resolution of the feature map. The activation layer ReLU is used to control the effect of the squashing nonlinearity. This triad of modules can be repeated in a sequence to increase the number of feature maps and to decrease the spatial resolution. The FC layer is used then to find the correlation between the label and the features contained in the feature maps.

A single step of convolution multiplies and sums the pixel values of an image with the values of a filter. This filter can be of shape $N \times N$. Next step, the filter is shifted to a different position, and the convolutional step is repeated until all pixels are processed at least once. In Fig. 3, there is an example of applying it to a pixel, with $N = 3$.

The resulting matrix eventually detects edges, or transitions between dark and light colors, and eventually more complex forms. The more filters are applied, the more details the CNN is able of recognizing. Moreover, the inclusion of the ReLU aims to

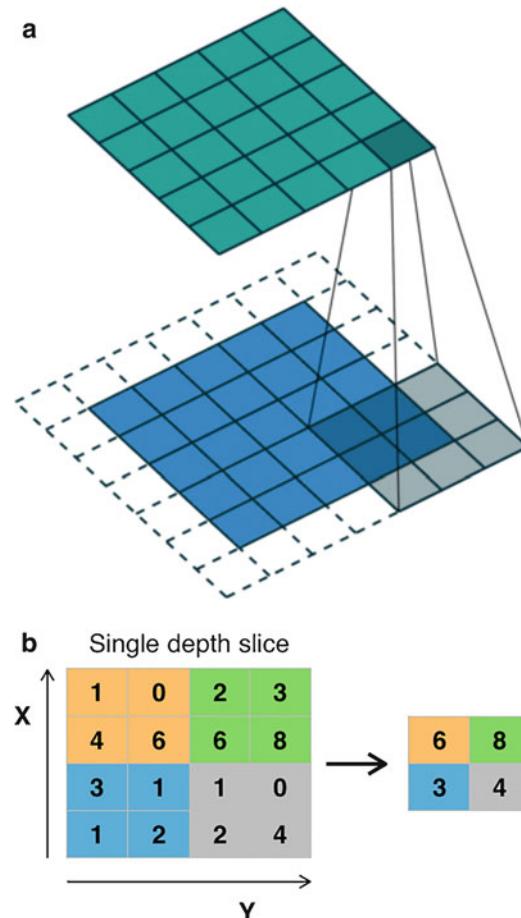


Fig. 3 (a) A step in the convolution process with a filter 3×3 . **(b)** An example of the max pooling operation with a filter of size 2

apply an element-wise activation function, such as sigmoid, to the output of the activation produced by the previous layer. This nonlinear function is necessary for the network to be able to represent nonlinear relationships between neurons.

3.2.2 Pooling Layer

Pool takes as parameter the dimension of the output mask. The most used are max pooling and average pooling; max pooling selects the maximum value of all selected squares to make feature detection more robust. Average pooling uses instead the average of all values. Neither of this two pooling methods requires parameters, so backpropagation also does not need to learn anything. Max pooling is generally preferred. Max pooling is a way for the network to ask whether a given feature is found anywhere in a region of the image. It then throws away the exact position, since once a feature has been found, only its rough location relative to other features is important. The advantage is that there are fewer pooled features; so

later layers will require fewer parameters. Figure 3 on the right shows an example of max pooling when a filter = 2 is used.

3.2.3 Fully Connected Layer

The final layer in the CNN network is a FC layer; this layer connects every neuron from the max-pooled layer to every one of the output neurons, as in regular feed forward NNs. Their activations can hence be computed with a matrix multiplication followed by a bias offset. It is also suggested by [52] that ReLU may be used between these layers to improve performance.

A distinguishing feature of CNNs is that many neurons share the same filter, so reducing memory needs, because a single bias and a single vector of weights are used. This leads to the deep architecture called Inception [44] that allows the net to be computationally feasible.

3.3 Inception Network

The first Inception network, the GoogleNet [53], introduced the concept of inception, which means to apply three different filters on the image: 1×1 , 3×3 , and 5×5 . Then, with the Microsoft network ResNet [54], the networks started going wider, with different modules in parallel, instead of going deeper, so reducing the computation time using GPU.

Inspired by a neuroscience model of the primate visual cortex [55], the Inception network uses a series of filters of different sizes to handle multiple scales. Moreover, this network includes the concept of Network-in-Network by [56] in order to increase the representation power; additional 1×1 convolutional layers are added as dimension reduction modules, allowing increasing both the depth and the width of the networks without a significant performance penalty. The main drawbacks are that computational resources and the number of parameters also increase, which make the network more prone to over fitting. A fundamental way of solving both issues is to introduce sparsity, replacing the fully connected layers by sparse ones, even inside the convolutions, so mimicking biological systems. This is done by building subgroups of convolutional blocks that form the units of the next layer and are connected to the units in the previous layer. We assume that each unit from an earlier layer corresponds to some region of the input image. These “Inception modules” are stacked on top of each other. As higher layers capture features of higher abstraction, their spatial concentration decreases. The “Inception module with dimensionality reduction” in Fig. 4 includes 1×1 convolutions that are used to compute reductions before the expensive 3×3 and 5×5 convolutions. The first Inception network was a combination of these blocks with an occasional max-pooling layer with stride two, leading to 22 deep layers.

According to [52], it is wise to balance width and depth of the network. This is obtained through factorization, i.e., decomposing larger filter sizes into more layers of filters: 5×5 convolutions are

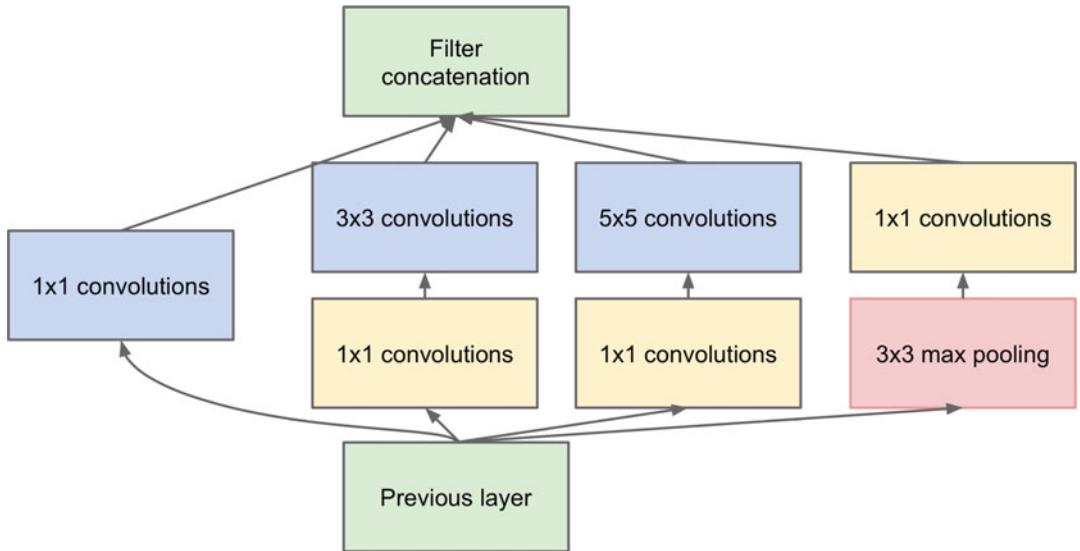


Fig. 4 Inception module with dimensionality reduction

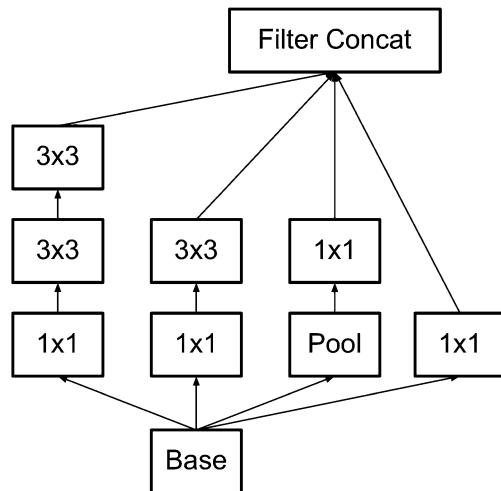


Fig. 5 Inception modules where each 5×5 convolution is replaced by two 3×3 convolutions. The 3×3 convolutions are then further decomposed in 1×3 and 3×1 convolutions

replaced by two layers of 3×3 convolution. Asymmetric convolutions, e.g., $n \times 1$, are used to push even further the factorization process. For example, a 3×1 convolution followed by a 1×3 convolution is equivalent to sliding a two-layer network as in a 3×3 convolution.

The factorization in Fig. 5 represents Inception-v3 [52], which has 42 layers.

3.3.1 Residual Network

Let us consider $H(x)$ as an underlying mapping to be fit by a few stacked layers, with x denotes the inputs to the first of these layers. Since multiple nonlinear layers can asymptotically approximate any complicated function, then they can asymptotically approximate the residual functions, i.e., $H(x) - x$. So we explicitly let these layers to approximate a residual function $F(x) := H(x) - x$. The original function thus becomes $F(x) + x$. In this reformulation, if identity mappings are optimal, the solvers may simply drive the weights of the multiple nonlinear layers toward zero to approach identity mappings.

ResNet [54] used this concept named residual learning; each building block is considered with residual learning, and it is expressed as in Eq. 3:

$$y = F(x, W_i) + x \quad (3)$$

where x is the input of the vector of the layers considered; y is the output of the vector of the layers considered; and $F(x, W_i)$ represents the residual mapping to be learned.

The ResNet network contains the residual connection as in Fig. 6, has lower complexity, and is 152-layer depth.

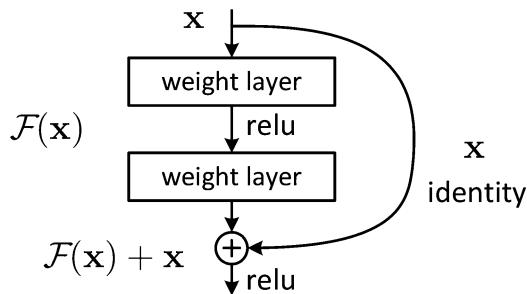
In conclusion, the final inception network [44] combines residual network [54] and inception with factorization [52]. Also, the stem block is introduced to replace all the linear layers of the precedent networks. The main advantage with respect to older CNN is that training time is highly reduced.

4 Dataset Construction and Preprocessing

The property of interest we explored for building (Q)SAR DNN is mutagenicity. Mutagenicity refers to a chemical or physical agent's capacity to cause mutations, i.e., genetic alterations. Agents that damage DNA causing lesions that result in cell death or mutations are genotoxins. All mutagens are genotoxic, but not all genotoxins are mutagens.

4.1 Ames Test for Mutagenicity and Its Models

The most adopted test for mutagenicity is the Ames test [57]. It uses bacteria *Salmonella typhimurium* and rat liver, and it is executed in a test tube. Different bacterial strains are sensitive to different types of mutations; so more than one strain should be used. This test has some limitations as *Salmonella typhimurium* is a prokaryote; therefore it is not a perfect model for humans, and rat liver is used to mimic the human metabolic conditions. Moreover, the reproducibility of this test is around 85% as reported by [58]. This implies that the accuracy of any computational model cannot reach 100% of accuracy. For instance, chemicals that contain the nitrate moiety sometimes are positive for Ames when they are indeed safe.

**Fig. 6** The building block for residual learning**Table 1**

Results of models tested on an external test set of about 2000 chemicals as reported in [59]

	TESTI	CAESAR	ISS	SARpy	KNN	SAm	Aim	AZAMES
True positives	110	208	187	182	163	178	126	116
False negatives	197	110	131	136	155	140	189	196
False positives	377	683	592	678	635	377	395	178
True negatives	1673	1426	1517	1431	1474	1732	1689	1925
Total predictions	2357	2427	2427	2427	2427	2427	2399	2415
Coverage	0.97	1.00	1.00	1.00	1.00	1.00	0.99	1.00
Positive pred. value	0.23	0.23	0.24	0.21	0.20	0.32	0.24	0.39
Negative pred. value	0.89	0.93	0.92	0.91	0.90	0.93	0.90	0.91
Balanced accuracy	0.59	0.67	0.65	0.63	0.61	0.69	0.61	0.64
Accuracy	0.76	0.67	0.70	0.66	0.67	0.79	0.76	0.85
Sensitivity	0.36	0.65	0.59	0.57	0.51	0.56	0.40	0.37
Specificity	0.82	0.68	0.72	0.68	0.70	0.82	0.81	0.92
Matthews correlation coefficient	0.14	0.23	0.22	0.18	0.15	0.31	0.17	0.29

We decided to model the Ames test for various reasons. The first is that this test is necessary for the registration of any chemical in different regulations and in the screening process. Moreover the cost the AMES test is quite high, making the model of interest in all the situations that require prioritization. The third and crucial fact that influenced our choice is the number of data available in literature and in databases. Of course, the choice of the Ames also implies some drawbacks. The most important is the different standardization of the available data, which requires data preprocessing.

Many (Q)SAR models for predicting the Ames test are in the literature and in use. In a recent study [59], whose results are

reported in Table 1, many models based on various techniques were considered, in particular: the US-EPA T.E.S.T. [60]; four models in VEGAhub, namely, CAESAR [20], SARpy [2], ISS implementing Toxtree [61], and KNN [62]; two models by Nestlé [63]; and one model from SweTox [64]. The results reported in Table 1 are on an external test set extracted from a confidential dataset of 18,338 compounds.

4.2 SMILES and Chemical Graphs

Simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemicals using short ASCII strings. This specification was first described in [65]. It allows to pass from a 2D representation of a chemical into a simple string and to do the inverse process from the string to the image.

Typically multiple SMILESs can be written in different ways for a single molecule, since there exist different algorithms to encode the structure into a string. The algorithm used for our data is canonicalization algorithm, and the SMILES derived from are called canonical SMILES. In terms of procedure, the algorithm is graph based: the string is obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. The chemical graph is first trimmed to remove hydrogen atoms, and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

The resultant SMILES string depends on some choices: the bonds chosen to break cycles, the starting atom used for the depth-first traversal, and the order in which branches are listed. To analyze a dataset, it is necessary that all the SMILES strings be obtained in a consistent way; in our case the SMILES taken from different libraries are all normalized using the SMILES generation algorithm of VEGA [66].

4.3 Images Generated from Smiles

Two modules compose the function realized by RDKit⁸ to convert a SMILES string to a two-dimensional drawing: a SMILES parser to convert the SMILES back to its parent spanning tree and a SMILES drawer to convert this spanning tree to a two-dimensional structure drawing.

- The parser module generates a tree from the input SMILES, in which each atom is encoded by a node object in a linked tree data structure. The topology of the parse tree is identical to the spanning tree used to generate the SMILES string. In practice, the parser uses a simple context-free grammar. In addition the parser can identify the location of an erroneous symbol.

⁸ Rdkit. URL <https://bit.ly/2OYLjj9>

- The SMILES drawer module converts the parse tree obtained from the SMILES to a 2D-structure drawing. The module positions acyclic atoms, atoms in fused rings, and atoms in Spiros based on Euclidean and molecular geometry according to the VSEPR model. The placement of bridged ring systems with n-rings ≥ 2 is treated as a two-dimensional graph embedding problem, solved using graph theoretic distances. The algorithm sets up a virtual dynamic system, where weighted topological distances between all vertices are modeled as springs. The spring introduces repulsive electrical forces between no connected vertices to keep them apart. The drawer, at the end of the process, saves the image to a PNG or SVG file.

4.4 The Dataset

As we said, the execution of the Ames test is quite expensive so the results are often proprietary. In order to collect a sufficient number of chemicals tested, we researched in the literature and on the web. We had to distinguish between trustworthy sources and suspicious sources. We defined a trusted source as “a source that is published from certified authorities or that has been used in computational model approved by the regulators.”

The main trusted source is created by National Institute of Health Sciences, Japan (DGM/NIHS) as part of their Ames/QSAR International Challenge Project [67]. It contains around 12 thousand compounds, pharmaceutical or industrial products. Between the trusted sources, we also included two databases created by the National Cancer Institute (NCI): GeneTox and CCRIS. We also used data from the VEGAhub that merges most of the cited database.

We also used as suspicious source CGX [68]. This source contains various Ames tests for each chemical and defines as positive to AMES test a chemical where more than three tests result positive. However, we wanted to keep a more cautious definition of mutagenic chemicals, so we defined positive to Ames test the chemical if just one of the tests resulted positive. We also found a list of really cited sources as the ones provided by [59].

All these different sources contain pharmaceuticals, pesticides, and industrial products. We paid a particular attention in the selection of the SMILES. In particular, excluding the trusted sources, for every result found, we compared it with the trusted databases in order to find duplicates, and we eliminated all the incoherent tuples. Moreover, we kept the database source during the whole preprocessing phase in order to eliminate at each step the duplicated derived from a dataset that was not marked as trusted. Unfortunately, it was not possible to collect also the full specifications of the conducted tests, since the different sources do not always report the standards applied.

Another important information we would like to have is the use of each compound, since there are different regulations for the toxicity test depending on the final usage the chemical. Unfortunately, in the used databases, there is no such information.

5 Toxception and Its Results for Mutagenicity

Toxception is based on Inception and on ideas proposed by Goh [45] who developed Chemception, a deep model about Tox21 database and HIV virus. The main differences between Goh's and our network are the absence of a priori chemical knowledge and the elimination of image transformations.

To adapt the inception network to our problem, we reduced the size of the input image from $299 \times 299 \times 3$ to $80 \times 80 \times 3$ changing also the number of nodes. Differently from Chemception we did not use data augmentation, and we passed the images to the network exactly as they exit from the generation function. In fact both the chemical knowledge passed to the network and the image preprocessing done in Chemception have not improved the performance of the model.

An abstract bottom-up scheme of the Inception-Toxception network is in Fig. 7; the input image goes to hidden convolutional layers and, after pooling, to the output layer. A dropout layer is added to prevent the net from overfitting [69]. Dropout means that, at each training stage, individual nodes are either dropped out of the network with probability $1 - p$ or kept with probability p , so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed. This regularization reduces over-fitting by adding a penalty to the loss function. However, it also increases the number of iterations required to converge. A detailed view of the inner organization is in Fig. 8. Globally the net has 650,881 parameters to optimize.

5.1 Learning and Optimization in Toxception

The optimization algorithms help to minimize the error function $E(x)$ which is a mathematical function dependent on the model internal learnable parameters. The different optimization algorithms are called optimizers; in this work we used stochastic gradient descent algorithms.

RMSProp (root mean square propagation) adapts the learning rate for each of the parameters on the average first moment, while the Adam algorithm [70] uses the average of the second moments of the gradients. It has three parameters that can be optimized: learning rate, exponential decay rate for the first moment estimates, and exponential decay rate for the second-moment estimates (this value should be close to 1.0).

We trained many times the network with Adam and a combination of RMSProp for half of the training epochs and SDG

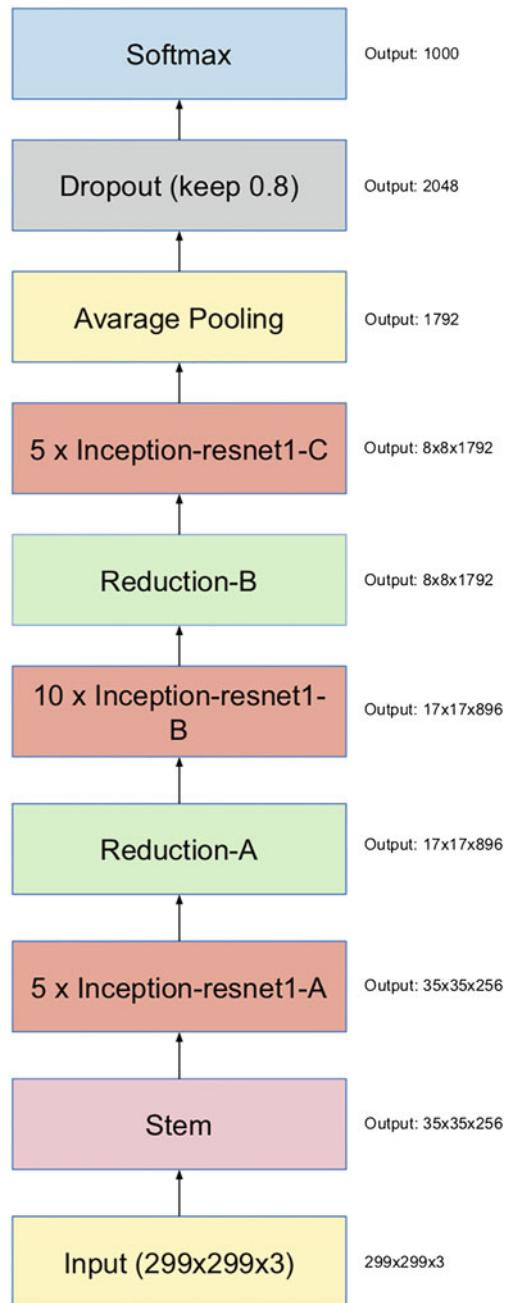


Fig. 7 Architecture of Toxception (based on Inception ResNet-v2)

(Stochastic gradient descent) for the rest. The parameters used are reported in Table 2.

Toxception is a big network, and this implies a high number of hyper parameters to manage and optimize. Testing all the possible combinations is impossible. We constructed a decision tree with the

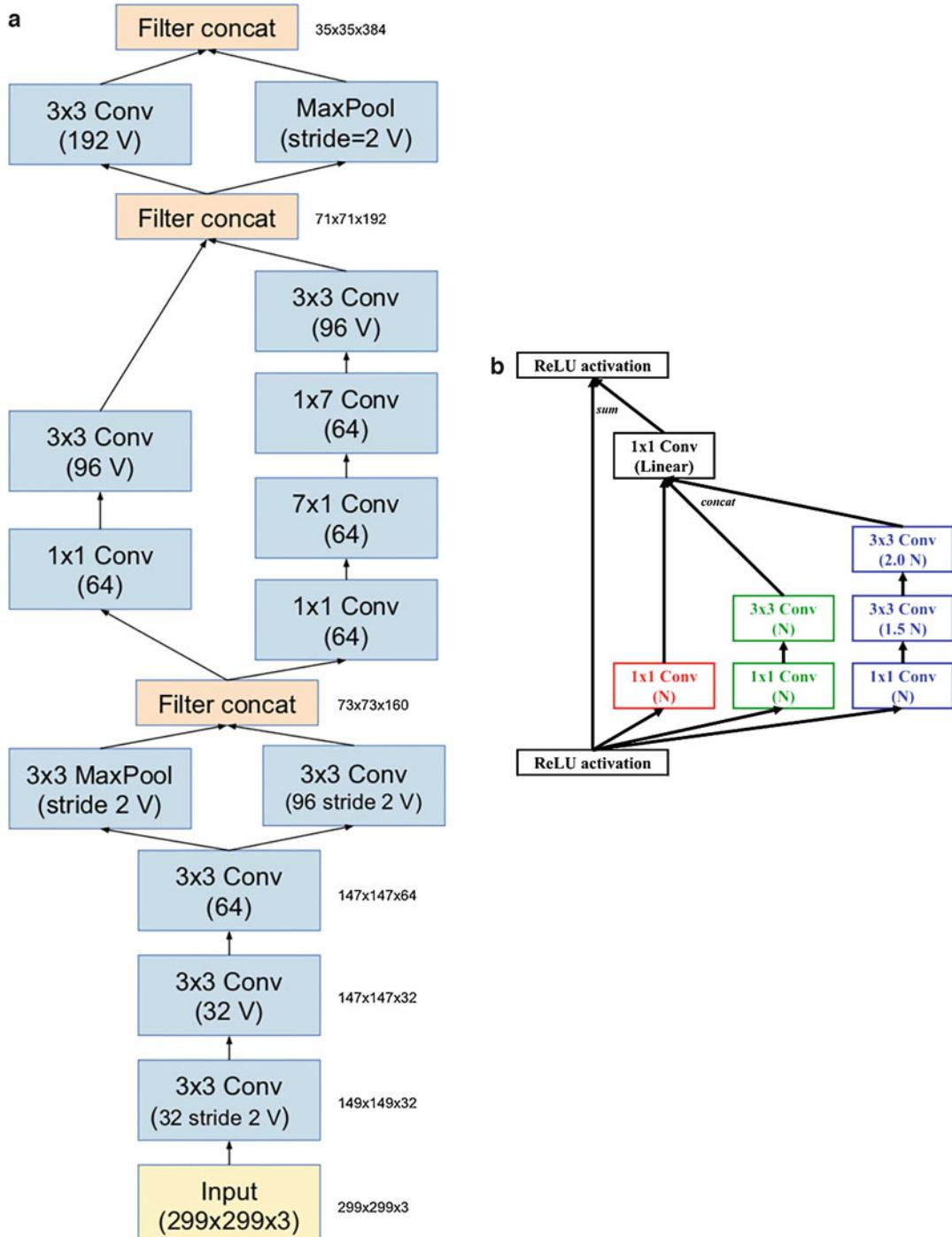


Fig. 8 The blocks of the network in Fig. 7. (a). Stem Block; (b). ResNet1A; (c). ResNet1B; (d). ResNet 1C; (e). Reduction A; (f). Reduction B. Layers (n colored boxes) have a ReLU activation layer after the specified convolution layer, with a stride of 1, and same padding unless otherwise noted. Each block has N convolutional filters for each layer, and the variations are indicated as multiples of N

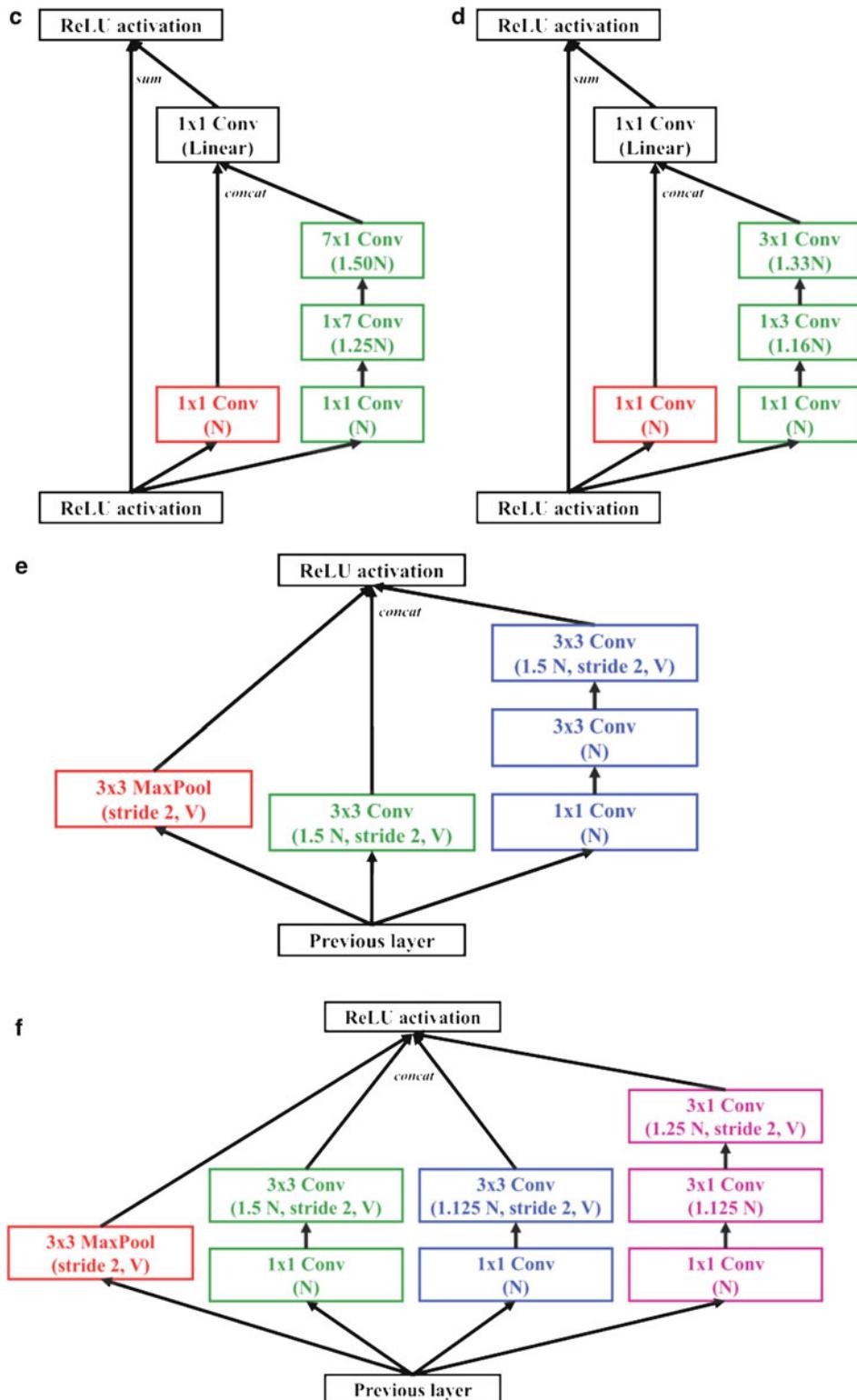
**Fig. 8** (continued)

Table 2
Parameters and optimization algorithms used

Optimizer	Parameter	Value
RMSProp	Initial lr	e^{-3}
RMSProp	Rho	0.9
RMSProp	Epsilon	e^{-8}
RMSProp	Decay	0
SDG	Initial lr	e^{-3}
SDG	Momentum	0.9
SDG	Gamma	e^{-8}

most important parameters that could affect accuracy and computation resources; a tool for this task is provided in Talos,⁹ which allows the user to set the parameters to change and run all the possible permutations of them.

We decided to leave unchanged the convolution layers in each block, taking as a variable only the number of neurons in the first layer. We tested different dropout values (0.4, 0.2, 0.1, 0) to check how to prevent over-fitting without affecting the performances. In order to reduce the complexity of the network, we decided to reduce the stem block, and we tested the different solutions to analyze if the stem block was needed. Finally, we tested different epochs and early stopping technique.

5.2 Results of Toxception

We applied Toxception to the dataset already described, transformed into images with resolution 80×80 pixels. We used a randomly selected 20% of the data collected as the validation set. In addition, we used as metrics in order to optimize the network the loss on the validation set and the accuracy on the validation set. Even though accuracy does not prove the consistency of the model, in image classification it is considered a good metric to evaluate the model. Moreover, we use sensitivity and specificity on the validation set. In order to discover the potentiality of our model, we tested both MSE and xentropy loss functions.

Also, we analyzed the probability distribution given by the network without rounding to 0 or 1 the final values. It is the best practice to analyze these values in order to avoid the phenomenon of false results, which happens when the classification results are correct, but the outcome of the network is actually too weak. Moreover, this allows extracting the probability for a particular compound to belong to a tox or no-tox class.

⁹Talos. URL <https://bit.ly/2yL9gQJ>

Table 3
Optimization results using MSE as loss function

Architecture	Parameters	Metrics					
		Acc	Loss	Val Acc	Val Loss	Val Spec	Val Sens
Tox-basic _{A1}	Epochs Optimizer Neurons	200 Adam 16	0.99 0.01	0.81	0.16	0.59	0.84
Tox-basic _{B1}	Epochs Optimizer Neurons	200 RMS 16	0.99 0.01	0.80	0.15	0.62	0.87
Tox _{A1}	Epochs Optimizer Neurons	200 Adam 32	0.85 0.12	0.79	0.17	0.67	0.87
	Epochs Optimizer Neurons	200 Adam 16	0.96 0.37	0.79	0.18	0.63	0.89
	Epochs Optimizer Neurons	200 RMS 32	0.82 0.14	0.77	0.18	0.38	0.93
	Epochs Optimizer Neurons	200 RMS 16	0.95 0.04	0.79	0.17	0.65	0.87

Using Talos, we ran multiple times the network on the training set applying the cross-validation method. In Tables 3 and 4, we report the best results. In particular, Table 3 reports results in terms of MSE, and Table 4 reports results with xentropy. This is the legenda:

- Tox_Basic indicates the Toxception network without stem block.
- Tox includes also the stem block.
- The subscript next to each name indicates as follows: _A, the network has been trained using RMSProp optimizer for 100 epochs and then re-trained with the SDG; _B, the Adam optimizer has been used or the number of epochs indicated in the tables.
- Neurons indicate the number of neurons in the first Inception block; the values of the others are not cited as they are strictly related to that one.

Figure 9 reports the learning curves for accuracy and loss using MSE without the stem block and using the Adam optimizer.

We start all the training with a learning rate of $1e^{-3}$, and we apply the early stopping technique. Looking at the tables, there are

Table 4
Optimization results using xentropy as loss function

Architecture	Parameters	Metrics					
		Acc	Loss	Val Acc	Val Loss	Vol Spec	Vol Sens
Tox-basic _A	Epochs Optimizer Neurons	200 Adam 16	0.99 0.02	0.80	0.17	0.62	0.88
Tox-basic _B	Epochs Optimizer Neurons	200 RMS 16	0.84 0.13	0.77	0.17	0.42	0.79
ToxA	Epochs Optimizer Neurons	200 Adam 16	0.96 0.09	0.78	0.38	—	—
ToxB	Epochs Optimizer Neurons	200 RMS 16	0.83 0.14	0.76	0.18	0.39	0.92

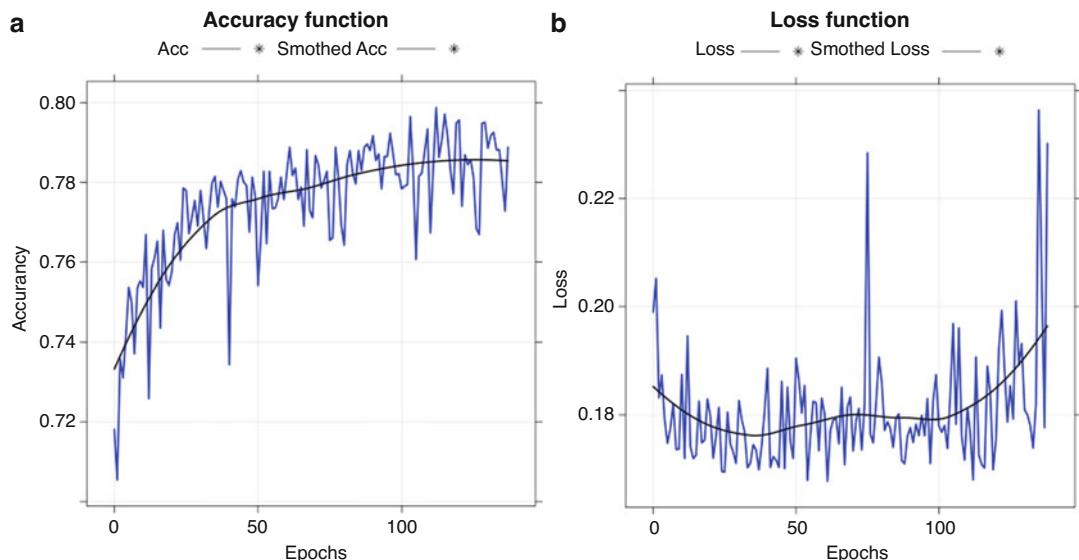


Fig. 9 In (a) accuracy, in (b) loss function learning curves using MSE and the Adam optimizer, without the stem block. We can observe how the accuracy and the loss function follow the normal training pattern with the increase of the epochs. All the metrics reported here are calculated on the validation set

different aspects to point out. First of all, it is clear the effect of the Stem on the network. In fact, in all the simulations, the usage of the stem block limits the learning. It can be explained by the simplicity of the data passed into Toxception. As already said, the images are composed by a lot of white space; this implies that the division of the input into small features can considerably increase the resources

needed from the network to converge. Observing Tox_B_1 and Tox_B_2, it is visible how increasing the training epochs allows the network to converge and to minimize the loss function. However, the results achieved by these networks in terms of accuracy are not so better than basic networks.

The second interesting point is about sensitivity and specificity. A low value of the sensitivity implies a high number of false negatives. The implications of a high number of false negatives are obvious. As we can see in the tables, this number in Toxception is small as the sensitivity is more than 85%. However, the values of specificity, in all the combinations, are not really high. It is due to precautionary choices we have made during data collection and data preprocessing. As already reported while collecting the data, we decided to include all the Ames test results in the literature but, in order to create a safe model, we considered as mutagenic all the compounds that appeared more than once with opposite results.

The third aspect is the comparison between RMSProp and Adam Optimizer. As we can see in Tables 3 and 4, the Adam optimizer seems to perform better than the RMS. That could be due to the particular learning rate schedule we implemented in RMS; we split the training epochs into two, and we trained the network for half using RMS and for another half using SDG.

With respect to Chemception, our model performs well without receiving any basic knowledge of chemistry. Our method also can adopt different loss functions, while Chemception gives low results if trained with crossentropy.

Another parameter that needs to be considered is the computation time needed to execute the training and the prediction. Toxception takes around 130 milliseconds per step, and each epoch is composed of 600 steps. This makes the time spent by the networks for the training process around 4 hours. The prediction instead is fast as we do not need to calculate the weight of each neuron. To evaluate the whole dataset, the net takes around 5 min; this means around 12 milliseconds per compound.

In order to better analyze the results, we run some other training for the most interesting networks. In particular, we report below some results of three main configurations:

- Tox_basic_A with Adam optimizer and MSE loss function, in Fig. 10
- Tox_basic_B with RMS optimizer and MSE loss function, in Fig. 11
- Tox with stem block and RMS optimizer and MSE loss function, in Fig. 12

The black line reported in all the graphs is the smoothed function obtained from the blue curve.

There are a few remarkable facts to comment. The first is that in all the configurations, the metrics tend to converge pretty fast.

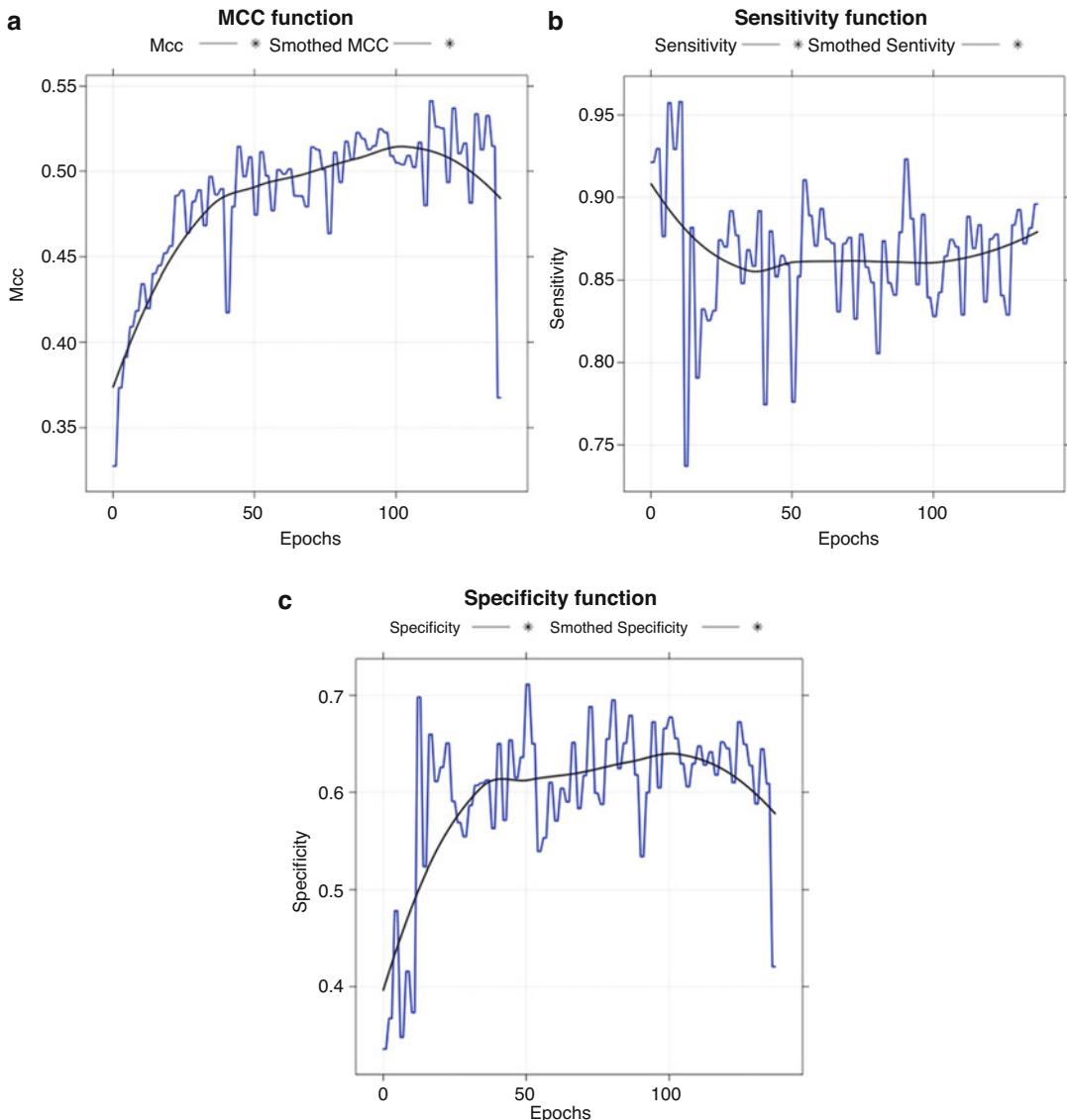


Fig. 10 In (a) MCC, in (b) sensitivity, and in (c) specificity of Tox_basic_A using MSE and the Adam optimizer. We can see how the sensitivity decreases in favor of the specificity. These two metrics are well summarized by the MCC. The precision instead stays stable after a few dozens of epochs. All the metrics reported here are calculated on the validation set

Another interesting information is that precision, sensitivity, and specificity reach fast their asymptotic value.

As we wanted to compare the best models with and without the stem block, we report in Fig. 12 the results obtained using the MSE loss function, the RMS optimizer, and the stem module just after the input. Indeed we can see that our initial hypothesis, that the stem block adds only complexity, is validated, as the learning curves in Fig. 12 are more unstable than in Fig. 11. We attribute this to the

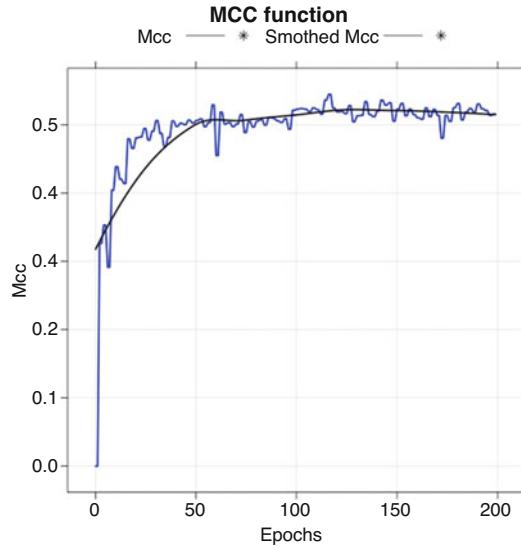


Fig. 11 MCC for regression of Tox_basic_B (without the stem block and using the RMS optimizer). All the metrics reported here are calculated on the validation set

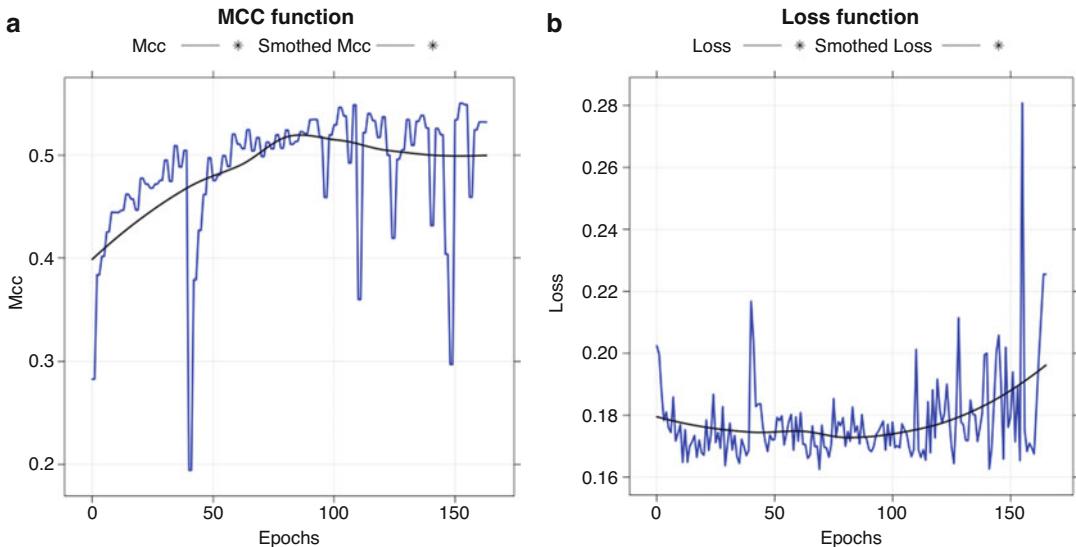


Fig. 12 In (a) accuracy, in (b) loss function learning curves of MCC with the stem block and the RMS optimizer. All the metrics reported here are calculated on the validation set

presence of the stem block, which introduces more sparse features to the convolutions layers.

In conclusion, as the results with Adams and RMS are very similar but RMS requires more computation, we take as the best network the one with the characteristics as in Table 5. Table 6 contains the parameters used.

Table 5
Results of Toxception on the test set

Epochs	Optimizer	Neurons	Batch size	Dropout	Learning rate
200	Adam	16	32	0.2	0.001

Table 6
Parameters used

Accuracy	MCC	Specificity	Sensitivity	MCC
0.80	0.53	0.62	0.87	0.53

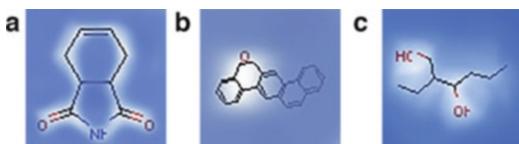


Fig. 13 In (a) MCC, in (b) sensitivity, and in (c) specificity of Tox_basic_A using MSE and the Adam optimizer. Examples of substructures of interest found by the network

6 Discussion

If we compare the results of our model with available models, as the ones reported in Table 1, we can observe that our results are not worse than their ones. So it is worth discussing Toxception according to its added value in terms of interpretability, uncertainty characterization, and knowledge acquired.

6.1 Interpretation and Knowledge Extraction

In order to extract knowledge from our model, we can observe that the final layer of the network gives a correlation matrix that we need to translate it into a human readable format. We indeed add an overlay layer to the picture of the investigated molecule so to visually indicate which part is taken by the net as important in giving the result. Figure 13 shows some examples of visualization results. We can see from the images that the network identifies a part of the chemicals as the important substructure that helped to classify the compound.

Other possible interpretation ways are to check the result of the network against the set of known mutagenicity SAs. In this case, the coincidence of the found area of interest with a known SA can improve the confidence in the model. In the reverse case, the absence of coincidence does not imply that the model is wrong.

6.2 Uncertainty Evaluation

About the interpretability of the results, we may observe that using the real value provided in output by the network is an indication about the uncertainty of the prediction. It cannot be interpreted as

a potency of the mutagen substance, since no doses have been used in training; however values in the middle of the interval have the highest uncertainty.

A deeper analysis of the network can also improve the uncertainty value. As demonstrated by [71], the dropout technique is a way to represent the model uncertainty in terms of Bayesian theory. The direct application of the Bayes formula is prohibitive due to the cost of computing all the conditional probabilities, but deep learning offers almost the same advantages, since dropout in NN is equivalent to the Bayesian approximation of the Gaussian process model. Dropout indeed is used in Toxception, and a possible post-processing can be added to show it.

6.3 Comparison with the State of the Art

In the before-mentioned paper [59], about 18 thousand molecules were predicted using 10 QSAR/SAR models from the literature. Two hundred molecules were wrongly predicted (resulting in either false positive or false negative) by all the 10 models. We predict these 200 compounds with our model, and we obtain an accuracy of 0.62 and 0.1 FP. The accuracy value is low. The causes can be various. First of all, some of the tested chemicals may be outside the applicability domain. It could also mean that the Ames tests must be repeated on these compounds, as the test results may be wrongly reported. However, the false negatives are few. This is coherent with the fact that in our dataset we preferred positive values when the experimental values disagree.

6.4 Pros and Cons of Toxception

In developing Toxception, we made a strong choice: avoid adding any external knowledge besides the SMILES and the value of the Ames test. This choice has advantages and drawbacks. The advantages are obvious, since the steps of feature computation and reduction are not needed, and no fingerprints with a priori choice of substructures are generated. The knowledge we extract from the network is completely new and not biased by a priori choices. Comparing Chemception and Toxception, indeed we can see that the introduction of some chemical properties to the network only negligibly improves the final result; instead it increases the computation burden.

The drawbacks are also obvious, since we need to extract the knowledge from the network. Another possible drawback is that the training of the network requires hours and dedicated hardware, so it cannot be done on the fly when data are available.

We can conclude this discussion with interesting theoretical results about the concept of models and learnability.

A family of models can be derived from the same dataset, and many of them can be valid models. Is there something to be considered as “the best possible model”? The question is raised by Wolpert’s “no free-lunch” (NFL) theorems [72]. In practice it means that “for any two learning algorithms A and B, there are just

as many situations in which algorithm A is superior to algorithm B as vice versa. So if we know that learning algorithm A is superior to B averaged over some set of targets F, then the NFL theorems tell us that B must be superior to A if one averages over all targets not in F.”

That something can be learned is again difficult to prove. What is learnable by a machine should be mathematically defined. As [73] discuss, learnability can be undecidable. It means that any mathematical formulation of what can be learned by any system cannot be demonstrated neither true nor false.

Of course those results do not have practical consequences, but they underline how difficult it is to give formal definitions and descriptions of models, even in case they are constructed with ML.

7 Conclusions

ML methods are already in use in various aspects of QSAR modeling. After reviewing important achievements in CMR and PBT assessment, we address the quite new research area of using deep learning methods to build QSARS.

We modified the most successful architecture, Inception, to build a model called Toxception, trained on data of the Ames test.

To reach this target, we built a dataset of about 22 thousand chemicals with their Ames test results. This dataset was extracted from different sources and curated to become the training and test sets of Toxception.

Toxception is a CNN network with 140 layers and over 1 thousand parameters. The introduction of the attention layer allows to analyze the important part of the images in the output. The performances of this model are around 80% of accuracy, which is almost the current state-of-the-art results. Moreover its sensitivity and specificity are quite high and balanced.

Toxception extended the ideas adopted in Chemception by eliminating the need of using any molecular descriptor. While the results are never worse than the results of available QSAR models for mutagenicity present in the literature, the main advantage of Toxception is that the model does not depend on a priori knowledge, and the crucial part of computing and selecting molecular descriptors is not needed. Moreover, it must be considered that most of the literature QSARs have a smaller applicability domain, since they are based on about 6000 data. Another advantage of Toxception is that it is also a regression model, where the output is in (0, 1) and can give a measure of the uncertainty. In practice this avoids the main problem with SA-based systems, which cannot classify a molecule as negative just when it does not contain a known SA.

Toxception is not ready for use by the regulatory bodies since we need to dedicate more research to fully define applicability domain and confidence values for the DNN models.

Further work is needed also to apply it to new endpoints. A possible improvement may derive from using images at higher resolution sizes, to see whether it can lead to a better prediction. About extending to other endpoints, we could think about constructing a large sparse matrix with all the available toxicity data of the 22 thousand molecules and check how the use of similar properties can improve the prediction of any property. Actually, the code proposed already contains the structure to predict other endpoints and is available from the authors.

References

- Judson J, Richard A, Dix DJ (2009) The toxicity data landscape for environmental chemicals. *Environ Health Perspect* 117(5):685–695
- Gini G, Ferrari T, Cattaneo D, Golbamaki N, Manganaro A, Benfenati E (2013) Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. *SAR QSAR Environ Res* 24(5):365–383. <https://doi.org/10.1080/1062936X.2013.773376>
- Collins FS, Gray GM, Bucher J (2008) Transforming environmental health protection. *Science* 319(5865):906–907. <https://doi.org/10.1126/science.1154619>
- Gini G, Katritzky A (eds) (1999) Predictive toxicology of chemicals: experiences and impact of AI tools, papers from the AAAI Spring Symposium on Predictive toxicology SS-99-01. AAAI Press, Menlo Park, 1999
- Lo Y-C, Rensi SE, Torgn W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 23(8):1538–1546
- Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). *Expert Opin Drug Discovery* 13(12):1075–1089. <https://doi.org/10.1080/17460441.2018.1542428>
- Devinyak OT, Lesyk RB (2016) 5-Year trends in QSAR and its machine learning methods. *Curr Comput Aided Drug Des*, Las Vegas, NV, USA. 12(4):265–271
- Zhang L, Tan J, Han D, Zhu H (2017) From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today* 22(1):1680–1685
- Lee Y, Buchanan BG, Mattison DM, Klopman G, Rosenkranz HS (1995) Learning rules to predict rodent carcinogenicity of non-genotoxic chemicals. *Mutat Res* 328:127–149
- Bradbury SP, Feijtel TCJ, Van Leeuwen CJ (2004) Meeting the scientific needs of ecological risk assessment in a regulatory context. *Environ Sci Technol* 38(23):463A–470A
- Mazzatorta P, Benfenati E, Lorenzini P, Vighi M (2004) QSAR in ecotoxicology: an overview of modern classification techniques. *J Chem Inf Comput Sci* 44:105–112
- Helma C, King RD, Kramer S, Srinivasan A (2001) The predictive toxicology challenge 2000–2001. <http://www.informatik.uni-freiburg.de/~rnl/ptc/>
- Gini G, Benfenati E, Lorenzini M, Bruschi M, Grasso P (1999) Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using artificial neural networks. *J Chem Inf Comput Sci* 39:1076–1080. <https://doi.org/10.1021/ci9903096>
- Gini G, Lorenzini M, Benfenati E, Brambilla R, Malve' L (2001) Mixing a symbolic and a sub-symbolic expert to improve carcinogenicity prediction of aromatic compounds. *Proceedings of second workshop on Multiple Classifier Systems (MCS 2001)*, Springer, pp 126–135
- Rallo R, Espinosa G, Giralt F (2005) Using an ensemble of neural based QSARs for the prediction of toxicological properties of chemical contaminants. *Process Saf Environ Prot* 83(B4):387–392
- Fjodorova N, Vračko M, Novič M, Roncaglioni A, Benfenati E (2010) New public QSAR model for carcinogenicity. *Chem Cent J* 4(Suppl 1):S3. <https://doi.org/10.1186/1752-153X-4-S1-S3>

17. Golbamaki A, Benfenati E, Golbamaki N, Manganaro A, Merdivan E, Gini G (2016) New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds. *J Environ Sci Health C* 34(2):97–113
18. Li N, Qi J, Wang P, Zhang X, Zhang T, Li H (2019, 2019) Quantitative structure–activity relationship (QSAR) study of carcinogenicity of polycyclic aromatic hydrocarbons (PAHs) in atmospheric particulate matter by random forest (RF). *Anal Methods*. <https://doi.org/10.1039/C8AY02720J>
19. Papamokos G, Silins I (2016) Combining QSAR modeling and text-mining techniques to link chemical structures and carcinogenic modes of action. *Front Pharmacol.* 30 Aug 2016. <https://doi.org/10.3389/fphar.2016.00284>
20. Ferrari T, Gini G (2010) An open source multistep model to predict mutagenicity from statistic analysis and relevant structural alerts. *Chem Cent J* 4(Suppl 1):S2. online <http://www.journal.chemistrycentral.com/>
21. Gini G, Franchi AM, Manganaro A, Golbamaki A, Benfenati E (2014) ToxRead: a tool to assist in read across and its use to assess mutagenicity of chemicals, SAR and QSAR in environmental research. <https://doi.org/10.1080/1062936X.2014.976267>, pp 1–13, online December 2014
22. Toropov AA, Toropova AP, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines. *Chemom Intell Lab Syst* 109:94–100
23. Maunz A, Gütlein M, Rautenberg M, Vorgrimmler D, Gebele D, Helma C (2013) Lazar: a modular predictive toxicology framework. *Front Pharmacol* 4:38. <https://doi.org/10.3389/fphar.2013.00038>
24. Zhang Q-Y, Aires-de-Sousa J (2007) Random forest prediction of mutagenicity from empirical physicochemical descriptors. *J Chem Inf Model* 47(1):1–8. <https://doi.org/10.1021/ci050520j>
25. Maran U, Sid S (2003) QSAR Modeling of genotoxicity on non-congeneric sets of organic compounds. *Artif Intell Rev* 20:13–38
26. Cronin MTD, Worth AP (2008) (Q)SARs for predicting effects relating to reproductive toxicity. *QSAR Comb Sci* 27(1):91–100
27. Cassano A, Manganaro A, Martin T, Young D, Piclin N, Pintore M, Bigoni D, Benfenati E (2010) CAESAR models for developmental toxicity. *Chem Cent J* 4(Suppl 1):S4. <http://www.journal.chemistrycentral.com/content/4/S1/S4Cassano>
28. Baker JR, Gamberger D, Mihelcic JR, Sabljic A (2004) Evaluation of artificial intelligence based models for chemical biodegradability prediction. *Molecules* 9(12):989–1003. <https://doi.org/10.3390/91200989>
29. Lombardo A, Pizzo F, Benfenati E, Manganaro A, Ferrari T, Gini G (2016) A new in silico classification model for ready biodegradability, based on molecular fragments. *Chemosphere* 108(2016):10–16
30. Miller TH, Gallidabino MD, MacRae JI, Owen SF, Bury NR, Barron LP (2019) Prediction of bioconcentration factors in fish and invertebrates using machine learning. *Sci Total Environ* 648:80–89
31. Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E (2010) Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem Cent J* 4(Suppl 1):S1
32. Valsecchi C, Grisoni F, Consonni V, Ballabio D (2019) Structural alerts for the identification of bioaccumulative compounds. *Integr Environ Assess Manag* 15(1):19–28
33. Benfenati E, Roncaglioni A, Petoumenou MI, Cappelli CI, Gini G (2015) Integrating QSAR and read-across for environmental assessment. *SAR QSAR Environ Res* 26(7–9):605–618
34. Benfenati E (ed) (2007) Quantitative structure-activity relationships (QSAR) for pesticide regulatory purposes. Amsterdam Elsevier Science
35. Gini G, Ferrari T, Lombardo A, Cassano A, Benfenati E (2019) A new QSAR model for acute fish toxicity based on mined structural alerts. *J Toxicol Risk Assess* 5(1):016. <https://doi.org/10.23937/2572-4061.1510016>
36. Gini G, Craciun M, Benfenati E (2004) Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *J Chem Inf Comput Sci* 44(6):1897–1902
37. Pintore M, Piclin N, Benfenati E, Gini G, Chretien JR (2003) Predicting toxicity against the fathead Minnow by Adaptive Fuzzy Partition. *QSAR Comb Sci* (Wiley-VCH) 22:210–219
38. Toropova A, Toropov A, Veselinovic A, Veselinović J, Leszczynska D, Leszczynski J (2016) Monte Carlo based QSAR models for toxicity of organic chemicals to *Daphnia magna*. *Environ Toxicol Chem* 35 (11):2691–2697
39. Xu Y, Pei J, Lai L (2017) Deep learning based regression and multi-class models for acute oral

- toxicity prediction with automatic chemical feature extraction. arXiv:1704.04718v3 [stat. ML]
40. Sayre R, Grulke C (2018) Universal LD50 predictions using deep learning. ICCVAM – Predictive models for acute oral systemic toxicity, Bethesda, 11–12 Apr 2018
 41. Benfenati E, Mazzatorta P, Neagu CD, Gini G (2002) Combining classifiers of pesticides toxicity through a neuro-fuzzy approach. Proceedings of 3rd international workshop on multiple classifier systems, MCS 2002, Springer, Cagliari, June 2002, pp 293–303
 42. Mazzatorta P, Cronin MTD, Benfenati E (2006) A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. Mol Inform 25(7):616–628
 43. Gini G, Garg T, Stefanelli M (2009) Ensembling regression models to improve their predictivity: a case study in QSAR (Quantitative Structure Activity Relationships) within computational chemometrics. Appl Artif Intell 23:261–281
 44. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv:1602.07261v2 [cs.CV]
 45. Goh G, Siegel C, Vishnu A, Hodas NO, Baker N (2017) Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. <https://arxiv.org/abs/1706.06689>
 46. McCulloch WS, Warren S, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. B Math Biophys 5(4):115–133. ISSN 1522-9602. <https://doi.org/10.1007/BF02478259>
 47. Werbos PJ (1994) The roots of backpropagation: from ordered derivatives to neural networks and political forecasting. Wiley, New York
 48. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Geoffrey G, David D, Miroslav D (eds) Proceedings of the fourteenth international conference on artificial intelligence and statistics, Fort Lauderdale, 11–13 Apr 2011; PMLR Proceedings of Machine Learning Research, pp 315–323
 49. Devillers J (ed) (1996) Neural networks in QSAR and drug design. Academic Press, San Diego
 50. O’Shea KT (2015) An introduction to convolutional neural networks. arXiv:1511.08458v2 [cs.NE]
 51. LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. <http://yann.lecun.com/exdb/publis/pdf/lecun-bengio-95.pdf>
 52. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA. [arXiv:1511.08458](https://arxiv.org/abs/1511.08458) [cs.NE]
 53. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2016) Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA. pp 1–9
 54. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. The IEEE conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. pp 770–778
 55. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 29(3):411–426
 56. Lin M, Chen Q, Yan S (2016) Network in network. arXiv preprint arXiv:1312.4400, 2013
 57. Ames BN (1984) The detection of environmental mutagens and potential. Cancer 53:2030–2040
 58. Piegorsch W W, Zeiger E (1991) Measuring intra-assay agreement for the Ames salmonella assay. In: Hotorn L (ed), Statistical methods in toxicology, Lecture Notes in Medical Informatics, Springer, Berlin-Heidelberg, pp 35–41
 59. Benfenati E, Golbamaki A, Raitano G, Roncaglioni A, Manganello S, Lemke F, Norinder U, Lo Piparo E, Honma M, Mangano A, Gini G (2018) A large comparison of integrated SAR/QSAR models of the Ames test for mutagenicity. SAR QSAR Environ Res 29(8):591–611
 60. Martin T (2016) User’s guide for T.E.S.T. (Toxicity Estimation Software Tool), U.S. EPA/National Risk Management Research Laboratory/Sustainable Technology Division, Cincinnati, OH (2016). Available at <https://www.epa.gov/sites/production/files/2016-05/documents/600rl6058.pdf>
 61. Benigni R, Netzeva T, Benfenati E, Bossa C (2007) The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. J Environ Sci Health C 25(1):53–97. <https://doi.org/10.1080/10590500701201828>

62. Manganaro A, Pizzo F, Lombardo A, Pogliaghi A, Benfenati E (2016) Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm. *Chemosphere* 144:1624–1630
63. Mazzatorta P, Tran LA, Schilter B, Grigorov M (2007) Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of Ames test mutagenicity. *J Chem Inf Model* 47:34–38. <https://doi.org/10.1021/ci600411v>
64. Norinder U, Ahlberg E, Carlsson L (2019) Predicting Ames mutagenicity using conformal prediction in the Ames/QSAR International challenge project mutagenesis 34:33–40. <https://doi.org/10.1093/mutage/gey038>
65. Weininger M, Weininger A, Weininger JL (1989) Smiles. Algorithm for generation of unique SMILES notation. *J Chem Inf Model* 29(2):97–101
66. Benfenati E, Manganaro A, Gini G (2013) VEGA-QSAR: Ai inside a platform for predictive toxicology, PAI@ AI* IA, pp 21–28
67. NIHS. Ames/QSAR international collaborative study. URL <https://bit.ly/2z7Rg2g>
68. Corvi R, Madia F (2018) EurI ECVAM genotoxicity and carcinogenicity consolidated database of Ames positive chemicals. European Commission, Joint Research Centre (JRC)
69. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958
70. Kingma DP, Lei Ba J (2017) Adam: a method for stochastic optimization, arXiv:1412.6980 [cs.LG]
71. Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning Bayesian in deep learning. arXiv:1506.02142v6 [stat.ML]
72. Wolpert D (1996) The lack of a priori distinctions between learning algorithms. *Neural Comput* 8:1341–1390
73. Ben-David S, Hribes P, Moran S, Shpilka A, Yehudayoff A (2019) Learnability can be undecidable. *Nat Mach Intell* 1:121



Chapter 7

Use of Machine Learning and Classical QSAR Methods in Computational Ecotoxicology

Renata P. C. Barros, Natália F. Sousa, Luciana Scotti, and Marcus T. Scotti

Abstract

In recent years, there has been an increase in concern about environmental pollution and human health, especially in the areas of manufacturing, storage, distribution, and release of hazardous substances after use. Several researchers have been dedicating studies to develop methods to identify and assess the toxicity of chemicals. Quantitative structure-activity relationship (QSAR) modeling has evolved a lot in recent years and also developed in the area of ecotoxicology. In the course of this evolution, there was the application of machine learning techniques in QSAR studies. The use of ML algorithms is a great approach for assessing toxicity to generate predictive models involving QSAR. Several studies are being conducted not only comparing ML techniques but applying them to generate potentially predictive models and excellent performances.

Key words Ecotoxicology, Machine learning, Quantitative structure-activity relationship (QSAR)

1 Ecotoxicology

The term ecotoxicology was introduced by Truhalt in 1969 and derives from the words ecology (a discipline that studies the relationships between living things and the environment) and toxicology (describes the adverse effects of a given substance on a given organism and seeks to establish the mechanism of toxic action in it). Its introduction reflects a growing concern about the effect of environmental chemical compounds on species, in addition to man [1].

Ecotoxicology describes the relationship between the chemical pollutants, the environment in which they are released, and the organisms that live there [2]. Thus, ecotoxicology is a tool to analyze the exposure of various xenobiotics to the environment in which they were inserted and the adverse effects of exogenous pollutants on the environment and aquatic organisms.

Many authors elucidate the emergence of toxicology, defending the hypothesis that it was born in the beginnings of mankind,

thus anticipating written history about the use of poisons and animals and plants for the purpose of assisting in hunting and fishing and with poisoning in war activities [3, 4].

According to Buikema and Voshel (1993) [5], there are reports that Aristotle (384–322 BC) submitted freshwater fish to seawater to gauge and identify the reactions that occurred. The first known aquatic organism's toxicity test was conducted around the year 1816 with aquatic insects [5, 6].

The first ecotoxicology book was published in 1977, defining it as the science that aims to study the modalities of contamination of the environment by natural or synthetic pollutants produced by human activities, their mechanisms of action, and their effects on the set of living beings that inhabit the biosphere [7]. In this way, ecotoxicology was born as an environmental monitoring tool, based mainly on the response of individual organisms to chemical stressors. Therefore, it is a science with its own study objective, which consists of the phenomenon of environmental intoxication in all its nuances and consequences, with the purpose of preventing certain intoxications or knowing how to stop it, reverse it, and remedy it with an appropriate method [8].

1.1 Ecotoxicological Tests

Ecotoxicological tests measure the effects of different concentrations of a sample on individuals of a given species. The EC₅₀ effect concentration or lethal concentration LC₅₀ corresponds to the concentration of the sample responsible for the effect in 50% of the organisms tested. These tests may be acute or chronic depending on their duration and observed effect. In the case of acute tests, the evaluated effect is related to mortality, immobilization, or growth inhibition rates, and the lower the value, the higher the toxicity of the sample, which often leads to erroneous interpretations of the results obtained. Thus, UT (toxic unit) corresponding to $(1/\text{EC}_{50} \times 100)$ can be used for expression of the results [8].

Ecotoxicological tests may be performed using aquatic or terrestrial organisms depending on the type of study to be performed. These studies can be elaborated at the level of the individual, the population, the community, and even the ecosystem and in some cases may last for several years (IAP) [9].

Toxicity tests are classified into two main groups [10]:

- Acute toxicity tests
- Chronic toxicity tests

1.1.1 Acute Toxicity Tests

Acute toxicity tests assess a rapid and rapid response of aquatic organisms to a stimulus which generally manifests in a range of 0–96 hours [11]. Usually, the observed effect is the lethality or other manifestation of the organism that precedes it, such as the state of immobility in invertebrates. These tests are designed to determine the mean lethal concentration (LC₅₀) or mean effective

concentration (EC_{50}), i.e., the concentration of the toxic agent causing mortality or immobility, respectively, to 50% of the test organisms after a given time of exposure [7].

1.1.2 Chronic Toxicity Tests

Chronic toxicity tests are directly dependent on the results of the acute toxicity tests because sublingual concentrations are calculated from the LC_{50} . Compared with the acute tests, these are more sensitive to the expected dilution in environmental samples and evaluate the action of the pollutants whose effect is translated by the response to a stimulus that continues for a long time, usually during a period that goes from 1/10 of the cycle vital to the whole life of the organism [11]. In general, however, these effects are sublethal and observed in situations where the concentrations of the toxic agent to which they are exposed to the organism allow their survival, but affect one or more of their biological functions, interfering, for example, with reproduction, egg development, and growth [7].

According to Chasin and Azevedo (2003) [8], chronic intoxication can occur for two reasons:

1. By xenobiotic accumulation in the body, which occurs when the amount of foreign agent eliminated is less than the amount absorbed. The concentration of the toxic agent in the body progressively increases until sufficient levels are obtained to generate adverse effects.
2. By adding the effects caused by repeated exposures, without the toxic compound accumulating in the organism.

Ecotoxicological tests are carried out with indicator organisms which, due to their small ecological tolerance to certain chemical substances, present some alteration, be it physiological, morphological, or behavioral when exposed to certain pollutants. Exposures are made in different concentrations of chemical substances and compounds, effluent samples, or raw water, for a certain period of time. The tests present a range of standards and standardized procedures that must be followed for responses to be considered valid [12].

1.2 Indicators

1.2.1 Soil Organisms

The organisms that compose the soil biota play important roles in the development and stability of the ecosystem as a whole. They are involved in soil formation and structuring processes, in the decomposition of organic matter, in regulation of microbial activity, and consequently in nutrient cycling. The best known organisms are Oligochaeta, Enchytraeidae, and Collembola [13].

1.2.2 Water Organisms

The toxicity tests involving algae species usually fit into the profile of the chronic tests, where the population increase of the organism exposed to different doses of the contaminant during a certain

number of days is evaluated, comparing to the performance in a control sample. The best known aquatic organisms are *Daphnia* and algae [14].

2 Standardization of Ecotoxicological Tests

The standardization of toxicity testing is based on the publications of the regulatory bodies, which consists of the Organisation for Economic Co-operation and Development (OECD) Guidelines. In Brazil, there is also the adoption of the protocols of the Brazilian Association of Technical Standards and State Environmental Companies, such as CETESB—Environmental Company of the State of São Paulo [7].

2.1 Data Sets

2.1.1 OCHEM: Online Chemical Database

The Online Chemical Database (OCHEM) is a platform that has 2,858,801 records for 636 properties, collected from 13,098 sources (Fig. 1). The existing collection consists of chemical and biological data, exposed to the scientific community by program users [15].

Sushko et al. (2011) [16] characterize this tool as an online environment that allows the search and execution of a quantitative structure-activity relationship (QSAR)/quantitative structure-property relationship (QSPR) cycle semiautomatically. The platform includes two main systems, namely, (1) the database of properties measured experimentally and (2) the modeling structure.

Structure of the Database

The databases contain biological and physicochemical properties of the molecules (it being possible to specify the experimental conditions for obtaining). Experimental measurements cover the information related to the experiment, and especially to the result of this, being numerical or qualitative, depending on the measured property [17].

For the storage of data in OCHEM, it is necessary to specify the data source, which may refer to publications in scientific journals and book chapters, as well as the complete work, but the source of achievement is mandatorily required. PubMed accesses link binding and ISBN insertion [18].

The addition of data in OCHEM can be performed manually, manually and individually, as well as by batch. The entry by the manual record corresponds to the insertion of experimental data separately. The second way is the batch upload feature that allows you to upload large amounts of Excel data, Comma-Separated Values (CSV), or Structure Data File (SDF). In addition, molecules can be drawn and imported in MOL 2 format or in the form of SMILES codes (canonical representations of the chemical structure of the molecule). After the introduction, the data are reviewed to

The screenshot shows the OCHEM platform interface. At the top, there's a navigation bar with 'Home', 'Database', 'Models', and user account options. Below the header, a sidebar on the left lists actions: 'Explore OCHEM data', 'Create QSAR models', 'Run predictions', 'Screen compounds with ToxAlerts', 'Optimise your molecules', and 'Check out the properties available on OCHEM'. The main content area displays a list of properties: Melting Point, logPow, logBB, LogP(water), LogD, logP(+), Water solubility, LogP(blood), LogP(oil), ER, Cbrain/Plasma, IC50, Papp(Caco-2), Papp(MDCK), Oral absorption, LIC 50, Papp ratio(Caco-2), Plasma protein binding, Papp ratio(MDCK-Mdr1), PIC50, %Human FA, Human IA, Human FA fraction unbound (fu), fraction ionized (%), pKa, VDss, LogIC50, LogPI, BBB permeability (qualitative), LogKoa, LogRBA, CYP450 modulation, CYP450 reaction, Vapor Pressure, EC50 aquatic, NOEC aquatic, LOEC aquatic, IC50 aquatic, LC50 aquatic, LD50 aquatic, and LD50 oral. On the right, there are sections for 'Latest active users' (with a list of users like napple-oleon, skaye, carson.britt, gsempie, johnycarbonillo, argonaut) and 'Latest published models' (with entries like Ld50 mouse oral model by Tinkov_Oleg, o_ggg_ort_LD model by pirotex, and Anti-TB activity IC50 (qualitative) model by vikovalishyn).

Fig. 1 Representation of the OCHEM platform

verify the occurrence of duplicates as well as errors. At the end of the data check in the Upload process, the data can be saved and registered in the OCHEM database. The user can thus access them [19].

Structure of the Modeling Process

The database is strongly integrated with the modeling framework; the data can be flexibly filtered and used for the training of predictive computational models. The OCHEM modeling framework supports all typical QSAR/QSPR modeling steps: data preparation, molecular descriptor calculation and filtering, application of machine learning methods (classification and regression), model analysis, modeling domain evaluation applicability, and finally using the model to predict target properties for new molecules. It is important to note that OCHEM allows the combination of data with different units of measurements, different conditions of experiments, and even different properties and activities. The complexity of the modeling process is hidden behind a convenient and well-documented user interface. Templates can be published on the Web and used publicly by others [16].

Regarding the training set, the OCHEM system allows the user to combine heterogeneous data reported in different units of measure into a single set of units. An example of this application refers to the work performed by Oprisiu and collaborators (2013) [20]. In their work, the authors performed the modeling of properties of nonadditive mixtures with the online environment OCHEM. It was realized the combination of data and the use of special descriptors, realizing that the modeling of mixtures requires an automatic calculation of the precision, use of a wide spectrum of learning of algorithms of machine learning and descriptors, and storage, publication, and application of models.



Fig. 2 ECETOC. (From ECETOC—<http://www.ecetoc.org/pt/> [21])

2.1.2 ECETOC: European Centre for Ecotoxicology and Toxicology of Chemicals (Fig. 2)

Since 1978, ECETOC, an industry-funded scientific and nonprofit think tank, has been working to improve the quality and reliability of the chemical risk assessment with scientific support. Topics covered include ecosystem services and risk assessment of chemicals, nanotechnology, epigenetics, and interpolation. Publications consist of technical and detailed reports, manuscripts, and publications [21].

The European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) has risk assessment tools, which are responsible for calculating the risk of exposure to chemicals by workers, consumers, and the environment. It has been identified by the European Commission Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as the preferred approach for the assessment of consumer and worker health risks [22, 23].

Bahhatarai et al. (2016) [24] evaluated the TOPKAT, Toxtree, and Derek Nexus methods for in silico models developed for ocular irritation and to develop a framework to improve the prediction of severe irritation. In this work, the authors used databases containing 1644 and 123 compounds belonging to the ECETOC. ECETOC has published data on 132 compounds in rabbit toxicity studies involving two phases. This research emphasizes the need for in silico models to address chemical reactivity and filtering based on the physicochemical characteristics of compounds as well as clustering based on compound mechanisms, which could also address the problem of the lack of predictive power of these tools. Even existing in silico models may be able to implement such filters for better categorization of irritation potential [24].

2.1.3 MOA: The Toxic Mode of Action

The toxic mode of action (MOA) is recognized as a major determinant of chemical toxicity and is an alternative to chemical class predictive toxicity modeling. However, MOA classification has never been standardized in ecotoxicology, and a comprehensive comparison of tools and classification approaches has never been reported [25].

MOAtox is composed of a database of MOA designations for 1208 chemicals, including metals, organometallics, pesticides, and other organic compounds. The categorization scheme was based on earlier work that determined chemical modes of acute toxic action in fish and covered six broad and specific MOAs. The resulting data set used a combination of high confidence MOA assignments, including biological responses in acute toxicity tests, 22 pesticide classification schemes [e.g., Insecticide Resistance Action Committee (IRAC)], predictions QSAR (e.g., ASTER), and weight of evidence of professional judgment incorporating an assessment of the chemical structure (e.g., analog structure, group presence/functional group) and available information on MOA, mechanism of action and toxicity pathways. Chemicals with an uncertain attribution of specific MOAs and MOAs for invertebrates were excluded. Specific MOAs were developed as subcategories of the broad MOAs, based on the known chemical structure or mechanism of action [26, 27].

2.1.4 ECHA: European Chemicals Agency

The European Chemicals Agency (ECHA) is the driving force between regulatory authorities in the implementation of innovative European Union legislation on chemicals for the benefit of human health and the environment, as well as for innovation and competitiveness. ECHA brings to its electronic address the latest published legislation, public consultation materials on regulatory procedures, and information on chemicals imported and marketed in Europe (Fig. 3). The main risk information contained in ECHA's portals is collected on account of the REACH processes, which is set out below [27].

2.1.5 REACH: Registration, Evaluation, Authorization and Restriction of Chemicals

REACH aims to improve the protection of human health and the environment through the best and earliest identification of the intrinsic properties of chemical substances (Fig. 4). This is done by the four REACH processes, namely, the registration, evaluation, authorization, and restriction of chemicals. REACH also aims to improve the innovation and competitiveness of the EU chemical industry. The REACH Regulation places the responsibility on the industry to manage the risks of chemicals and to provide safety information about the substances. Manufacturers and importers are required to collect information on the properties of their chemical substances, which will allow their safe handling, and to record the information in a central database of the ECHA in Helsinki. The agency is at the heart of the REACH system: it manages the

The screenshot shows the ECHA (European Chemicals Agency) website. At the top, there's a navigation bar with links for 'Sign In', 'English (en)', and a search bar. Below the header, there are four main menu categories: 'LEGISLATION', 'PUBLIC CONSULTATIONS', 'INFORMATION ON CHEMICALS', and 'SUPPORT'. A 'Search for Chemicals' section is prominently displayed, featuring a search input field, a 'Search' button, and a link to 'ADVANCED SEARCH'. To the right of the search area, there's a 'Hot topics' section with a question mark icon and a link to 'Are you a consumer?'. Below the search section, there's a 'News' section with a thumbnail image of a computer screen showing the ECHA submission portal. The news item is titled 'ECHA releases new submission portal for poison centres' and includes a date (24/04/2019) and a brief description about the new portal for preparing, submitting, and tracking information for poison centres. To the right of the news section, there are two boxes: one for 'Substances in focus' and another for an 'ECHA CONFERENCE SAFER CHEMICALS' event in Helsinki, Finland. At the bottom right, there's a small graphic related to the UK's withdrawal from the EU.

Fig. 3 ECHA. (From ECHA—<https://echa.europa.eu/home> [27])

The screenshot shows the REACH section of the European Commission Environment website. The top navigation bar includes links for 'About this site', 'Contact', 'Sitemap', 'Search', 'Legal notice', and 'Cookies', along with a language selector for 'English (en)'. The main header features the European Commission logo and the word 'ENVIRONMENT'. Below the header, there's a breadcrumb navigation path: 'European Commission > Environment > Chemicals > REACH >'. The main menu has several dropdowns: 'Home', 'About us', 'Policies', 'Funding', 'Legal compliance', 'News & outreach', and 'Chemicals Home'. On the left, a sidebar menu for 'REACH' is expanded, showing sub-links like 'Introduction', 'Legislation', 'Implementation', 'Member States reports on the operation of REACH', 'Review REACH annexes', 'REACH Review 2017', 'Enforcement', and 'Competent Authorities'. The main content area features a large image of blue industrial drums. Below the image, the word 'REACH' is written in green, followed by a detailed description of the REACH Regulation (EC 1907/2006), which aims to improve protection of human health and the environment through better identification of chemical substances and their properties. The text also mentions the four processes of REACH: registration, evaluation, authorisation, and restriction, and its goal to enhance innovation and competitiveness of the EU chemicals industry.

Fig. 4 REACH. (From REACH—http://ec.europa.eu/environment/chemicals/reach/reach_en.htm [28])

databases required to operate the system, coordinates the in-depth evaluation of suspect chemicals, and is building a public database in which consumers and practitioners can find risk information. The Regulation also requires the phasing out of the most dangerous chemicals (referred to as “substances of very high concern”) when appropriate alternatives are identified [28].

2.1.6 ECOSAR: Ecological Structure- Activity Relationships Predictive Model

The ecological structure-activity relationships predictive model (ECOSAR) is a computerized predictive system that estimates aquatic toxicity. The program estimates the acute (short-term) chemical toxicity and chronic (long-term or delayed) toxicity to aquatic organisms, such as fish, aquatic invertebrates, and aquatic plants, using computerized structure-activity relationships (SARs). Key features of the program include grouping structurally similar organic chemicals with available experimental effect levels that correlate with physicochemical properties to predict the toxicity of new or untested industrial chemicals, programming a classification scheme to identify the most representative class for new or untested chemicals, and continuous updating of aquatic QSARs based on experimental studies collected or sent from public and confidential sources [29].

The ECOSAR program is available for download under the version: ECOSAR V2.0.

2.1.7 OECD: The Organisation for Economic Co-operation and Development (Fig. 5)

This is an international organization of 36 countries that accept the principles of representative democracy and the market economy, which seeks to provide a platform for comparing economic policies, solving common problems, and coordinating domestic and international countries. Most OECD members are composed of economies with high GDP per capita and Human Development Index and are considered developed countries [30].

From the point of view of ecotoxicology, the OECD presents publications that are a reference in this field. The OECD Guidelines are a unique tool to assess the potential effects of chemicals on human health and the environment. Accepted internationally as standard methods for safety testing, the Guidelines are used by industry, academia, and government professionals involved in the testing and evaluation of chemicals (industrial chemicals, pesticides, cosmetics, etc.) [31]. These Guidelines are regularly updated with the assistance of hundreds of national experts from OECD member countries. Currently, the OECD Guidelines are distributed as follows:

- Section 1: Physical Chemical Properties
- Section 2: Effects on Biotic Systems (Software for TG 223)
- Section 3: Environmental Fate and Behavior (Software for TG 305 and TG 318)
- Section 4: Health Effects (Software for TG 455, TG 432, and TG 425)
- Section 5: Other Test Guidelines

The OECD Guidelines were used as a dataset in the studies by Das et al. (2013) [32] for the development of models for rodent toxicity and their interspecific correlation with aquatic toxicity of

The screenshot shows the official website of the Organisation for Economic Co-operation and Development (OECD). At the top, there's a navigation bar with links for 'OECD.org', 'Data', 'Publications', 'More sites', 'News', and 'Job vacancies'. Below the navigation is the OECD logo with the tagline 'BETTER POLICIES FOR BETTER LIVES'. A green banner at the top right includes a 'Google Custom search' bar and a magnifying glass icon. The main content area has a blue header with 'OECD Home', 'About', 'Countries', 'Topics', and a link to 'Français'. On the left, there's a sidebar with a list of links under 'About the OECD' such as 'Who does what', 'What we do and how', 'Members and partners', 'History', 'Budget', 'Publishing', and 'OECD Careers'. In the center, there's a large image of many international flags. To the right, there's a 'Quick facts' box with information like 'History: established in 1961', 'Headquarters: Paris, France', and 'Membership: 36 countries'. On the far right, there are social media icons for Twitter, Facebook, LinkedIn, YouTube, and Email.

Fig. 5 OECD. (From OECD—<http://www.oecd.org/about/> [30])

The screenshot shows the eChemPortal homepage, which is described as 'The Global Portal to Information on Chemical Substances'. The top navigation bar includes a 'Print' button and a language selector set to 'English'. The main content area features the eChemPortal logo and a sub-header 'eChemPortal'. On the left, there's a sidebar with a dropdown menu titled 'eChemPortal' containing links for 'Home', 'Substance Search', 'Property Search', 'GHS Search', 'What's new?', 'General Information', 'Participating Databases', 'Roles & Responsibilities', 'Linking to eChemPortal', 'Schedules of Assessments', 'Structure Search', 'GHS Classifications', 'Useful links', 'FAQ', 'How to search for information', 'Contact form', and 'Disclaimer'. The central part of the page has three search buttons: 'Chemical Substance Search', 'Chemical Property Data Search', and 'GHS Search'. To the right of these buttons, there's a box containing text about data sources participating in the portal. Below this, there's a section about eChemPortal providing free public access to chemical properties and a list of physical chemical properties. Further down, there's a detailed description of how eChemPortal allows simultaneous searching of reports and datasets by chemical name and number, by chemical property, and by GHS classification. On the far right, there's a 'Latest news' box listing several updates with their respective dates.

Fig. 6 eChemPortal. (From eChemPortal [33])

pharmaceuticals. In this work, the authors sought acute toxicity data using OECD toxicity test guidelines and were reported in mg/kg (rodent) and mg/L (algae and fish), which were transformed into their corresponding molar concentration and finally converted to their logarithmic negative values ($-\log LD_{50}$ or pLD_{50} and $-\log LC_{50}$ or pLC_{50}) for the development of the QSAR models. The authors state that more work is needed to demonstrate robust mechanistic interpolated models between rodents and aquatic species [31, 32].

The OECD Guidelines provide a portal for the provision of chemical information; this is called the eChemPortal (Fig. 6).

This platform provides free public access to information on chemical properties, with reference to physicochemical properties, ecotoxicity, fate and environmental behavior, and toxicity, allowing the simultaneous search of reports and datasets by name and number of chemicals, by chemical properties, and by the classification of the Globally Harmonized System of Classification and Labeling of Chemicals (GHS). Direct links to the collection of information on chemical risks and risks prepared for government chemical programs at the national, regional, and international levels are obtained. The classification results according to the national/regional or national hazard classification schemes (GHS) are provided when available. In addition, eChemPortal also provides information on exposure and use of chemicals [33].

2.1.8 US EPA: US Environmental Protection Agency

The US Environmental Protection Agency (EPA) is a federal agency of the US government charged with protecting human health and the environment: air, water, and land (Fig. 7). The EPA began operating on December 2, 1970, when it was instituted by President Richard Nixon. It is headed by an administrator, appointed by the president [34].

The EPA presents a database called ECOTOX, which corresponds to a comprehensive, publicly available knowledge base that provides unique chemical environmental toxicity data on aquatic life, terrestrial plants, and wildlife that features 11,695 registered chemical compounds, 12,713 species, and 48,464 references [35] (Fig. 8).

3 Machine Learning

Nowadays, there are several in silico tools that integrate the processes of discoveries and help in several studies performing various types of functions [36, 37]. Machine learning methods and techniques (ML) is a subfield of computer science that has evolved a lot in recent years and emerged from the study of pattern recognition and the theory of computational learning in artificial intelligence,



Fig. 7 U.S. EPA. (From *U.S. EPA*—<https://www.epa.gov/aboutepa> [34])

Fig. 8 ECOTOX. (From ECOTOX (<https://cfpub.epa.gov/ecotox/> [35])

involving areas such as computational statistics and mathematics [38, 39].

One of the first definitions of ML, made by Arthur Samuel (1959) [37, 39], says that ML is the field of study that gives the computer the ability to learn without being explicitly programmed. Another definition made by Mitchell (1997) defines it as the area of study that is performed with mathematical and computational properties of algorithms (code) that can learn meaningful and complex patterns from a given set of input samples and a set of

labels to predict correctly the label or response of new examples [36, 40–42]. In short, ML investigates and explores the study and development of algorithms capable of learning from its errors and generating predictions about data, constructing models from sample input to generate predictions [38, 43–46].

Since the early 1990s, ML has become quite popular, and with the development of computer technology in the last decade, ML techniques have been applied in several areas such as in the field of computer science, social network analysis, data mining, facial recognition, drug discovery, biology, and ecology [40, 46–52].

ML techniques can be classified into three common types: supervised, unsupervised, and reinforcement learning. In supervised learning, a desired input and output sample files are presented to the computer, where the goal is to learn a general rule that includes the inputs and outputs. Because unsupervised learning does not provide the computer with any output sample file, allowing the algorithm to find a pattern in the input file alone, unsupervised learning can be understood as being by itself an objective (discover patterns in the data) or a way to reach an end. In reinforcement learning, the computer program will interact in a dynamic environment whose purpose is to accomplish some task, and to this end feedback will be provided as to awards and punishments as one navigates through the problem space [53–55].

Another classification of ML is with respect to the desired output in an ML system. There is the classification category whose entries are divided into two or more classes, and the model being produced is trained from data with previously known classes to be able to respond correctly to data that were not in the data training. A good classification model will be one that has a good generalization, that is, it has the ability to respond correctly to the examples contained in the training base but also to other examples contained in a test basis. Classification models are obtained through supervised learning [53, 56, 57]. Another type is regression models, which are also solved by supervised learning and the outputs are continuous or discrete. In clustering, the input set is divided into groups, but different from the classification category, the groups here are not previously known, which makes clustering a characteristic task of unsupervised learning [53, 55].

3.1 Machine Learning and QSAR

The use of ML techniques in QSAR studies has been progressively evolving in the last 60 years. To understand how ML and QSAR studies are now closely related, we need to go back in history and observe the emergence of QSAR studies.

The historical milestone of the QSAR studies was Hammett's equation in his classic work "Physical Organic Chemistry: Reactions, Balances and Mechanisms," published in 1940, in which Hammett studied the ionization of meta- and para-substituted benzoic acids in water at 25 °C and pioneered the linear relationship between the logarithm of the ionization constant of the

substituted benzoic acid and the logarithm of the ionization constant of benzoic acid [58–61].

Through Hammett's study, Hansch and co-workers became interested in creating a mathematical model that correlates chemical structure with biological activity. His pioneering work published in 1964 generated a mathematical equation, called the Hansch equation, which demonstrated a correlation between the biological activity of a chemical compound and its physicochemical properties [58–64].

Since the establishment of the Hansch equation, many researchers have carried out work proving and recognizing the discovery of Hansch and collaborators, which led to the development of quantitative in silico methods for modeling the structure-activity relationship [58–62, 64–66].

The QSAR studies start with the representation of the chemical structure, and it is fundamental to perform a good description of these molecules, from which the molecular descriptors emerge, defined as the final result of a logical and mathematical procedure that transforms the chemical information codified within the symbolic representation of a structure chemistry in numbers or the result of some standardized experiment [67].

Initially, the QSAR models were limited to small series of congeners and simple regression methods. Over the years, QSAR has undergone several transformations, varying from the dimensionality of the molecular descriptors (1D to nD) and different methods to establish a correlation between chemical structures and biological property [68–72].

Many different approaches to building QSAR models have been developed over time [64]. These models became primordial and effective tools for computational prediction of the biological activity of a biological compound, where in the construction of these models there was the insertion of ML algorithms [64–66].

Today, QSAR studies have undergone more growth and diversification and evolved into virtual screening (VS) modeling whose task is to screen large databases, comprising thousands of molecules, and find probabilities of the molecules to have activity against a particular target and/or biological activity, using a wide variety of ML techniques [58–75].

The VS classifies itself into two approaches: structure-based techniques (SBVS) and techniques based on the ligand structure (LBVS). For the SBVS approach, it is necessary to know about the three-dimensional structure of the target protein, whereas the LBVS approaches make use of the information of at least one known ligand and its biological activity [76, 77].

In VS, ML techniques use information from the biological activities of the molecules of the training set, both active and inactive molecules, and have achieved great successes. In general and simplified, ML techniques in VS use compound banks with

their respective values of biological activity (be they active or inactive) applying different algorithms and generating mathematical models from the molecular descriptors capable of triaging banks of molecules of the series test and select compounds most likely to present the biological activity in question [77, 78].

There are several studies reporting that ML methods outperform other methodologies in QSAR studies, such as empirical scoring methodology and knowledge-based functions of datasets [75–79]. However, the ML methods do not have simple interpretability, requiring a good knowledge from the researcher [75, 79].

3.2 ML Algorithms Applied to QSAR

Several ML algorithms have already been considered useful for the establishment of structure-biological activity relations, such as the support vector machine (SVM), decision tree (DT), random forest (RF), K-nearest neighbor (K-NN), naive Bayes (NB), neural network (NN), and ensemble learning (EL), which will be briefly explained below [80].

3.2.1 Support Vector Machine (SVM)

The SVM is a very effective ML technique and has comparable and sometimes superior results to those obtained by other ML algorithms such as neural networks [81–83]. It is a supervised method that uses associated algorithms that analyze data from both the classification category and the regression category. Given a classification data set, the SVM will construct a model that assigns new examples to one category or another, making it a binary and non-probabilistic linear classifier, but there are methods such as the Platt scale that was developed to use SVM in a probabilistic classification configuration [84, 85].

An SVM model will represent the examples of the training flock in points in space so that the points of different categories are separated by a clear gap that is as wide as possible. The new examples will then be mapped in the same space and predicted to belong to a category according to which side of the gap they fall [84, 85].

In addition to linear classification, SVM can still perform a nonlinear classification through the kernel trick, mapping implicitly into high-resource spaces. And when the data are not labeled, the SVM can still perform the unsupervised ML through the support vectoring algorithm, created by Hava Siegelman and Vladimir Vapnik, which will apply support vector statistics, developed in the SVM, to categorize unlabeled data, being one of the grouping algorithms most widely used in several applications [83–85].

3.2.2 Decision Tree and Random Forest

DT learning uses DTs that are a representation of a decision table, in the form of a tree, for generating predictive models [86]. The tree is constructed from the use of diagrams to map the various possibilities and results of decisions of a particular item as well as the probabilities of occurring [86, 87]. The result of each course is

weighted by the associated probability, whose result will be summed and the value of each course is then determined. The course that provides the highest expected value will be the preferred course. In a simpler way, DTs are diagrams that allow the representation and evaluation of problems involving sequential decisions, highlighting the risks and results identified in each decision and course taken [86–89].

When using DT learning, it is possible to use two types of data, one in which the final variable is a discrete set of values called classification trees and those in which the final variable is a set of continuous values called regression trees [89–91].

The algorithms used in learning DTs generally work from top to bottom, that is, they choose a variable in each step that best divides the data set [86–89].

In RF learning, a large number of DTs are generated, and at the end a vote for the most popular case will be held [77, 92]. In general, the RF is classified that consists of a set of classifiers structured in trees $\{h(x, k), k = 1, \dots\}$, where $\{k\}$ are independently identically distributed random vectors and each tree throws a unit vote for the most popular class in the x entry [77, 92].

3.2.3 K-Nearest Neighbor

K-NN is a type of ML based on an instance whose function is to approximate locally, and all computation is postponed until classification. This algorithm is one of the simplest of all ML algorithms [93, 94]. Through pattern recognition, K-NN is a nonparametric method used in both classification and regression data [93–95].

In classification data, the output is an association of classes, so an object is classified by a plurality of votes of its neighbors, with the object being assigned to the most common class among its nearest k -neighbors, where k is an integer value and positive. In regression data, the output is the value of the property under analysis for the object, where this value will be the average of the nearest k -neighbors values [94, 95].

3.2.4 Naive Bayes

The NB classifiers in ML are a simple family of probabilistic classifiers based on Bayes' theorem with a strong (naive) independence between characteristics, created by Thomas Bayes (1701–1761) [96, 97].

There is no single algorithm to train classifiers, but an algorithm family based on a single principle. All NB classifiers assume that the value of a new characteristic is independent of the value of any other characteristic, given the class variable [97, 98].

It is a simple and fast classifier, which has a relatively higher performance than other classifiers, and to use this classifier, only a small number of test data are required to complete classifications with good accuracy [96–98].

3.2.5 Neural Networks

NNs are computational models inspired by the central nervous system (in the brain) and can perform ML as well as pattern recognition [99]. The most important property of NN is the ability to learn from its environment and thereby improve its performance. The final learning occurs when the NN reaches a generalized solution to a class of problems [99–101].

NN can be both supervised and unsupervised. The NN usually presents a system of interconnected artificial neurons that can compute input values, always aiming to simulate the behavior of biological neural networks [99–101].

The artificial neuron is composed of three basic elements: the first is the set of $\{n\}$ input connections characterized by weights, the second is an adder to accumulate the input signals, and the third element is an activation function that will limit the permissible range of output signal amplitude at a fixed value [100, 101].

3.2.6 Ensemble Learning

In ML, the EL methods use various learning algorithms to achieve better predictive performance when compared with using just one of any underlying learning algorithms [102–104].

EL is an algorithm of the supervised learning category, i.e., an ensemble can be trained and used to make predictions. Empirically, ensembles usually produce better results when there is a good diversity among the models. Thus, many EL methods seek to promote diversity among the models they combine [103–105]. It has been proven that the use of a variety of ML algorithms is more effective than using techniques that attempt to simulate models to promote diversity, such as RF [105, 106].

3.3 Machine Learning and Ecotoxicology

In recent years, there has been an increase in concern about environmental pollution and human health, especially in the areas of manufacturing, storage, distribution, and release of hazardous substances after use, and it is therefore regulated and controlled at various levels by different governments and regulatory agencies all around the world [107, 108].

In this situation, one can apply QSAR models to predict toxicity in an organism based on the physicochemical properties of the chemical [109–112]. There are two basic objectives in toxicological-based QSAR analysis, where the first objective is to determine with greater precision the limits of variation in molecular structure that can produce a specific toxicological effect. The second goal is to define which structural changes will influence compound potency [113].

When conducting QSAR studies, it is of utmost importance to define the application limits, which should be considered which types of molecules, thus delimiting the molecular domain, and the range of descriptor values (the domain of the descriptor) that can have predicted toxicity with confidence. Besides these, of course, it has an adequate, significant, and robust statistical measure. Another

factor of great importance is the validation of the model, in which this validation needs to test the predictive capacity of the structure-activity relationship, to explore the application limits of the model, and to challenge the mechanistic hypotheses of the model [111–113].

ML and classification algorithms today represent a powerful way to extract relevant information from large data sets. But it is necessary to be used with caution and the correct interpretation of the results because the supervised training is prone to overfitting, resulting in excellent classification success [114].

The use of ML algorithms is a great approach for assessing toxicity to generate predictive models relating to QSAR. However, even though QSAR modeling is a powerful technique, there are two major problems faced by researchers, which are the domain of applicability and the interpretability of the models, mainly due to the use of hundreds of molecular descriptors [17].

The classical linear and multilinear models of QSAR, which use multiple linear regression (MLR), for example, have been supplanted with the new ML modeling techniques, especially the NNs, whose main advantage is their ability to process nonlinear QSARs [115].

A study by Hansen et al. (2009) [116] used a set of Ames mutagenicity data to compare the predictive performances of three commercial tools with four implementations of ML, SVM, RF, K-NN, and Gaussian processors. During their study, it was revealed that the difference between the best business model and the best ML approach using SVM is a sensitivity of only a few percent, demonstrating the power of ML approaches [116].

Several studies have shown that NNs have several important advantages over MLR, such as NN being able to cope more efficiently with higher-order interactions between descriptors because no proper model knowledge is required *a priori* because there is no need to classify chemical substances and because NN can improve predictive power by taking advantage of the information contained in the descriptors, rather than relying only on some specific descriptors, such as in the MLR [113, 117].

Samghani and Fatemi (2016) [118] analyzed the half-life of 58 herbicides using QSPR analysis using the SVM and MLR methods to map the traits and predict the half-life. The proposed models could identify and provide information on what structural features were related to the half-life of these compounds. However, the result showed that the SVM model exhibits a more reliable prediction and statistical performance than the MLR model, proving once again the importance and applicability of ML in ecotoxicology studies [118, 119].

Du et al. (2008) [120] used three MLs, GA-MLR, least squares SVM (LS-SVM), and PPR methods, to develop linear and nonlinear QSAR models to predict the fungicidal activity of 100 thiazoline

derivatives against the blister caused by *M. grisea*. The linear and nonlinear models obtained good prediction results, but the nonlinear ones had the better predictive capacity, showing that the LS-SVM and PPR methods can simulate the relation of the molecular descriptors and fungicidal activities more accurately than the GA-MLR. Then, Song et al. (2008) [121] used the same bank of molecules as Du et al. (2008) [120] and developed MLR and NN models to study the effects of substituents at the R1 site at three sites (ortho, meta, and para) of aromatic rings and observed that correlations between descriptors and activities were improved with the NN method [119, 120].

Oprisiu and collaborators (2013) [20] developed a publicly accessible system for storing binary mixes and developing models to predict their nonadditive properties. Its main objective is to contribute with publicly available tools for modeling and forecasting chemical compounds. They used the same database modeled with OCHEM by Oprisiu et al. (2012) [122] using the SVM ML and showed that the performances of the developed models were superior to those of Oprisiu et al. (2012) [122], demonstrating the capacity and usefulness of the tool that they developed. In addition, the results obtained by Oprisiu et al. (2012) [122] were based on a consensus forecast of three models using ML of NN, SVM, and RF, and the model developed by Oprisiu et al. (2013) [20] achieved similar performance without having to build a consensus model [20, 122].

Tekto et al. (2013) [123] developed and analyzed QSPR models to predict DMSO solubility of chemical compounds using various data sets. The ML algorithms J48 (Java implementations in the WEKA C4.5 DT), RF, and ASNN (Neural Network Associations) provided greater precision than the other methods analyzed, such as libSVM, K-NN, MLRA, FSMLR, and PLS. It was notorious that simple classification algorithms such as J48 and RF obtained a much higher accuracy than predictions when using the bagging approach. From a practical point of view, the J48 and RF methods were faster to calculate and required a smaller size to store the models [123].

Several studies are being conducted not only comparing ML techniques but applying them to generate potentially predictive models and excellent performances. For example, Michielan et al. applied single- and multiple-label classification tactics to a set of 580 CYP450 substrates. Models were generated with ML, SVM, K-NN, and artificial neural network (ANN) algorithms. The percentage of correct predictions for all classes was over 80% for multi-marker models and over 70% for single-label models when evaluated in an external set [124, 125].

4 Conclusion and Perspectives

In this work, we present an overview of ML methods applied to ecotoxicity in QSAR studies, namely, artificial neural networks, support vector machines, random forest, and decision tree, among others.

The use of ML techniques in the evaluation of ecotoxicology has increased considerably in recent decades, demonstrating the need for such approaches in this area. The use of these techniques has proved very useful, generating predictive models of excellent performances.

New methods and new algorithms are being applied to QSAR studies in the field of ecotoxicology, and improvements are being made to existing methods. However, the difficulty of interpretability is still a challenge and obstacle in the application of ML to QSAR studies in ecotoxicology.

Glossary

ABNT	Brazilian Association of Technical Standards
CETESB	Environmental Company of the State of São Paulo
CSV	Comma-Separated Values
DT	Decision tree
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
ECHA	European Chemicals Agency
ECOSAR	Ecological structure-activity relationships predictive model
EC ₅₀	Effect concentration 50%
EL	Ensemble learning
EL ₅₀	Lethal concentration 50%
K-NN	K-nearest neighbor
ML	Machine learning
MLR	Multiple linear regression
MOA	The toxic mode of action
NN	Neural networks
OCHEM	Online Chemical Database
OECD	The Organisation for Economic Co-operation and Development
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship (QSPR)
REACH	Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals
RF	Random forest
SARs	Structure-activity relationships
SDF	Structure Data File

SVM	Support vector machine
US EPA	US Environmental Protection Agency
UT	Toxicity unit

References

1. Silva DCVR, Pompeo M, Paiva TCB (2015) A ecotoxicologia no contexto atual no Brasil, vol 1. Instituto de Biociências: Universidade de São Paulo – USP, São Paulo, pp 340–351
2. Silva JS, Rocha IKBS, Freitas LC, Pereira NJ, Carvalho Neta RN (2015) Princípios bioéticos aplicados aos estudos toxicológicos aquáticos. *Rev Bioética* 23:409–118
3. Lombardi JV, Ferreira CM Rodrigues EL (2004) Toxicologia aquática. In: Ranzani-Paiva MJT, Takemoto RM, Lizama MAP (eds) Sanidade de organismos aquáticos, vol 1. Varela, São Paulo, pp 262–297
4. Fukushima AR, Azevedo FA (2008) História da toxicologia. Parte I – Breve panorama brasileiro. Ver InterTox Tox Risc Amb Soc 1:2–32
5. Buikema AL, Voshell JR (1993) Toxicity studies using freshwater benthic macroinvertebrates. In: Rosenberg DM, Resh VH (eds) Freshwater biomonitoring and benthic macroinvertebrates, vol 1. Chapman and Hall, New York pp 344–398
6. Rosenberg DM, Resh VH (eds) (1993) Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York, pp 344–398
7. Magalhães DP, Ferrão Filho AS (2008) A ecotoxicologia como ferramenta no biomonitoramento de ecossistemas aquáticos. *Oecol Bras* 12:355–381
8. Azevedo FA, Chasin AAM (2003) As bases toxicológicas da ecotoxicologia, vol 1. Editora Rima, São Carlos, São Paulo, p 40
9. Costa CR, Olivi P, Botta CMR, Espíndola ELGA (2008) Toxicidade em ambientes aquáticos: discussão e métodos de avaliação. *Química Nova* 1:1820–1830
10. IAP – Instituto Ambiental do Paraná (1997) Manual de métodos para avaliação de toxicidade, vol 1. Curitiba, p 101
11. Rand GM, Petrocelli SR (1985) Fundamentals of aquatic toxicology. Hemisphere Publishing Corporation, Washington, D.C
12. Schwartsman S (1991) Intoxicações agudas. Sarvier. Edição 4, 355p
13. Cortet J, De Vauflery AG, Balaguer NP, Gomot L, Cluzeau D (1999) The use of invertebrate soil fauna in monitoring pollutant effects. *Eur J Soil Biol* 35:115–134
14. Bianchi MO, Correia MEF, Resende AS, Campello EFC (2010) Importância de estudos ecotoxicológicos com invertebrados do solo. Embrapa Agrobiologia, Documento 266. ISSN 1517-8498
15. OCHEM – The Online Chemical Database. <https://occhem.eu/home/show.do>. Accessed 29 Apr 2019
16. Sushko I, Novotarski S, Koner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-souza J, Zang Q-I, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko I (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25:533–554
17. Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV (2012) ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J Chem Inf Model* 52:2310–2316
18. Bolton E, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem* 4:217–241
19. Ertl P (2010) Molecular structure input on the web. *J Chemin* 2:1
20. Oprisius I, Novotarski S, Tetko IV (2013) Modeling of non-additive mixture properties using the online chemical database and modeling environment (OCHEM). *J Chem Inf* 5:1–7
21. ECETOC – European Center for toxicology and ecotoxicology of chemicals. <http://www.ecetoc.org/pt/>. Accessed 29 Apr 2019
22. ECHA (a). REACH – Technical Guidelines for information requirements and safety assessment of chemicals. Chapter R14: Estimate of occupational exposure. European Chemicals Agency, Helsinki, 2010

23. ECHA (b) Guidelines on information requirements and safety assessment of chemicals. Chapter R15: Estimation of consumer exposure (2nd Version, April 2010). European Chemicals Agency, Helsinki, Finland. Addendum to TR114: Technical Base for TRA v3.1 (June 2014) ECETOC, 2010
24. Bhattacharai B, Wilson DM, Parks AK, Carney EW (2016) Evaluation of TOPKAT, Txtree, and Derek Nexus *in silico* models for ocular irritation and development of a knowledge-based framework to improve the prediction of severe irritation. *Chem Res Toxicol* 1:810–822
25. Kienzler A, Barron MG, Belanger SE, Beasley A, Embry MR (2017) Mode of action (MOA) assignment classifications for ecotoxicology. *Environ Sci Technol* 1:1203–1211
26. Russom CL, Bradbury SP, Briderius SJ, Hammermeister DE, Drummond R (1997) A predicting model of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ Toxicol Chem* 5:948–967
27. ECHA – European Chemical Agency. <https://echa.europa.eu/home>. Accessed 29 Apr 2019
28. REACH – Registration, Evaluation, Authorization, Evaluation and Restriction of Chemicals. European Commission. http://ec.europa.eu/environment/chemicals/reach/reach_en.htm. Accessed 29 Apr 2019
29. ECOSAR – Ecological Structure Activity Relationships Predictive Model. <https://www.epa.gov/tcsa-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model>. Accessed 29 Apr 2019
30. OECD – The Organization for Economic Cooperation and Development. <http://www.oecd.org/about/>. Accessed 29 Apr 2019
31. OECD – The Organization for Economic Cooperation and Development – Guidelines. <https://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm>. Accessed 29 Apr 2019
32. Das RN, Sanderson H, Mwambo AE, Roy K (2013) Preliminary studies on model development for rodent toxicity and ITS interspecies correlation with aquatic toxicities of pharmaceuticals. *Bull Environ Contam Toxicol* 90:375–383
33. eChemPortal. <https://www.echemporal.org/echemportal/index.action>. Accessed 29 Apr 2019
34. U.S. EPA – United States Environmental Protection Agency. <https://www.epa.gov/aboutepa>. Accessed 29 Apr 2019
35. EPA – United States Environmental Protection Agency/ECOTOX. <https://cfpub.epa.gov/ecotox/>. Accessed 29 Apr 2019
36. Kosala R, Blockeel H (2000) Web mining research: a survey. *SIGKDD Explorat* 2:1–15
37. Wale N (2011) Machine learning in drug discovery and development. *Drug Develop Res* 72:112–119
38. Wiley and SAS Business Series. Too big to ignore: the business case for big data. 256 paginas, ISBN 1118642104, 9781118642108, 2013
39. Kohavi R, Provost F (1998) Glossary of terms. *Machine Learning* 30:271–274
40. Mitchell TM (1997) Machine learning, vol 49. WCB/McGraw Hill, New York
41. Engelbrecht AP (2003) Computational intelligence: an introduction. Wiley, New York
42. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
43. Konar A (2005) Computational intelligence; principles, techniques and applications. Springer, Berlin
44. Webb A (2002) Statistical pattern recognition. Wiley, New York
45. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, Berline
46. Dobchev DA, Pillai GG, Karelson M (2014) In silico machine learning methods in drug development. *Curr Top Med Chem* 14:1913–1922
47. Agarwal S, Dugar D, Sengupta S (2010) Ranking chemical structures for drug discovery: a new machine learning approach. *J Chem Inf Model* 50:716–731
48. Geppert H, Horvath T, Gartner T, Wrobel S, Baorath J (2008) Support vector machine based ranking significantly improves the effectiveness of similarity searching using 2d fingerprints and multiple reference compounds. *J Chem Inf Model* 48:742–746
49. Rangwala H, Karypis G (2007) fRMSDPred: predicting local RMSD between structural fragments using sequence information. *Comput Syst Bioinformatics Conf* 6:311–322
50. Katritzky A, Kuanar M, Slavov S, Hall C, Karelson M, Kahn I, Dobchev D (2010) Quantitative correlations of physical and chemical properties with chemical structure; utility for prediction. *Chem Rev* 110:5714–5789
51. Sharma OP, Saini NK, Gupta V, Sachdeva K, Arya H (2011) Evolutionary history of QSAR: a review. *J Natur Cons* 1:266–272

52. Berhanu WM, Pillai GG, Oliferenko AA, Katritzky AR (2012) Quantitative structure–activity/property relationships: the ubiquitous links between cause and effect. *ChemPlusChem* 77:507–517
53. Russel S, Norvig P (2003) Artificial intelligence: a modern approach. 3rd ed. Prentice Hall, Upper Saddle River, NJ
54. Garcia I, Fall Y, Gomez G, Gonzalez-Diaz H (2011) First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors *in vitro*, *in vivo*, and in different cellular lines. *Mol Divers* 15:561–567
55. Gertrudes JC, Maltrarollo VG, Silva RA, Oliveira PR, Honório KM, Silva ABF (2012) Machine learning techniques and drug design. *Curr Med Chem* 19:4289–4297
56. Salzberg SL. Book review: C4.5: Programs for machine learning Morgan Kaufmann Publishers by J. Ross Quinlan. Inc., 1993. Machine Learning, © 1994 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, vol 16, pp 235–240, 1994
57. Livingstone D (1995) Data analysis for chemists. Oxford Science Publications, New York
58. Hansch C, Sammes PG, Taylor JB (1990) Comprehensive medicinal chemistry: the rational design, mechanistic study & therapeutic application of chemical compounds, vol 4. Pergamon Press, Oxford
59. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR: hydrophobic, electronic and steric constants. ACS, Washington, D.C.
60. Hansch C, Leo A (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. ACS, Washington, D.C.
61. Tavares LC (2004) QSAR: the Hansch's approach. *Quimera* 27:631–639
62. Hansch C, Fujita T (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J Am Chem Soc* 85:2817–2824
63. Hammet LP (1937) The effect of structures upon the reactions of organic compounds benzene derivatives. *J Am Chem Soc* 59:96–103
64. Ning X, Karypis G (2011) In silico structure-activity-relationship (SAR) models from machine learning: a review. *Drug Develop Res* 72:138–146
65. Agrafiotis D, Bandyopadhyay D, Wegner J, van Vlijmen H (2007) Recent advances in chemoinformatics. *J Chem Inf Model* 47:1279–1293
66. Bravi G, Green EGD, Hann V, Mike M (2000) Modeling structure-activity relationship. In: Bohm H, Schneider G (eds) Virtual screening for bioactive molecules, vol 10. Wiley-VCH, Weinheim, pp 81–116
67. Todeschini R, Consoni V (2008) Handbook of molecular descriptors. Wiley-VCH, Weinheim
68. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M et al (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010
69. Mitchell JBO (2014) Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* 4:468–481
70. Ekins S, Lage de Siqueira-Neto J, McCall L-I, Sarker M, Yadav M, Ponder EL et al (2015) Machine learning models and pathway genome data base for *Trypanosoma cruzi* drug discovery. *PLoS Negl Trop Dis* 9: e0003878
71. Goh GB, Hodas NO, Vishnu A (2017) Deep learning for computational chemistry. *J Comput Chem* 38:1291–1307
72. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH (2018) QSAR-based virtual screening: advances and applications in drug discovery. *Front Pharmacol* 9:1275
73. Warren GL (2012) Special issue: a snapshot in time: docking challenge. *J Comput Aided Mol Des* 26:675–799
74. Carlson HA, Smith RD, Damm-Ganamet KL, Stuckey JA, Ahmed A, Convery MA, Somers DO, Kranz M, Elkins PA, Cui G, Peishoff CE, Lambert MH, Dunbar JB (2016) CSAR 2014: a benchmark exercise using unpublished data from pharma. *J Chem Inf Model* 56:1063–1077
75. Sieg J, Flachsenberg F, Rarey M (2019) In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J Chem Inf Model* 59:947–961
76. Svensson F, Karlén A, Sköld C (2012) Virtual screening data fusion using both structure- and ligand-based methods. *J Chem Inf Model* 52:225–232
77. Barros RPC (2017) Triagem virtual de metabolitos secundários com potencial atividade antimicrobiana do gênero *Solanum* e estudo fitoquímico de *Solanum capsicoides* All. Dissertação de Mestrado. 216 p
78. Acevedo CAH (2018) Estudo quimiotaxonómico e triagem virtual de sesquiterpenos lactonizados isolados da família Asteraceae com

- potencial atividade leishmanicida e tripanocida. Dissertação de Mestrado, 271 p
79. Chuang KV, Keiser MJ (2018) Comment on “Predicting reaction performance in C-N cross-coupling using machine learning”. *Science* 362:eaat8603
80. Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, Liu H (2018) Applications of machine learning methods in drug toxicity prediction. *Curr Top Med Chem* 18:987–997
81. Braga A, Carvalho CPLF, Ludemir TB (2000) Redes Neurais Artificiais: teoria e aplicações. Editora LTC, Rio de Janeiro
82. Haykin S (1999) Neural networks – a comprehensive foundation, 2nd edn. Prentice-hall, New Jersey
83. Lorena AC, Carvalho ACPLF (2007) Uma introdução às support vector machines. *RITA* 14(2):43–67
84. Corinna C, Vapnik VN (1995) Support vector networks. *Mach Lear* 20:273–297
85. Ben-Hur A, Horn D, Siegelmann H, Vapnik VN (2001) Support vector clustering. *J Mach Learn Res* 2:125–137
86. Rokach L, Maimon O (2008) Data mining with decision trees: theory and applications. World Scientific Pub Co Inc., Singapore. ISBN 978-9812771711
87. Quinlan JR (1986) Introduction of decision trees. *Mach Lear* 1:81–106
88. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Books & Software, Monterey. ISBN 978-0-412-04841-8
89. Friedman JH (1999) Stochastic gradient boosting. Technical Report, Stanford University, Stanford
90. Wang F, Rudin C (2015) Falling rule lists. *J Mach Lear* 38:1013–1022
91. Ben-Gal I, Dana A, Shkolnik N, Singer G (2014) Efficient constructions of decision trees by the dual information distance method. *Qual Technol Quant M* 11:133–147
92. Breiman L (2001) Random forests. *Mach Lear* 45:5–32
93. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 46:175–185
94. Everitt BS, Landau S, Leese M, Stahlh D (2011) Miscellaneous clustering methods. In: Cluster analysis, 5th edn. Wiley, Chichester
95. Samworth RJ (2012) Optimal weighted nearest neighbor classifiers. *Ann Stat* 40:2733–2763
96. Rennie J, Shih L, Teevan J, Karger D (2003) Talking the poor assumptions of Naïve Bayes classifiers. *ICML*
97. Maron ME (1961) Automatic indexing: an experimental inquiry. *J ACM* 8:404–417
98. Narasimha Murty M, Susheela Devi V (2011) Pattern recognition: an algorithmic approach. ISBN 978-0857294944
99. Donalek C (2011) Supervised and unsupervised learning. In: *Astronomy colloquia*, USA
100. Masson E, Wang YJ (1990) Introduction to computation and learning in artificial neural networks. *Eur J Oper Res* 47:1–28
101. Siegelmann HT, Sontag ED (1991) Turing computability with neural nets. *Appl Math Let* 4:77–80
102. Optiz D, Maclin R (1999) Popular ensemble methods: an empirical study. *J Art Int Res* 11:169–198
103. Polikar R (2006) Ensemble based systems in decision making. *IEEE Circ Syst Mag* 6:21–45
104. Rokach L (2010) Ensemble-based classifiers. *Art Int Rev* 33:1–39
105. Brown G, Wyatt J, Harris R, Yao X (2005) Diversity creation methods: a survey and categorization. *Inform Fusion* 6:5–20
106. Gashler M, Giraud-Carrier C, Martinez T (2008) Decision tree ensemble: small heterogeneous is better than large homogeneous. The seventh international conference on machine learning and applications, pp 900–905. San Diego, CA, USA
107. Fjodorova N, Novich M, Vrachko N, Smirnov B, Kharchenikova N, Zholdakova Z, Novikov S, Skvortsova N, Filimonov D, Poroikov V, Benfenati E (2008) Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. *J Environ Sci Health C* 26:201–236
108. Roy K, Kar S (2016) In silico models for ecotoxicity of pharmaceuticals. Chapter 12, book in silico methods for predicting drug toxicity. *Methods Mol Biol* 1425:237–304
109. Kluver N, Vogs C, Altenburger R, Escher BI, Scholz S (2016) Development of a general baseline toxicity QSAR model for the fish embryo acute toxicity test. *Chemosphere* 164:164–173
110. Embry MR, Belanger SE, Braunbeck TA, Galay-Burgos M, Halder M, HInton DE, Leonard MA, Lilicrap A, Noberg-king T, Ehale G (2010) The fish embryo toxicity test as an animal alternative method in hazard and risk assessment and scientific research. *Aquat Toxicol* 97:79–87

111. Halder M, Leonard M, Iguchi T, Oris JT, Ryder K, Belanger SE, Braunbeck TA, Embry MR, Whale G, Noberg-king T, Lili-crap A (2010) Regulatory aspects on the use of fish embryos in environmental toxicology. *Integr Environ Assess Manag* 6:484–491
112. Belanger SE, Rawlings JM, Carr GJ (2013) Use of embryo toxicity tests for the prediction of acute fish toxicity to chemicals. *Environ Toxicol Chem* 32:1768–1783
113. Schultz TW, Cronin MTD, Netzeva TI (2003) The present status of QSAR in toxicology. *J Mol Struc-Theochem* 622:23–38
114. Monsinjon T, Andersen OK, Leboulenger F, Knigge T (2006) Data processing and classification analysis of proteomic changes: a case study of oil pollution in the mussel, *Mytilus edulis*. *Prot Sci* 4:2–13
115. Kaiser KLE (2007) Evolution of the international workshops on quantitative structure-activity relationships (QSARs) in environmental toxicology. *SAR QSAR Environ Res* 18:3–20
116. Hansen K, Mika S, Schroeter T, Sutter A, Laak A, Steger-Hartmann T, Heinrich N, Muller K-R (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *J Chem Inf Mod* 16:1567–1580
117. Aoyama T, Suzuki Y, Ichikawa H (1990) Neural networks applied to structure-activity relationship. *J Med Chem* 33:905–908
118. Samghani K, HosseinFatemi M (2016) Developing a support vector machine based QSPR model for prediction of half-life of some herbicides. *Ecotox Environ Safe* 129:10–15
119. Kar S, Roy K, Leszczynski J (2017) On applications of QSARs in food and agricultural sciences: history and critical review of recent developments. *Adv QSAR Mod*:203–302
120. Du H, Wang J, Hu Z, Yao X, Zhang X (2008) Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. *J Agric Food Chem* 56:10785–10792
121. Song JS, Moon T, Nam KD, Lee JK, Hahn H-G, Choi E-J (2008) Quantitative structural-activity relationship (QSAR) study for fungicidal activities of thiazoline derivatives against rice blast. *Bioorg Med Chem Lett* 18:2133–2142
122. Oprisiu I, Varlamova E et al (2012) QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. *Mol Inform* 6:491–502
123. Tetko IV, Novotarskyi S, Sushko I, Ivanov V, Petrenko AE, Deiden R, Lebon F, Mathieu B (2013) Development of dimethyl sulfoxide models using 163 000 molecules: using a domain applicability metric to select more reliable predictions. *J Chem Inf Model* 53:1990–2000
124. Braga RC, Alves VM, Silva FC, Andrade CH (2015) QSAR and molecular modeling approaches for prediction of drug metabolism. In: *Encyclopedia of drug metabolism and interactions*. Wiley, Hoboken, pp 1–28
125. Michielan L, Terfloth L, Gasteiger J et al (2009) Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J Chem Inf Model* 49:2588–2605



Chapter 8

On the Relevance of Feature Selection Algorithms While Developing Non-linear QSARs

Riccardo Concu and M. Natália Dias Soeiro Cordeiro

Abstract

Quantitative structure-activity relationships (QSARs) are mathematical models aimed at finding a quantitative relationship between a set of chemical compounds and a specific activity or endpoint, such as toxicity, chemical or physical property, biological activity, and so on. In order to find out the correlation between the chemicals and the selected endpoints, QSAR models use the so-called molecular descriptors (MDs) which encode specific chemical information or features of the molecules. The early QSAR models were based on a small set of MDs and a specific endpoint, and the correlation was usually a linear mathematical correlation. However, nowadays, QSAR models are usually non-linear and made up by thousands of chemicals and hundreds of MDs. In addition, novel QSAR models are also aimed at the prediction of different endpoints with the same model, the so-called multi-target QSAR (MT-QSAR). Due to this, nowadays many QSARs are usually developed using machine learning approaches which can model a dataset with different endpoints. Although these approaches have demonstrated to be able to solve MT-QSAR models, feature selection (FS) in these cases is a challenging task and a main point in the QSAR field. Considering these aspects, the main aim of this chapter is to analyze feature selection methods while developing non-linear QSAR models.

Key words QSAR, Molecular descriptors, Feature selection, Neural networks, Filter methods, Wrapper methods, Machine learning, Linear models, Non-linear models

1 Introduction

In 1963 Hansch et al. [1] for the first time introduced the concept of quantitative structure-activity relationship (QSAR) developing a linear equation that was able to correlate chemical features of molecules, the so-called molecular descriptors (MDs), to a specific activity, also called an endpoint, like toxicity, drug activity, physical properties, and so on. In that first work, the authors developed a simple linear regression model with a small dataset composed by 35 compounds. Starting from that first model, thousands of different QSAR models have been developed later, using hundreds of diverse approaches. However, the main idea behind a QSAR model is still the same: changes in molecular structure are reflected in

proportional changes in the observed response or biological activity. Although at the very beginning, QSAR was basically used for drug design and drug development, nowadays it has been applied to a lot of different areas. In fact, the main QSAR concept has been extended to other fields such as quantitative structure-toxicity relationship (QSTR) [2, 3], quantitative structure-property relationship (QSPR) [4–6], and quantitative structure-nanomaterial relationship (QNAR) [7, 8]. In doing so, all these different approaches have been used to predict toxicity, property, etc. of different chemicals, and because of this, QSAR modelling is widely employed in academic, industry, and government institutions around the world [9]. In addition, also government agencies are supporting the implementation of QSAR tools and QSAR models for regulatory purposes. The main advantages of these approaches are that they are time and money saving. In fact, using these techniques it is possible to predict a desired property or function avoiding chemical and/or animal testing [10, 11].

Although there are thousands of QSAR models aimed at predicting very different activities, properties, etc., the general workflow is always the same. The first step is aimed at building up and curating a dataset of compounds with their biological activity, physical property, toxicity, etc. [12]. This task is usually achieved by retrieving molecules from online databases such as ChEMBL [13], PubChem [14], or ChemSpider [15]. It is also common that medicinal chemists integrate these data with novel compounds to validate the model and new drugs [16]. The second step is to calculate the MDs, which are the independent variables in the QSAR model. The MDs are the nucleus of QSAR modelling, and thousands of them have been developed in order to codify very different aspects and features of chemicals compounds. Regardless the type of information they are encoding, MDs are numerical representations of a specific feature of the compound. Hundreds of software have been developed to calculate MDs, for instance, Dragon [17], CORINA [18], CODESSA [19], etc. In any case, there are mainly four different MD classes: 1D, which encodes the molecular formula; 2D, which represents the structural formula; 3D, which codifies the three-dimensional structure of the compound; and 4D, which are multidimensional MDs. Due to this, feature selection (FS) is one of the most important and relevant steps while developing a QSAR model, since the performances, predictions, and reliability of the model are strictly correlated with the MDs and the information they are encoding.

The third step is to develop the QSAR model using a linear or non-linear approach able to correlate the MDs with the endpoint. While at the very beginning, the QSAR models were usually based on linear correlations, nowadays there is a general trend to employ non-linear models based on different machine learning algorithms. Linear models are completely reproducible, easier to interpret, and

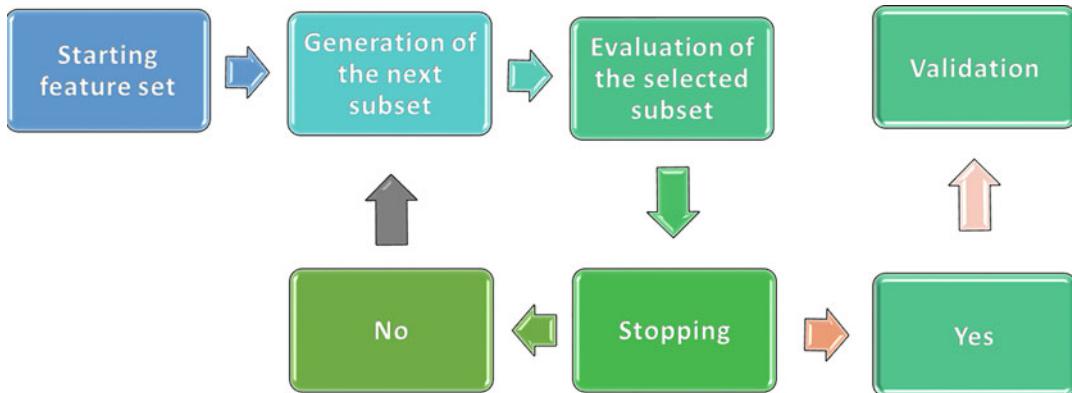


Fig. 1 The FS process

computationally less demanding and thus, in some cases, may be a better choice. However, linear models may fail when inputs in the dataset show high diversity and great complexity and when independent variables have high correlation among each other. In these cases, non-linear models are usually a better choice due to the fact that they can model very complex datasets giving robust and reliable predictions. Several methods have been applied while developing QSAR models; the selection of the proper non-linear technique may have an influence on the final result of the QSAR model. Mostly used non-linear methods are artificial neural networks (ANN) [20–22], support vector machine [23–25], partial least squares (PLS) [26–28], and multivariate adaptive regression splines (MARS) [29, 30]. Even if these are powerful and reliable techniques, they also have several drawbacks: usually they are not reproducible, since they are based on semiempirical algorithms, are more complex and more difficult to interpret, and may suffer from the overfitting problem [31]. Finally, the models should be validated using appropriate methodologies; this is usually done with diverse techniques such as leave one out (LOO), cross-validation, leave many out (LMO), etc. [32]. This general workflow is also described in Fig. 1.

In any case, one of the key points while developing linear and/or non-linear QSAR models is the FS. Although a lot of substantial development of QSAR methods have been done, FS is still a challenging process that is fraught with pitfalls. Since the number of MDs that can be calculated and used in a model is very huge, a proper FS method is essential to develop robust and reliable QSAR models. Even though a QSAR model might be developed using all the calculated MDs, this is usually not recommended for several reasons. First of all, a QSAR model developed with a large set of MDs is really complex, and a mechanistic interpretation is almost impossible. Reduction of the number of MDs reduces the noise in the model, improves the overall accuracy, and

eliminates redundant or collinear variables which are encoding the same information [9, 33]. Moreover, a small set of variables is also one of the most important points to avoid overfitting. As a general rule, the number of MDs should be always less than the compounds in the dataset [34, 35]. Topliss and Castello [36] developed a basic rule that states the ratio between the number of training set compounds and descriptors should be at least 5:1 when developing linear correlations. This rule is not directly applicable for non-linear models; however, there is a general consensus that this ratio should be as higher as possible. In fact, models with high number of features are usually keen to be overfitted and fail in the validation or prediction of external compounds. Finally, a small set of descriptors usually means that the model is also less computational demanding and easier to interpret.

2 Feature Selection

The main aim of each FS process is to identify the best subset of features that best describes the dataset in use and improves the performances of a learning algorithm. Nowadays, the identification and elimination of irrelevant features is one of the biggest challenges in the machine learning field since databases and available features are huge. Many authors provided various definitions for FS; Dash and Liu [37] generalized these definitions as:

- (a) *Idealized*: finds the minimally sized feature subset that is necessary and sufficient to the target concept [38].
- (b) *Classical*: selects a subset of M features from a set of N features, $M < N$, such that the value of a criterion function is optimized over all subsets of size M [39].
- (c) *Improving prediction accuracy*: the aim of FS is to choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [40].
- (d) *Approximating original class distribution*: the goal of FS is to select a small subset such that the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values [40].

In addition, Dash and Liu [37] also standardized the two main criteria while performing a FS:

- The classification accuracy does not significantly decrease.
- The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

In any case, from a general point of view, as reported by Dash and Liu [41], each FS protocol should consist of four basic steps:

1. *Generation*: to generate the next subset for evaluation
2. *Evaluation*: to evaluate the candidate subset
3. *Stopping*: to select a criterion to decide when stopping
4. *Validation*: to validate the selected subset

This process is also reported in the Fig. 1.

The first step may start with three different scenarios: (1) with all features, (2) with no features, and (3) with a random subset. While in the first two cases the features are iteratively added/removed during the process, in the case of the random subset, features are randomly generated or added/removed during the task. The second step, the evaluation of the generated subset, is aimed at evaluating the previous subset with the latest one. If the new subset performs better, then the previous is replaced; otherwise the latest one is discarded. The stopping procedure is crucial since without a rationale stopping criterion, this step may run unnecessarily long. The stopping criterion is usually decided by the researcher and should be based on the two previous steps. For instance, one could consider the number of iterations or the number of features selected if the generation procedure is used to define the stopping point. In the case of reliability on the validation procedure, we may consider if the addition of features generates a more performing subset. The validation procedure basically works comparing the validity and results obtained with the new one in comparison to the previous subset.

There are basically three different methods to efficiently perform the FS procedure. The first one is called filter method, which is based on the reduction of the entire set of descriptors using statistical measures to rank available features; then those achieving scores below a predetermined threshold are automatically rejected. The main advantage of this method is that it is usually easy to design and not computational demanding. The second one, called wrapper, is based on the employment of machine learning algorithms to evaluate the accuracy performance of a large feature subset to find a better correlation between compounds and features. These methods perform better than the filter method; however, they are usually computationally more expensive. Finally, hybrid methods are approaches based on both filter and wrapper methods. Usually, these methods firstly apply a filter method to reduce the dimensionality of the dataset and then a wrapper method.

The main aim of this chapter is to analyze and discuss the most relevant and reliable methods used to perform an efficient FS while developing non-linear QSAR models.

3 Filter Methods

Filter methods are the simplest ones used to select features. These approaches use the training data to select features without applying any kind of algorithm or machine learning technique. There are several filter methods such as consistency methods [42], information methods [43], dependency methods [44], distance methods [45], forward selection [46], backward elimination [47], and step-wise selection [48].

3.1 Consistency Methods

Consistency methods, also called consistency evaluation measures, are methods heavily based on the robustness of the training dataset. In addition, this method also uses the Min-Features bias in selecting a subset of features. In any case, this method basically evaluates if the set of selected features is consistent or not. The output of this procedure is a Boolean value. The FOCUS [49] method and its improved version FOCUS2 [50] applies this selecting measure to stop the procedure while evaluating the consistency of the selected features.

Even if this procedure is simple allowing to achieve small subsets, it has several drawbacks. This method can be applied only with discrete features; if the subset consists of continuous features, this should be first discretized. Otherwise other developments of FOCUS methods may be used, such as CFOCUS [51]. In addition, this method has low noise tolerance since the selected subset may turn inconsistent just shuffling only one feature. Finally, this method is unable to build a subset by itself since it requires supporting tools like the Min-Features bias. Some authors tried to improve this method curtailing its drawbacks [52–54].

3.2 Information Methods

This method basically compares the information gained by the new feature with respect to the previous one. Based on this concept, Bell and Wang [55] developed an algorithm in order to perform a feature subset selection (FSS). They then evaluated the algorithm using 23 public datasets, improving the prediction accuracy of 16 and losing accuracy on only 1 dataset using non-linear algorithms. Other authors applied this method such as Cardie [56] which developed a so-called hybrid approach that outperforms both the decision tree and case-based approaches and two case-based systems that incorporate expert knowledge into their case retrieval algorithms.

3.3 Dependency Methods

Dependency methods evaluate how the value of one variable could be predicted using the value of another variable. In this case, the method selects the feature with the high correlation with the target class selected. Peng et al. [57] developed the minimal-redundancy-maximal-relevance criterion (mRMR) for the first-order incremental FS that was subsequently combined with a wrapper method. In order to evaluate their approach, the authors compared four different public datasets (handwritten digits, arrhythmia, NCI cancer cell lines, and lymphoma tissues) [58–61] and three different classifiers (naive Bayes, support vector machine, and LDA). This approach led the authors to obtain good results with an initial set of features relatively large. In addition, the accuracy of the classifiers was also improved.

Claypo et al. [62] proposed a method which adopts the mutual information to determine significant features from probability between feature and class using a class dependency and feature dissimilarity (CDFD). In addition, the authors also consider the dissimilarity between features based on Euclidean distances in order to consider the differences between features. The authors tested their approach on five datasets [63] using multilayer perceptron (MLP) neural network, decision tree, radial basis function (RBF) neural network, and probabilistic neural network (PNN); in addition, the approach was tested also against Fisher-Markov feature selector (MRF) and genetic algorithm (GA) FS methods [64, 65]. The results reported by the authors show that the CDFD algorithm can produce the lower classification errors in many classifiers.

3.4 Distance Methods

Distance methods are a huge class of FS methods. From a general point of view, they uses conventional distances (e.g., Euclidean distance) to measure the similarity between two samples. One of the most used distance methods is the Bhattacharyya distance [66, 67], which basically measures the similarity of two probability distributions. Piramuthu evaluated and compared several inter-class as well as probabilistic distance-based FS methods as to their effectiveness in preprocessing input data for inducing decision trees [45]. After evaluating these methods on five real-world datasets, they concluded that the non-linear measure is one of the choices in most cases since the reduction of the features was effective without loss of performances. Other authors also developed robust FS method using Bhattacharyya distance [68, 69].

3.5 Forward Selection

Forward selection is a method widely used for FS. This procedure is a specific type of stepwise regression which begins with an empty variable subset and adds in features one by one at each step. The feature selected is the one which allows the best improvement in the model. This procedure continues until no more features able to improve the model are found. The main drawback of this method is that it tends toward overfitting, due to which, it is important to

have a strictly stopping criteria. In addition, when a variable is added, it cannot be excluded later, even if that variable might be redundant or is not improving the performance the model due to other features that are added. In any case, this method has been used by many authors for feature selection [70].

3.6 Backward Elimination

This method works in an opposite way compared with the forward selection. In fact, it starts including all the features and starts eliminating one by one at each step evaluating the contribution of the feature to the improvement of the model. This model is not widely used since it may produce overfitted models. In any case, an interesting approach has been developed by Akhlaghi while investigating A series of 1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine] (HEPT) derivatives, as non-nucleoside reverse transcriptase inhibitors (NNRTIs) [71]. Thorough this approach, the authors were able to identify 11 relevant descriptors from a large set in order to develop an RBF neural network which presents an overall accuracy of 90%.

3.7 Stepwise Selection

This method is probably the most used one for FS in the QSAR area. This is a hybrid method based on both the forward and backward algorithms. The main advantage of this method is that a variable which enters in the model can be then deleted if it is found to be irrelevant. In fact, the process starts adding the variable with the highest correlation with the selected endpoint. At each step, the variable with the highest correlation is added until no more variables with significance are found among the whole set of features. In addition, in each step, all the features included are analyzed, and if a variable previously added is found insignificant, it is deleted from the set of features. The significance criteria used for the forward selection and backward elimination are 0.25 and 0.1, respectively. Based on this criterion, the predictors are selected or eliminated from the analysis. In any case, this method as well as the forward and backward are prone to entrapment in local minima which means that a set of features that cannot be improved in the next step may be found. Shanableh and Assaleh used a stepwise approach to reduce dimensionality [72]. In order to test their approach, two application scenarios were used to test the proposed solution, namely, image-based hand recognition and video-based recognition of isolated sign language gestures. Other authors used this approach to efficiently reduce the number of features while developing non-linear models [7].

4 Wrapper Methods

Wrapper methods are based on greedy search algorithms as they evaluate all possible combinations of the features and select the combination that produces the best result for a specific machine

learning algorithm. They usually perform better than filter methods, although the computational cost is usually higher and thus also the time needed to check all the feature combinations is longer [73]. In addition, wrapper algorithms usually do not suffer from overfitting since typically a cross-validation procedure is applied in order to avoid the problem [74].

In this section, we will review the most relevant wrapper methods used to select the best feature subset in order to develop non-linear QSAR models. There are a huge number of algorithms and approaches that have been developed; here we will review only the most relevant and used in the QSAR field.

4.1 Evolutionary Algorithms

Evolutionary algorithms (EA) are metaheuristic optimization algorithms used for selection features and based on the biological evolution. This means that fitter features will survive and replicate, while unfit features will be discarded, like genes in the evolutionary selection. This class includes genetic algorithm (GA), genetic programming (GP), evolutionary programming (EP), and other related approaches [75–78]. These algorithms are widely used for FS in very different fields [79–83]. Due to this, also in the QSAR field, EA are widely used since they are considered one of the best methods for FS [84–86].

From a general point of view, all the evolutionary algorithms are developed using four steps: initialization, selection, genetic operators, and termination. The procedure starts generating an initial population which are likely solutions for the problem. This first generation is usually randomly created; however, in some cases some prior requirements may be used. The second step is focused on evaluating the generated population according to a fitness function. This function evaluates the fitness of each individual in that population and usually selects only the top two. Once the two top members are selected, then the genetic operators (GO) may be applied. The GO step is performed thorough three different steps: reproduction, crossover, and mutation. Using the two top members that we may consider as “parents,” two new are created crossing the features of the parents; this procedure represents the reproduction and crossover. Subsequently, a mutation is performed randomly changing some part of the feature previously generated. Finally, the termination step may occur. There are two cases in which this usually occurs: either if the procedure has reached a defined runtime or it has reached some threshold of performance. There is also a variation of the classical EA which is called multi-objective evolutionary algorithms (MOA) [87]. These algorithms are useful while solving problems with multiple fitness functions which require a set of optimal solutions points. The set of optimal solutions is called the Pareto frontier.

In any case, many authors have used EA to select best features while developing non-linear QSAR models. Ozdemir et al. [88] applied a novel GA to select the most relevant features among an

initial set of 160. The algorithm was able to select 40 features with low intercorrelation and high correlation with the endpoint. A QSAR model was developed using feedforward ANN; the sensitivity of the model was higher with only 40 features.

Another interesting application of the application of GA to FS was presented by Bahmani et al. [89]. In this case, the authors developed a back-propagation ANN (BP-ANN) for modelling the retention time of 57 morphine and its derivatives. In this case, the GA was able to select only 3 descriptors among the 200 calculated.

Mizera et al. developed an innovative quantitative structure-retention relationship (QSRR) for analysis of triptans, selective serotonin 5-HT1 receptor agonists used for the treatment of acute headache [90]. Also in this case, the GA was able to select the most relevant features in order to develop ANN robust model substantially reducing the initial number of descriptors. Other relevant works of combining GA with ANN could be found here [91–95].

4.2 Ant Colony Optimization (ACO)

Ant colony optimization (ACO) algorithm is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. Even though this algorithm is based on a natural phenomenon, in fact it mimics the behavior of an ant colony while looking for the shortest path between nest and food sources. When ants travel, they deposit pheromone, this means that as more ants travel thorough a route richer in pheromone that will be the primary route. In addition, pheromone tends to decay during time, and as more ants use a specific route to food source, the trail becomes richer in pheromone concentration, and as a final result, ants will follow the final route. On the other hand, routes less used will lose their pheromone concentration due to evaporation, and ants will not follow this route. This algorithm was firstly introduced by Dorigo et al. [96], and a general workflow for solving a problem using an ACO approach was proposed by Mullen et al. [97]. In any case, at each iteration, every ant constructs a solution to the problem by moving on a graph. Each edge of the graph represents the possible steps the ant can make; in addition, two kinds of information that can guide ant movements are assigned to each edge:

1. Heuristic information, which measures the heuristic preference of moving from a node a to a node b , of traveling the edge a_{ab} . It is denoted by η_{ab} . This information is preserved during all the process.
2. Artificial pheromone deposition, which measures the “learned desirability” of the movement and mimics the real pheromone that natural ants deposit. This information is modified at each iteration depending on how many ants move through each edge. It is denoted by τ_{ab} .

Considering the above general scheme, ACO algorithm is usually defined as follow:

1. Represent the problem by a defined graph where ants can move and so find a solution.
2. Define the meaning of the pheromone trails.
3. Define the heuristic preference for the ant while constructing a solution.
4. Select an appropriate ACO algorithm and apply to solve the problem.
5. Tune the parameter of the ACO algorithm.

An interesting application of ACO to FS in non-linear regression methods was presented by Goodarzi et al. [98]. In this case, the authors first calculate 1457 MDs using the Dragon package. Only 16 descriptors from the total pool were independently selected by suitable selection methods including GA, successive projections algorithm (SPA), and fuzzy rough set ACO. Finally, they tested the set of selected features using MLR, artificial neural network (ANN), and support vector machine (SVM).

4.3 Sensitivity Analysis

Sensitivity analysis (SA) generally refers to the assessment of the importance of features in the respective models. In short, given a fitted model with certain model parameters for each predictor, what would be the effect from varying the parameters of the model (for each variable) on the overall model fit? Therefore, the predictors can be sorted by their importance or relevance for the specific neural net, and thus, the features with less importance can be eliminated. This approach has been adopted by several authors and has the great advantage of checking the relevance of the features directly for the model.

Bing et al. developed a SVM model to discriminate 32 phenethyl-amines between antagonists and agonists and predict the activities of these compounds [99].

Embrechts et al. developed a new sensitivity analysis approach for FS using multiple ensemble neural networks and then applied it to in silico drug design with QSAR. Using this innovative approach, the authors were able to reduce the initial set of 160 descriptors to 35 improving the accuracy of the resulting neural network [100].

Another relevant approach was developed by Kurita et al. [101]. In this case, the authors built a QSAR model based on SVM to predict carcinogenicity. They also developed a new SA method which improves the overall accuracy of the model obtained with a correlation coefficient and F-score-based FS.

4.4 Particle Swarm Optimization

The particle swarm optimization (PSO) is another swarm intelligence algorithm based on the simulating the natural behavior of bird flocking. The main idea of this algorithm is that particle

swarms could explore the search space through a population of particles, which adapt by returning to previously successful regions. The particles have random position in the variable space; the movement of the particles is stochastic, and it is also influenced by the particles' and peers' memories. Each particle keeps track of its coordinates in the problem space, which are associated with the best solution (fitness) it has achieved so far. At the end of each iteration, PSO will change the velocity of each particle and update the best solution. This procedure was presented in 1995 by Kennedy and Erberhart [102] and has been successfully applied as FS method in non-linear QSAR.

In this sense, Agrafiotis and Cedeño checked the efficiency of this method on three different datasets developing feedforward neural networks [103]. The results reported by authors demonstrate better selectivity against other approaches identifying better and more diverse set of features.

On the basis of this approach, Wang et al. developed an interesting binary PSO (BPSO) combined with a back-propagation ANN. The authors tested this approach on four different datasets [104].

4.5 Other Methods

Xue et al. developed a FS method called recursive feature elimination (RFE) based on SVM. At each step, a SVM is trained, and the worst variable is identified by the absolute weight of the feature in the model and subsequently eliminated [105]. A total of 22 MDs was selected from an initial pool of 159, reducing thus the noise and the dimensionality of the model.

Another relevant approach was presented by Soto et al. [106]. The authors described a two-phase methodology for FS that could be applied to both linear and non-linear QSAR. The first step is based on a multi-objective evolutionary technique which allows several advantages compared to mono-objective methods. The second step complements the first one and was developed to refine and improve the confidence in the chosen subsets of features.

5 Conclusion

FS is an essential and fundamental step while developing non-linear QSAR since the number of descriptors that actually can be calculated and included in a model is huge. In fact, software like Dragon, CODESSA, PaDEL-Descriptor, etc. can calculate thousands of MDs. While great improvements have been achieved developing new MDs and MDs computing software, QSAR algorithms, and dataset curation, FS is still a challenging task, and no great improvements have been achieved. FS is needed for some very important reasons. First, it is essential to reduce the number of features in a

model in order to avoid overfitting. Second, in order to perform a mechanistic interpretation of the model, it is essential to have a small set of features. Third, reducing the number of features avoids collinearity between variables and reduces the noise of the model. Fourth, a model with a small set of MDs is usual more robust and stable and is of higher quality. Finally, if a model is built with a huge number of MDs, it is fairly impossible to know what we are modelling.

There are a lot of FS methods and algorithms, and there is no consensus about which method is better and should be preferred. This is due to the fact that each dataset is different with respect to number, type of MDs, biological activity, and so on. Filter methods are usually faster and less computational expensive and time consuming. On the other hand, wrapper methods are usually more robust but time and computational expensive. In this sense, a good approach may be first applying a filter method and then a wrapper algorithm. In any case, a proper FS method should be always applied while developing a non-linear QSAR model.

Acknowledgments

This work received financial support from Fundação para a Ciência e a Tecnologia (FCT/MEC) through national funds and co-financed by the European Union (FEDER funds) under the Partnership Agreement PT2020, through projects UID/QUI/50006/2013, POCI/01/0145/FEDER/007265, NORTE-01-0145-FEDER-000011 (LAQV@REQUIMTE), and the Interreg SUDOE NanoDesk (SOE1/P1/E0215; UP). RC acknowledges also FCT and the European Social Fund for financial support (Grant SFRH/BPD/80605/2011). To all financing sources, the authors are greatly indebted.

The authors declare no competing financial interest.

References

- Hansch C, Muir RM, Fujita T, Maloney PP, Geiger F, Streich M (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with hammett constants and partition coefficients. *J Am Chem Soc* 85(18):2817–2824
- Gombar VK, Enslein K, Blake BW (1995) Assessment of developmental toxicity potential of chemicals by quantitative structure-toxicity relationship models. *Chemosphere* 31(1):2499–2510
- Roy K, Ghosh G (2004) QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. *J Chem Inf Comput Sci* 44(2):559–567
- Basak SC, Nikolic S, Trinajstic N, Amic D, Beslo D (2000) QSPR modeling: graph connectivity indices versus line graph connectivity indices. *J Chem Inf Comput Sci* 40(4):927–933
- Grover II, Singh II, Bakshi II (2000) Quantitative structure-property relationships in pharmaceutical research – part 2. *Pharm Sci Technolo Today* 3(2):50–57
- Grover II, Singh II, Bakshi II (2000) Quantitative structure-property relationships in

- pharmaceutical research – part 1. *Pharm Sci Technolo Today* 3(1):28–35
7. Concu R, Kleandrova VV, Speck-Planche A, Cordeiro M (2017) Probing the toxicity of nanoparticles: a unified *in silico* machine learning model based on perturbation theory. *Nanotoxicology* 11(7):891–906
 8. Burello E, Worth AP (2011) QSAR modeling of nanomaterials. *Wiley Interdiscip Rev Nanomed Nanobiotechnol* 3(3):298–306
 9. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M et al (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57 (12):4977–5010
 10. Wilm A, Kuhn J, Kirchmair J (2018) Computational approaches for skin sensitization prediction. *Crit Rev Toxicol* 48(9):738–760
 11. Ford KA (2016) Refinement, reduction, and replacement of animal toxicity tests by computational methods. *ILAR J* 57 (2):226–233
 12. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29(6–7):476–488
 13. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–DD54
 14. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 47(D1):D1102–D11D9
 15. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87(11):1123–1124
 16. Jabeen I, Wetwitayaklung P, Chiba P, Pastor M, Ecker GF (2013) 2D- and 3D-QSAR studies of a series of benzopyranes and benzopyrano[3,4b][1,4]-oxazines as inhibitors of the multidrug transporter P-glycoprotein. *J Comput Aided Mol Des* 27(2):161–171
 17. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: an easy approach to molecular descriptor calculations. *Match Commun Math Comput Chem* 56 (2):237–248
 18. Sadowski J, Gasteiger J, Klebe G (1994) Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J Chem Inf Comput Sci* 34(4):1000–1008
 19. Ignatz-Hoover F, Petrukhin R, Karelson M, Katritzky AR (2001) QSRR correlation of free-radical polymerization chain-transfer constants for styrene. *J Chem Inf Comput Sci* 41(2):295–299
 20. Roy K, Pratim RP (2009) Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur J Med Chem* 44(7):2913–2922
 21. Baskin II, Palyulin VA, Zefirov NS (2008) Neural networks in building QSAR models. *Methods Mol Biol* 458:137–158
 22. Wiese M, Schaper KJ (1993) Application of neural networks in the QSAR analysis of percent effect biological data: comparison with adaptive least squares and nonlinear regression analysis. *SAR QSAR Environ Res* 1 (2–3):137–152
 23. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV (2003) Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* 43(6):2048–2056
 24. Li S, Fedorowicz A, Andrew ME (2007) A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset. *SAR QSAR Environ Res* 18(5–6):423–441
 25. Shi W, Zhang X, Shen Q (2010) Quantitative structure-activity relationships studies of CCR5 inhibitors and toxicity of aromatic compounds using gene expression programming. *Eur J Med Chem* 45(1):49–54
 26. Stoyanova-Slavova IB, Slavov SH, Pearce B, Buzatu DA, Beger RD, Wilkes JG (2014) Partial least square and k-nearest neighbor algorithms for improved 3D quantitative spectral data-activity relationship consensus modeling of acute toxicity. *Environ Toxicol Chem* 33(6):1271–1282
 27. Nikolic K, Filipic S, Smolinski A, Kaliszan R, Agbaba D (2013) Partial least square and hierarchical clustering in ADMET modeling: prediction of blood-brain barrier permeation of alpha-adrenergic and imidazoline receptor ligands. *J Pharm Pharm Sci* 16(4):622–647
 28. Brandmaier S, Sahlin U, Tetko IV, Oberg T (2012) PLS-optimal: a stepwise D-optimal design based on latent variables. *J Chem Inf Model* 52(4):975–983
 29. Koba M, Baczek T (2013) The evaluation of multivariate adaptive regression splines for the prediction of antitumor activity of acridinone derivatives. *Med Chem* 9(8):1041–1050
 30. Put R, Xu QS, Massart DL, Vander HY (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure-retention relationship studies. *J Chromatogr A* 1055(1–2):11–19

31. Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, Bernard P (2009) How to recognize and work-around pitfalls in QSAR studies: a critical review. *Curr Med Chem* 16(32):4297–4313
32. Gramatica P (2013) On the development and validation of QSAR models. *Methods Mol Biol* 930:499–526
33. Basak SC, Natarajan R, Mills D, Hawkins DM, Kraker JJ (2006) Quantitative structure-activity relationship modeling of juvenile hormone mimetic compounds for *Culex pipiens* larvae, with a discussion of descriptor-thinning methods. *J Chem Inf Model* 46(1):65–77
34. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). *Expert Opin Drug Dis* 13(12):1075–1089
35. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E et al (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 48(9):1733–1746
36. Topliss JG (1972) Utilization of operational schemes for analog synthesis in drug design. *J Med Chem* 15(10):1006–1011
37. Dash M, Liu H (1997) Feature selection for classification. *Intell Data Anal* 1(1):131–156
38. Kira K, Rendell LA (1992) The feature selection problem: traditional methods and a new algorithm. Proceedings of the tenth national conference on artificial intelligence, San Jose, 1867155, AAAI Press, pp 129–134
39. Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. *IEEE Trans Comput* 26(9):917–922
40. Koller D, Sahami M (1996) Toward optimal feature selection. Proceedings of the thirteenth international conference on machine learning, Bari, 3091731, Morgan Kaufmann Publishers Inc., pp 284–292
41. Dash M, Liu H (2003) Consistency-based search in feature selection. *Artif Intell* 151(1):155–176
42. Araujo-Azofra A, Benitez JM, Castro JL (2008) Consistency measures for feature selection. *J Intell Inf Syst* 30(3):273–292
43. Jun BH, Kim CS, Song H, Kim J (1997) A new criterion in selection and discretization of attributes for the generation of decision trees. *IEEE Trans Pattern Anal Mach Intell* 19(12):1371–1375
44. Chandrashekhar G, Sahin F (2014) A survey on feature selection methods. *Comp Electr Eng* 40(1):16–28
45. Piramuthu S (2004) Evaluating feature selection methods for learning in data mining applications. *Eur J Oper Res* 156(2):483–494
46. Whitley DC, Ford MG, Livingstone DJ (2000) Unsupervised forward selection: a method for eliminating redundant variables. *J Chem Inf Comput Sci* 40(5):1160–1168
47. Sutter JM, Kalivas JH (1993) Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchem J* 47(1):60–66
48. Livingstone DJ, Salt DW (2005) Variable selection—Spoilt for choice? Reviews in Computational Chemistry. In: Lipkowitz KB, Larter R, Cundari TR (eds) John Wiley & Sons, Inc., chap.4, vol 21, pp. 287–348
49. Almuallim H, Dietterich TG (1991) Learning with many irrelevant features. Proceedings of the ninth National conference on Artificial intelligence, vol 2, Anaheim, 1865761, AAAI Press, pp 547–552
50. Almuallim H, Dietterich TG (1994) Learning Boolean concepts in the presence of many irrelevant features. *Artif Intell* 69(1):279–305
51. Araujo A, Benítez JM, Castro JL (eds) C-FOCUS: a continuous extension of FOCUS2003. Springer, London
52. Tay FEH, Lixiang S (2002) A modified Chi² algorithm for discretization. *IEEE Trans Knowl Data Eng* 14(3):666–670
53. Boros E, Hammer PL, Ibaraki T, Kogan A, Mayoraz E, Muchnik I (2000) An implementation of logical analysis of data. *IEEE Trans Knowl Data Eng* 12(2):292–306
54. Demšar J, Zupan B, Leban G, Curk T (eds) Orange: from experimental machine learning to interactive data mining 2004. Springer Berlin Heidelberg, Berlin, Heidelberg
55. Bell DA, Wang H (2000) A formalism for relevance and its application in feature subset selection. *Mach Learn* 41(2):175–195
56. Cardie C (1993) Using decision trees to improve case-based learning, in machine learning proceedings. Morgan Kaufmann, San Francisco (CA), pp 25–32
57. Hanchuan P, Fuhui L, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238
58. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A et al (2000) Distinct types of diffuse large B-cell lymphoma

- identified by gene expression profiling. *Nature* 403(6769):503–511
59. Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell* 19(2):153–158
60. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P et al (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 24:227
61. Ding C, Peng H (eds) (2003) Minimum redundancy feature selection from microarray gene expression data. Computational systems bioinformatics CSB2003 proceedings of the 2003 IEEE bioinformatics conference CSB2003, 11–14 Aug 2003
62. Claypo N, Jaiyen S (eds) (2015) A new feature selection based on class dependency and feature dissimilarity. 2015 2nd international conference on advanced informatics: concepts, theory and applications (ICAICTA), 19–22 Aug 2015
63. Yu-Shuen T, Ueng-Cheng Y, Chung IF, Chuen-Der H (eds) (2013) A comparison of mutual and fuzzy-mutual information-based feature selection strategies. 2013 IEEE international conference on fuzzy systems (FUZZ-IEEE), 7–10 July 2013
64. Cheng Q, Zhou H, Cheng J (2011) The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data 2011, pp 1217–1233
65. Aalaei S, Shahraki H, Rowhani manesh A, Eslami S (2016) Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran J Basic Med Sci* 19(5):476–482
66. Fukunaga K (1990) Chapter 10 – feature extraction and linear mapping for classification. In: Fukunaga K (ed) Introduction to statistical pattern recognition, 2nd edn. Academic Press, Boston, pp 441–507
67. Fukunaga K (1990) Chapter 9 – feature extraction and linear mapping for signal representation. In: Fukunaga K (ed) Introduction to statistical pattern recognition, 2nd edn. Academic Press, Boston, pp 399–440
68. Choi E, Lee C (2003) Feature extraction based on the Bhattacharyya distance. *Pattern Recogn* 36(8):1703–1709
69. Drotár P, Gazda J, Sméká Z (2015) An experimental comparison of feature selection methods on two-class biomedical datasets. *Comput Biol Med* 66:1–10
70. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46(1):389–422
71. Akhlaghi Y, Kompany-Zareh M (2006) Application of radial basis function networks and successive projections algorithm in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J Chemom* 20 (1–2):1–12
72. Shanableh T, Assaleh K (2010) Feature modeling using polynomial classifiers and stepwise regression. *Neurocomputing* 73 (10):1752–1759
73. Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artif Intell* 97 (1):273–324
74. Nasariparsa M, Bidgoli A-M, Varae T (2013) A hybrid feature selection method to improve performance of a group of classification algorithms. *CoRR;abs/1403.2372*
75. Nicolotti O, Carotti A (2006) QSAR and QSPR studies of a highly structured physico-chemical domain. *J Chem Inf Model* 46 (1):264–276
76. Yang J, Honavar V (1998) Feature subset selection using a genetic algorithm. In: Liu H, Motoda H (eds) Feature extraction, construction and selection: a data mining perspective. Springer US, Boston, pp 117–136
77. Wang XZ, Buontempo FV, Young A, Osborn D (2006) Induction of decision trees using genetic programming for modelling ecotoxicity data: adaptive discretization of real-valued endpoints. *SAR QSAR Environ Res* 17 (5):451–471
78. Fjell CD, Jenssen H, Cheung WA, Hancock RE, Cherkasov A (2011) Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chem Biol Drug Des* 77(1):48–56
79. Kumar M, Husain M, Upadhyay N, Gupta D (2010) Genetic algorithm: review and application. *IJITM* 2(2):451–454
80. Weile DS, Michielssen E (1997) Genetic algorithm optimization applied to electromagnetics: a review. *IEEE Trans Antennas Propag* 45(3):343–353
81. Hopper E, Turton B (eds) (1998) Application of genetic algorithms to packing problems — a review. Springer, London
82. Hussein F, Khurana N, Ward R (eds) (2001) Genetic algorithms for feature selection and weighting, a review and study. Proceedings of Sixth International Conference on Document Analysis and Recognition. 13 Sept 2001

83. Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. *J Chemom* 15(7):559–569
84. Fernandez M, Caballero J, Fernandez L, Sarai A (2011) Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Mol Divers* 15 (1):269–289
85. Niculescu SP (2003) Artificial neural networks and genetic algorithms in QSAR. *J Mol Struct THEOCHEM* 622(1):71–83
86. Venkatraman V, Dalby AR, Yang ZR (2004) Evaluation of mutual information and genetic programming for feature selection in QSAR. *J Chem Inf Comput Sci* 44(5):1686–1692
87. Zhou A, Qu B-Y, Li H, Zhao S-Z, Suganthan PN, Zhang Q (2011) Multiobjective evolutionary algorithms: a survey of the state of the art. *Swarm Evolutionary Comput* 1(1):32–49
88. Ozdemir M, Embrechts MJ, Arciniegas F, Breneman CM, Lockwood L, Bennett KP (eds) (2001) Feature selection for in-silico drug design using genetic algorithms and neural networks. SMCia/01 proceedings of the 2001 IEEE mountain workshop on soft computing in industrial applications (Cat No01EX504), 27 June 2001
89. Bahmani A, Saaidpour S, Rostami A (2017) Quantitative structure–retention relationship modeling of morphine and its derivatives on OV-1 column in gas–liquid chromatography using genetic algorithm. *Chromatographia* 80 (4):629–636
90. Mizera M, Krause A, Zalewski P, Skibiński R, Cielecka-Piontek J (2017) Quantitative structure–retention relationship model for the determination of naratriptan hydrochloride and its impurities based on artificial neural networks coupled with genetic algorithm. *Talanta* 164:164–174
91. Ghasemi G, Nirouei M, Shariati S, Abdolmaleki P, Rastgoo Z (2016) A quantitative structure–activity relationship study on HIV-1 integrase inhibitors using genetic algorithm, artificial neural networks and different statistical methods. *Arab J Chem* 9: S185–SS90
92. Velásco-Mejía A, Vallejo-Becerra V, Chávez-Ramírez AU, Torres-González J, Reyes-Vidal Y, Castañeda-Zaldivar F (2016) Modeling and optimization of a pharmaceutical crystallization process by using neural networks and genetic algorithms. *Powder Technol* 292:122–128
93. Li Y, Abbaspour MR, Grootendorst PV, Rauth AM, Wu XY (2015) Optimization of controlled release nanoparticle formulation of verapamil hydrochloride using artificial neural networks with genetic algorithm and response surface methodology. *Eur J Pharm Biopharm* 94:170–179
94. Noorizadeh H, Farmany A, Noorizadeh M (2011) Application of GA-KPLS and L-M ANN calculations for the prediction of the capacity factor of hazardous psychoactive designer drugs. *Med Chem Res* 21:2680–2688
95. Sukumar N, Prabhu G, Saha P (2014) Applications of genetic algorithms in QSAR/QSPR modeling. In: Valadi J, Siarry P (eds) Applications of metaheuristics in process engineering. Springer International Publishing, Cham, pp 315–324
96. Dorigo M, Maniezzo V, Colorni A (1996) Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern B Cybern* 26(1):29–41
97. Mullen RJ, Monekosso D, Barman S, Remagnino P (2009) A review of ant algorithms. *Expert Syst Appl* 36(6):9608–9617
98. Goodarzi M, Freitas MP, Jensen R (2009) Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3 beta inhibitory activities. *J Chem Inf Model* 49(4):824–832
99. Niu B, Lu W-C, Yang S-S, Cai Y-D, Li G-Z (2007) Support vector machine for SAR/QSAR of phenethyl-amines1. *Acta Pharmacol Sin* 28(7):1075–1086
100. Embrechts MJ, Arciniegas F, Ozdemir M, Breneman CM, Bennett K, Lockwood L (eds) (2001) Bagging neural network sensitivity analysis for feature reduction for in-silico drug design. IJCNN'01 international joint conference on neural networks proceedings (Cat No01CH37222), 15–19 July 2001
101. Tanabe K, Kurita T, Nishida K, Lučić B, Amić D, Suzuki T (2013) Improvement of carcinogenicity prediction performances based on sensitivity analysis in variable selection of SVM models. *SAR QSAR Environ Res* 24(7):565–580
102. Kennedy J, Eberhart R (eds) (1995) Particle swarm optimization. Proceedings of ICNN'95 – international conference on neural networks. 27 Nov–1 Dec. 1995
103. Agrafiotis DK, Cedeño W (2002) Feature selection for structure–activity correlation using binary particle swarms. *J Med Chem* 45(5):1098–1107

104. Wang Z, Durst GL, Eberhart RC, Boyd DB, Miled ZB (eds) Particle swarm optimization and neural network application for QSAR. 18th international parallel and distributed processing symposium, 2004 proceedings, 26–30 Apr 2004
105. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* 44 (5):1630–1638
106. Soto AJ, Cecchini RL, Vazquez GE, Ponzoni I (2009) Multi-objective feature selection in QSAR using a machine learning approach. *QSAR Comb Sci* 28(11–12):1509–1523



Chapter 9

Got to Write a Classic: Classical and Perturbation-Based QSAR Methods, Machine Learning, and the Monitoring of Nanoparticle Ecotoxicity

Ana S. Moura and M. Natália D. S. Cordeiro

Abstract

Machine learning has become a central feature in the development or refinement of in silico methodologies and techniques. Quantitative structure-activity relationship (QSAR) models are no exception. In fact, one can consider there is a renaissance of QSAR techniques and respective reliability as there is a greater synergy between the two of them. Further, this new wave of *QSAR + machine learning* (ML) techniques allows new avenues in several fields of application, namely, when regarding cytotoxicity and/or ecotoxicity monitoring of nanoparticles (NPs). The latter is of major importance, as the challenges brought by environment management and the increasing concern it has on the food chain are met with expensive and overall slow experimental answers. Within this context, and alongside *classical QSAR + machine learning* techniques, recent QSAR perturbation-based models join methods with ML as well. The QSAR perturbation models feature the possibility of simultaneous modeling multi bio-targets versus NPs in different experimental conditions, thus offering practical solutions to classical QSAR + ML limitations. The use of in silico models could be the most feasible answer to the present and future scenarios of mandatory ecotoxicity monitorization for nanotechnology by-products. This chapter approaches the methodologies and fundamentals of classical and perturbation-based QSAR models within the environmental risk assessment framework, as scaffold to develop novel in silico techniques.

Key words QSAR, QSTR, Machine learning, Ecotoxicity, Environmental monitorization, Nanoparticles, Hybrid in silico models

1 Introduction

Quantitative structure-activity relationship (QSAR) models, whatever their classification might be, are rooted on the concept presented by Crum and Fraser relating the chemical structure of a compound and its chemical activity [1, 2]. This assumption of synergetic relation between chemical structure and biochemical consequential properties has now been established after more than a hundred years of research, namely, with the publication, by Corwin Hansch, of a free-energy model which correlated

psychochemical properties and biological activities, enabling the postulate of QSAR methodology as a sound research technique [3].

As computational capacity and specific software packages were developed, QSAR models gained popularity among researchers. However, a trend in fewer QSAR-related publications led to a published analysis, which concluded that this slight decrease of published QSAR studies could be related with the present matured classical QSAR techniques [4]. Further, the authors suggested the decreasing trend would be reverted as machine learning methods entered a synergetic mode with QSAR models, thus reaching a higher plateau in drug design productivity, for example.

Machine learning (ML) refers to the algorithms and statistical methodologies that allow computer systems to emulate, to put it in simple terms, the human process of adjustment to a task [5]. One of the consequences is the higher autonomy in the process, as the foundation of machine learning rests on inference and pattern recognition rather than explicit instructions regarding a particular task. This upgrades QSAR models as the predictions can attain a refinement and degree of accuracy as well as correlation of excellence level, while decreasing the time period of accomplishing the task.

Focusing on the ecotoxicity context, alongside with QSAR models, one also considers quantitative structure-toxicity models (QSTR). The environmental concerns are an essential feature, on survival lining in many occasions, and within that scope, the nanoparticles (NPs) appear as an increasingly focal point. Recent years have witnessed the emergence of nanotechnology as one of the most promising scientific fields, with applications in areas such as electronics, catalysis, magnetism, optics, photonics, and biomedicine [6–19]. Albeit such interesting scenario, the fact remains NPs are not entirely monitored regarding toxic effects on human and biosystems, including aqueous environment. The plethora of high expensive laboratory assays, as well as a holistic approach to the development of an efficient monitor, prompted the quest to research the possibility of *in silico* models providing a plateau for initial monitorization that could be effective, reliable, and swift while being economically competitive.

Why initial? Because *in silico* models live in close synergy with experimental techniques, through either input data or external validation, and are sustained and complemented by the laboratory validation, models, and methods. As such, the presented techniques never lose sight of the physical reality and comprise validation not as exterior to the protocols but as integrant of the methodologies.

The present chapter is divided in the following subsections: Subheading 2, which summarizes the interfaces for input and output data of the *in silico* models; Subheading 3, comprising three subsections (the first, a brief introduction to *machine learning methods*; the second, detailing the protocol of *classical QSAR + ML*

techniques; and the last, exploring the fundamentals of *QSAR perturbation-based + ML models*); and a final section for the concluding remarks, Subheading 4.

2 Materials

Materials regarding in silico techniques fall heavily on the “descriptor” class. This is the situation for classical QSARs and most QSAR-based models. The following subsections address descriptors within in silico modeling and ecotoxicity context.

2.1 Input Data: Collecting and Calculating

The notion of a descriptor is the cornerstone of input data in this type of models. A descriptor is a property, such as molar volume, or a particular aspect, such as coating agent, encoded in such a manner as to permit the original chemical information to be used in in silico techniques [20, 21]. Descriptors can be collected through available experimental data from literature; public sources, such as the Chemcool Periodic Table source tool; or commercial and noncommercial software for calculating molecular descriptors for QSAR models, such as the ADMET Predictor [22, 23].

However, it is pertinent to introduce and summarize several of the chemical information, apart from the chemical composition, which might be of interest when developing an in silico model within nanoparticle environmental risk assessment, namely, NP ecotoxicity monitoring. Industrial nanoparticles with reckoning ecotoxic effects can be classified into six categories, as it follows: fullerenes, such as nanocones; metal nanoparticles, as elemental silver, oxides and/or binary compounds, as when including carbides; complex compounds of two or more elements, such as alloys; quantum dots; and organic polymers, such as polystyrene [24]. The mobility factor of a NP is also important, as a NP with unstable suspension will follow the natural tendency to aggregate, leading to massive deposition and the formation of much larger particles [25, 26]. To avoid agglomeration context, the manufacturer may use coating agents on the surface of the NPs, which leads to superficial changes, eventually impacting the ecotoxicity of the NP [27]. Furthermore, other important factors might affect the NP superficial properties and therefore its potential toxicity, such as pH changes [28].

Finally, in the context of *QSAR + ML models*, a very recent work exploring the possibility of generating universal nanodescriptors is noteworthy [29]. Aiming at establishing a new methodology to develop universal nanodescriptors, the authors used the Pauling electronegativity as empirical information on the definition of the descriptor characters and the Delaunay tessellation approach in the simulation of the nanostructures [30]. This latter decomposes the nanostructure by a given set of points, which in turn decompose the nanosurface into tetrahedrons, with the atoms located in the

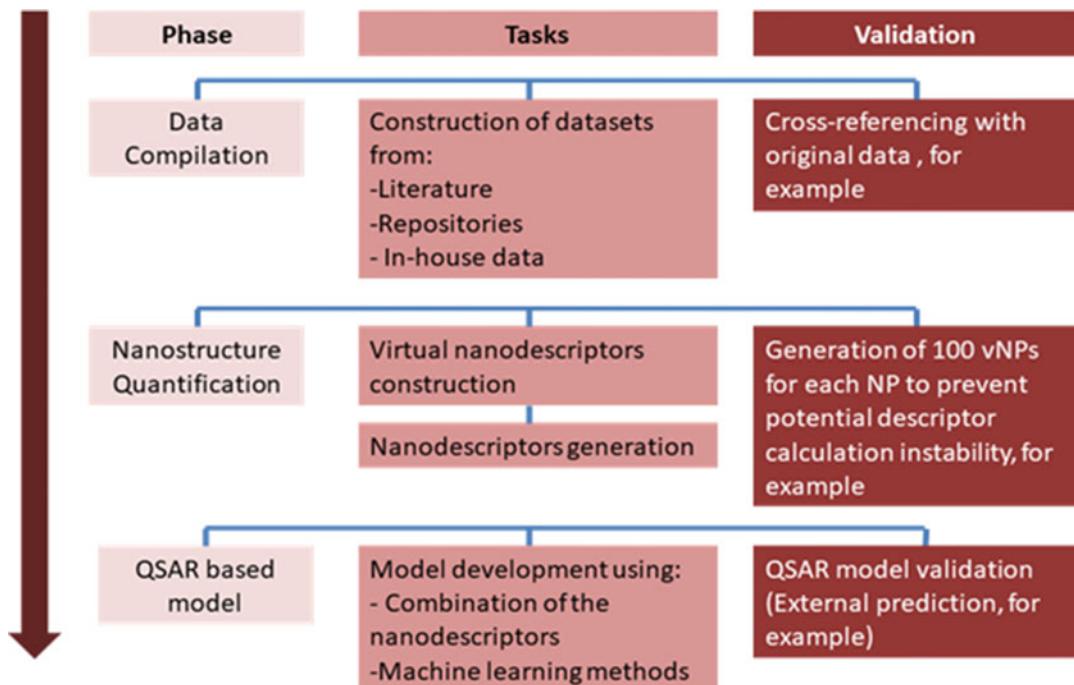


Fig. 1 Scheme for developing universal nanodescriptors

vertices. The selection of atoms for a tetrahedron is not fortuitous, as they are selected in such a manner that their circumscribing sphere is absent of any other atoms. After generating the nanodescriptors, a validation stage ensued, through quantitative nanostructure activity relationship (QNAR) models developed by the new nanodescriptors sets, each set representing different properties versus biological activities.

2.2 Protocols for Developing Universal Nanodescriptors

The protocols for the novel technique for development of universal nanodescriptors comprise three stages, illustrated by Fig. 1 [29]. The details of the protocols will follow this phase categorization as well.

1. Data compilation stage
 - (a) Curation of data for the new datasets from original data sources and/or data repositories, in such a manner that each dataset should model a specific bioactivity/physico-chemical property data, as NP-enzyme binding affinities, developed through intrinsic fluorescence intensities versus NP presence or absence, or log *P* values, as examples
 - (b) Constitution of the new modeling datasets, with appropriate number of NPs
2. Nanostructure quantification
 - (a) Virtual nanodescriptor construction

- (i) Each NP from the new dataset must be virtual represented by a corresponding virtual NP (vNP) through specific software packages, as the GNPrep program, which assemble the NP atoms as a sphere core based on the particle size information for each vNP, following with random placement of associated surface ligands (with ligand density information) on the core surface through NP-surface bond attachment [31].
- (ii) Save the vNP as Protein Data Bank (PDB) files for all NPs.
- (iii) Generation of 100 vNPs for each NP in all datasets to prevent potential instability of the calculation of the descriptor due to the operation on the item 2.(a).(ii).
- (b) Nanodescriptor generation
 - (i) Categorical classification of the vNPs according to atomic element (e.g., C for carbon or H for halogens).
 - (ii) Identification of four nearest neighboring atoms, through Delaunay tessellation, which can form a vNP structure tetrahedron (e.g., cutoff distance, excluding a tetrahedron when the distance between two atoms is higher than such value, can be employed).
 - (iii) Tetrahedrons of identical compositions are counted as similar descriptor, and the value of each nanodescriptor for each vNP is established as the summation of the electronegativity values of all atoms at the vertices of the tetrahedron multiplied by its occurrences in the vNP.
 - (iv) As a measure to assure descriptor value consistency obtained for each NP, each nanodescriptor value is the average of the results for the 100 vNPs constructed in item 2.(a).(iii), followed by normalization, which ranges from 0 to 1.

3. QSAR-based model to test nanodescriptors

- (a) Development of QSAR or QSAR-based models, according to usual protocols, taking into consideration the type of bioactivity and physicochemical properties at the foundation of the nanodescriptors
- (b) Refinement of the QSAR or QSAR-based models through machine learning techniques, such as random forest

3 Methods

When selecting the methods to include regarding protocols of *QSAR-based + ML* models, two major categories were felt as

pertinent. One consists the recent developments of classical QSARs joining forces with machine learning tools, which refine the predictive power of the *in silico* techniques. The second, very recent and promising, consists *QSAR perturbation-based + ML* models, which not only refine the predictive capacity of the QSAR technique but propose a solution to the one-to-one usual limitation of QSAR models, i.e., one bio-target versus one chemical compound, through an interesting mathematical approach.

This section is divided in three subsections, Subheading 3.1, which briefly introduces the concepts of the most common machine learning techniques; Subheading 3.2, approaching the first major category mentioned above; and Subheading 3.3, where the multi bio-target in different experimental condition *QSAR + ML* approach is detailed.

3.1 Machine Learning Tools

One of the most quoted phrases when addressing the subject of machine learning is the operational definition of these techniques, “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E ,” i.e., a machine learning technique is able to adjust itself to better performance through interaction with the data [5]. Within the QSAR-based model contexts, machine learning tools are usually employed as foundation for regression or classification mathematical models, being the most common of these machine learning tools artificial neural networks (ANN), deep neural networks (DNN), or others [32, 33] (Table 1).

3.2 Classical QSAR + ML Models

As mentioned in the *Introduction*, the quantified decrease of QSAR publications might be overcome as novel approaches to QSAR are made by employing ML [4]. In fact, just in the last 2 years, interesting results on the context of environmental risk assessment and prediction of ecotoxicity have been published [34–36].

A published work in 2017 approached the investigation regarding risk assessment of ionic liquids, which can potentially present a danger to the environment, namely, aquatic organisms such as green algae, *Vibrio fischeri*, and fish, through such a hybrid *in silico* methodology of *QSAR + ML* model [34, 37–44]. Interestingly, the σ -profile descriptors were the focus of the ML techniques. These descriptors translated the structural information of the studied ionic liquids into numerical variables, using their anionic and cationic σ -profile as source data. Previously, the ions present in the ionic liquids had already been structurally optimized through the principles of density functional theory (DFT). It was the optimized geometry file of every single ion that was transformed in the σ -profile function, as the latter represents a surface composition function. Five of such descriptors were used for developing a linear model, while the k -fold cross validation method explored the

Table 1
Summary of the concepts of the most common machine learning tools

Machine learning tool	Concept
Multiple linear regression (MLR)	Models linear relationship between independent, or explanatory, variables to predict the outcome of a response variable
Partial least squares (PLS)	Extension of MLR, determines the regression coefficients from the several independent variables and the intercept of explanatory and response data
Linear discriminant analysis (LDA)	Reduces dimensionality through removal of redundant and depending features, i.e., transforms higher-dimensional space features into lower-dimensional space features
Support vector machines (SVM)	Discriminative classifier technique which separates labeled training data or supervised data, outputting a separating hyperplane to establish new categories
Artificial neural networks (ANN)	Computational model inspired by the biological neural network structure and functioning, which are changed as information (input and output) flows through the network
Deep neural networks (DPN)	A multilayered ANN
Genetic algorithms (GA)	Search algorithm with capacity to adapt, exploiting information to focus the search on the subregion of better performance from the initial search space
Random forest (RF)	Generates decision trees, with great visual simplicity, and searches for the best feature among a random subset of features
Bayesian modeling (BM)	Also designated belief networks, is a segmentation technique which makes decision on how to interpret probabilistic evidence in two classes, supporting a hypothesis or rejecting it, with optimal solutions presenting the highest expected utility

performance and stability of the model. Further, the selected descriptors were then used to develop a nonlinear model, through the resourcing to multilayer perceptron (MLP), which further enhanced the accuracy of the model. The choice befell on three-layered feed-forward networks with back-propagation training function, using the Neural Network Toolbox of MATLAB. As the weights controlling the back connections modulate the output of the neuron before transmitting information to the following layer, they were optimized to enable a more accurate prediction. This calculation process was executed through activation and transfer functions in the hidden and output layers. The five employed descriptors permitted the MLR-based model to present a correlation coefficient of 0.906 regarding the toxicity-structure relationship, while the model also indicated, in concurrence with experimental evidence, that the increase of ionic liquid toxicity is proportionally related with the length of the alkyl chain [34].

In 2018, a published investigation regarding the development of a model for predicting aquatic toxicity opted for a *QSAR + ML* model where the chosen ML technique was the classification wrapper feature elimination approach, which is a support vector machine pairwise recursive feature extraction (RFE) method [35]. The authors employed such strategy to find the most relevant pairs of molecular feature within that context while intending to derive from them chemical frontiers between the chemical properties of toxic and nontoxic organic chemicals, in order to provide a framework for the design of less toxic chemicals. The descriptors were constructed taking into consideration 36 physicochemical properties, such as hydrogen bond acceptors or solvent accessible surface area. Furthermore, they also took into consideration previous published calculations for the highest occupied molecular orbital HOMO and the lowest unoccupied molecular orbital LUMO by semiempirical AM1 methodologies. To validate the model, the choice fell on the fivefold cross validation, which was employed in the feature selection, in the model parameter optimization, and in establishing the hyperplanes from the support vector classification. To validate the model, the chosen methods were external validation with new data as validation sets and testing versus a different machine learning technique, a decision tree model. The authors concluded that, within the aquatic environment context, nontoxic chemicals presented aqueous solubility $QPlogS > 1$ and a $LUMO > 1$ or $QPLogo.w > 1$ and $\Delta LUMO - HOMO$, i.e., the energetic difference between LUMO and HOMO, greater than 9, and such values could be used as thresholds for assessing aqueous ecotoxicity of chemical compounds.

A work published in 2018 proposed a solution to the repetitive nature of the QSAR/QSPR life cycle nature, by approaching as answer the automation of several of the typical in silico technique stages [36]. As such, they presented a workflow where tasks, such data curation, data set characteristic evaluation, variable selection, and validation, were fully automated, testing the workflow versus 30 different problems. At its most optimal, the methodology removed a percentage between 62 and 99 of redundant data and almost less of a fifth of prediction error. For feature selection, the authors implemented a random forest (RF) voting procedure where each importance score of the variables is calculated by several available importance's measures, in a hybrid approach that allows these measures to be input data for any machine learning algorithm in stepwise predictive model development.

3.3 QSAR/QSTR Perturbation Theory Models and Machine Learning

Within the search to develop in silico models able to present robust results for NP risk assessment in ecosystems, the question for overcoming the economic and celerity impairments of experimental techniques is not the solo objective to be met. In fact, the reason for the mentioned economic and time-consuming disadvantages of

experimental techniques, with an already established role in assessing the toxicity of NPs, is positively correlated with the impressive amount of experimental assays necessary to cover the reasonable spectrum of biological behaviors [45].

Notwithstanding, this diversity of bioindicators and NP characteristics is also a challenge for *in silico* models. Classical QSAR/QSTR models base their prediction in a one-to-one strategy, i.e., assessing the toxicity of the NPs against a single bioindicator. Further, the main descriptors tend to be chemical composition and size, while other important properties and factors are not integrated in the techniques.

Therefore, the need for a model that could comprise such complexity in a unified manner has prompted researchers to include perturbation theory in QSAR/QSTR techniques [46–50]. Perturbation theory is a designation for general mathematical methods concerned in determining a quasi-solution for any problem which is mathematically challenged when it comes to determine the exact solution. The quasi-solution, obtained by adding small terms to the original problem and constructing a formal power series, designated perturbation series, which includes the exact solution and the deviation terms due to the approximation. Such a strategy also allows the interface of the problem with numerical methods and algorithms to be far more feasible [51].

In the following subsections, the several aspects of how such a strategy approaches the challenges of risk assessment of NPs in ecosystems are discussed and explored.

3.3.1 Mathematics Sustaining QSAR Perturbation Models

When attempting to develop a QSAR perturbation model, there are two main aspects to consider, the first being the classical QSAR principle of one-to-one biological effect versus reference chemical should be replaced by the new QSAR perturbation paradigm of including more than one chemical as reference, while the second focuses on descriptor sensitivity, i.e., generating new descriptor dynamics regarding the chemical compositions of NPs and the experimental conditions [47–50].

As the experimental conditions vary, an ontological form presenting dependence of several experimental factors should be developed, avoiding the static nature of classical one-to-one QSAR principles. The ontological form, c_j , represents a unique experimental condition sensitive to the variability of the conditions where NPs interface bioindicators. The unique ontological experimental condition departs from experimental assays of NPs for toxicity against an array of biological entities, such as bacteria, algae, fish, and others, comprising the measures of toxicity (m_e) against the bioindicator (b_t), i.e., m_e vs. b_t ; the nanoparticle shape labels, (n_s), measured under specific conditions (d_m), i.e., (n_s vs. d_m); and the assay times during which the biological entities were interfacing the NPs (t_a) [47–50]. The ontological experimental condition is thus

Table 2**Parameters for constructing unique ontological experimental condition c_j under which an NP is tested**

Parameter	Concept	Interface
m_e	Measures of toxicity ^a	m_e vs. b_t
b_t	Biological targets	
n_s	Shape labels	n_s vs. d_m
d_m	Measurement conditions	
t_a	Assay times ^b	

^aPossible toxicity measures: EC₅₀, IC₅₀, LC₅₀, TC₅₀, etc.^bAssay times during which the biological targets have been exposed to NPs**Table 3****Measure of toxicity with mandatory cutoff values in QSAR perturbation models**

Measure of ecotoxic effect (units)	Concept
IC ₅₀ (μM)	Concentration of the nanoparticle that inhibits the root elongation of the living system (plant) at 50%
EC ₅₀ (μM)	Effective concentration of the nanoparticle that inhibits at 50% the growth of the living system
CC ₅₀ (μM)	Cytotoxic concentration of the nanoparticle leading to 50% reduction in cell viability assays
TC ₅₀ (μM)	Concentration that causes toxic effects in 50% of the living system
LC ₅₀ (μM)	Lethal concentration that causes mortality in 50% of the living system

defined as $c_j = > (m_e, b_t, n_s, d_m, t_a)$. Summary of the parameters and corresponding concepts for constructing the unique ontological experimental condition is presented on Table 2. It should also be referred that the quantitative values and curtailing proportions, i.e., if one uses 1 out 5 measures of toxic effects against 1 of 50 biological targets when considering m_e vs. b_t , for example, are gathered from published experimental data for NPs/cases.

Once c_j is defined, another function needs to be defined and assembled. Discriminating each of the NPs/cases into two classes, or groups, designated “positive” and “negative,” an experimental condition regarding the toxic effect of a given nanoparticle, NP_j, can be constructed from the ontological unique experimental condition and its value allocated to the value one if the situation corresponded to a nontoxic situation and minus one otherwise. This new function, Tox_j(c_j), thus makes a correspondence of high toxicity, i.e., low values of measures of ecotoxicity—as “toxic” nanoparticles need small concentrations to inhibit or cause

mortality to living organisms—to the ecotoxic group, or class, “negative” [$\text{Tox}_i(c_j) = -1$], and low toxicity, i.e., high values of measures of ecotoxicity, to the non-ecotoxic class, “positive” [$\text{Tox}_i(c_j) = 1$]. This new function belongs to a categorical variable, based on the ontological experimental condition of published data, and therefore, to effect the categorization, the measures of biological effects need to be assigned to all the cases taking into account cutoff values of ecotoxicity regarding m_c . These values are rigorously established after analysis of the published data. Summary of the diverse measures of biological effects are presented on Table 3.

After defining an ontological unique experimental condition, c_j , which takes into account several experimental variables from experimental data published in toxicity assays research, and a categorical variable, $\text{Tox}_i(c_j)$, dependent on those very published results to discriminate NP/cases in toxic and nontoxic, the concern should be the sensitivity of the QSAR perturbation model to the changes in both the NPs and distinct sets at the basis of the quantification of c_j .

To assure such thing, one starts by choosing a small integer number of molecular descriptors, between three and five, for example, which can be obtained through standard physicochemical properties, such as molar volume, or available experimental data, as the NP size. In the eventuality of the NP being formed by more than one element, the physicochemical properties should be normalized, i.e., for each physicochemical property, the normalized physicochemical properties are the division quotient of the sum of the properties of all atoms forming the molecule by the total number of atoms.

After reaching this stage, new descriptors must be generated by applying the moving average approach (MAA) in order to create a new set of descriptors, which incorporate both the molecular structure and the ontological unique experimental condition, c_j , therefore being able to discriminate the ecotoxicological effect of a given NP as the different elements composing c_j are varied [52–56]. Defining D_i as the original descriptor/property, $D_i(c_j)_{\text{aver}}$ as the same experimental condition c_j calculated as the average of all D_i values for NPs/cases in a subset of n_j NPs/cases considered as non-ecotoxic cases as defined above for the categorical variable in the same element of ontological c_j , then the following equation indicates how the moving average approach descriptor, $\Delta D_i(c_j)$, is generated:

$$\Delta D_i(c_j) = \Delta D_i - D_i(c_j)_{\text{aver}} \quad (1)$$

Notwithstanding the reasoning followed in developing this new descriptor function, it must be adverted, it does not include the descriptor regarding coating agents, s_c , as coating agents have their own chemical structures. Thus, a new function is developed to

emulate the characteristics of coating agents when they are present, or absent, on NPs. If one considers PP to represent a given physicochemical property, such as hydrophobicity, NMU as the number of molecular units of a coating agent, and $\mu_k(\text{PP})$ as the spectral moment of order k from the bond adjacency matrix and $G\mu_k(\text{PP})$ represents the descriptor of the coating agent structure, a general spectral moment of order k of the bond adjacency matrix, with null $G\mu_k(\text{PP})$ for uncoated NPs, then the relation between these variables can be described as in Eq. 2.

$$G\mu_k(\text{PP}) = \mu_k(\text{PP}) \times (\text{NMU})^{1/2} \quad (2)$$

After enabling the sensitivity of the model to the varying specifics of a context, i.e., the particulars of an ecosystem and the pharmacological identity of an NP, the second aspect to be considered is how the model may overcome the classical limitation of a QSAR/QSTR technique of one-to-one structure, or chemical composition, versus toxicological magnitude. Aiming at being able to incorporate a multireference of chemicals versus toxicological magnitude, thus adding the complexity of NPs and their inner synergetic consequences, the perturbation series is constructed [46–50]. Considering the combination of the data set original cases as case-case pairs, one can define one of the cases as the reference state, ref, and the other as the new or output case, new, which represents the prediction. With this concept, not only a new prediction can be made out of the total of the other cases, but also it can participate as a reference state in the model. Interfacing this concept with Eqs. 1 and 2, new equations, Eqs. 3 and 4, indicating the differences between the NPs cases can be defined, and their output allows the identification of possible deviations, or perturbations, within the pairs, as the differences are dependent on their chemical composition and the ontological unique experimental condition.

$$\Delta \text{DD}_i(c_j) = \text{DD}_i(c_j)_{\text{new}} - \text{DD}_i(c_j)_{\text{ref}} \quad (3)$$

$$\Delta \text{DG}\mu_k(\text{PP}) = G\mu_k(\text{PP})_{\text{new}} - G\mu_k(\text{PP})_{\text{ref}} \quad (4)$$

In Eqs. 3 and 4, the terms $\Delta \text{DG}\mu_k(\text{PP})$ stand for the moving average descriptor between the new and reference state coating agent structure descriptors, $G\mu_k(\text{PP})_{\text{new}}$ and $G\mu_k(\text{PP})_{\text{ref}}$, respectively, while $\text{DD}_i(c_j)_{\text{new}}$ and $\text{DD}_i(c_j)_{\text{ref}}$ are similar in definition to the terms of Eq. 1 but now with adjusted symbology to represent the new and reference case. With all these elements, it is now possible to define the categorical variable, $\text{Tox}_i(c_j)$, as a general expression able to predict ecotoxicity of a given NP under an array of experimental conditions, with one NP always used as reference, or initial, state and the other as the new, or output, case, which is the prediction, as follows:

$$\text{Tox}_i(c_j)_{\text{new}} = f [\text{Tox}_i(c_j)_{\text{ref}}, \Delta\text{DD}_i(c_j), \Delta\text{DG}\mu_k(\text{PP})] \quad (5)$$

As Eq. 5 clearly illustrates, the toxicity of a given NP in the new state of the categorical variable $\text{Tox}_i(c_j)$ depends on the toxicity of the same NP on the reference state of the categorical variable $\text{Tox}_i(c_j)$ and on the perturbation terms $\Delta\text{DD}_i(c_j)$ and $\Delta\text{DG}\mu_k(\text{PP})$, which are sensitive to the variations of chemical characteristics and coating agents of the NPs.

Further, the $\text{Tox}_i(c_j)_{\text{new}}$ and $\text{Tox}_i(c_j)_{\text{ref}}$ also have a categorical variable nature, with similar meaning to the original definition of $\text{Tox}_i(c_j)$. Finally, as the new function f is a nonlinear function, machine learning presents itself as the most suited methodology to determine it. The particulars of such determination are described on the following subsection.

3.3.2 Protocols for QSAR Perturbation + ML Models

The protocols for QSAR/QSTR perturbation model development and implementation, described in this subsection, are illustrated on Fig. 2.

1. Selection of data regarding the nanoparticles. It comprises usual an input set of initial data gathered from both the literature and the experimental results (*vide* section Subheading 2). In this phase, the initial descriptors regarding the NPs are chosen, being common examples the molar volume, V ; the electronegativity, E ; the polarizability, P ; and the size of the NP, L , or coating agent types.
2. Selection of experimental conditions to construct the ontological unique experimental function, c_j . This comprises, as described on the previous subsection, different measures of toxicity, such as IC_{50} ; the presence of coating agents; endpoint, i.e., type of bio-target and complexity of the bio-target; and assay exposure time.
3. Construction of the ontological unique experimental condition, c_j , from the selected experimental conditions (*vide* Subsection 3.3.1).
4. Clustering of the original NPs in two classes, toxic and non-toxic, through the $\text{Tox}(c_j)$ categorical variable (*vide* Subsection 3.3.1).
5. Calculation of descriptors such as spectral moments of adjacency matrix and the categorical function $G\mu_k(\text{PP})$ regarding the coating agents of the NPs (*vide* Subsection 3.3.1).
6. Determination of the new set of descriptors by applying the moving average approach (MAA), according to Eq. 1.
7. Calculation of the descriptors $\Delta\text{DD}_i(c_j)$ and $\Delta\text{DG}\mu_k(\text{PP})$ as perturbation terms to be used on the general expression of the QSAR/QSTR model as depicted on Eq. 5.

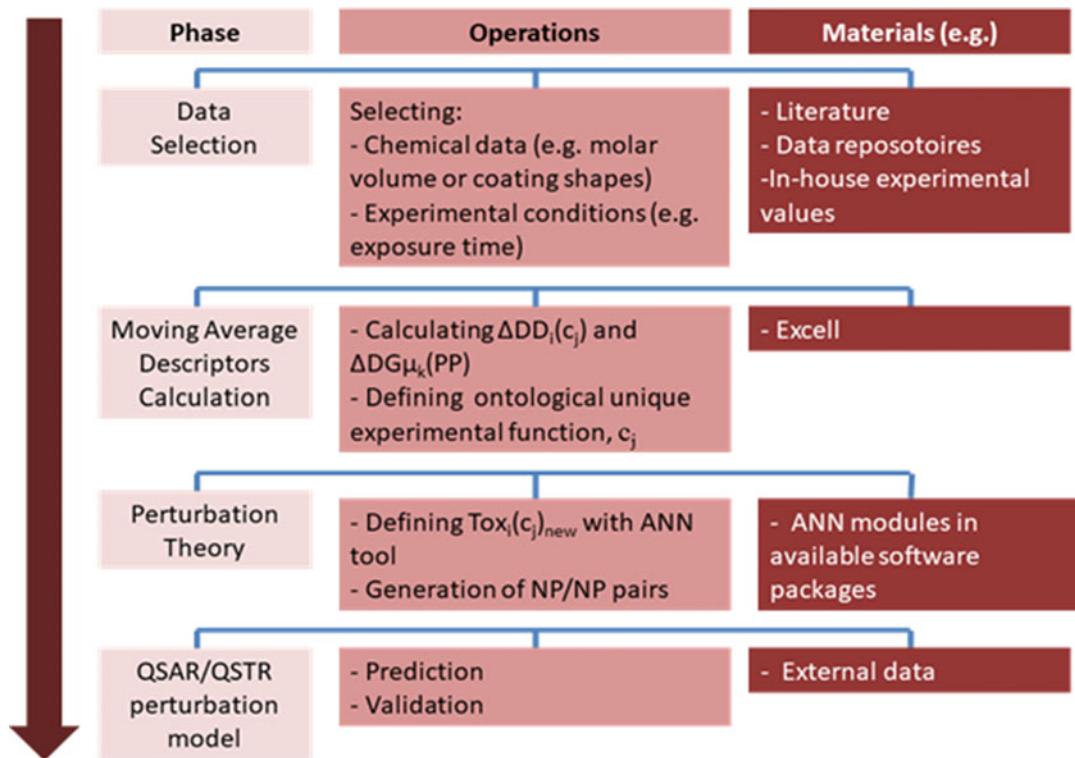


Fig. 2 Scheme for developing QSAR perturbation-based models

8. Determine which machine learning (ML) methodology is adequate for the chosen scenario of dataset. For example, if the choice should be artificial neurons network (ANN) as the proper nonlinear analysis method, several different ANN architectures and topologies, such as linear neural network (LNN), radial basis function (RBF), multilayer perceptron (MLP), or probabilistic neural network (PNN), must be assessed as to prevent underfitting and/or overfitting problems.
9. Conducting a sensitivity analysis, if possible, aiming to identify through the chosen ML method which are the most significant descriptors to be included in the QSAR/QSTR perturbation model.
10. Resorting to the adequate machine learning methodology in order to determine the parameters of the function $f[\text{Tox}_i(c_j)_{\text{ref}}, \Delta\text{DD}_i(c_j), \Delta\text{DG}\mu_k(\text{PP})]$.
11. Dividing the QSAR/QSTR perturbation model based on ML into two sets, a training set, for model development and indicating statistical quality, and a test set, for internal validation and predictive power.

Table 4
Descriptors produced by QSAR perturbation model

Descriptor	Type	Concept
Tox _i (c_j) _{ref}	Dummy classifier	Describes the toxic effect of the NP used in the reference state
DDV(m_e)	Perturbation term	Indicates the change of the molar volume between the NPs used in the new and reference states, being dependent on the measures of the toxic effects
DDL(m_e)	Perturbation term	Accounts for the variation of the size between the NPs used in the new and reference states, being dependent on the measures of the toxic effects
DDμ ₁ (ATO) b_t	Perturbation term	Describes the difference of the spectral moment of order 1 (weighted by the atomic weight) between the NPs used in the new and reference states, being dependent on the bio-target
DD μ ₃ (POL) n_s	Perturbation term	Characterizes the change of the spectral moment of order 3 (weighted by the polarizability) between the NPs used in the new and reference states, being dependent on the shapes of the NP
DDE(d_m)	Perturbation term	Accounts for the variation of the electronegativity between the NPs used in the new and reference states, being dependent on the conditions under which the sizes of the NPs were measured
DDμ ₃ (VAN) t_a	Perturbation term	Describes the difference of the spectral moment of order 3 (weighted by the atomic van der Waals radius) between the NPs used in the new and reference states, being dependent on the exposure times
DDμ ₂ (ATO) t_a	Perturbation term	Characterizes the change of the spectral moment of order 2 (weighted by the atomic weight) between the NPs used in the new and reference states, being dependent on the exposure times
DGμ ₂ (Hyd) s_c	Perturbation general spectral moment of order 2 (weighted by the hydrophobicity) term	Accounts for the difference between the chemical structures of the coating agents used in the new and reference states
DGμ ₅ (PSA) s_c	Perturbation general spectral moment of order 5 (weighted by the polar surface area) term	Characterizes the difference between the chemical structures of the coating agents used in the new and reference states

12. Confirmation of the predictive power of the QSAR/QSTR perturbation model through blind external validation, i.e., new external data.

3.3.3 QSAR Perturbation Models and Ecotoxicity Assessment

Recent results regarding risk assessment in ecosystems resorting to QSAR perturbation + ML models have displayed excellent capacity of the in silico technique to model toxicological profiles of NPs in multi-experimental conditions [50]. The model presented ten descriptors, displayed on Table 4, obtained after deriving the original and MAA descriptors from an ANN analysis.

These descriptors were calculated resourcing to the data analysis method ANN and, in particular, employing the specific module for ANN designated Intelligent Problem Solver in the STATISTICA® package [57]. As mentioned on Subsection 3.3.2, there was a preliminary phase where the most adequate ANN architecture was investigated. When considering neural networks, there are three layers, the hidden, input, and output layers, and the number of neurons on the hidden layer should be between the number of neurons of the other two layers. However, that is not always the situation, as the degree of complexity of the problem at hand, or number of training sets, is not taken into account for this empirical rule. In fact, the only seemingly assumption to be made is that the number of neurons on the hidden layer should be as low as possible.

In the published work, the best model was found to present an ANN profile of multilayer perceptron (MLP) 10:10-44-1:1. The explanation for the expression 10:10-44-1:1 is that 10 descriptor variables (**10:10-44-1:1**) generated 10 neurons on the first layer (10:**10-44-1:1**); the second, and hidden, layer had 44 neurons (10:10-**44-1:1**); and the output layer had only 1 neuron (10:10-44-**1:1**), which predicted the response variable, $\text{Tox}_i(c_j)_{\text{new}}$ (10:10-44-1:1). The MLP topology was found through careful analysis of 50 ANN runs.

To further guarantee the correct selection of descriptors, sensitivity test through model misclassification rates or sum of square residuals was made by the ANN module. Within internal correlation, several statistical indices were computed, namely, the overall percentage of correct classification, i.e., model accuracy; the percentage of correct classification for nontoxic and toxic cases, i.e., the sensitivity (SENS) and specificity (SPEC), respectively; the Matthews correlation coefficient (MCC); and the areas under the receiver operating characteristic (ROC) [58, 59]. The validation yielded for a number of 54,371 NP/NP cases, a computational accuracy circa 97%, while the minimum percentage for correct classification with new external data was as follows: for several sizes of silver-based NPs versus RAW 264.7 mouse cells, and CC₅₀ as measure of toxicity, values higher than 85%; for nickel ferrite NPs versus A549 human cells, and CC₅₀ as measure of

toxicity, values higher than 95%; for iron(II) oxide NPs versus *D. rerio* (zebrafish) and LC₅₀ as measure of toxicity, values higher than 95%; for silver-based NPs, of 34 nm, versus microalgae, and EC₅₀ as measure of toxicity, values higher than 91%; and, finally, for platinum-based NPs, of 51 nm, and EC₅₀ as measure of toxicity, values higher than 72%.

4 Conclusions

The recent years have presented challenges to the established classical QSAR model and techniques. As the *in silico* method reached a plateau of maturity, also it faced the challenge of its limitations. Several paths for overcoming this state of affairs are presented and reviewed in this chapter, alongside with the protocols of selected aspects of these novel techniques. One path consist on the conjoining forces between the QSAR models with machine learning methods, especially on the steps regarding descriptor construction and selection or validation of the model. The other consist in facing the classical limitation of QSAR models of one-to-one approach with new mathematical techniques and establishing QSAR-based models able to predict multiconditions versus multi bio-targets.

Acknowledgments

This work was supported by UID/QUI/50006/2019, contract IF CEECIND/03631/2017, and project PTDC/QUI-QIN/30649/2017 with funding from FCT/MCTES through national funds.

References

1. Crum-Brown A, Fraser TR (1865) The connection of chemical constitution and physiological action. *Trans R Soc Edinb* 25 (1968–1969):257
2. Crum-Brown A, Fraser TR (1868) On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *J Anat Physiol* 2(2):224–242
3. Hansch C (1969) A quantitative approach to biochemical structure-activity relationships. *Acc Chem Res* 2:232–239
4. Devinyak OT, Lesyk RB (2016) 5-year trends in QSAR and its machine learning methods. *Curr Comput Aided Drug Des* 12(4):265–271
5. Mitchell TM (1997) Machine learning, vol 45. McGraw Hill, Ridge, pp 870–877
6. Kim BJ, Ko Y, Cho JH (2013) Organic field-effect transistor memory devices using discrete ferritin nanoparticle-based gate dielectrics. *Small* 9(22):3784–3791
7. Liz-Marzán LM, Kamat PV (2004) Nanoscale materials. Kluwer Academic Publishers, New York
8. Chen CY, Retamal JR, Wu IW et al (2012) Probing surface band bending of surface-engineered metal oxide nanowires. *ACS Nano* 6(11):9366–9372
9. Biffis A, Králik M (2001) Catalysis by metal nanoparticles supported on functional organic polymers. *J Mol Catal A* 177(1):113–138
10. Chan NY, Zhao M, Wang N et al (2013) Palladium nanoparticle enhanced giant

- photoresponse at LaAlO/SrTiO two-dimensional electron gas heterostructures. *ACS Nano* 7(10):8673–8679
11. Lu P, Campbell CT, Xia Y (2013) A sinter-resistant catalytic system fabricated by maneuvering the selectivity of SiO₂ deposition onto TiO₂ surface versus Pt nanoparticle surface. *Nano Lett* 13(10):4957–4962
 12. Yang B, Zhao C, Xiao M et al (2013) Loading metal nanostructures on cotton fabrics as recyclable catalysts. *Small* 9(7):1003–1007
 13. Moseler M, Walter M, Yoon B et al (2012) Oxidation state and symmetry of magnesia-supported Pd₁₃O(x) nanocatalysts influence activation barriers of CO oxidation. *J Am Chem Soc* 134(18):7690–7699
 14. Corchero JL, Villaverde A (2009) Biomedical applications of distally controlled magnetic nanoparticles. *Trends Biotechnol* 27(8):468–476
 15. Zhang Z, Wang J, Chen C (2013) Near-infrared light-mediated nanoplatforms for cancer thermo-chemotherapy and optical imaging. *Adv Mater* 25(28):3869–3880
 16. Schoen DT, Coenen T, Garcia de Abajo FJ et al (2013) The planar parabolic optical antenna. *Nano Lett* 13(1):188–193
 17. Liao L, Liu J, Dreaden EC et al (2014) A convergent synthetic platform for single-nanoparticle combination cancer therapy: ratiometric loading and controlled release of cisplatin, doxorubicin, and camptothecin. *J Am Chem Soc* 136(16):5896–5899
 18. Lu CH, Willner B, Willner I (2013) DNA nanotechnology: from sensing and DNA machines to drug-delivery systems. *ACS Nano* 7(10):8320–8332
 19. Brigger I, Dubernet C, Couvreur P (2002) Nanoparticles in cancer therapy and diagnosis. *Adv Drug Deliv Rev* 54(5):631–651
 20. Todeschini R, Consonni V (2000) Handbook of molecular descriptors, vol 11. Wiley VCH, Weinheim
 21. Halder AK, Moura AS, Cordeiro MNDS (2018) Advanced chemometric modeling approaches for the design of multitarget drugs against neurodegenerative diseases. In: Roy K (ed) Multi-target drug design using chem-bioinformatic approaches. Methods in pharmacology and toxicology. Humana Press, New York
 22. Hsu DD, Chemicool Periodic Table, <http://www.chemicool.com/>, Accessed April 4, 2019
 23. Simulations Plus, Inc [US]. <http://www.simulations-plus.com/>. Accessed 10 Apr 2019
 24. Rana S, Kalaichelvan PT (2013) Ecotoxicity of nanoparticles. *ISRN Toxicology* 2013:1. <https://doi.org/10.1155/2013/574648>
 25. Ostiguy C, Lapointe G, Ménard L, et al. (2006) Les nanoparticules: Etat des connaissances sur les risques en santé et sécurité du travail, Rapport IRSST Soumis, IRSST, Montréal
 26. Gimbert LJ, Hamon RE, Casey PS, Worsfold PJ (2007) Partitioning and stability of engineered ZnO nanoparticles in soil suspensions using flow field-flow fractionation. *Environ Chem* 4(1):8–10
 27. Born PJA (2003) Toxicology of ultrafine particles. Rapport d'un atelier du BIA sur ultrafine aerosols at workplaces. BIA Report, Berufsgenossenschaftliches Institut für Arbeitsschutz
 28. Guzman KAD, Finnegan MP, Banfield JF (2006) Influence of surface potential on aggregation and transport of titania nanoparticles. *Environ Sci Technol* 40(24):7688–7693
 29. Yan X, Sedykh A, Wang W et al (2019) In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* 11:8352–8362. <https://doi.org/10.1039/C9NR00844F>
 30. Delaunay M (1924) Sur la sphère vide. Congrès International des Mathématiciens, Toronto, Canada pp 695–700
 31. Wang W, Sedykh A, Sun H et al (2017) Predicting nano-bio interactions by integrating nanoparticle libraries and quantitative nanostucture activity relationship modeling. *ACS Nano* 11:12641–12649
 32. Quintero FA, Patel SJ, Muñoz F et al (2012) Review of existing QSAR/QSPR models developed for properties used in hazardous chemicals classification system. *Ind Eng Chem Res* 51(49):16101–16115
 33. Lewis RA, Wood D (2014) Modern 2D QSAR for drug discovery. *Wiley Interdiscip Rev Comput Mol Sci* 4(6):505–522
 34. Ghanem OB, Mutualib MIA, Lévéque J-M et al (2017) Development of QSAR model to predict the ecotoxicity of *Vibrio fischeri* using COSMO-RS descriptors. *Chemosphere* 170:242–250
 35. Husowitz B, Sanchez-Arias R (2017) A machine learning approach to designing guidelines for acute aquatic toxicity. *J Biom Biostat* 8:385. <https://doi.org/10.4172/2155-6180.1000385>
 36. Kausar S, Falcao AO (2018) An automated framework for QSAR model building. *J Cheminform* 10(1). <https://doi.org/10.1186/s13321-017-0256-5>

37. Kulacki KJ, Lamberti GA (2008) Toxicity of imidazolium ionic liquids to fresh water algae. *Green Chem* 10:104–110
38. Latala A, Nedzi M, Stepnowski P (2009) Toxicity of imidazolium and pyridinium based ionic liquids towards algae, *Chlorella vulgaris*, *Oocystis submarina* (green algae) and *Cyclotella meneghiniana*, *Skeletonema marinoi* (diatoms). *Green Chem* 11:580–588
39. Pretti C, Chiappe C, Baldetti L et al (2009) Acute toxicity of ionic liquids for three freshwater organisms: *pseudokirchneriella subcapitata*, *Daphnia magna* and *Dario rerio*. *Ecotoxicol Environ Saf* 72:1170–1176
40. Costa SP, Justina VD, Bica K et al (2014) Automated evaluation of pharmaceutically active ionic liquids' (eco) toxicity through inhibition of human carboxylesterase and *Vibrio fischeri*. *J Hazard Mater* 265:133–141
41. Viboud S, Papaiconomou N, Cortsei A et al (2012) Correlating the structure and composition of ionic liquids with their toxicity in *Vibrio fischeri*: a systematic study. *J Hazard Mater* 215:40–48
42. Stolte S, Matzke M, Arning J et al (2007) Effects of different head groups and functionalized side chains on the aquatic toxicity of ionic liquids. *Green Chem* 9:1170–1179
43. Radosevic K, Cvjetko M, Kopjar M et al (2013) In vitro cytotoxicity assessment of imidazolium ionic liquids: biological effects in fish channel catfish ovary (CCO) cell line. *Ecotoxicol Environ Saf* 92:112–118
44. Dong M, Zhu S, Wang J et al (2013) Toxic effects of 1-decyl-3-methylimidazolium bromide ionic liquid on the antioxidantenzyme system and DNA in zebrafish (*Danio rerio*) livers. *Chemosphere* 91:1107–1112
45. Holden PA, Nisbet RM, Lenihan HS et al (2013) Ecological nanotoxicology: integrating nanomaterial hazard considerations across the subcellular, population, community, and ecosystems levels. *Acc Chem Res* 46:813–822
46. Gonzalez-Diaz H, Arrasate S, Gomez-SanJuan A et al (2013) General theory for multiple input-output perturbations in complex molecular systems. I. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Curr Top Med Chem* 13:1713–1741
47. Kleandrova VV, Luan F, Gonzalez-Diaz H et al (2014) Computational ecotoxicology: simultaneous prediction of Ecotoxic effects of nanoparticles under different experimental conditions. *Environ Int* 73C:288–294
48. Kleandrova VV, Luan F, Gonzalez-Diaz H et al (2014) Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of Ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ Sci Technol* 48:14686–14694
49. Luan F, Kleandrova VV, Gonzalez-Diaz H et al (2014) Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* 6:10623–10630
50. Concu R, Kleandrova VV, Speck-Planche A et al (2017) Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology* 11(7):891–906
51. Kato T (1995) Perturbation theory in a finite-dimensional space. In: *Perturbation theory for linear operators*, (Reprint of the 1980 edn). Springer, Berlin
52. Concu R, Dea-Ayuela MA, Perez-Montoto LG et al (2009) 3D entropy and moments prediction of enzyme classes and experimental – theoretic study of peptide fingerprints in *Leishmania* parasites. *Biochim Biophys Acta* 1794:1784–1794
53. Concu R, Podda G, Uriarte E et al (2009) Computational chemistry study of 3Dstructure–function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials. *J Comput Chem* 30:1510–1520
54. García A, Espinosa R, Delgado L et al (2011) Acute toxicity of cerium oxide, titanium oxide and iron oxide nanoparticles using standardized tests. *Desalination* 269:136–141
55. Hill T, Lewicki P (2006) Statistics methods and applications. A comprehensive reference for science, industry and data mining. StatSoft, Tulsa
56. Tenorio-Borroto E, Garcia-Mera X, Penuelas-Rivas CG et al (2013) Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. *Curr Top Med Chem* 13:1636–1649
57. Statsoft-Team (2001) Statistica. Data analysis software system. v6.0. Tulsa
58. González-Díaz H, Pérez-Bello A, Cruz-Monteagudo M et al (2007) Chemometrics for QSAR with low sequence homology: mycobacterial promoter sequences recognition with 2DRNA entropies. *Chemom Intell Lab Syst* 85:20–26
59. Hanczar B, Hua J, Sima C et al (2010) Small-sample precision of ROC-related estimates. *Bioinformatics* 26:822–830



Chapter 10

Ecotoxicological QSAR Modeling of Nanomaterials: Methods in 3D-QSARs and Combined Docking Studies for Carbon Nanostructures

Bakhtiyor Rasulev

Abstract

One of the main approaches in cheminformatics, so-called a quantitative structure-activity relationship (QSAR) approach, nowadays plays an important role in lead structure optimization, as well as in prediction of various physicochemical properties, biological activity, and environmental toxicology. One of the recent developments in QSAR approaches for nanostructures is a three-dimensional QSAR. For the last two decades, 3D-QSAR has already been successfully applied to various datasets, especially of enzyme and receptor ligands. The application of 3D-QSAR for nanostructured materials is still at early stage. Often, 3D-QSAR studies are going together with protein-ligand docking studies, and this combination works synergistically, improving the accuracy of prediction. Carbon nanostructures, such as fullerenes, and carbon nanotubes are nanomaterials with specific properties that make them useful in pharmacological applications. In this methodological review, we outline recent advances in development and application of 3D-QSAR and protein-ligand docking approaches in the studies of nanostructured materials, such as fullerenes and carbon nanotubes.

Key words 3D-QSAR, Carbon nanostructure, Nanomaterials, Docking, Toxicity, Biological activity

1 Introduction

One of the main approaches in cheminformatics, so-called quantitative structure-activity relationship (QSAR), is described in a number of publications [1–3]. QSAR methods allow cheminformatics scientists to find correlations and mathematical models for congeneric series of compounds and to predict such properties as affinities of ligands to their binding sites, rate constants, inhibition constants, toxicological effect, electronic properties, steric properties, and so on, based on structural features [1–9]. For example, QSAR approaches have been used for many types of biological activities to describe correlations for series of drugs and drug candidates [2, 10–12].

In case of available crystallographic data on proteins, QSAR models can be developed with the use of the additional information from three-dimensional (3D) structures of these proteins. Such information as interaction with drug candidates can be revealed by applying protein-ligand docking data. However, if there is no data on 3D structure of protein, then the QSAR approach can be applied, and structure-activity models may be developed based on three-dimensional features of investigated molecules [13–18]. The latter approach is named as 3D-QSAR [19–22]. Lately, several other multidimensional approaches were developed, such as 4D-QSAR and others; however, these methods are just extension of a QSAR approach, by considering only a few conformations (including orientations, tautomers, stereoisomers, or protonation states) per molecule and number of concentrations (dosages) per compound [23]. Often, by the term “3D-QSAR,” computational chemists usually assume a QSAR analysis that considers a 3D structure of the compound in a minimal energy conformation, where QSAR model is built based on various 3D fields generated [2, 3]. A first approach of 3D-QSAR was developed by Cramer in 1983, which was the predecessor of 3D approaches called dynamic lattice-oriented molecular modeling system (DYLOMMS) that involves the use of principal component analysis (PCA) to extract vectors from the molecular interaction fields, which are then correlated with biological activities [19]. The same authors later improved this approach, and by combining the two existing techniques, GRID and PLS, they developed a powerful 3D-QSAR methodology, so-called a Comparative Molecular Field Analysis (CoMFA) [21, 22]. Right after that, CoMFA has become a prototype of 3D-QSAR methods [24–26]. Later, a powerful CoMFA approach was implemented in the Sybyl software [27] from Tripos Inc.

It is worth to note that a great and fruitful approach in pharmacological properties prediction is a combination of molecular docking and 3D-QSAR pharmacophore methods [16–18, 28]. Nowadays, molecular docking and 3D-QSAR are two important and powerful approaches in drug discovery process, which are heavily utilized in pharmaceutical companies. Thus, virtual screening using 3D-QSAR approaches followed by docking has become one of the reputable combinations of methods for drug discovery enhancing the efficiency in lead optimization [29, 30]. The main advantage of this combined approach is to focus on specific key interactions in protein-ligand binding to improve drug candidates and ameliorate the selection of active compounds. Thus, it is optimal to use both these methods synergistically in drug design [31–34].

A number of QSAR studies and methods' developments were published covering 3D-QSAR methodology. However, not much attention was given to application of 3D-QSAR and protein-ligand

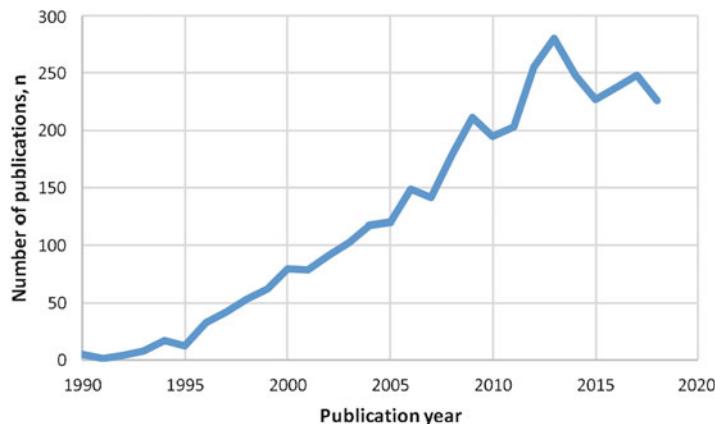


Fig. 1 A plot representing the number of papers that cover 3D-QSAR-related papers per year (keyword—“3D-QSAR”), for a period of 1990–2018. (Source—Scopus)

docking approaches to nanostructured materials. In this review, we briefly list and explain 3D-QSAR-related methods and then discuss recent developments and applications of these 3D-QSARs in the assessment of the properties of biologically active carbon nanostructures. As can be seen from Fig. 1, the number of publications related to 3D-QSAR approach increases, starting from very few publications in the beginning of the 1990s to about 226 publications per year in 2018, with a peak in 2013 (280 publications). It confirms the increasing importance of 3D-QSAR and successful application in drug design.

Another plot (Fig. 2) shows a number of papers that cover nanoparticles/nanostructures’ QSAR/QSPR-related papers per year, for a period of 2003–2018. It can be seen that the number of QSAR studies for nanoparticles/nanomaterials increases dramatically for the last 5–6 years. This confirms one more time the importance of this area of research.

2 Methods for 3D-QSAR: Overview

In this section will be given a detailed information on the methods and a list of 3D-QSAR methods developed. Each of these methods was developed within the last three decades. Only short description is given, to show only an essence of the method, while details can be found in the references cited.

One of the first methods which then called as a 3D-QSAR method is Comparative Molecular Field Analysis method. Comparative Molecular Field Analysis (CoMFA) is the method which correlates the field values of the structure with biological activities. Generally, CoMFA generates an equation correlating the biological

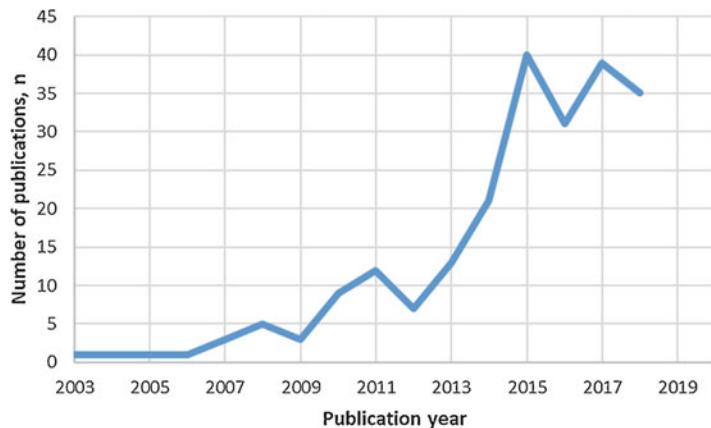


Fig. 2 A plot representing the number of papers that cover nanoparticles/nanostructures QSAR/QSPR-related papers per year (keywords—"QSAR," "QSPR," "nanomaterials," "nanoparticles," "nanostructures"), for a period of 2003–2018. (Source—Scopus)

activity with the contribution of interaction energy fields at every grid point [21]. This method was developed in the 1988 and still can be called as one of the most popular 3D-QSAR methods.

The second one, comparative molecular similarity indices analysis (CoMSIA) method, is another popular method, where the molecular similarity indices are calculated from steric and electrostatic alignment (SEAL) similarity fields and applied as descriptors to encode steric, electrostatic, hydrophobic, and hydrogen bonding properties [4]. CoMSIA is a development of CoMFA method and also gets very popular in drug design.

The next method was designed as an alternative to the original CoMFA approach and known as GRID. It is actually a force field which calculates the interaction energy fields in molecular-field analysis, and it determines the energetically favorable binding sites on molecules of known structure. This method is similar to CoMFA to some extent, and it computes explicit non-bonded (or non-covalent) interactions between a molecule of known 3D structure and a probe (i.e., a chemical group with certain user-defined properties). The probe is located at the sample positions on a lattice throughout and around the macromolecule. Thus, the method offers two distinct advantages—one of them is the use of a 6–4 potential function for calculation of the interaction energies, which is smoother than the 6–12 form of the Lennard-Jones type in CoMFA, and another advantage is the availability of different types of probes [35]. Moreover, the program, in addition to computing the regular steric and electrostatic potentials, also calculates the hydrogen bonding potential using a hydrogen bond donor and acceptor, as well as the hydrophobic potential using a "DRY

probe.” In the next versions, a water probe was included to calculate hydrophobic interactions [24, 36].

Another interesting method is MSA—molecular shape analysis (MSA), which is a ligand-based 3D-QSAR method that attempts to merge conformational analysis with the classical Hansch approach. MSA deals with the quantitative characterization, representation, and manipulation of molecular shape in the construction of a QSAR model [37].

One more method that based on grid technique is HASL—inverse grid-based approach that represents the shapes of the molecules inside an active site as a collection of grid points [38]. The methodology of this approach begins with the intermediate conversion of the Cartesian coordinates (x, y, z) for superposed set of molecules to a 3D grid consisting of the regularly spaced points that are (1) arranged orthogonally to each other, (2) separated by a certain distance termed as the resolution (which determines the number of grid points representing a molecule), and (3) all sprawl within the van der Waals radii of the atoms in the molecules. Here, the resulting set of points is referred to as the molecular lattice which represents the receptor active site map (like in CoMFA). Importantly, the overall lattice dimensions are dependent on the size of the molecules and the resolution chosen.

Another interesting method is GERM, which is an abbreviation of Genetically Evolved Receptor Model. GERM is a method for 3D-QSAR, which can also be used for constructing 3D models of protein-binding sites in the absence of a crystallographically established or homology-modeled structure of the receptor [39]. Similar to many 3D-QSAR datasets, the primary requirement for GERM is a structure-activity set for which a sensible alignment of realistic conformers has been determined. The implemented methodology is the following: it encloses the superimposed set of molecules in a shell of atoms (analogous to the first layer of atoms in the active site) and allocates these atoms with explicit atom types (aliphatic H, polar H, etc. to match the types of atoms found in the investigated proteins).

The next method is GRIND; it uses *grid-independent descriptors* (GRIND), which encodes the spatial distribution of the molecular interaction fields (MIF) of the studied compounds [40]. In another development of GRIND methods, the anchor-GRIND method [41], to compare the MIF distribution of different compounds, the user defines a single common position in the structure of all the compounds in the series, so-called anchor point. Importantly, the anchor point does not provide enough geometrical constraints to align the compounds studied. However, it is used by the method as a common reference point, making it possible to describe the geometry of the MIF fields in a more precise way than GRIND does. Thus, the anchor point can be easily assigned in datasets having some chemical substituents that are well known as

being crucial for the activity. In the anchor-GRIND approach, the R groups are described with two blocks of variables: the anchor-MIF and the MIF-MIF blocks. The first one describes the geometrical distribution of the R MIF relative to the anchor point, while the second one describes the geometrical distribution of the MIF within each R group. The described blocks are obtained by conducting the following steps: (1) every R group is considered as attached to the scaffold, (2) the anchor point is set automatically on an atom of the scaffold, (3) a set of MIFs are calculated with the program GRID [35], and (4), as a final step, the blocks of descriptors are computed from the anchor point and the filtered MIF. Thus, authors also incorporated a molecular shape into the GRIND descriptors [42].

The next interesting 3D-QSAR technique is CoMMA, Comparative Molecular Moment Analysis, which is one of the unique alignment-independent 3D-QSAR methods that involves the computation of molecular similarity descriptors (like in CoMSIA), based on the spatial moments of molecular mass (i.e., shape) and charge distributions up to second-order as well as related quantities [43].

Ortiz et al. [44] in 1995 developed a technique called COMBINE—Comparative Binding Energy Analysis—which was developed to make the use of the structural data from ligand-protein complexes, within a 3D-QSAR methodology. The authors developed this method based on the hypothesis where free energy of binding can be correlated with a subset of energy components calculated from the structures of receptors and ligands in bound and unbound forms.

The next method, Comparative Molecular Surface Analysis (CoMSA), is a non-grid 3D-QSAR method that utilizes the molecular surface to define the regions of the compounds which are required to be compared using the mean electrostatic potentials (MEPs) [45, 46]. In general, the methodology proceeds by subjecting the molecules in the dataset to geometry optimization and assigning them with partial atomic charges.

Another interesting 3D-QSAR technique is CoRIA, Comparative Residue Interaction Analysis, which utilizes descriptors that describe the thermodynamic events involved in ligand binding, to explore both the qualitative and the quantitative features of the ligand-receptor recognition process. In general, the CoRIA methodology is the following: initially it calculates the non-bonded (van der Waals and coulombic) interaction energies between the ligand and the individual active site residues of the receptor that are involved in interaction with the ligand [47, 48]. Then, by employing the genetic algorithm-supported PLS technique (GA-PLS), these energies are then correlated with the biological activities of molecules, together with other physicochemical variables describing the thermodynamics of binding, such as molar refractivity, surface

area, lipophilicity, molecular volume, polar surface area, strain energy, etc.

Robinson with co-authors [49] in 1999 developed the Self-Organizing Molecular-Field Analysis (SOMFA) as a similar method to 3D-QSAR, where they implemented a technique in which initially the mean activity of training set is subtracted from the activity of each molecule to get their mean-centered activity values. The steps in the methodology are the following:

1. A 3D grid around the molecules with values at the grid points signifying the shape or electrostatic potential is generated.
2. The shape or electrostatic potential value at every grid point for each molecule is multiplied by its mean-centered activity
3. The grid values for each molecule are summed up to give the master grids for each property.
4. Finally the so-called SOMFA_{property,i} descriptors from the master grid values are calculated and correlated with the log-transformed molecular activities [49].

The next method, 3D-HoVAIFA, is based on three-dimensional holographic vector of atomic interaction field analysis [50]. The holographic vector for 3D-QSAR methods was developed initially by Zhou et al. in 2007 [50]. In general, the method proceeds from two spatial invariants, namely, atom relative distance and atomic properties on the bases of three common non-bonded (electrostatic, van der Waals, and hydrophobic) interactions that are directly associated with bioactivities. Thus, 3D-HoVAIFA method derives multidimensional vectors to represent molecular steric structural characteristics.

One of the relatively new 3D-QSAR methods, kNN-MFA, was developed and reported in 2006 by Ajmani and co-authors [51]. kNN-MFA is an abbreviation of k-Nearest Neighbor Molecular-Field Analysis. In general, kNN-MFA adopts a k-nearest neighbor principle for generating relationships of molecular fields with the experimentally reported activity. The method utilizes an active analogue principle that lies at the foundation of medicinal chemistry. As a 3D-QSAR method, kNN-MFA requires suitable alignment of a given set of molecules. This step is followed by generation of a common rectangular grid around the molecules. In addition, the steric and electrostatic interaction energies are computed at the lattice points of the grid using a methyl probe of charge +1. Finally, the obtained interaction energy values are considered for relationship generation and utilized as descriptors to decide nearness between molecules.

The next method, a recently introduced continuous molecular-field approach, is CMF [52]. This is a novel approach that involves encapsulating continuous molecular fields into specially constructed kernels. The method is based on the application of

continuous functions for the description of molecular fields instead of finite sets of molecular descriptors (such as interaction energies computed at grid nodes) commonly used for this purpose. The feasibility of using molecular-field kernels in combination with the support vector regression (SVR) machine learning method to build 3D-QSAR models has been demonstrated by the same authors earlier [53]. Another important method is PHASE, a flexible system (engine) [54] for common pharmacophore identification and assessment, 3D-QSAR model development, and 3D database creation and searching (within Schrodinger Suite, Schrodinger, LLC). It includes some subprograms, for example, LigPrep, which attaches hydrogens, converts 2D structures to 3D, generates stereoisomers, and neutralizes charged structures or determines the most probable ionization state at a user-defined pH. It also includes MacroModel conformational search engine to generate a series of 3D structures that sample the thermally accessible conformational states. For purposes of 3D modeling and pharmacophore model development, each ligand structure is represented by a set of points in 3D space, which coincide with various chemical features that may facilitate non-covalent binding between the ligand and its target receptor. PHASE provides six built-in types of pharmacophore features: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic (H), negative ionizable (N), positive ionizable (P), and aromatic ring (R). In addition, users may define up to three custom feature types (x , y , z) to account for characteristics that do not fit clearly into any of the six built-in categories. To construct a 3D-QSAR model, a rectangular grid is defined to encompass the space occupied by the aligned training set molecules. This grid divides space into uniformly sized cubes, typically 1 Å on each side, which are occupied by the atoms or pharmacophore sites that define each molecule.

The latest developed 3D-QSAR method is APF, which was developed in 2008 by Totrov [55] introduced atomic property fields (APF) for 3D-QSAR analysis. The APF concept is introduced as a continuous, multicomponent 3D potential that reflects preferences for various atomic properties at each point in space. In addition, the approach is extended to multiple flexible ligand alignments using an iterative procedure, Self-Consistent Atomic Property Fields by Optimization (SCAPFOld). Thus, the application of atomic property fields and SCAPFOld for virtual ligand screening and 3D-QSAR is tested on published benchmarks. Interestingly, the new method is shown to perform competitively in comparison to the current state-of-the-art methods (CoMFA and CoMSIA). There are studies with comparative analysis of these two methods, PHASE and Catalyst (HypoGen) [56]. Importantly, in 2007 Evans and co-authors [57] provided a comparative study of PHASE and Catalyst methods for their performance in determining 3D-QSARs and concluded that the performance of PHASE is better than or

equal to that of Catalyst HypoGen, with the datasets and parameters used. The authors found that within PHASE, the atom-based grid QSAR model generally performed better than the pharmacophore-based grid, and by using overlays from Catalyst to the built PHASE grid QSAR models, they found evidence that better performance of PHASE on these datasets was due to the use of the grid technique.

Next will be discussed an application of 3D-QSAR methods to study carbon nanostructures, fullerenes, and carbon nanotubes.

3 3D-QSARs and Combined Docking Studies of Carbon Nanostructured Materials

Nanomaterials are getting increasing attention due to their peculiar properties and wide application in various industries, including medicine and pharmaceuticals. At the same time, theoretical modeling of physicochemical and biological activity of these species is still at the initial step. While the prediction of properties and activities of “classical” substances is well-developed with the use of QSAR and 3D-QSAR methods, the application of QSAR for the nanomaterials is a very new and complicated task, since “nonclassical” structure of nanomaterials is not easy to investigate. Here, will be shown a few very first applications of the 3D-QSAR and docking methods for carbon nanostructured materials, which can be useful in predicting various properties and activities of these materials.

Since carbon nanostructures are chemical systems that mainly consist of carbon atoms, the methods of QSAR and molecular docking can be still applicable in this case. In this regard, one of the first 3D-QSAR studies for nanostructured materials was provided by Durdagi and co-authors [58], where the authors have investigated novel fullerene analogues as potential HIV PR inhibitors. It was the first work where authors analyzed nanostructured compounds by combination of 3D-QSAR and protein-ligand docking. In addition, the authors conducted molecular dynamics (MD) simulations of ligand-free and the inhibitor bound HIV-1 PR systems to provide a proper input structure of HIV-1 PR for further docking simulations. Thus, authors developed five different combinations of 3D-QSAR/CoMSIA models based on stereoelectronic fields, which were obtained from the set of biologically evaluated and computationally designed fullerene derivatives (training = 43, test = 6 compounds). The best 3D-QSAR/CoMSIA model yielded a cross-validated r^2 value of 0.739 and a non-cross-validated r^2 value of 0.993. In conclusions, the authors stated that the derived model indicated the importance of steric (42.6%), electrostatic (12.7%), H-bond donor (16.7%), and H-bond acceptor (28.0%) contributions (Fig. 3). Moreover, the derived contour plots together with applied de novo drug design were then used as pilot models for proposing novel analogues with

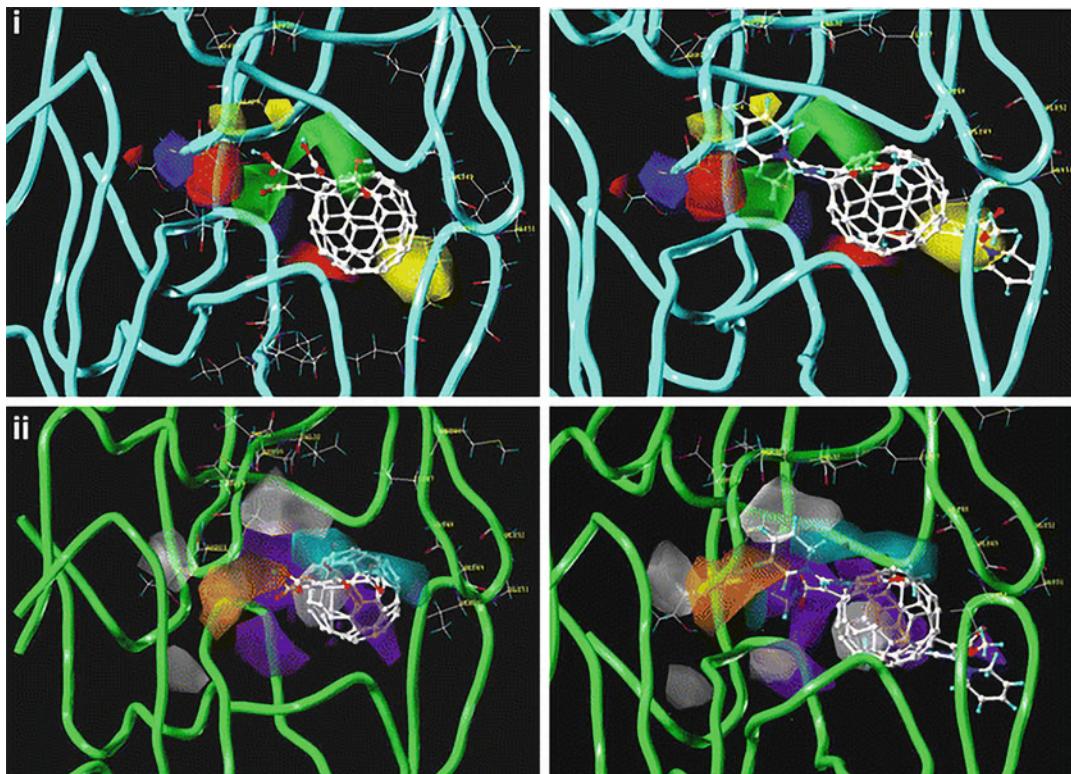


Fig. 3 (i) CoMSIA steric/electrostatic contour maps of template compound **23** (template compound has best binding affinity in the training set, *left* on the figure) and compound **36** (has worst binding affinity in the training set, *right* on the figure). Sterically favored areas are shown in green color (contribution level of 80%). Sterically unfavored areas are shown in yellow color (contribution level of 20%). Positive potential favored areas are shown in blue color (contribution level of 80%). Positive potential unfavored areas are shown in red color (contribution level of 20%). (ii) CoMSIA H-bond donor/H-bond acceptor contour maps of compounds **23** and **36** (on the *left* and *right* of the figure, correspondingly). The individual contributions from the H-bond donor and H-bond acceptor favored and disfavored levels are fixed at 80% and 20%, respectively. The contours for H-bond donor favored fields have been shown in cyan color, while its disfavored fields have been shown in purple color. H-bond acceptor favored fields have been shown in orange color, while its disfavored fields have been shown in white color. (Reproduced with Permission from Durdagi et al. [58]. Copyright, Elsevier)

enhanced binding affinities. Interestingly, the investigated nanostructured compounds have triggered the interest of medicinal chemists to look for novel fullerene-type HIV-1 PR inhibitors possessing higher bioactivity. Later this year, the authors published a second study for the same type of fullerene-based nanomaterials [59].

Next, the same group published later another study based on fullerene derivatives, functionalized by amino acids [60]. The authors used *in silico* screening approach in order to propose potent fullerene analogues as anti-HIV drugs. As a result, two of the most promising derivatives showing significant binding scores were subjected to biological studies that confirmed the efficacy of

the new compounds. The results showed that using combined computational approach, new leads can be discovered possessing higher bioactivity. The authors used docking approach together with MD simulations to get the best hits during the virtual screening.

Later, in 2011, the same group conducted one more study to design better anti-HIV fullerene-based inhibitors [61]. In this study, authors employed a protein-ligand docking technique, two 3D-QSAR models, MD simulations, and the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) calculations. The authors investigated (1) hydrogen bonding (H-bond) interactions between specific fullerene derivatives and the protease; (2) the regions of HIV-1 PR that play a significant role in binding; (3) protease changes upon binding; and (4) various contributions to the binding-free energy, in order to identify the most significant of them. To build 3D-QSAR models, the CoMFA and CoMSIA methods were applied, with results that showed good correlation coefficients, for both methods, $r^2 = 0.842$ and 0.928, respectively. In conclusion, the authors stated that the computed binding free energies are in satisfactory agreement with the experimental results.

Another group published in 2013 a study that conducted a comprehensive investigation of fullerene analogues by combined computational approach including quantum chemical, molecular docking, and 3D descriptor-based QSAR [17]. In this work, the authors stated that the protein-ligand docking studies and structure-activity QSAR models have been able both to predict binding affinities for the set of fullerene-C₆₀ derivatives and to assist in finding mechanisms of functionalized fullerene interactions with human immunodeficiency virus type 1 aspartic protease, HIV-1 PR. The authors concluded that protein-ligand docking revealed several important molecular fragments that are responsible for the interaction with HIV-1 PR (Fig. 4). In parallel, the authors utilized a density functional theory (DFT) to identify optimal geometries and predict physicochemical parameters of 49 fullerene derivatives. In this study, a five-variable genetic algorithm-multiple linear regression (GA-MLR)-based model has been developed, which showed a good predictive ability, with correlation coefficient $r^2_{\text{train}} = 0.882$ for training set and $r^2_{\text{test}} = 0.738$ for the test.

In 2010, Calvaresi and Zerbetto [62] published an interesting study where the authors investigated a pristine fullerene binding with a set of proteins, to find potentially toxic ones and potentially highly selective “drug-like” ones. In this study, the authors investigated about 20 proteins that are known to modify their activity upon interaction with C₆₀. It is worth to note that for C₆₀-protein system investigations, the authors applied a relatively new docking software—PatchDock [63] which can handle such large ligand systems as fullerenes and utilize an algorithm that appraises quantitatively the interaction of C₆₀ and the surface of each protein. The

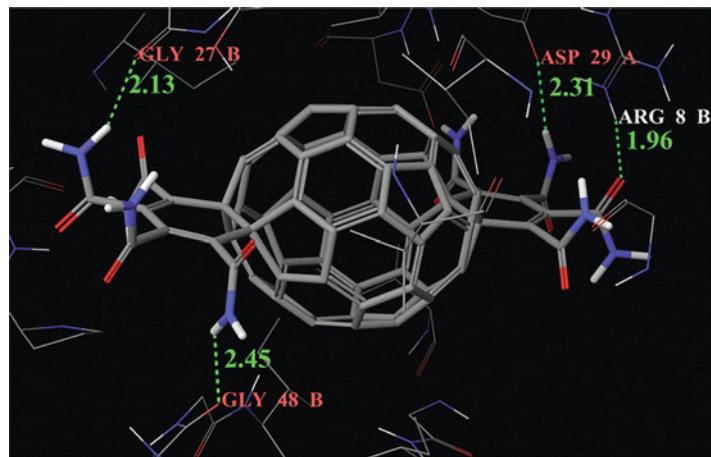


Fig. 4 The binding site interactions. H-bonds formed by the ligand **42** in the binding site (Glide). (Reproduced with Permission from reference Ahmed et al. [17]. Copyright, Royal Society of Chemistry)

authors claim that the redundancy of the set allowed them to establish the predictive power of the approach that finds explicitly the most probable site where C₆₀ docks on each protein. Interestingly, about 80% of the known fullerene-binding proteins fall in the top 10% of scorers. The authors identified the sites of docking and discussed them in view of the existing experimental data available for protein-C₆₀ interactions. Moreover, the authors identified new proteins that can interact with C₆₀ and discussed for possible future applications as drug targets and fullerene derivative bioconjugate materials.

Later, the same authors, Calvaresi and Zerbetto [64], published another study, where they investigated a larger dataset, i.e., binding of fullerene C₆₀ with 1099 proteins. In this study, the authors one more time confirmed that hydrophobic pockets of certain proteins can accommodate a carbon cage either in full or in part. In this regard, since the identification of proteins that are able to discriminate between different cages is still an open issue, they were interested in investigating a significantly larger library than in the previous paper [62]. Importantly, in this work, the prediction of candidates is achieved with an inverse docking procedure, which is able to accurately account for (1) van der Waals interactions between the cage and the protein surface, (2) desolvation free energy, (3) shape complementarity, and (4) minimization of the number of steric clashes through conformational variations. The authors divided a set of 1099 protein structures into four categories that either select C₆₀ or C₇₀ (p-C₆₀ or p-C₇₀) and either accommodate the cages in the same pocket or in different pockets. In overall, the authors were able to confirm the agreement of obtained computational results with the experiment, where the

KcsA potassium channel is predicted to have one of the best performances for both cages.

Next, in a relatively recent work done by Ghasemi and co-authors [65], the authors used two molecular interaction field (MIF)-based descriptors, VolSurf and GRIND, as alignment-independent three-dimensional quantitative structure-activity relationship (3D-QSAR) approaches to predict C60 solubility in a diverse set of 132 organic solvents. The authors applied a GRIND methodology with fractional factorial design and then applied a PLS analysis, which yielded a highly descriptive and predictive model. Moreover, the authors applied a genetic algorithm (GA) and successive projection algorithm (SPA) to feature selection and extract more informative VolSurf descriptors. In addition, a support vector machine (SVM) was used to develop a model, where SPA-SVM-based VolSurf descriptors showed an excellent performance in predicting the C60 solubility for fullerene. The authors conducted a validation, reliability, and robustness analysis of the obtained models, as well as evaluation of the prediction ability of external test sets, by applying leave-one-out and progressive scrambling approach. Thus, the results of this study confirmed that hydrophobic interactions besides steric effects are main factors influencing solubility of C60 in different organic solvents.

In another study, Rofouei and co-authors [66] used an alignment free, 3D-QSAR approach to investigate a dispersibility of single-walled carbon nanotubes (SWNTs) in a diverse set of organic solvents. In this work, again the GRIND methodology was applied, where the descriptors are derived from GRID molecular interaction fields, MIFs. In this comprehensive study, the authors applied different variable selection procedures including fractional factorial design (FFD), stepwise multiple linear regression (SW-MLR), successive projection algorithm (SPA), genetic algorithm (GA), and enhanced replacement method (ERM), to extract the more informative factors from exported GRIND descriptors and generate a predictive model. The PLS method was applied for model development, where ERM-PLS-based GRIND descriptors showed an excellent performance in predicting SWNT dispersibility values. The authors stated that the obtained ERM-PLS model satisfied a set of rigorous validation criteria and performed well in the prediction of an external test set. The authors also stated that from the GRIND variables involved in ERM-PLS model, it is possible to identify some key molecular features/fragments and their position in a solvent structure, which are responsible for a SWNT dispersibility. In overall, the obtained results in this study confirmed the importance of hydrophobic interactions, size, and steric hindrance of hydrophobic part of solvent molecule. Moreover, the authors stated that the effect of presence of a hydrogen bond donor or polar group in a structure of solvent molecule with a large size couldn't be neglected.

In the recent paper, Esposito and co-authors [67] published a QSAR study, where authors studied decorated carbon nanotubes (CNTs) to predict toxicity using 4D fingerprints. Thus, the authors proposed detailed mechanisms of action that relate to nanotoxicity, for a series of functionalized CNT complexes based on previously reported QSAR models. Moreover, the authors proposed possible mechanisms of nanotoxicity for six endpoints (bovine serum albumin, carbonic anhydrase, chymotrypsin, hemoglobin along with cell viability, and nitrogen oxide production) based on optimized QSAR models. The molecular features relevant to each of the endpoint specific mechanisms of action for investigated decorated CNTs are discussed in the paper. The following responsible factors were revealed by the authors—either the decorator attached to the nanotube is directly responsible for the expression of certain activity, irrespective of the decorator's 3D geometry and independent of the CNT or those decorators having structures that place the functional groups of the decorators as far as possible from the CNT's surface most strongly influence the biological activity.

In the next study, a combined docking and comprehensive DFT analysis was conducted by Saikia and co-authors [68]. The authors conducted a modeling study to analyze the interaction of carbon nanomaterials with biomolecular systems, where the DFT calculations on the interaction of pyrazinamide (PZA) drug with functionalized single-wall CNT (f-SWCNT) were made. The analysis is based on CNT properties mainly as a function of nanotube chirality and length, followed by docking simulation of f-SWCNT with pncA protein. The authors stated that the functionalization of pristine SWCNT that facilitates in enhancing the reactivity of the nanotube and formation of such type of nanotube-drug conjugate is thermodynamically feasible. The conducted docking studies predicted the plausible binding mechanism and suggested that PZA loaded f-SWCNT facilitates in the target-specific binding of PZA within the protein, following a lock-and-key mechanism. In this study, the authors pointed out that no major structural deformation in the protein was observed after binding with CNT and the interaction between ligand and receptor is mainly hydrophobic in nature. In overall, the authors anticipate that these findings may provide new routes toward the drug delivery mechanism by CNTs with long-term practical implications in tuberculosis chemotherapy.

In another study, Turabekova et al. [69] published a comprehensive study of CNT and pristine fullerene interactions with Toll-like receptors (TLRs), where the latter are responsible for immune response, i.e., researchers investigated the immunotoxicity of fullerenes and CNTs. The authors performed a cytokine expression experimental analysis and conducted a comprehensive computational protein-ligand investigation, where the authors showed that CNTs and fullerenes can bind to certain TLRs. The authors suggested a hypothetical model providing the potential mechanistic

explanation for immune and inflammatory responses observed upon exposure to carbon-nanostructured materials. Using computational approaches, the authors performed a theoretical study to analyze CNT and C₆₀ fullerene interactions with the available X-ray structures of TLR homo- and heterodimer extracellular domains. The computational investigation was based on the fact that both CNT and C₆₀ are similar to the known TLR ligands and in cells they induce a secretion of certain inflammatory protein mediators, such as interleukins and chemokines. Signal proteins are observed within inflammation downstream processes resulting from the ligand molecule-dependent inhibition or activation of TLR-induced signal transduction. Thus, the computational studies have shown that the internal hydrophobic pockets of some TLRs might be capable of binding small-sized carbon nanostructures, SWCNTs, and C₆₀. The obtained high binding scores and minor structural alterations induced in TLR ectodomains upon binding C₆₀ and CNTs further supported the proposed hypothesis (Fig. 5). The proposed hypothesis is confirmed by the conducted experimental study indicating that CNTs and fullerenes induce an excessive expression of such cytokines as IL-8 and MCP1.

Interestingly, Mozolewska and co-authors in a follow-up study have confirmed this kind of interactions of CNT and C₆₀ with TLR [70]. In this study, the authors made an attempt to determine if the CNTs could interfere with the innate immune system by interacting with TLRs. For this purpose, authors used the following TLR structures, obtained from the RCSB Protein Data Bank, TLR2 (3A7C), TLR4/MD (3FXI), TLR5 (3V47), and TLR3 (2A0Z), and the complexes of TLR1/TLR2 (2Z7X) and TLR2/TLR6 (3A79). In result, based on steered molecular dynamics (SMD) simulations, the authors showed that certain size CNTs interact very strongly with the binding pockets of some receptors (e.g., TLR2), which results in their binding to these sites without substantial use of the external force.

4 Notes and Concluding Remarks

In this review paper, we have discussed various 3D-QSAR methods and their applications with and without a combination with protein-ligand docking studies, to investigate and support a design of carbon nanostructured materials. Despite of large size of carbon nanomaterials, 3D-QSAR and protein-ligand docking have confirmed the feasibility of these methods to investigate carbon-nanostructured materials and importance of combination of these methods in further assessment of interaction of CNTs and fullerenes with biological molecules. Thus, the development of techniques for 3D-QSAR methodology is continuing, giving scientists better accuracy in predictions. We believe that 3D-QSAR methods,

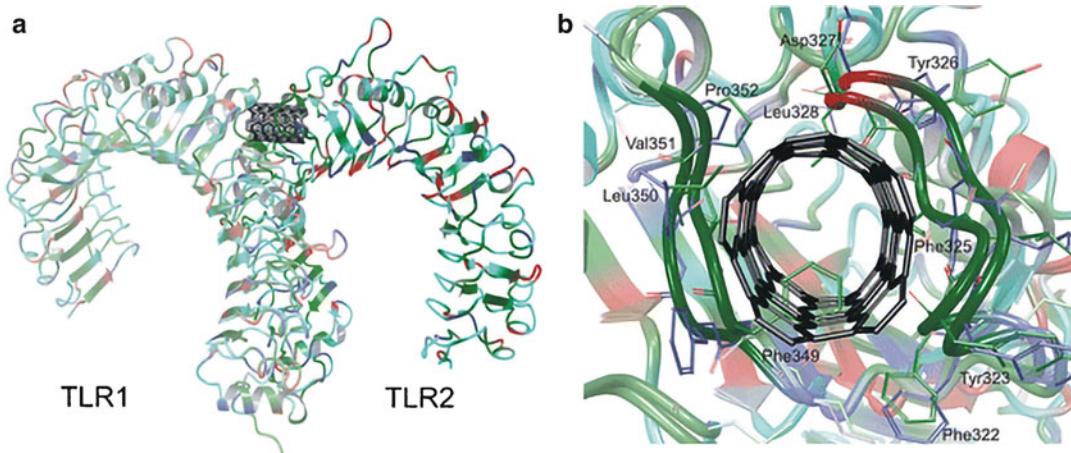


Fig. 5 5,5 CNT-bound TLR1/TLR2 ECDs: **(a)** 5,5 CNT is bound to the TLR1 and TLR2 ECD interface dimerization area, **(b)** aligned structures of TLR2 ECDs before (*green* carbon atoms) and after (*blue* carbon atoms) the impact OPLS2005 refinement upon binding 5,5 CNTs. The orientation of two parallel entrance loops and the side chains of hydrophobic Phe349, Phe325, and Leu328 preventing the nanotube from intrusion is shown to be optimized. (Reproduced with Permission from reference Turabekova et al. [69]. Copyright, Royal Society of Chemistry)

and especially their combination with protein-ligand docking analysis, soon will be able to model various large nanomaterials which can help scientists to investigate important biological and physicochemical properties of them.

References

- Hansch C, Leo A, Hoekman D, Leo A (1995) Exploring QSAR. American Chemical Society, Washington, DC
- Kubinyi H (1997) QSAR and 3D QSAR in drug design part 1: methodology. *Drug Discov Today* 2(11):457–467
- Kubinyi H (1997) QSAR and 3D QSAR in drug design part 2: applications and problems. *Drug Discov Today* 2(12):538–546
- Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 37(24):4130–4146
- Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 96(3):1027–1044
- Isayev O, Rasulev B, Gorb L, Leszczynski J (2006) Structure-toxicity relationships of nitroaromatic compounds. *Mol Divers* 10(2):233–245
- Rasulev B, Toropov AA, Hamme AT II, Leszczynski J (2008) Multiple linear regression analysis and optimal descriptors: predicting the cholesterlyl ester transfer protein inhibition activity. *QSAR Comb Sci* 27(5):595–606
- Rasulev B, Kušić H, Leszczynska D, Leszczynski J, Koprivanac N (2010) QSAR modeling of acute toxicity on mammals caused by aromatic compounds: the case study using oral LD₅₀ for rats. *J Environ Monitor* 12(5):1037–1044
- Turabekova MA, Rasulev B, Dzhakhangirov FN, Leszczynska D, Leszczynski J (2010) Aconitum and Delphinium alkaloids of curare-like activity. QSAR analysis and molecular docking of alkaloids into AChBP. *Eur J Med Chem* 45(9):3885–3894
- Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 45(12):2615–2623
- Turabekova MA, Rasulev B, Dzhakhangirov FN, Salikhov SI (2008) Aconitum and Delphinium alkaloids. “Drug-likeness” descriptors

- related to toxic mode of action. *Environ Toxicol Pharmacol* 25:310–320
12. Toropov AA, Toropova AP, Rasulev B, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2008) CORAL: binary classifications (active/inactive) for liver-related adverse effects of drugs. *Curr Drug Saf* 7(4):257–261
 13. Rago R, Artico M, De Martino G, La Regina G, Coluccia A, Di Pasquali A, Silvestri R (2005) Docking and 3-D QSAR studies on indolyl aryl sulfones. Binding mode exploration at the HIV-1 reverse transcriptase non-nucleoside binding site and design of highly active N-(2-hydroxyethyl) carboxamide and N-(2-hydroxyethyl) carbohydrazide derivatives. *J Med Chem* 48(1):213–223
 14. Hu R, Barbault F, Delamar M, Zhang R (2009) Receptor-and ligand-based 3D-QSAR study for a series of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Bioorg Med Chem* 17(6):2400–2409
 15. Sun J, Cai S, Yan N, Mei H (2010) Docking and 3D-QSAR studies of influenza neuraminidase inhibitors using three-dimensional holographic vector of atomic interaction field analysis. *Eur J Med Chem* 45(3):1008–1014
 16. Araújo JQ, de Brito MA, Hoelz LVB, de Alencastro RB, Castro HC, Rodrigues CR, Albuquerque MG (2011) Receptor-dependent (RD) 3D-QSAR approach of a series of benzyl-piperidine inhibitors of human acetylcholinesterase (HuAChE). *Eur J Med Chem* 46 (1):39–51
 17. Ahmed L, Rasulev B, Turabekova M, Leszczynska D, Leszczynski J (2013) Receptor-and ligand-based study of fullerene analogues: comprehensive computational approach including quantum-chemical, QSAR and molecular docking simulations. *Org Biomol Chem* 11(35):5798–5808
 18. Jagiello K, Grzonkowska M, Swirot M, Ahmed L, Rasulev B, Avramopoulos A, Papadopoulos MG, Leszczynski J, Puzyn T (2016) Advantages and limitations of classic and 3D QSAR approaches in nano-QSAR studies based on biological activity of fullerene derivatives. *J Nanopart Res* 18(9):256
 19. Wise M, Cramer RD, Smith D, Exman I (1983) Progress in three-dimensional drug design: the use of real-time colour graphics and computer postulation of bioactive molecules in DYLOMMS. Elsevier, Amsterdam
 20. Cramer R, Bunce JD (1987) The DYLOMMS method: initial results from a comparative study of approaches to 3D QSAR. In: Hadzi D, Jerman-Blazic B (eds) QSAR in drug design and toxicology. Elsevier Science, Amsterdam, pp 3–12
 21. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110(18):5959–5967
 22. Clark M, Cramer RD, Jones DM, Patterson DE, Simeroth PE (1990) Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. *Tetrahedron Comput Methodol* 3(1):47–59
 23. Lill MA (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov Today* 12 (23):1013–1017
 24. Kim KH, Greco G, Novellino E (1998) A critical review of recent CoMFA applications. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in drug design. Springer, Dordrecht, pp 257–315
 25. Todeschini R, Gramatica P (1998) 3D QSAR in drug design, vol 2. Kluwer/ESCOM, Dordrecht, pp 355–360
 26. Podlogar BL, Ferguson DM (2000) QSAR and CoMFA: a perspective on the practical application to drug discovery. *Drug Des Discov* 17 (1):4
 27. Tripos (2006) SYBYL, version 7.3, 2006, St. Louis
 28. Patel PD, Patel MR, Kaushik-Basu N, Talele TT (2008) 3D QSAR and molecular docking studies of benzimidazole derivatives as hepatitis C virus NS5B polymerase inhibitors. *J Chem Inf Model* 48(1):42–55
 29. Oprea TI, Matter H (2004) Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* 8(4):349–358
 30. Ahmed L, Rasulev B, Kar S, Krupa P, Mozolewska M, Leszczynski J (2017) Inhibitors or toxins? Large library target-specific screening of fullerene-based nanoparticles for drug design purpose. *Nanoscale* 9 (29):10263–10276
 31. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 11(13):580–594
 32. Perola E (2006) Minimizing false positives in kinase virtual screens. *Proteins: Struc Func Bioinf* 64(2):422–435
 33. Pajeva IK, Globisch C, Wiese M (2009) Combined pharmacophore modeling, docking, and 3D QSAR studies of ABCB1 and ABCC1 transporter inhibitors. *ChemMedChem* 4 (11):1883–1896
 34. Yang S-Y (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today* 15 (11):444–450

35. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28(7):849–857
36. Kim KH (2001) Thermodynamic aspects of hydrophobicity and biological QSAR. *J Comput Aid Mol Des* 15(4):367–380
37. Hopfinger AJ (1980) A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J Am Chem Soc* 102(24):7196–7206
38. Doweyko AM (1988) The hypothetical active site lattice. An approach to modelling active sites from data on inhibitor molecules. *J Med Chem* 31(7):1396–1406
39. Walters DE, Hinds RM (1994) Genetically evolved receptor models: a computational approach to construction of receptor models. *J Med Chem* 37(16):2527–2536
40. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 43:3233–3243
41. Fontaine F, Pastor M, Zamora I, Sanz F (2005) Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *J Med Chem* 48(7):2687–2694
42. Fontaine F, Pastor M, Sanz F (2004) Incorporating molecular shape into the alignment-free GRid-INdependent descriptors. *J Med Chem* 47(11):2805–2815
43. Silverman B, Platt DE (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J Med Chem* 39(11):2129–2140
44. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem* 38(14):2681–2691
45. Polanski J, Gieleciak R, Bak A (2002) The comparative molecular surface analysis (CoMSA)-a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting pKa values of benzoic and alkanoic acids. *J Chem Inf Comput Sci* 42(2):184–191
46. Polanski J, Bak A, Gieleciak R, Magdziarz T (2006) Modeling robust QSAR. *J Chem Inf Model* 46(6):2310–2318
47. Datar PA, Khedkar SA, Malde AK, Coutinho EC (2006) Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *J Comput Aid Mol Des* 20(6):343–360
48. Dhaked DK, Verma J, Saran A, Coutinho EC (2009) Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). *J Mol Model* 15 (3):233–245
49. Robinson DD, Winn PJ, Lyne PD, Richards WG (1999) Self-organizing molecular field analysis: a tool for structure-activity studies. *J Med Chem* 42(4):573–583
50. Zhou P, Tian F, Li Z (2007) Three-dimensional holographic vector of atomic interaction field (3D-HoVAIF). *Chemom Intell Lab Syst* 87(1):88–94
51. Ajmani S, Jadhav K, Kulkarni SA (2006) Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J Chem Inf Model* 46(1):24–31
52. Baskin II, Zhokhova NI (2013) The continuous molecular fields approach to building 3D-QSAR models. *J Comput Aid Mol Des* 27(5):427–442
53. Zhokhova NI, Baskin II, Bakhronov DK, Palyulin VA, Zefirov NS (2009) Method of continuous molecular fields in the search for quantitative structure-activity relationships. *Dokl Chem* 429(1):273–276
54. Dixon SL, Smolyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development and 3D database screening. 1. Methodology and preliminary results. *J Comput Aid Mol Des* 20:647–671
55. Totrov M (2008) Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chem Biol Drug Des* 71(1):15–27
56. Li H, Sutter J, Hoffmann R (2000) HypoGen: an automated system for generating 3D predictive pharmacophore models. In: Güner OF (ed) Pharmacophore perception, development, and use in drug design (pp. 171–189). International University Line, La Jolla, Calif, USA
57. Evans DA, Doman TN, Thorner DA, Bodkin MJ (2007) 3D QSAR methods: phase and catalyst compared. *J Chem Inf Model* 47 (3):1248–1257
58. Durdagı S, Mavromoustakos T, Chronakis N, Papadopoulos MG (2008a) Computational design of novel fullerene analogues as potential HIV-1 PR inhibitors: analysis of the binding interactions between fullerene inhibitors and HIV-1 PR residues using 3D QSAR, molecular docking and molecular dynamics simulations. *Bioorg Med Chem* 16(23):9957–9974
59. Durdagı S, Mavromoustakos T, Papadopoulos MG (2008b) 3D QSAR CoMFA/CoMSIA,

- molecular docking and molecular dynamics studies of fullerene-based HIV-1 PR inhibitors. *Bioorg Med Chem Lett* 18(23):6283–6289
60. Durdagi S, Supuran CT, Strom TA, Doostdar N, Kumar MK, Barron AR, Mavromoustakos T, Papadopoulos MG (2009) In silico drug screening approach for the design of magic bullets: a successful example with anti-HIV fullerene derivatized amino acids. *J Chem Inf Model* 49(5):1139–1143
61. Tzoupis H, Leonis G, Durdagi S, Mouchlis V, Mavromoustakos T, Papadopoulos MG (2011) Binding of novel fullerene inhibitors to HIV-1 protease: insight through molecular dynamics and molecular mechanics Poisson–Boltzmann surface area calculations. *J Comput Aid Mol Des* 25(10):959–976
62. Calvaresi M, Zerbetto F (2010) Baiting proteins with C60. *ACS Nano* 4(4):2283–2299
63. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and Symm-Dock: servers for rigid and symmetric docking. *Nucleic Acid Res* 33:W363–W367
64. Calvaresi M, Zerbetto F (2011) Fullerene sorting proteins. *Nanoscale* 3(7):2873–2881
65. Ghasemi JB, Salahinejad M, Rofouei MK (2013) Alignment independent 3D-QSAR modeling of fullerene (C₆₀) solubility in different organic solvents. *Fuller Nanotub Car N* 21(5):367–380
66. Rofouei MK, Salahinejad M, Ghasemi JB (2014) An alignment independent 3D-QSAR modeling of dispersibility of single-walled carbon nanotubes in different organic solvents. *Fuller Nanotub Car N* 22(7):605–617
67. Esposito EX, Hopfinger AJ, Shao CY, Su BH, Chen SZ, Tseng YJ (2015) Exploring possible mechanisms of action for the nanotoxicity and protein binding of decorated nanotubes: interpretation of physicochemical properties from optimal QSAR models. *Toxicol Appl Pharmacol* 288(1):52–62
68. Saikia N, Rajkhowa S, Deka RC (2013) Density functional and molecular docking studies towards investigating the role of single-wall carbon nanotubes as nanocarrier for loading and delivery of pyrazinamide antitubercular drug onto pncA protein. *J Comput Aid Mol Des* 27(3):257–276
69. Turabekova M, Rasulev B, Theodore M, Jackman J, Leszczynska D, Leszczynski J (2014) Immunotoxicity of nanoparticles: a computational study suggests that CNTs and C₆₀ fullerenes might be recognized as pathogens by Toll-like receptors. *Nanoscale* 6 (7):3488–3495
70. Mozolewska MA, Krupa P, Rasulev B, Liwo A, Leszczynski J (2014) Preliminary studies of interaction between nanotubes and toll-like receptors. *Task Quarterly* 18(4):351–355



Chapter 11

Early Prediction of Ecotoxicological Side Effects of Pharmaceutical Impurities Based on Open-Source Non-testing Approaches

Anna Rita Tondo, Michele Montaruli, Giuseppe Felice Mangiatordi, and Orazio Nicolotti

Abstract

Despite the increasing efforts to limit waste and avoid environmental contaminants, a large number of compounds used in the pharmaceutical field may have an ecotoxicological impact. Nevertheless, a complete overview of all possible ecotoxicological effects of pharmaceuticals is missing; that is especially true for chemical impurities. The lacking information regarding environmental behavior of impurities could be faced by computational techniques: the ability to predict the unknown toxicity of a compound can reduce uncertainties regarding possible negative effects on the environment of pharmaceutical impurities. In the current scenario, non-testing methods may answer to the requirement of assessing the ecotoxicological impact of chemicals in a more affordable way. For this purpose, in the first part of the review, definition and classification of chemical impurities are proposed, while in the second part, a description of four open-source computational tools (T.E.S.T., VEGA, LAZAR, and QSAR Toolbox) is provided after a brief survey of the computational methods. The paper also shows the advantages of combining individual test methods in order to increase confidence in the predictive results.

Key words Impurities, QSAR Toolbox, Ecotoxicity, LAZAR, VEGA, T.E.S.T., Non-testing approaches, QSAR models

1 Introduction

The current era is characterized by the continuous improvements not only for care but also for wellness with an increasing request of new pharmaceuticals. The invention of new medicines and the improvement of existing drugs constitute a never-ending process for pharmaceutical industry whose business is always addressed toward new generation of safe drugs. The drug manufacturing is a multistep sophisticated process starting from the synthesis of active pharmaceutical ingredients (APIs) until the final packaging of the finished product. Each manufacturing step entails the use of several chemical substances for obtaining the final medicinal drug.

Among different commercial sectors, the pharmaceutical industry is regarded as the most wasteful, having the highest *E factors*, which are measures of the amount of waste produced compared to the yield of useful material obtained [1, 2]. Despite the increasing efforts to limit waste and avoid pollution (air and water) and accidents, by applying the concept of “green chemistry,” a large number of compounds used in the pharmaceutical field may spread through the environmental media. Such contaminants are not only the APIs and their metabolites released in the environment after their excretion from humans or animals via urine or feces but also unwanted residues of APIs and other chemical entities (i.e., pharmaceutical impurities) developed during drug manufacturing.

Despite the high quality and purity standard of the raw materials used for preparation, a drug substance typically contains a range of low-level impurities, for example, arising as residues of starting materials, reagents, and intermediates or as side products generated by the synthetic processes or degradation reactions. Residues of pharmaceuticals at trace quantities are widespread in aquatic systems [3]. Therefore, the potential impact of pharmaceutical residues has recently become a worrying environmental concern, due to growing and, sometimes uncontrolled, of human, veterinary, and agriculture pharmaceuticals [4].

In recent years, many scientific research programs have been oriented to monitoring pharmaceuticals in various aqueous matrices (i.e., water and/or wastewater), as several studies [5–7] demonstrated. Nowadays, there is a public concern that the cascade of unforeseeable effects may be responsible for spread in the environment even in trace concentrations [8–11]. In this respect, the persistence of medicinal products against degradation is another issue: drugs and their manufacturing process-related substances (i.e., chemical impurities) could remain in the environment for a long time, and their presence is considered dangerous in both low and high concentrations [12, 13].

Pharmaceuticals and chemical impurities of drug development process may have potential toxic effects virtually at any level of the biological hierarchy, i.e., cells, organs, organisms, population, and ecosystems. There is thus the need to set an environmentally friendly development of novel and cost-effective industrial approaches. The green chemistry approach was developed to meet this expectation. However, a full understanding of the toxicological effects of pharmaceuticals on the environment is still far from being reached. This gap of information is particularly relevant for those unwanted organic compounds arising from industrial drug development process that are the pharmaceutical impurities. Despite current analytical methodologies make possible to reveal and reduce impurities of pharmaceutical industrial processes, it is very difficult to assess the toxicological impact of the impurities included in the API. The need of making early predictions of the

ecotoxicological side effects of pharmaceutical impurities has promoted the development of non-testing computational methods which are cheaper and ethically acceptable compared to in vitro and in vivo experimental approaches. Non-testing methods not only answer to the economic pressure to reduce the huge cost of pharmaceutical industry but also to the regulatory purposes of the 3R principle (replacement, reduction, and refinement) [14] aimed at safeguarding animal welfare. In this context, computational methods have been largely employed for the early identification and assessment of human health risks associated with the exposure to pharmaceutical impurities resulting from drug preparation. In particular, a key aspect is the prediction of genotoxic risk of drug impurities.

2 Impurity Definition

Pharmaceutical industry produces drug substances which can be extracted from natural products or chemically prepared. In this diversified scenario, there is a precept that drug manufacturers must follow that the final product should be as pure as possible since purity is an essential factor for ensuring drug quality. In this respect, raw materials, manufacturing method, crystallization, and purification process are of utmost importance. However, the complexity of drug development and manufacturing process and the high number of components required to prepare pharmaceutical products make the definition of “pharmaceutical impurity” difficult. To the best of our knowledge, the simplest definition of pharmaceutical impurity stands from drug definition. Substantially, a drug is composed of a drug substance—which is the only chemical component accountable for the therapeutic effect—and of one or more excipient(s), i.e., the inactive constituent(s), which are normally important for a satisfying pharmacokinetics. Therefore, an impurity is defined as any component present in the drug product that is neither an active substance nor an excipient [15]. The presence of pharmaceutical impurities, even in very small amounts, may influence the efficacy and safety of the medicinal products [15]: they are unwanted chemicals that could be derived from APIs or that could be developed during drug formulation. APIs invariably contain impurities: the latter may be residues of starting materials and intermediates used in the manufacturing process, as well as products of degradation of chemicals. It must be remarked that any extraneous material present in the drug substance has to be considered an impurity even if it has totally inert properties. Most APIs are produced by organic chemical synthesis, and many components can be generated during such a process. Those components remaining in the final API are considered as impurities. Under this aspect, considering the quality control checks

conducted by each industry in order to detect the quality profile of pharmaceutical impurities, the latter could be defined in a manner that is depending from the adopted analytical method used for their identification. In this respect, an identified impurity could be defined depending on the availability of structural information based on quantitative or qualitative analytical values.

3 Classification of Impurities

Most pharmaceutical products are manufactured either by applying a total synthesis approach or by modifying a naturally occurring product. In both cases, a wide range of reactive reagents is used. Therefore, it is natural that low levels of such reagents or side products are present in the final API or drug product as impurities. Such impurities may have unwanted toxicities, including genotoxicity and carcinogenicity. The risk for patient's health caused by the presence of small molecules as impurities in APIs has become an increasing concern of pharmaceutical companies, regulatory authorities, patients, and doctors alike. According to FDA/ICH guidelines, three attributes define the drug quality of a pharmaceutical product: identity, strength, and purity. If identity is of greater importance in the preliminary phases of pharmaceutical analysis and strength safeguards the maximum efficacy of a drug, purity is the only crucial attribute assuring the maximum safety of drug therapy. This is the reason why regulatory agencies pay attention to listing and catalogue all various types of impurities in several categories, characterizing them in all ecotoxicological and non-ecotoxicological aspects, in order to guarantee that these cannot contribute to the side effect profile of the drugs. There are many ways to classify pharmaceutical impurities associated with APIs. For regulatory purpose, in 1994, ICH guidelines [15] distinguished pharmaceutical impurities in three big branches that are reported in Fig. 1. While several impurities, such as heavy metals, can be avoided or held in reduced levels by using particular manufacturing technique, trace presence of some pharmaceutical impurities, like residual solvents above all, could be inescapable. Regardless, it is an undeniable fact that both controlled process-related impurities and uncontrolled process-related impurities could reach the pharmaceutical wastewater and present an environmental concern. Bearing in mind that traces of residual solvents occasionally can be surveyed during the manufacture of drug products (since residual solvents also arise in excipients) and that some solvents are known to be toxic, careful attention must be addressed for minimizing risks of detecting pharmaceutical impurities in wastewaters and in the environment. In fact, it is worth annotating that a separate, specific guideline for residual solvents is available [15]. Residual solvents are divided into three classes depending

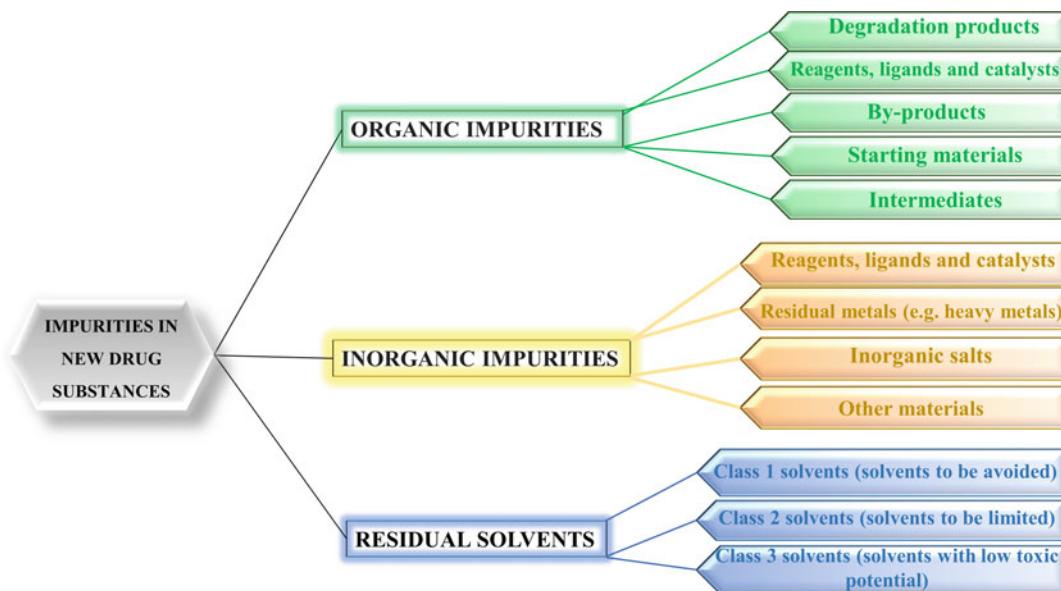


Fig. 1 Classification of pharmaceutical impurities according to ICH guidelines

on the possible risk to human health. Moreover, the pharmaceutical companies themselves move to a more ecological-sustainable manufacturing process just through the construction and adoption of solvent selection guides, which address chemists to the selection of the more sustainable solvents [16].

4 Legislation of Impurities

Identification, quantification, and control of impurities in the drug substances and drug products are an important part of drug development, mostly in terms of regulatory assessment: even if there is a lack of a consolidated (united, affiliated) legal system, several documents outline concepts and principles for the regimentation of pharmaceutical impurities (listed in Table 1). Moreover, for medicinal products, only in the 1990s, regulatory agencies have issued detailed guidelines for possible unwanted effects on the environment of pharmaceuticals [17].

The US Food and Drug Administration (FDA or US-FDA), the agency of the United States Department of Health and Human Services, responsible for supervising the safety of foods, dietary supplements, and drugs, issued three guidance—named NDAs—for industry about impurities, separating ones limited to drug substances produced by chemical synthesis and ones reserved to drug products that are manufactured from drug substances. However, USP recognizes that other impurities may come from a variety of situations, such as a change in processing or

Table 1
Principal guidelines for regulation of pharmaceutical impurities

Guideline	Document title
<i>ICH guideline Q1A(R)</i>	Stability Testing of New Drug Substances and Products
<i>ICH guideline Q3A(R)</i>	Impurities in Drug Substances
<i>ICH guideline Q3B</i>	Impurities in Drug Products
<i>ICH guideline Q3C</i>	Impurities: Residual Solvents
<i>ICH guideline Q6A</i>	Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances
<i>ICH guidelines Q3A(R2)</i>	Impurities in New Drug Substances
<i>ICH guidelines Q3B(R2)</i>	Impurities in New Drug Products
<i>ICH guidelines Q3D(R1)</i>	Guideline for Elemental Impurities
<i>FDA guidelines</i>	NDAs: Impurities in Drug Substances
<i>FDA guidelines</i>	ANDAs: Impurities in Drug Products

extraneous sources, so that many USP monographs contain tests for specific impurities, recommending that all impurities including the monograph-specified impurities should not exceed 2.0% [18].

5 Non-testing Predictive Models

Many pharmaceutical R&D programs focus on optimization of lead compounds in terms of potency and selectivity with respect to the biological targets and on the safety of all chemicals involved in the actual manufacturing process of the drug product. The safety must be guaranteed by a drug manufacturer through toxicological assessments and by following GMPs. All the chemicals used during manufacturing must undertake toxicological evaluations, according to the “*primum non nocere*” (which stands for “first to do no harm”) principle. Such paradigm embraces human health but is also of primary importance for drug process-related impurities released in environment as pharmaceutical waste. Most of the traditional methods to determine the toxicological safety of chemicals rely predominantly on experimental work often involving the use of animals. Being this approach too demanding in terms of time and costs, it would be impossible to experimentally assess the hazards and risks for every single pharmaceutical used in R&D programs. One solution to this problem is to establish the lowest threshold level of contaminants (i.e., impurities) above which the impurity can be tolerated, as contemplated by ICH guidelines. Provided that impurities can never be completely removed, purification

techniques (HPLC-MS, TLC) [19] are useful to detect their amount at harmless levels. Such policies are pursued to assess the genotoxic potential of pharmaceutical impurities. However, ecotoxicological evaluation concerning environmental impact of chemical impurities released from pharmaceutical wastes is still lacking. A more all-embracing solution is represented by a prioritization approach useful to identify those compounds that are likely to pose high ecotoxicological risk and, therefore, need further attention. This alternative involves the use of computer-aided methods in order to clarify how drugs, and chemicals in general, may adversely affect functions of the organisms. Broadly speaking, the term *in silico* toxicology refers to the application of computer technologies to predict toxicological activity of a substance making use of existing data and mathematical models. The increasing number of data analysis tools allows the predictions of toxicity based on a query chemical. As a new emerging scientific discipline, toxicoinformatics exploits bioinformatic methods and computer-based analyses with the goal of unveiling the relationships between a chemical structure and a toxicological endpoint. The underlying and intimate concept of computational toxicology is that it is possible to retrieve and analyze existing and relevant data of many toxic (and nontoxic) chemicals and to relate the structure of compounds to their fate. Such relationships, in turn, can be employed for making predictions of the toxicity of untested compounds. In other words, the purpose of *in silico* toxicology is to make predictions regarding the fate and effects of chemicals starting from what is known about the similar structures. Among the many advantages of *in silico* techniques (cost-effectiveness, time-saving, and reduction in animal use), the most noteworthy benefit lies in its full complementary to the standard testing approaches (i.e., *in vitro* and *in vivo* testing) [20]. This synergy has been also claimed in the regulatory context [21, 22].

6 Classification of Predictive Models

In the field of predictive toxicology, several *in silico* approaches have been developed considering the scientific and economical driving forces in recent years (including governments, academia, and industry), which promote the use of *in silico* methods in toxicology as alternative of *in vivo* methodologies. In their huge number, *in silico* tools vary in complexity and performance. Figure 2 highlights the three major categories of *in silico* toxicology techniques. Depending on the adopted computational tool, and above all on the toxic effect to be investigated, toxicological predictions may be globally subdivided into binary predictions (like those for mutagenic potential which is expressed by a “yes/mutagenic” or “no/not mutagenic”) and into continuous predictions

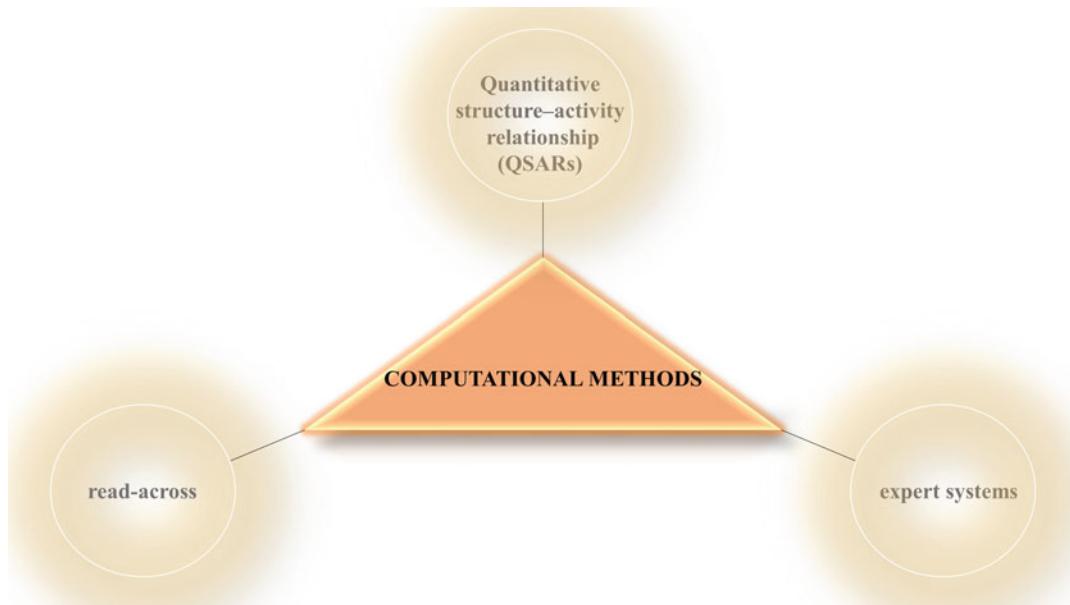


Fig. 2 Principal commonly used in silico toxicology techniques

(like those for quantitatively predicting the toxic dose or potency, such as TD₅₀). The three approaches (QSAR, read-across, and expert systems) belong to the wider group of the so-called non-testing methods, the use of which is nowadays strongly promoted. Non-testing methods are substantially based on the similarity principle, i.e., the hypothesis that similar compounds should have similar biological activities. Before a concise overview of the main characteristic of such computational techniques and a characterization of the main open-source software substantial for computational toxicology, it should be emphasized that the central points on which software lies—irrespective of their typology and characteristic—are databases and molecular descriptors. The former is essential for gathering and storing biological data that contain information on chemicals and their toxicity that are later used for building a prediction model. The latter encodes the chemical information contained in a molecule in order to reduce chemical and biological complexity into an expression useful for the computational prediction. For an in-depth definition about molecular descriptors, the reader is referred to the work of Todeschini et al. [23].

6.1 QSAR

The term Quantitative Structure-Activity Relationship (QSAR) stands for models that predict toxicity and fate of a chemical on the basis of its physicochemical properties. These statistical models return structure-activity predictions based on the belief that the biological activity of a chemical is substantially related to its structure. For instance, QSAR properties such as oil/water partition

coefficient, water solubility, or volatility can quantitatively explain the toxicological potential of a chemical [24–26]. In other words, a QSAR model is a mathematical function (f) that calculates the toxicity (T) of a chemical based on its physicochemical properties (P).

$$T = f(P)$$

These structure-related physicochemical properties are the molecular descriptors that can be calculated through many algorithms. The first step to predict toxicity of a chemical through this approach is building a QSAR model. Initially, information about a toxicological endpoint is collected for a specific group of chemicals (training set). Later, an investigation to establish which chemical property is responsible for toxicity is carried out. This step can be computer-based (generation and calculation of molecular descriptors), or it can be determined experimentally. After choosing descriptors that can properly describe the training set and relate chemical structure to toxicity of interest, several types of algorithms are used to generate QSAR models. In particular, depending on the type of the used algorithm, QSAR models are distinguished into linear and nonlinear models. Successively, the QSAR model is validated externally by a set of compounds with experimentally measured properties that were not used to build the model (external set). Once validated, the QSAR model is considered useful for making predictions for untested chemicals. Since they are developed from a set of compounds, it is important to say that QSARs are local models, which means that the prediction is accepted only when the test compound is similar enough to the training set compounds. If this condition is satisfied, the untested compound is considered within the applicability domain of the QSAR model, i.e., the chemical space within which a QSAR model can be applied to make reliable predictions. Just to mention a few, QSAR models are used for the ecotoxicological prediction of the bioconcentration factor (BCF) [27–29].

6.2 Expert Systems

An expert system has been defined as “any formal system, not necessarily computer-based, which enables a user to obtain rational predictions about the toxicity of chemicals. All expert systems for the prediction of toxicity are built upon experimental data representing one or more toxic manifestations of chemicals in biological systems (the database), and/or rules derived from such data (the rulebase)” [30]. An example of an expert rule might be: IF a compound contains aniline AND metabolic activation is present, THEN the compound is genotoxic, or ELSE non-genotoxic. As a result, these models provide a binary prediction: either the molecule has a toxic fragment—and the compound is predicted as toxic—or it does not have a toxic fragment, and in this case, the compound is predicted as nontoxic. It is evident that expert rules

are based on information from the collective experiences of experts working within the field. Having gathered a list of toxicological data and chemical structures that have been linked by cause-and-effect relationships, toxicologists create rules based on this information. The most known activity rules are the Ashby-Tennant rules [31]. Since expert systems operate with human knowledge stored in the form of rules [32], it is crucial the availability of sufficient data to develop the predictive model. As expert systems are rule-based models, one concept strongly related to this computational approach is that of structural alerts (SAs) [33, 34] because predictions made by this methodology for specific toxicological effects or mechanisms of toxicity take advantages from these rules. SAs (or toxicophores) are defined as molecular substructures that can activate the toxicological effect or mechanism. The concept of SAs follows the same logic flow scheme before mentioned (IF, THEN), and in this way, the prediction of an expert system is built based on a binary classification: when a test compound triggers a structural alert, a positive prediction is generally made; otherwise, a negative prediction may be generated. It is clear that SAs do not attribute to the prediction a mechanism-based rationale of toxicity, but it has the benefit of assigning the toxicity to a compound from knowledge of its chemical structure alone. Toxicology assessment based on the presence or absence of SAs is recommended by the FDA guideline, for instance, in the case of evaluation of the potential genotoxicity and carcinogenicity of impurities which acts via a non-threshold-related mode of action [35]. Pharmaceutical impurities that require control at low levels can be readily identified based on observation of SAs. This is however considered an initial screen because impurities with an identified alert can be prioritized for *in vitro* mutation assay (Ames test). The same filter can be applied for the assessment of ecotoxicological endpoints.

6.3 Read-Across

A concept closely related to the SAs is that of chemical category. This is the cornerstone of read-across method. In fact, the intimate rationale of read-across is that it is possible to make a prediction about a query compound (named the “target compound,” TC) using toxicological data from one or more analogues (named the “source compound(s)”, SCs) provided that they share the same structural feature and the same toxicological pathways (i.e., toxic mechanisms). The group of chemicals formed by the TC and the SCs with these shared properties is named chemical category. Once grouped, read-across can then be applied to arrive at a toxicity prediction of the TC just making usage of experimental data of the SCs. This step is properly called “data gap filling,” and it avoids the experimental studies of the TC based on the evidence that a structurally or toxicologically meaningful relationship between TC and SCs exists. When considering read-across, the similarity between TC and SCs for having a well-founded prediction is not

limited only to the structural resemblance [36] assessed, for instance, by Tanimoto index [37]. Actually, within a chemical category, a chemical is similar to another one by considering also reactivity, toxicokinetics, mechanism/mode of action, physico-chemical properties, and metabolic profile [38–40]. Read-across is formally the simplest method for data gap filling based on the chemical grouping approach. It goes without saying that the chance of making a reliable prediction depends on the available number of SCs used and, above all, on the available number of toxicological-associated data. From this perspective, grouping approaches can be split into analogue and category approaches. When read-across is performed using a very limited number of substances (usually 2), the term analogue approach is used; otherwise, the term category approach refers to a more consistent chemical category with three or more SCs. It must be underlined that also an analogue approach can attribute robustness to read-across when there are high-quality experimental data. For instance, when there is only one SC with data for the specific endpoint of interest, then the read-across may simply be a substitution with the same information if toxicological data of SC are of high quality and complete. In general terms, data gap filling can predict both categorical and continuous endpoints. In this respect and depending upon the mathematical formalism employed, read-across is typically utilized with categorical endpoints, while trend analysis is the method to be applied when dealing with continuous endpoints. The toxicological prediction of chemicals via read-across assessment has been strongly promoted by OECD [41] and ECHA [42] by providing guidelines on the process of performing a valid read-across. This approach, in fact, turns out to be one of the first choices among in silico toxicological assessment for regulatory purposes. Beyond the above-quoted computational methods, also machine-learning techniques are widespread for the prediction of toxicity. The reader is referred to [43–46] for a detailed study. In certain cases, other in silico approaches—typically used in drug discovery programs—are applied for toxicological purposes. Examples include the use of structure-based approaches such as molecular docking (e.g., estrogen receptor-ligand docking) [33, 47–50]. Irrespective of the choice of the in silico approaches to employ, several assessments of key computational aspects and specific issues must be taken into account. Two of these relevant issues, argued later in this review, encompass the data quality of the source compounds used for making prediction and a statistical analysis of the prediction model(s), from which confidence in the predictions depends.

7 T.E.S.T.

Toxicity Estimation Software (T.E.S.T.) is a freely available software tool that has been developed by US EPA. T.E.S.T. allows to estimate toxicity values using multiple advanced QSAR methodologies, listed in Fig. 3: the particularity of T.E.S.T. is that before building the QSAR model from a training set in a classic way, a procedure of hierarchical clustering [51] anticipates this step; hence, each one of six QSAR methodologies is based on a cluster of the training set (clustering methods), so giving the advantage of increasing accuracy to the prediction when they are combined—as it occurs in the consensus method. When the consensus method is selected for the prediction, an average of the predicted toxicities of the other QSAR methods is used for the estimation of toxicity, while for each method, the uncertainty in the overall prediction is always calculated. In addition to toxicological endpoints, T.E.S.T. gives the opportunity to estimate also physicochemical properties, as shown in Fig. 3. The ecotoxicological endpoint implemented in T.E.S.T. is the *Daphnia magna* LC₅₀ endpoint which is model built from a data set obtained by a refined ECO-TOX aquatic toxicity database [52]. The software gives also the opportunity to calculate the bioconcentration factor (BCF) using a dataset of chemicals taken from various filtered databases [53–55]. It should be pointed out that the software gives back a prediction report only if the query compound is into the AD of the selected model.

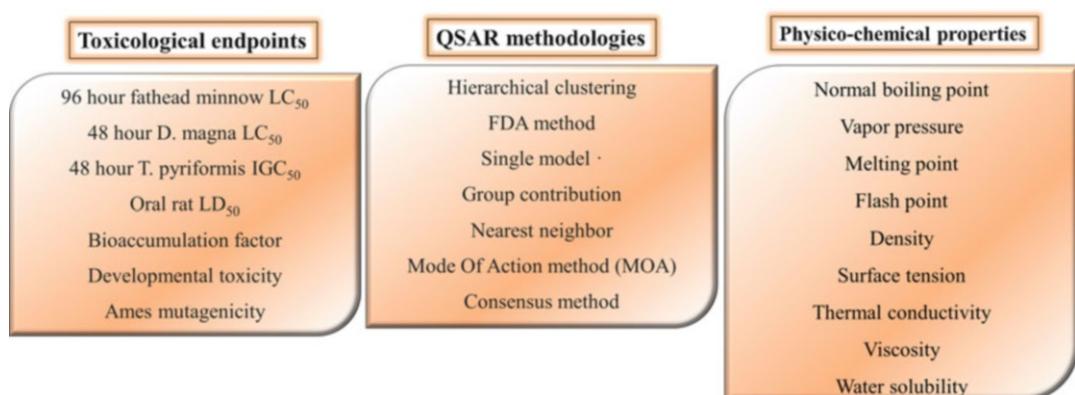


Fig. 3 List of toxicological endpoints, QSAR methodologies, and physicochemical properties implemented in the T.E.S.T. software

8 VEGA Platform

The VEGA platform is a free-available software that arises from an EC-funded project called CAESAR [56]. Besides inheriting the works of CAESAR project, the platform implements also models derived from other sources, such as ISS (Istituto Superiore di Sanità) and US EPA (US Environmental Protection Agency). The VEGA platform encases different series of QSAR models suitable for regulatory purposes: these, listed in Fig. 4, are able to earn biological, environmental, and physicochemical properties from



Fig. 4 List of toxicological endpoints, QSAR methodologies, physicochemical properties, environmental QSAR models, and physicochemical QSAR models implemented in the VEGA software

the structure of the query compound. Juxtaposed to other platforms, the robustness of VEGA mostly relies in the completeness of detailed information included in the format of the predictive dossier, structured in order to apply for regulatory purposes. This can be attractive for stakeholders who must submit information about the safety of the chemical substances before marketing, as the REACH regulation requires [57, 58]. The comparison and statistical analysis between commercial and freely available software conducted by Golbamaki et al. [59] revealed that for the prediction of acute toxicity in *Daphnia magna*, the performance of VEGA and T.E.S.T. in making predictions is parallel to those of the commercial software despite the improvement of the QSARs is however recommended.

9 LAZAR

LAZAR (Lazy Structure-Activity Relationships) is a freely available tool that uses a lazy machine-learning technique for the prediction of several toxicological endpoints. The lazy machine-learning technique is a specific technique whose main characteristic is the no need of continuously updating the predictive model. LAZAR resembles a read-across procedure and combines it with a QSAR model. On one side, it utilizes a local QSAR model using only similar compounds of the query from the training set. Based on statistical criteria, LAZAR derives its prediction specifically for a query structure using a modified k-nearest neighbor (k-NN) algorithm. The advantage of LAZAR is that it provides an easy to interpret and complete prediction report, with a detailed prediction result on the top and an intelligible list of similar compounds with relative experimental activity on the bottom. Another advantage of LAZAR is its possibility to predict more (eco)toxicological endpoints together, embracing both those regarding human health effect and those regarding environmental fate. The toxicological endpoint incorporated in LAZAR is listed in Table 2. About the ecotoxicological evaluation of a substance, LAZAR gives the opportunity of predicting the acute toxicity both in *Daphnia magna* and fathead minnow.

10 QSAR Toolbox

QSAR Toolbox, developed by the Laboratory of Mathematical Chemistry (LMC, Burgas University, Bulgaria), is a free-available software arising from a project developed under the guide of the Organisation for Economic Co-operation and Development (OECD, Paris) in collaboration with the European Chemicals Agency (ECHA, Helsinki). QSAR Toolbox (referred to hereafter

Table 2
(Eco)toxicological endpoints available in LAZAR

(Eco)toxicological endpoints	Species
Acute toxicity	<i>Daphnia magna</i>
Acute toxicity	<i>Fathead minnow</i>
Blood-brain barrier penetration	<i>Human</i>
Carcinogenicity	<i>Rat</i>
Carcinogenicity	<i>Rodents</i>
Carcinogenicity	<i>Mouse</i>
Carcinogenicity (TD50)	<i>Mouse</i>
Carcinogenicity (TD50)	<i>Rat</i>
Lowest observed adverse effect level (LOAEL)	<i>Rat</i>
Maximum recommended daily dose	<i>Human</i>
Mutagenicity	<i>Salmonella typhimurium</i>

as Toolbox, TB) is used by governments, chemical industry, and other stakeholders in filling gaps in (eco)toxicity data needed for assessing the hazards of chemicals by the mean of read-across. In terms of existing tools to assist current read-across approaches for regulatory purposes, TB (current version 4.3) is perhaps the most widely used. The TB workflow follows the grouping concept. The latter term indicates the operation of grouping chemical into chemical categories in order to predict (eco)toxicity of the target chemical. To reach this aim, TB follows a workflow schematized in six steps that are all included in the GUI as six modules. The software identifies the relevant structural characteristics and potential mechanism or mode of action of a target chemical. Subsequently, it finds other chemicals that have the same structural characteristics and/or mechanism or mode of action, and finally, it uses the existing experimental data of the similar chemicals to fill the data gap by read-across and to predict the toxicological potential of query chemicals. The six modules of the TB software application are below described.

10.1 Input Module

In the input module, the user specifies the target chemical that will be the object of the prediction. For this procedure, the software gives the possibility to specify the identity of the target chemical by several options, such as:

- Specify target chemical by entering the chemical name.
- Specify target chemical by entering the registry number (CAS number).

- Specify target chemical by drawing the molecule.
- Specify target chemical by entering the SMILES.
- Specify target chemical by selecting it from a list, a database, an inventory, or a file).

Such opportunities, already, underline the regulatory purpose for which the software was developed: in fact, the CAS number is often used by regulatory agencies, in addition to the chemical names. However, for the correct workflow of TB, it is very important that not so much the chemical name but the chemical structure is appropriate for the target chemical; for the prediction, precise structural information is needed. In some cases, indeed, there is discrepancy in regulatory inventories because not always structural information is provided; it is not rare that the same structure may be associated with several CAS numbers. Supposing that the user has entered the target chemical by submitting the SMILES, the quality of substance identification, based on this two-dimensional connectivity, may be affected by several factors (such as the trouble of a good representation of salts by SMILES). To face this obstacle, the software provides information on the quality of the structure generated. In fact, if the target chemical is listed in the chemical inventories/databases implemented in the TB and if different CAS numbers are associated with it, TB analyzes all possible combinations between the SMILES entered and the different CAS numbers discovered in the databases. These combinations are called CAS-SMILES relations, which qualitatively rate the chemical structures identified by the software (starting from the SMILES notation). The chemical structures supplied with their CAS-SMILES relations are to be chosen by the user in the input module. The rating of the chemical structure is based on the quality of the data sources (i.e., databases or inventories) in which TB has found the target chemical. Figure 5 shows the CAS-SMILES relations that are presented by the software and that will have to be selected by the user. At this regard, TB is mostly a user-dependent predictive tool. From the choice of the chemical category to the method of data gap filling, the prediction within QSAR Toolbox is user-dependent, in such a manner that the predictive result may be affected by the level of experience in the (eco)toxicology field and software knowledge of the user. This aspect involves that the quality and reliability of the results of the read-across rely on the knowledge, experience, and skills of the user, giving a certain degree of uncertainty to prediction. Such matter was handled with the standardized and automated workflows, new features available from version 4.0 of TB. The standardized and automated workflows [60] assist the user with the prediction of TB. At the moment, such features are available for two (eco)toxicological endpoints (aquatic toxicity and skin sensitization). The standardized and automated workflows

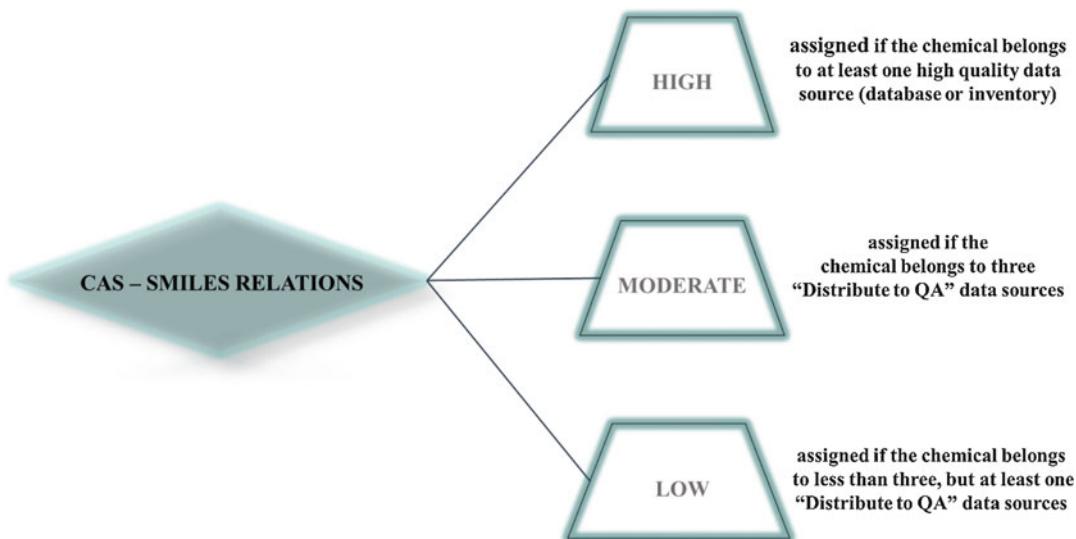


Fig. 5 QSAR Toolbox's establishment of correlation between CAS number and SMILES notation based on the quality of data sources with which the chemical is affiliated

provide the opportunity to pass directly from input module to the data gap-filling module with few (standardized) or none (automated) user intervention. Eventually the user can enter only one structure for the subsequent workflow and a group of chemicals from specialized databases implemented in the software or from a user list/inventory.

10.2 Profiling Module

After entering the target chemical in the input module, TB collects as many information as possible about the substance in the profiling module: this is a mandatory step to properly find similar source chemicals related to a given query. The "profiling module" identifies the main characteristics of the target chemical. In other words, they consist of chemical or biological properties of the target chemical that the software calculates starting from the chemical structure. It should be noted that the results of the profiling are not intended directly for a toxicological prediction. The outcome of the profiling module represents a rule-based addressing for the building of the chemical category. Profilers are effectively rulebases of SARs, and each profiler consists of a system of rules that serve as criteria in the "category definition" module. These sets of rules are predefined categories, established on chemical substructures (the toxicity alerts), developed by recognized institutions or organizations. According to the underlying data from which they were developed, profilers are globally divided into:

- Chemical profilers, which describe general chemical properties of the molecule

- Mechanistic profilers, which are related to specific modes of action or molecular initiating event responsible for a toxicological effect
- Endpoint specific profilers, which collect structural alerts that have been shown to be associated with specific toxicological endpoints

Table 3 lists the profilers implemented in QSAR TB and highlights the most relevant ones for ecotoxicological evaluation. The importance of the outcomes of the profilers relies in its guidance for the user to the subsequent steps of the workflow, in particular for category definition. The outcomes, indeed, help the user to correctly choose the chemical category to which the target chemical will belong. Stated otherwise, the outcome of this module also determines the most appropriate way to search for analogues.

10.3 Data Module

In the *data module*, the user can retrieve experimental data from different databases that are available in the toolbox. Based on the assumption that TB stores several databases donated from organizations, the information collected within this module should be checked by the user aiming at verifying the quality of data. Like in the profiling module, the latest version of the TB (starting from v4.0) facilitates the user's decision by highlighting databases with three different colors (green, orange, and white).

In accordance with the type of information provided, databases in TB are collected and showed to the user by grouping them in four sections, listed in Table 4.

- Physical-chemical properties
- Environmental fate and transport
- Ecotoxicological information
- Human health hazards

Together with the database, TB provides a list of inventories in which, however, no experimental data are reachable. The aim of this module is twofold. On one side, it collects all available information for the target chemical from the databases selected by the user; on the other side, databases from which source chemicals—that further form the chemical category—will be picked up are here selected. In fact, the same databases here chosen will be the ones from which the software will search the source chemicals in the “category definition” module. Once the databases are selected, the user has the possibility to collect all available experimental data for the target chemical or to retrieve only the experimental data concerning the endpoint of interest.

10.4 Category Definition Module

The “category definition” module is the most important step in TB workflow, because in this section, the user groups substances into chemical categories. It is a crucial step since it will affect the final

Table 3

Profiling schemes implemented in the “profiling module” of QSAR Toolbox. Relevant profilers for ecotoxicological endpoints are reported in italic

Predefined profilers	General mechanistic profilers	Endpoint-specific profilers	Empirical profilers	Toxicological profilers
Database affiliation	<i>Biodegradation probability (Biowin 1)</i>	<i>Acute aquatic toxicity classification by Verhaar</i>	Chemical elements	Repeated dose (HESS)
Inventory affiliation	<i>Biodegradation probability (Biowin 2)</i>	<i>Acute aquatic toxicity MOA by OASIS</i>	Groups of elements	
OECD HPV Chemicals categories	<i>Biodegradation probability (Biowin 5)</i>	<i>Aquatic toxicity classification by ECOSAR</i>	Lipinski Rule OASIS	
Substance type	<i>Biodegradation probability (Biowin 6)</i>	<i>Bioaccumulation—metabolism alerts</i>	Organic functional groups	
US EPA new chemical categories	<i>Biodegradation probability (Biowin 7)</i>	<i>Bioaccumulation—metabolism half-lives</i>	Organic functional groups (nested)	
	DNA binding by OECD	<i>Biodegradation fragments (Iowan MITI)</i>	Organic functional groups (US EPA)	
	Estrogen receptor binding	DNA alerts for AMES, MN, and CA by OASIS	Organic functional groups, Norbert Haider (checkmol)	
	Primary Biodeg (Biowin 3)	DPRA cysteine peptide depletion	Tautomers unstable	
	Protein binding by OASIS	DPRA lysine peptide depletion		
	Protein binding by OECD	Keratinocyte gene expression		
	Ultimate Biodeg	Eye irritation/corrosion Exclusion rules by BfR		
	Ultimate Biodeg (Biowin 4)	Eye irritation/corrosion Inclusion rules by BfR		
	DNA binding by OASIS	Carcinogenicity (genotox and nongenotox) alerts by ISS		
	<i>Hydrolysis half-life (Ka, pH 7)</i>	In vitro mutagenicity (Ames test) alerts by ISS		

(continued)

Table 3
(continued)

Predefined profilers	General mechanistic profilers	Endpoint-specific profilers	Empirical profilers	Toxicological profilers
	<i>Hydrolysis half-life</i> (K_a , pH 8)	In vivo mutagenicity (micronucleus) alerts by ISS		
	<i>Hydrolysis half-life</i> (K_b , pH 7)	Oncologic primary classification		
	<i>Hydrolysis half-life</i> (K_b , pH 8)	Skin irritation/ corrosion Exclusion rules by BfR		
	<i>Hydrolysis half-life</i> (pH = 6.5–7.4)	Skin irritation/ corrosion Inclusion rules by BfR		
	Ionization at pH = 1			
	Ionization at pH = 4			
	Ionization at pH = 7.4			
	Ionization at pH = 9			
	Protein binding potency			
	Toxic hazard classification by Cramer (original)			
	Toxic hazard classification by Cramer (with extension)			

result of the prediction. The building of chemical category in this module is user-dependent, so there is no a universal approach, but the user chooses the most appropriate principal chemical category according to the profiler outcomes. It must be noted, indeed, that profilers and chemical categories are the same: the rationale is that when a chemical category (i.e., a profiler)—which has depicted a toxic alert within the target chemical in the profiling module—is chosen, the software will search analogues having the same structural alert (without prejudicing the picked grouping options). Considering the complexity of the operation (which is case-by-case different), TB has adopted a color coding system as new functionality from version 4.0 and later. When the endpoint of interest is selected in the data matrix, the most suitable profilers

Table 4

Databases implemented in the “data module” of QSAR Toolbox. Relevant databases for ecotoxicological endpoints are reported in italic

Databases			
Physical-chemical properties	Environmental fate and transport	Ecotoxicological information	Human health hazards
Chemical reactivity COLIPA	Bioaccumulation Canada	Aquatic ECETOC	Acute oral toxicity
ECHA CHEM	Bioaccumulation Fish CEFIC LRI	Aquatic Japan MoE	Bacterial mutagenicity ISSTY
Experimental pKa	Bioconcentration NITE	Aquatic OASIS	Biocides and plant protection ISSBIOC
GSH Experimental RC50	Biodegradation in soil OASIS	ECHA CHEM	Carcinogenic Potency Database (CPDB)
Phys-Chem EPISUITE	Biodegradation NITE	Ecotox	Carcinogenicity and mutagenicity ISSCAN
	Biota-sediment Accumulation Factor US-EPA	Food Tox Hazard EFSA	Cell transformation assay ISSCTA
	ECHA CHEM		Dendritic cells COLIPA
	ECOTOX		Developmental and reproductive toxicity (DART)
	Hydrolysis rate constant OASIS		Developmental toxicity database (CAESAR)
	kM database environment Canada		Developmental toxicity ILSI
	Phys-Chem EPISUITE		ECHA CHEM
	REACH Bioaccumulation database (normalized)		ECOTOX
		Eye irritation ECETOC	
		Food Tox Hazard EFSA	
		GARD Skin sensitization	
		Genotoxicity and carcinogenicity ECVAM	
		Genotoxicity OASIS	
		Genotoxicity pesticides EFSA	
		Human half-life	

(continued)

Table 4
(continued)

Databases			
Physical-chemical properties	Environmental fate and transport	Ecotoxicological information	Human health hazards
		Keratinocyte gene expression Givaudan	
		Keratinocyte gene expression LuSens	
		Micronucleus ISSMIC	
		Micronucleus OASIS	
		MUNRO non-cancer EFSA	
		REACH skin sensitization database (normalized)	
		Receptor-mediated effects	
		RepDose Tox Fraunhofer ITEM	
		Repeated dose toxicity HESS	
		Rodent inhalation toxicity Database	
		Skin irritation	
		Skin sensitization	
		Skin sensitization ECETOC	
		ToxCast DB	
		Toxicity Japan MHLN	
		Toxicity to Reproduction (ER)	
		ToxRefDB US EPA	
		Transgenic rodent database	
		Yeast estrogen assay database	
		ZEBET database	
<i>Inventories</i>			
Canada DSL			
COSING			
DSSTOX			
ECHA PR			

(continued)

Table 4
(continued)

Databases	Physical-chemical properties	Environmental fate and transport	Ecotoxicological information	Human health hazards
EINECS				
HPVC OECD				
METI Japan				
NICNAS				
REACH ECB				
TSCA				
US HPV Challenge Program				

for the target endpoint that will be used for grouping chemicals are suggested by the software by highlighting them with three different colors:

- Green (suitable): the category is appropriate for the endpoint of interest.
- Orange (plausible): the category is somehow related to the endpoint of interest.
- White (unclassified): the category is not related to the endpoint of interest.

Given the user-dependent building of the chemical category, it is important that a consistent and reproducible approach is applied. To achieve this, several recommended approaches are suggested [41, 61] in order to reach a meaningful chemical category that will furnish a consistent predictive result. The most important recommendation is the subcategorization of the chemical category. Initially, in fact, a category is built in a stepwise manner, with no restricted conditions. Generally, as a good starting point, two wide groups of chemical profilers (US EPA new chemical categories and the organic functional groups) can be adopted. This will allow an initial collection of a large amount of non-endpoint-specific similar compounds and the construction of the so-called primary category. With the term sub-categorization, it is meant the electronic process of consecutive narrowing down and refinement of the retrieved analogues. First of all, a preliminary exclusion of compounds having additional functionalities/structural alerts different from those of the target chemical is possible. After the elimination of first analogues, a subsequent refinement by applying mechanistic and

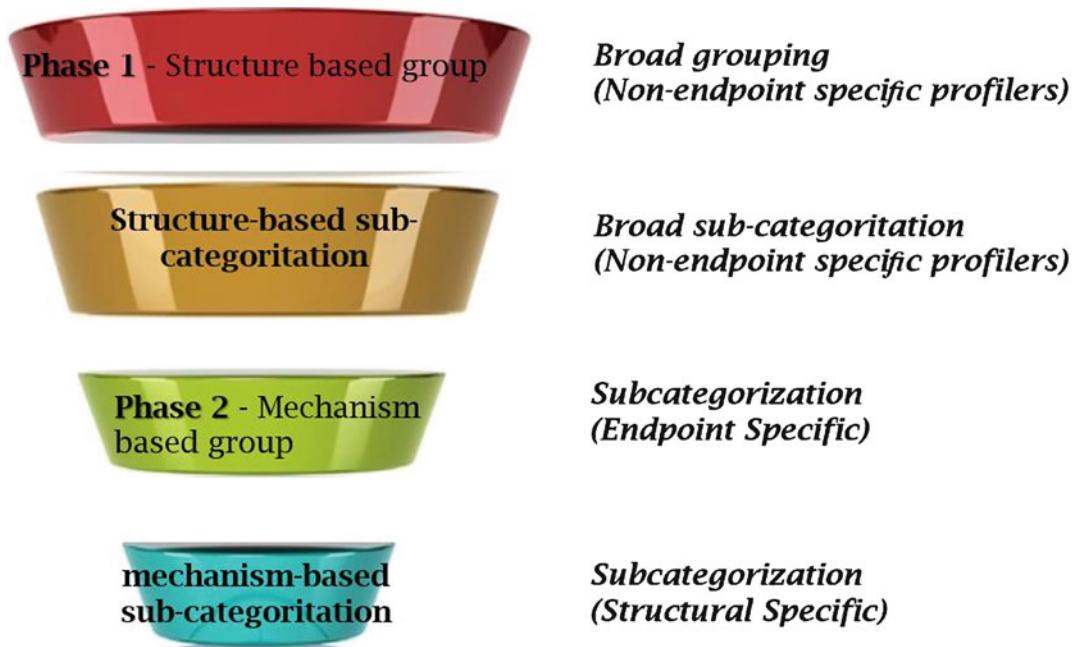


Fig. 6 General scheme of subcategorizations of chemical categories required for a more accurate QSAR Toolbox's (eco)toxicological prediction

endpoint-specific profilers is advised. It will consent to form the so-called secondary category, which is a more robust and consistent group of chemicals. The process of recommended subcategorization steps is illustrated in Fig. 6. It is noteworthy that the subcategorization process is also available in the “data gap-filling” module, so the user carries out this operation both in this module and later on. For the sake of clarity, performing the subcategorization in the “data gap-filling” module gives the chance to graphically visualize step-by-step the chemicals removed, in order to monitor step-by-step how the prediction improves/worsens with subcategorization. In connection with this last point, not always the subcategorization is necessary, for the reason that a first broad category might be sufficient to read-across. When the user has selected the principal chemical category and filters of subcategorization have been applied, analogues of the query (i.e., the source chemicals) are searched in the previously selected databases, and experimental data are automatically retrieved and added into the data matrix.

10.5 Data Gap-Filling Module

The “data gap-filling” module, among the software workflow, is the step that gives back the result of the prediction in TB. In fact, based on the experimental results of the members of the chemical category, it is possible to “fill the gap” of the sole chemical for which test results are not available (i.e., the target chemical). Three approaches of data gap filling are given to the user: Fig. 7 shows

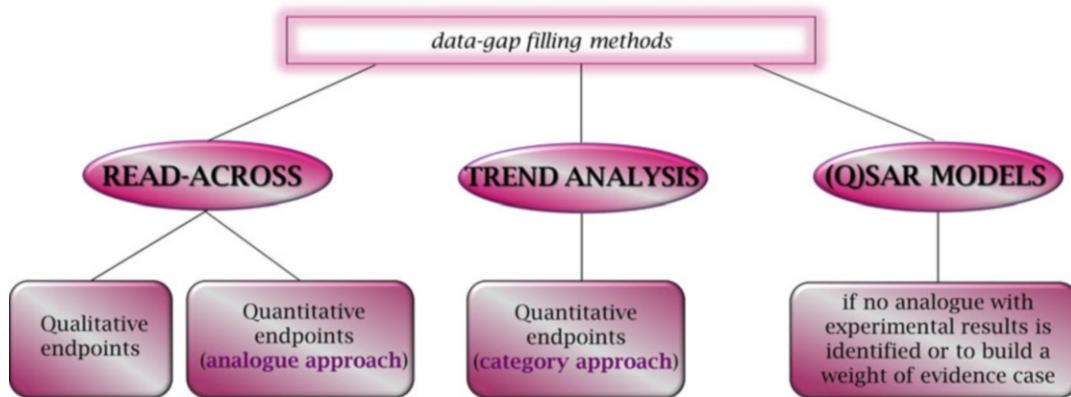


Fig. 7 The three principal data gap-filling methods available in the “data gap module” of QSAR Toolbox software. Conditions under which the methods are to be used are also reported

the different data gap-filling methods. When prediction is made by applying read-across or trend analysis, the graphical user interface of TB returns a plot in which the appropriate descriptor of the category members (K_{ow} by default) is stored in the x-axis and the endpoint object of the in silico evaluation is stored in the y-axis. Moreover, the prediction result is fulfilled with all statistical parameters (as, for instance, confidence interval correlation, regression analysis, and so on). A relevant operation that is included in this module is the data transformation, which is the operation of converting all experimental data in one reference unit/scale. This is included in TB for two principal reasons. First, certain values, rather than others, are statistically more appropriate (i.e., logarithmic scales for regression models) [62]. Second, this is made because experimental data in different databases could be reported with different units (since a toxic endpoint can be measured by different units).

10.6 Report Module

Once the prediction is accepted in the “data gap module,” the software produces in the “report module” a report file of the predictive result in a semiautomatically way. It means that many fields in the reports are automatically filled by TB with a predefined template, while there are other manually editable fields that the user should complete to justify the procedure. The standard formats of the reports can be divided into three types:

- The Chemical Category Reporting Format (CCRF)
- The Toolbox Prediction Reporting Format (TPRF)
- The Toolbox Model Reporting Format (TMRF)

For the completeness of information providing the TB reports (applicability domain, information about the group members, and so on), the prediction reports are promptly employable for

regulatory submissions. This is also due to the standard format with which the document of prediction is made, a characteristic that makes TB reports suitable for regulatory purposes.

11 Key Factors for a Reliable In Silico Prediction

Of utmost importance in the regulatory context, prediction via computational tools needs several assessments and considerations. Reliability is used to support any hazard assessment, and it must be an essential feature of any experimental study and/or in silico analysis. In this regard, it is worth mentioning the reliability score (RS) or Klimisch score, introduced by Klimisch and co-authors [63]. Ranging from 1 to 4, the Klimisch score attributes to a toxicological study several degrees of reliability (a Klimisch score of 1 means that the study is reliable without restriction because it has been generated according to generally valid and/or internationally accepted testing guidelines; otherwise, a Klimisch score of 4 represents the worst case since the conducted study does not give sufficient experimental details). Even if this kind of classification of testing data does not directly concern computational tools, they could be conceptually translated into in silico predictions, being a measure of assessment about their reliability. In fact, a prediction carried with the use of experimental data having a good RS could be considered of good reliability. The Klimisch score can be found as a standard field within the IUCLID database: the latter is implemented in the TB software. Apart from the Klimisch score, another important issue for a model validation (in special case of toxicological assessment through QSAR and read-across) is the reliability of prediction which is assessed on the basis of the applicability domain (AD) representing the physicochemical, structural, or biological space, within which all chemicals of the TS are enclosed. The AD of a model is very useful to define boundaries, whereby the obtained predicted values can be trusted with confidence [64]. It goes without saying that a QSAR is considered valid if it has a defined AD. In other words, the reliability of a prediction—both using QSAR models and read-across—increases if the compound for which a computer-aided toxicological evaluation is needed falls within the AD of the model.

12 Problem of False Negatives (FNs)

The quality of predictions is assessed by considering some statistic parameters. Among them, the most popular are the sensitivity and specificity.

$$\text{sensitivity (SE)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \cdot 100$$

$$\text{specificity (SP)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \cdot 100$$

In these expressions—taking, for example, a categorization of compounds into toxic and not toxic by means of in silico tools—TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively, whereas:

- TP are chemicals predicted as toxic and that really act as toxic compound during in vitro/in vivo experiments.
- TN are chemicals predicted as not toxic and that really act as not toxic compound during in vitro/in vivo experiments.
- FP are chemicals predicted as toxic and that really act as not toxic compound during in vitro/in vivo experiments.
- FN are chemicals predicted as not toxic and that really act as toxic compound during in vitro/in vivo experiments.

Sensitivity and specificity represent the statistical measures indicating the goodness of fit of an in silico model [65]. Having said that, the major concern about all predictive computational models, especially if their relevance in regulatory toxicology is recognized, is the need to minimize the number of FNs [25, 26, 66, 67].

For example, if a rule-based expert system (such as Derek®, COMPACT®, or ToxAlert®) is used for a toxicological prediction, the risk in which one could incur is that the number of FNs may be large because of incomplete list of SAs and rules [68].

Looking at this matter, computational modelers should pay attention to prevent the number of FNs that are hazardous molecules misclassified in the first instance as safe. This may be seen as the most severe challenge within in silico world, mainly for two interrelated reasons. First, biological entities—being nonlinear systems—show a chaotic behavior that cannot be fully represented in a binary way; second, biological systems are algorithmically incompressible, meaning that they cannot be properly modeled by a single algorithm.

13 Limits of Predictive Models

In the last years, scientific world—including academic, industrial, and regulatory fields—has turned to a new direction through a changing mind from the recourse of only in vitro and in vivo methods to an increasingly consolidated use of in silico alternatives: the focus is to find new testing strategies that attain the goal of protecting human health and the environment from toxic effects. Despite the significant advancements, the use of in silico tools is still

taken cautiously. There might be several reasons why a single predictive approach becomes unusable, for instance:

- The forced passage from the complexity of a living system to the “simplicity” of a computer simulation that cannot always portray the biological reality in an exhaustive manner
- The restricted applicability domain of training data sets as intrinsic limit of *in silico* models
- The availability of complete and curated databases and of high-quality bioassay data
- The availability of well-established free-of-charge computer-aided tools

For brevity, in this review, only two aspects are taken into consideration: the issue of data sharing and that of commercial software packages. Furthermore, a way to face the problems is proposed.

14 Open-Source Computational Tools

“Open data” represents the philosophy and practice requiring that certain data are freely available without restrictions from copyright or patents. The term “open source” is used to denote software whose source code is published and made available and which grants the rights to copy, modify, and redistribute the source code without fees. A wide variety of both publicly available and commercial computational tools has been developed in years, and nowadays, open-source software are widely used in both academic and commercial environments. Such tools include methods for data management and data mining, descriptor generation, molecular similarity analysis, and hazard assessment. Despite the different open-source solutions proposed, which extended from simple stand-alone and web applications to full-defined tools, there is still the need to develop a range of open-source software, in particular of those which could be fitting in the regulatory process. A wide variety of publicly available and commercial computational tools has been developed that are suitable for the development and application of QSARs. Due to the limited availability of freely accessible *in silico* software, there is a need to develop a plethora of open-source tools, which should be employed in the regulatory process. A quick-fix strategic approach tailored to this purpose is represented by the QSAR Application Toolbox and the VEGA platform introduced in this review.

15 Data Sharing

One important component of modern scientific work is the collection, analysis, and sharing of data. Unquestionably, within in silico predictive toxicology, a starting point to build predictive models is given by a collection of adequate in vivo/in vitro assays [69, 70] as well as chemical data (like molecular structure and physicochemical properties). It is undeniable a fact that for the in silico toxicological assessment, the availability of curated data is the basis to build a predictive model along with ensuring accuracy and completeness of prediction. In this sense, data sharing is always a growing issue. Indeed, given also the recent trends in academic research [71], demands for data-sharing systems are becoming widespread. Many efforts have been made for building high-quality databases publicly available, such as the Structure-Activity Relationships Database Project Committee [72] and the IUCLID database system [73]: the latter is also implemented in TB software. Table 5 shows some of publicly available databases befitting for computer-aided toxicological assessment. However, despite significant progress in recent years with respect to the promotion of publicly available toxicity data, much high-quality toxicity information continues to be locked away within organizations. The availability of high-quality toxicity data is of important concern for ensuring the reliability of computational predictive tools. Anyway, also contemplating that data sharing is anything but a simple process [74, 75], many data remain largely inaccessible, also due to confidentiality restrictions. The problem of data sharing could be also a critical issue for the toxicological assessment of a chemical. For example, the weight of evidence (WOE) approach [76–78] makes use of existing data; if suitable data exist for a compound—and they are enough for an exhaustive toxicological evaluation—there should be no requirement to initiate a new test or make a new prediction.

Table 5
Freely available databases useful for (eco)toxicological assessment

Name	Link	Ref.
Comparative Toxicogenomics Database (CTD)	http://ctdbase.org/	[89]
ChemSpider	http://www.chemspider.com/	[90]
ChEMBL	https://www.ebi.ac.uk/chembl/	[91]
Distributed Structure-Searchable Toxicity Database (DSSTox)	https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database	[92]

16 Consensus Predictions

Using a computational tool for making a toxicological prediction, independent *in silico* models may be selected. They may be different for the predictive method or for the data used. Currently, there are no perfect *in silico* alternatives as full replacement of all testing method is employable for the assessment of a specific hazard. In regulatory context, in fact, computer-aided approach is intended mainly to prioritize substances to be tested for the hazard characterization [79]. The consensus approach is intended as the strategy of combining multiple individual models in order to improve the single final prediction [80–82]. A consensus prediction can be helpful even when different predictive models reach an ambiguous conclusion [83, 84] (as in the case of predictions outside the AD) in order to aid the assessment of the chemical toxicity within an integrated testing strategy. To cite just an example, in the evaluation of carcinogenic potential, Lewis et al. [85] verified that the combined use of two software (COMPACT and HazardExpert, both expert systems) reaches the 100% of correct predictions, while alone, the software get the 71% and 57%, respectively. Consensus positive or negative results can be considered high confidence predictions. Moreover, consensus approach can reduce the appeal to an additional human expert evaluation. This consideration is also highlighted in the ICH guidelines [86] for the safety of impurities which classify them just in view of consensus approach. According to the guidelines, in fact, the range spaces from Class 5 impurities (i.e., non-mutagenic impurities) (if the consensus negative prediction is resulted) to Class 1 impurities (i.e., known mutagenic impurities) (if a consensus positive prediction is resulted) [86, 87]. On the other hand, compounds predicted with a lack of consensus would demand an in-depth evaluation using human expert knowledge [88].

17 Conclusion

In the last years, chemicals have been widely detected at trace levels in various aquatic environments. Because they might be biologically active compounds, designed to interact with specific pathways/processes in target humans and other animals, concerns have been raised over the potential side effects of these substances in the environment. This consideration is meaningful also for impurities arising from pharmaceutical manufacturing process and in particular for some environmental-problematic solvents. Despite the significant advances of several techniques for the quality and safety monitoring of impurities, new methods can be explored to identify a possible toxicological potential of pharmaceutical impurities. The

field of in silico toxicology has been in a continuous development, and it represents a key strategy to reduce the long timelines and spiraling cost since they are able to reliably estimate toxicity of chemicals. The use of an implementation and integration of different computational methods within a consensus approach can improve prediction and be applied for regulatory purposes. This review briefly surveys different free-available in silico methods that in the future can replace battery of in vitro and in vivo toxicity tests, and they can be seen as an alternative approach to protect ecosystems from the threat posed by the presence of chemicals in the environment.

Dedication

To the memory of Michele Montaruli, exceptionally gifted PhD student who has always devoted his life to serving others. To you, Michele, our huge embrace.

References

- Sheldon RA (Delft U of T (Netherlands)) (1994) Consider the environmental quotient. *CHEMTECH U S* 24:3
- Sheldon RA (2005) Green solvents for sustainable organic synthesis: state of the art. *Green Chem* 7:267–278. <https://doi.org/10.1039/B418069K>
- Halling-Sørensen B, Nors Nielsen S, Lanzky PF et al (1998) Occurrence, fate and effects of pharmaceutical substances in the environment- a review. *Chemosphere* 36:357–393. [https://doi.org/10.1016/S0045-6535\(97\)00354-8](https://doi.org/10.1016/S0045-6535(97)00354-8)
- Dietrich DR, Webb SF, Petry T (2002) Hot spot pollutants: pharmaceuticals in the environment. *Toxicol Lett* 131:1–3. [https://doi.org/10.1016/S0378-4274\(02\)00062-0](https://doi.org/10.1016/S0378-4274(02)00062-0)
- Kot-Wasik A, Jakimska A, Śliwka-Kaszyńska M (2016) Occurrence and seasonal variations of 25 pharmaceutical residues in wastewater and drinking water treatment plants. *Environ Monit Assess* 188. <https://doi.org/10.1007/s10661-016-5637-0>
- Ying G-G, Zhao J-L, Zhou L-J, Liu S (2013) Fate and occurrence of pharmaceuticals in the aquatic environment (surface water and sediment). In: *Comprehensive Analytical Chemistry*. Elsevier, pp 453–557
- Ferrari B, Paxéus N, Giudice RL et al (2003) Ecotoxicological impact of pharmaceuticals found in treated wastewaters: study of carbamazepine, clofibrate acid, and diclofenac. *Ecotoxicol Environ Saf* 55:359–370. [https://doi.org/10.1016/S0147-6513\(02\)00082-9](https://doi.org/10.1016/S0147-6513(02)00082-9)
- Barra Caracciolo A, Topp E, Grenni P (2015) Pharmaceuticals in the environment: biodegradation and effects on natural microbial communities. *Rev J Pharm Biomed Anal* 106:25–36. <https://doi.org/10.1016/j.jpba.2014.11.040>
- aus der Beek T, Weber F-A, Bergmann A et al (2016) Pharmaceuticals in the environment-global occurrences and perspectives: pharmaceuticals in the global environment. *Environ Toxicol Chem* 35:823–835. <https://doi.org/10.1002/etc.3339>
- Henschel K-P, Wenzel A, Diedrich M, Fliedner A (1997) Environmental hazard assessment of pharmaceuticals. *Regul Toxicol Pharmacol* 25:220–225. <https://doi.org/10.1006/rtpb.1997.1102>
- Cleuvers M (2003) Aquatic ecotoxicity of pharmaceuticals including the assessment of combination effects. *Toxicol Lett* 142:185–194. [https://doi.org/10.1016/S0378-4274\(03\)00068-7](https://doi.org/10.1016/S0378-4274(03)00068-7)
- Chatzitakis A, Berberidou C, Paspaltsis I et al (2008) Photocatalytic degradation and drug activity reduction of chloramphenicol. *Water Res* 42:386–394. <https://doi.org/10.1016/j.watres.2007.07.030>
- Méndez-Arriaga F, Esplugas S, Giménez J (2008) Photocatalytic degradation of non-steroidal anti-inflammatory drugs with

- TiO₂ and simulated solar irradiation. *Water Res* 42:585–594. <https://doi.org/10.1016/j.watres.2007.08.002>
14. Hughes Mike, Health JBS of P. The principles of humane experimental technique: preface. In: Johns Hopkins Bloom. Sch. Public Health. http://altweb.jhsph.edu/pubs/books/humane_exp/addendum. Accessed 28 May 2019
 15. Abraham J (2009) International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. In: Brouder A, Tietje C (eds) Handbook of transnational economic governance regimes. Brill, pp 1041–1054
 16. Prat D, Pardigon O, Flemming H-W et al (2013) Sanofi's solvent selection guide: a step toward more sustainable processes. *Org Process Res Dev* 17:1517–1525. <https://doi.org/10.1021/op4002565>
 17. Fontaine N, Reynders D (2001) Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the community code relating to medicinal products for human use. *Off J Eur Commun L* 311:67–128
 18. Convention USP (2009) USP NF 2009. United States Pharmacopeial Convention. https://www.uspnf.com/sites/default/files/usp_pdf/EN/USPNF/usp-nf-notices/usp38_nf33_gn.pdf
 19. Alsante KM, et al (2001) Isolation and identification of process related impurities and degradation products from pharmaceutical drug candidates, Part I. *Am Pharm Rev* 4:70–78
 20. Mangiatordi GF, Alberga D, Altomare CD et al (2016) Mind the gap! A journey towards computational toxicology. *Mol Inform* 35:294–308. <https://doi.org/10.1002/minf.201501017>
 21. Nicolotti O, Benfenati E, Carotti A et al (2014) REACH and in silico methods: an attractive opportunity for medicinal chemists. *Drug Discov Today* 19:1757–1768. <https://doi.org/10.1016/j.drudis.2014.06.027>
 22. Gissi A, Mangiatordi GF, Sobański T et al (2017) Nontest methods for REACH legislation. In: Comprehensive medicinal chemistry III. Elsevier, pp 472–490
 23. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. John Wiley & Sons, Weinheim/New York
 24. McKinney JD, Richard A, Waller C et al (2000) The practice of structure activity relationships (SAR) in toxicology. *Toxicol Sci* 56:8–17. <https://doi.org/10.1093/toxsci/56.1.8>
 25. Cronin Mark TD, Jaworska Joanna S, Walker John D et al (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ Health Perspect* 111:1391–1401. <https://doi.org/10.1289/ehp.5760>
 26. Cronin Mark TD, Walker John D, Jaworska Joanna S et al (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ Health Perspect* 111:1376–1390. <https://doi.org/10.1289/ehp.5759>
 27. Gissi A, Nicolotti O, Carotti A et al (2013) Integration of QSAR models for bioconcentration suitable for REACH. *Sci Total Environ* 456–457:325–332. <https://doi.org/10.1016/j.scitotenv.2013.03.104>
 28. Gissi A, Lombardo A, Roncaglioni A et al (2015) Evaluation and comparison of benchmark QSAR models to predict a relevant REACH endpoint: the bioconcentration factor (BCF). *Environ Res* 137:398–409. <https://doi.org/10.1016/j.envres.2014.12.019>
 29. Gissi A, Gadaleta D, Floris M et al (2014) An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes. *ALTEX - Altern Anim Exp* 31:23–36. <https://doi.org/10.14573/altex.1305221>
 30. Dearden JC, Barratt MD, Benigni R, et al (1997) The development and validation of expert systems for predicting toxicity: the report and recommendations of an ECVAM/ECB workshop (ECVAM Workshop 24)
 31. Ashby J, Tennant RW (1988) Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat Res Toxicol* 204:17–115. [https://doi.org/10.1016/0165-1218\(88\)90114-0](https://doi.org/10.1016/0165-1218(88)90114-0)
 32. Mentzas G (1994) A functional taxonomy of computer-based information systems. *Int J Inf Manag* 14:397–410. [https://doi.org/10.1016/0268-4012\(94\)90015-9](https://doi.org/10.1016/0268-4012(94)90015-9)
 33. Benfenati E (2016) In Silico methods for predicting drug toxicity. Springer New York, New York
 34. Pizzo F, Gadaleta D, Lombardo A et al (2015) Identification of structural alerts for liver and kidney toxicity using repeated dose toxicity data. *Chem Cent J* 9:62. <https://doi.org/10.1186/s13065-015-0139-7>
 35. Dobo KL, Greene N, Cyr MO et al (2006) The application of structure-based assessment to support safety and chemistry diligence to

- manage genotoxic impurities in active pharmaceutical ingredients during drug development. *Regul Toxicol Pharmacol* 44:282–293. <https://doi.org/10.1016/j.yrtph.2006.01.004>
36. Floris M, Manganaro A, Nicolotti O et al (2014) A generalizable definition of chemical similarity for read-across. *J Cheminform* 6:39. <https://doi.org/10.1186/s13321-014-0039-1>
37. Willett P, Barnard JM, Downs GM (1998) Chemical Similarity Searching. *J Chem Inf Comput Sci* 38:983–996. <https://doi.org/10.1021/ci9800211>
38. Blackburn K, Stuard SB (2014) A framework to facilitate consistent characterization of read across uncertainty. *Regul Toxicol Pharmacol* 68:353–362. <https://doi.org/10.1016/j.yrtph.2014.01.004>
39. Patlewicz G, Ball N, Booth ED et al (2013) Use of category approaches, read-across and (Q)SAR: general considerations. *Regul Toxicol Pharmacol* 67:1–12. <https://doi.org/10.1016/j.yrtph.2013.06.002>
40. Wu S, Blackburn K, Amburgey J et al (2010) A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments. *Regul Toxicol Pharmacol* 56:67–81. <https://doi.org/10.1016/j.yrtph.2009.09.006>
41. OECD (2017) Guidance on grouping of chemicals, Second edition. OECD, Paris
42. European Chemicals Agency (2017) Guidance on information requirements and chemical safety assessment chapter R.7b: endpoint specific guidance
43. Baldi P, Brunak S, Bach F (2001) Bioinformatics: the machine learning approach. MIT press
44. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20:318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>
45. Marchant CA (2012) Computational toxicology: a tool for all industries: computational toxicology. *Wiley Interdiscip Rev Comput Mol Sci* 2:424–434. <https://doi.org/10.1002/wcms.100>
46. Michie D (1968) “Memo” functions and machine learning. *Nature* 218:5136:19
47. Trisciuzzi D, Alberga D, Mansouri K et al (2015) Docking-based classification models for exploratory toxicology studies on high-quality estrogenic experimental data. *Future Med Chem* 7:1921–1936. <https://doi.org/10.4155/fmc.15.103>
48. Trisciuzzi D, Alberga D, Mansouri K et al (2017) Predictive structure-based toxicology approaches to assess the androgenic potential of chemicals. *J Chem Inf Model* 57:2874–2884. <https://doi.org/10.1021/acs.jcim.7b00420>
49. Kamel M, Ahmed A, Aleksandra R et al (2016) CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 124:1023–1033. <https://doi.org/10.1289/ehp.1510267>
50. Trisciuzzi D, Alberga D, Leonetti F et al (2018) Molecular docking for predictive toxicology. In: Nicolotti O (ed) Computational toxicology. Springer New York, New York, pp 181–197
51. Martin TM, Harten P, Venkatapathy R et al (2008) A hierarchical clustering methodology for the estimation of toxicity. *Toxicol Mech Methods* 18:251–266. <https://doi.org/10.1080/15376510701857353>
52. US EPA O (2015) Ecotoxicology database. In: US EPA. <https://www.epa.gov/chemical-research/ecotoxicology-database>. Accessed 29 May 2019
53. Dimitrov S, Dimitrova N, Parkerton T et al (2005) Base-line model for identifying the bioaccumulation potential of chemicals. *SAR QSAR Environ Res* 16:531–554. <https://doi.org/10.1080/10659360500474623>
54. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms - Environmental Reviews. <https://www.nrcresearchpress.com/doi/abs/10.1139/a06-005#XO5Rq4gzbIU>. Accessed 29 May 2019
55. Zhao C, Boriani E, Chana A et al (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73:1701–1707. <https://doi.org/10.1016/j.chemosphere.2008.09.033>
56. Benfenati E (2010) The CAESAR project for in silico models for the REACH legislation. *Chem Cent J* 4:11. <https://doi.org/10.1186/1752-153X-4-S1-11>
57. Commission E (2006) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives

- 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Off J 3961 30122006
58. Williams ES, Panko J, Paustenbach DJ (2009) The European Union's REACH regulation: a review of its history and requirements. Crit Rev Toxicol 39:553–575. <https://doi.org/10.1080/10408440903036056>
59. Golbamaki A, Cassano A, Lombardo A et al (2014) Comparison of *in silico* models for prediction of *Daphnia magna* acute toxicity. SAR QSAR Environ Res 25:673–694. <https://doi.org/10.1080/1062936X.2014.923041>
60. Yordanova D, Schultz TW, Kuseva C et al (2019) Automated and standardized workflows in the OECD QSAR toolbox. Comput Toxicol 10:89–104. <https://doi.org/10.1016/j.comtox.2019.01.006>
61. Enoch S j. (2010) Chemical category formation and read-across for the prediction of toxicity. In: Puzyn T, Leszczynski J, Cronin MT (eds) Recent advances in QSAR studies. Springer Netherlands, Dordrecht, pp 209–219
62. Devillers J (2013) Methods for building QSARs. In: Reisfeld B, Mayeno AN (eds) Computational toxicology: Volume II. Humana Press, Totowa, pp 3–27
63. Klimisch H-J, Andreae M, Tillmann U (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol 25:1–5. <https://doi.org/10.1006/ftp.1996.1076>
64. Gadaleta D, Mangiatordi GF, Catto M et al (2016) Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct-Prop Relatsh IJQSPR 1:45–63. <https://doi.org/10.4018/IJQSPR.2016010102>
65. Walker J, Carlsen L, Jaworska J (2003) Improving opportunities for regulatory acceptance of QSARs: the importance of model domain, uncertainty, validity and predictability. QSAR Comb Sci 22:346–350. <https://doi.org/10.1002/qsar.200390024>
66. Russom CL, Breton RL, Walker JD, Bradbury SP (2003) An overview of the use of quantitative structure–activity relationships for ranking and prioritizing large chemical inventories for environmental risk assessments. Environ Toxicol Chem 22:1810. <https://doi.org/10.1897/01-194>
67. Eriksson L, Jaworska J, Worth AP et al (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375. <https://doi.org/10.1289/ehp.5758>
68. Roncaglioni A, Toropov AA, Toropova AP, Benfenati E (2013) In silico methods to predict drug toxicity. Curr Opin Pharmacol 13:802–806. <https://doi.org/10.1016/j.coph.2013.06.001>
69. Helma C (2005) Predictive toxicology. CRC Press, Boca Raton, Florida, USA
70. Judson R (2010) Public databases supporting computational toxicology. J Toxicol Environ Health Part B 13:218–231. <https://doi.org/10.1080/10937404.2010.483937>
71. Atkins D (2003) Revolutionizing science and engineering through cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure
72. Judson PN, Cooke PA, Doerrer NG et al (2005) Towards the creation of an international toxicology information Centre. Toxicology 213:117–128. <https://doi.org/10.1016/j.tox.2005.05.014>
73. Heidorn CJA, Rasmussen K, Hansen BG et al (2003) IUCLID: an information management tool for existing chemicals and biocides. J Chem Inf Comput Sci 43:779–786. <https://doi.org/10.1021/ci0202786>
74. Louis KS, Jones LM, Campbell EG (2002) MacroScope: sharing in science. Am Sci 90:304–307
75. Hilgartner S, Brandt-Rauf SI (1994) Data access, ownership, and control: toward empirical studies of access practices. Knowledge 15:355–372. <https://doi.org/10.1177/107554709401500401>
76. Staples CA, Woodburn K, Caspers N et al (2002) A weight of evidence approach to the aquatic hazard assessment of bisphenol A. Hum Ecol Risk Assess Int J 8:1083–1105. <https://doi.org/10.1080/1080700291905837>
77. Benedetti M, Ciaprini F, Piva F et al (2012) A multidisciplinary weight of evidence approach for classifying polluted sediments: integrating sediment chemistry, bioavailability, biomarkers responses and bioassays. Environ Int 38:17–28. <https://doi.org/10.1016/j.envint.2011.08.003>
78. Piva F, Ciaprini F, Onorati F et al (2011) Assessing sediment hazard through a weight of evidence approach with bioindicator organisms: a practical model to elaborate data from sediment chemistry, bioavailability, biomarkers and ecotoxicological bioassays. Chemosphere 83:475–485. <https://doi.org/10.1016/j.chemosphere.2010.12.064>
79. Hartung T (2009) Food for thought ... on *in silico* methods in toxicology. ALTEX

- 26:155–166. <https://doi.org/10.14573/altex.2009.3.155>
80. Bunn DW (1988) Combining forecasts. *Eur J Oper Res* 33:223–229. [https://doi.org/10.1016/0377-2217\(88\)90165-8](https://doi.org/10.1016/0377-2217(88)90165-8)
81. Clemen RT (1989) Combining forecasts: a review and annotated bibliography. *Int J Forecast* 5:559–583. [https://doi.org/10.1016/0169-2070\(89\)90012-5](https://doi.org/10.1016/0169-2070(89)90012-5)
82. The Combination of Forecasts: Journal of the Operational Research Society: Vol 20, No 4. <https://www.tandfonline.com/doi/abs/10.1057/jors.1969.103>. Accessed 29 May 2019
83. Hewitt M, Cronin MTD, Madden JC et al (2007) Consensus QSAR models: do the benefits outweigh the complexity? *J Chem Inf Model* 47:1460–1468. <https://doi.org/10.1021/ci700016d>
84. Votano JR, Parham M, Hall LH et al (2004) Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* 19:365–377. <https://doi.org/10.1093/mutage/geh043>
85. Lewis DF, Bird MG, Jacobs MN (2002) Human carcinogens: an evaluation study via the COMPACT and HazardExpert procedures. *Hum Exp Toxicol* 21:115–122. <https://doi.org/10.1191/0960327102ht233oa>
86. Research C for DE and (2019) M7 (R1) assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk. In: US Food Drug Adm. <http://www.fda.gov/regulatory-information/search-fda-guidance-documents/m7r1-assessment-and-control-dna-reactive-mutagenic-impurities-pharmaceuticals-limit-potential>. Accessed 29 May 2019
87. Müller L, Mauthé RJ, Riley CM et al (2006) A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity. *Regul Toxicol Pharmacol* 44:198–211. <https://doi.org/10.1016/j.yrtph.2005.12.001>
88. Sutter A, Amberg A, Boyer S et al (2013) Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities. *Regul Toxicol Pharmacol* 67:39–52. <https://doi.org/10.1016/j.yrtph.2013.05.001>
89. Mattingly Carolyn J, Colby Glenn T, Forrest John N, Boyer James L (2003) The comparative toxicogenomics database (CTD). *Environ Health Perspect* 111:793–795. <https://doi.org/10.1289/ehp.6028>
90. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87:1123–1124. <https://doi.org/10.1021/ed100697w>
91. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40: D1100–D1107. <https://doi.org/10.1093/nar/gkr777>
92. Richard AM, Williams CR (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res Mol Mech Mutagen* 499:27–52. [https://doi.org/10.1016/S0027-5107\(01\)00289-5](https://doi.org/10.1016/S0027-5107(01)00289-5)



Chapter 12

Conformal Prediction for Ecotoxicology and Implications for Regulatory Decision-Making

Fredrik Svensson and Ulf Norinder

Abstract

Computational methods can be valuable tools for safety prediction of chemicals and have potential to play a role in the regulatory decision-making if the results are transparent and reliable. In this chapter, we discuss a type of confidence predictor called conformal prediction that can be used to generate predictions with a guaranteed error rate. We describe the underlying theory in an informal fashion and exemplify the method on a dataset of chronic toxicity of compounds to *Daphnia magna* and *Pseudokirchneriella subcapitata*.

Key words QSAR, Confidence, Uncertainty, Conformal prediction

1 Introduction

Today's society produces a myriad of different chemicals for all kinds of purposes. The sheer number of chemicals on the market and under development makes the safety testing of each and every one of these an insurmountable task [1]. Computational methods on the other hand can handle a high number of chemical compounds in a short period of time and are often used to prioritize compounds for further testing. This way the chemicals that are most likely to cause harm can be tested and regulated, but it still leaves a large corpus of chemicals without sufficient data for action. However, if sufficient confidence can be put into a prediction of a compound's safety, the prediction itself might be used as a basis for decision-making and safety regulations. Associating predictions with well-defined uncertainties also facilitates the communication of risks.

The use of computational methods for regulatory purposes has been discussed previously [2, 3], and the consensus is that these methods will have to be scientifically sound, ideally mechanistically based, not overly complex, and associated with a clear domain of applicability and measure of its predictive power on compounds

that were not used in the training of the model [4]. Attempts have also been made to standardize the requirements on predictive models, and OECD has published a guideline entitled “Guidance Document on the Validation of (Q)SAR Models” [5].

Clearly, a key concept for the acceptance of predictive models for regulatory purposes is the quantification of the uncertainty associated with the model. The definition of a model’s reliability is closely associated with the models applicability domain (AD), i.e., the domain where the model can be used to make reliable predictions [6]. In general terms this can be illustrated by a model trained for predicting the safety of a certain class of chemical not necessarily being predictive for another. Although a very useful concept, the precise definition of a model’s AD can be difficult and has been the subject of much debate. How similar is similar enough to be within the space where the model can be trusted to make reliable predictions?

Hanser and co-workers [7] introduce a useful way to reason about a model’s AD by dividing it into three different sub-domains: applicability, reliability, and decidability. These can be used to answer the three questions “Is my model applicable for this case?”, “Is my prediction reliable enough?”, and “Can I base my decision on the prediction?”. They also highlight important considerations when evaluating a predictive model, and these include the descriptor ranges seen by the model, the density of information and label distribution, and model consensus.

Reliability predictions and AD assessments tend to focus on the positive label, i.e., toxic label in this work. However, it is equally important to consider the reliability of the negative compounds, i.e., nontoxic compounds in this work, in order to derive meaningful negative predictions—something that is paramount in safety-related applications. It is therefore important to make and evaluate also negative predictions [8].

Other methods to estimate the reliability of computational models [9] that are not directly linked to the AD include Bayesian models [10] such as Gaussian processes [11], the comparison of experimental and predictive distributions [12], reliability-density neighborhoods [13], and confidence predictors [14].

In this chapter we will focus on a confidence predictor framework called conformal prediction. Conformal predictors address several of the important considerations for model reliability. Most importantly they allow the user to set the required confidence level, something that will also intrinsically determine many aspects of the AD. Especially the second domain defined by Hanser et al., reliability, is intrinsically handled by a conformal predictor as the error rate is guaranteed to reflect the set confidence level. Using Mondrian conformal predictors, the error rate is guaranteed for each class that is being considered, allowing confidence in both toxic and nontoxic predictions.

Confidence predictors output prediction regions rather than single labels for each compound. The details of conformal prediction are described in detail in the next section, but the main principle is that every new prediction is compared to the results of a calibration set to determine the predicted labels. A conformal predictor will always produce predictions with a guaranteed error rate as long as data is exchangeable. A conformal predictor is said to be valid if the error rate of the predictions does not exceed the set confidence level, i.e., at the 80% confidence level, the predictor is valid if the error rate is equal to or below 20%.

Being a flexible framework means that any model can be converted to a conformal predictor as long as the model gives a measure that can be used to rank the predictions. This is true both for classification and regression models. Examples in the literature include conformal predictors based on random forest (RF) [15], support vector machines [16], deep neural networks [17], and more [18].

We believe that for many safety outcomes, there is now sufficient data that with the use of conformal prediction, meaningful models with clear associated confidence can be derived. These should fulfil the criteria necessary to be useful both for the users and for regulatory purposes.

In the next section, we will discuss the principle behind a conformal predictor and following that demonstrate the application of a conformal predictor on a dataset from Ding et al. [19] on the chronic toxicity of compounds to *Daphnia magna* and *Pseudokirchneriella subcapitata*.

2 The Workings of a Conformal Predictor

Conformal prediction was introduced by Vovk and Gammerman [14] and is proven to always generate valid models as long as the data is exchangeable. In this section we will look at the principles for how this is achieved and focus on providing an informal explanation of the principle. The type of conformal predictor considered in this chapter is called an inductive conformal predictor [20].

The output from a conformal predictor is a prediction range that will encompass the correct value in at least the fraction of predictions corresponding to the confidence level (at a confidence level of 80%, 80% of the predictions will include the correct label). For a binary prediction problem, a compound can be associated with any of four different outcomes, either of the two classes, both classes, or no classes. In a regression setting, a conformal regressor will output a range of values for each compound.

If the set of labels predicted for a compound includes the correct label, that prediction is counted as being *valid*. In addition to the *validity*, we also define *efficiency* as the number of single-label

predictions. Thus, we would like for our conformal predictor to be as *efficient* as possible while maintaining validity.

In order to assign the set of predicted labels, the outcome of the predictor for a particular compound is compared with the predictor outcomes on a set of compounds with known labels called the calibration set. The comparison is done using a conformity or nonconformity score (α) that is used to rank the predicted compounds as well as the calibration set. In this investigation we use the probabilities from the RF ensemble as the conformity score (α). The relative position of a compound being predicted within the calibration compounds is then used to assign a *p-value* to the compound. If the *p-value* exceeds the significance level (defined as 1—confidence level), the considered label is assigned to the compound. In Fig. 1, and in the description below, we provide an illustrative example of how labels are assigned to new compounds given two sets of calibration compounds—one set for each of the two classes (toxic and nontoxic).

The α -values in each calibration set are sorted in ascending order. For example, a prediction from the RF ensemble, containing 100 decision trees, for a new test compound results in probabilities (α -values) of 0.8 and 0.2 for the toxic and nontoxic class, respectively. The first value (0.8) is then compared to the corresponding calibration set (toxic compounds) containing eight compounds with the values 0.14, 0.31, 0.36, 0.53, 0.71, 0.77, 0.85, and 0.90. The value of 0.8 will be positioned in between 0.77 and 0.85 with six compounds with lower probabilities out of now nine compounds (the original eight and the new test compound). The *p-value* is then calculated as $6/9 \approx 0.67$. If we are using a confidence level of 75%, the corresponding significance would be 0.25 (1–0.75), and since $0.67 > 0.25$, we would assign a toxic class label to the test compound. These second value (0.2) is then compared to the corresponding calibration set of nontoxic compounds (ten compounds with the values 0.18, 0.33, 0.53, 0.58, 0.6, 0.69, 0.72, 0.78, 0.86, and 0.93). In this case the new value will be positioned in between 0.18 and 0.33, and the *p-value* would be calculated as $1/11 \approx 0.09$. Since $0.09 < 0.25$, we cannot assign the nontoxic label, and the test compound is predicted to be toxic. A second new test compound with predicted probabilities of 0.33 and 0.67 (toxic and nontoxic) would in a similar manner have *p*-values of $2/9 \approx 0.22 < 0.25$ and $5/11 \approx 0.45 > 0.25$, respectively, and be assigned only a nontoxic label and predicted as nontoxic. A third new test compound with predicted probabilities of 0.44 and 0.56 (toxic and nontoxic) would have *p*-values of $4/9 \approx 0.44 > 0.25$ and $3/11 \approx 0.27 > 0.25$, respectively, and be assigned both a toxic and a nontoxic label and predicted as belonging to the *both* class. A compound can also have both *p*-values below 0.25 and not be assigned a class label at all thus belonging to the *empty* class. Compounds belonging to this class are too dissimilar to

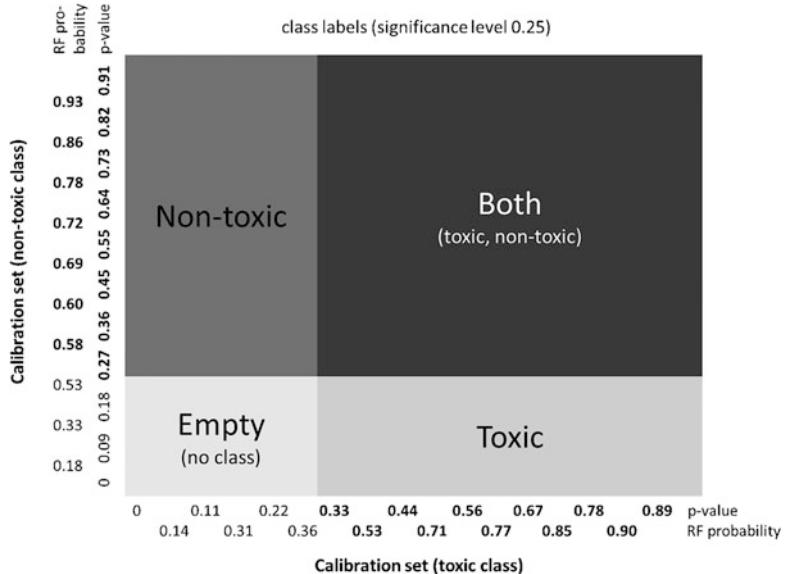


Fig. 1 P -value estimation in Mondrian conformal prediction. The α -values (in this case RF probability) are used to derive a conformal p -value that determines the class labels

all calibration set compounds and consider to be outside the applicability domain of the model where reliable predictions can be expected.

The procedure described above using separate calibration sets for each class is called class-conditional or Mondrian conformal prediction and ensures that the predictions are valid for each class independently. Mondrian conformal predictors are also an excellent way to handle imbalanced data, something that is often encountered when predicting safety endpoints since safe compounds are typically more prevalent. This is normally problematic for many machine learning algorithms and has to be addressed in model building [21]. For a Mondrian conformal predictor, the individual class consideration will automatically set a cutoff for the decision boundary that will balance the results [22]. This has been shown to work well in practice also for very highly imbalanced datasets (in the range of 1:1000) [15, 23].

Conformal prediction, as mentioned above, requires a conformity (sometimes referred to as nonconformity) measure which is, in many aspects, similar to a similarity (non-similarity) measure commonly used in QSAR modelling. In principle any such measure can be selected, but the more appropriate (effective) the conformity measure is in producing meaningful outputs for calculation of CP p -values, the more efficient predictions will result. In the examples presented in this chapter, the class probability from the RF classifier was chosen as conformity score. Studies have shown that the

efficiency of conformal predictors can be strongly influenced by the way the conformity score is calculated [18, 24, 25].

Since conformal prediction compared to more standard machine learning techniques require the additional partitioning of the available data into both proper training set, used to train the underlying algorithm, and calibration set, methods have been developed that still can make maximal use of the available data through the training of multiple predictors using different splits. This can be done through the random resampling of the calibration test, called aggregated conformal prediction [26], or through the division of data in severalfold where one is left out for calibration in each iteration, called cross-conformal prediction [27, 28].

A schematic of the conformal prediction procedure using aggregated conformal prediction is shown in Fig. 2. For a more in-depth example of the conformal prediction algorithm, please refer to Norinder et al. [29]

3 Example of a Conformal Predictor Applied to the Prediction of Chronic Toxicity to *Daphnia magna* and *Pseudokirchneriella subcapitata*

To exemplify the application of a conformal predictor, we use datasets from Ding et al. [19] on the chronic toxicity of compounds to *Daphnia magna* and *Pseudokirchneriella subcapitata*. An overview of the two datasets is shown in Table 1.

For the predictions we used an aggregated conformal predictor based on 100 aggregates, using 70% of the training data as proper training set and 30% as calibration set in each loop. The *p*-values were derived using the median value from the aggregations. The underlying model was a RF [31] based on 100 trees. All calculations were performed using Python, scikit-learn, and the nonconformist package.

Compound structures were standardization using the IMI eTOX project standardizer [32] in combination with the MolVS standardizer [33] for tautomer standardization. Compounds were represented using RDKit [34] physiochemical and structural descriptors as well as Morgan fingerprints (radius 4, hashed to 1024 bits). Data was split into 50 pairs of random training set (80%) and test set (20%). We present the overall predictions from the 50 test sets. Median CP *p*-values for each compound in each test set (50) were treated as a unique prediction.

The resulting validities from the models are shown in Figs. 3, 4, 5, and 6. Gratifyingly, the observed validities in general correspond closely to the set confidence levels. However, they are often slightly higher than the defined confidence level of the predictions, something that is often observed for aggregated conformal predictors [35].

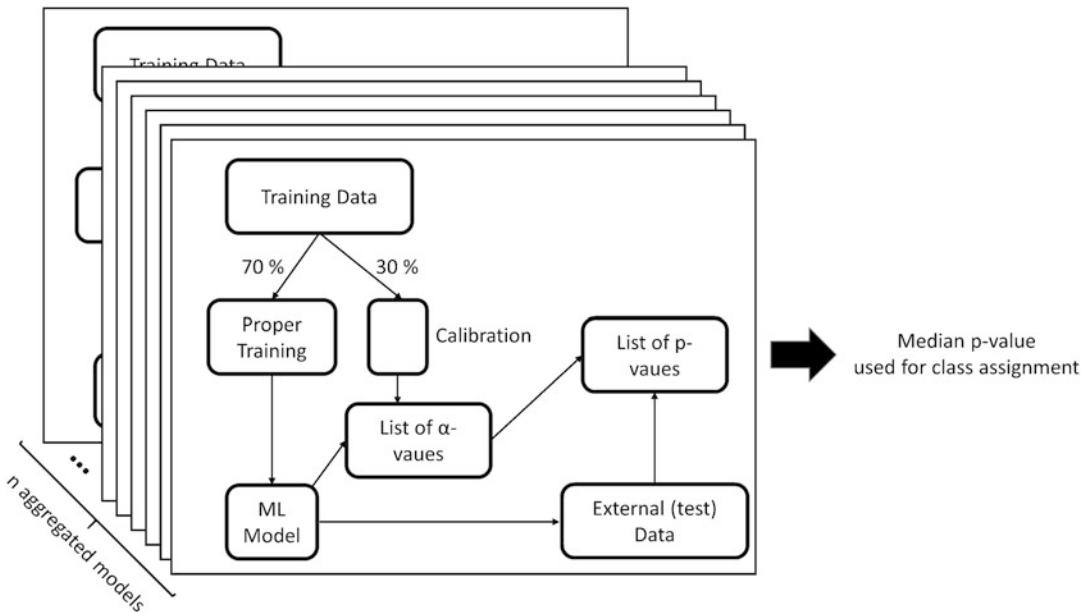


Fig. 2 Schematic representation of the aggregated conformal procedure. The practical implementation of conformal prediction is relatively straight forward, requiring only small modifications to existing prediction setups while maintaining the original model as a basis for the predictions. Conformal prediction is also readily available in the scikit-learn [30] compatible Python package nonconformist. (<https://github.com/donlnz/nonconformist>)

Table 1
The number of total and toxic compounds in the respective datasets

Dataset	# Compounds	# Toxic
<i>Daphnia magna</i>	403	237
<i>Pseudokirchneriella subcapitata</i>	551	272

In the case of *P. subcapitata* (Fig. 5) when the models were trained using the Morgan fingerprints as input, the observed validities are often well above the set confidence level. This is often an indication that the model is having difficulties to differentiate between the two labels for these compounds and is generating a high level of *both* (toxic and nontoxic label) predictions.

Ding et al. report the performance of the *P. subcapitata* models as MCC values for the training set and validation set of 0.683 and 0.511, and AUC values for the training set and validation set equal to 0.910 and 0.810. For the models on *D. magna*, they report MCC values for the training set and validation set of 0.772 and 0.603 and AUC values for the training set and validation set of 0.950 and 0.800. Using only the single-label predictions at the 80% confidence level, our RDKit descriptor-based models achieve a

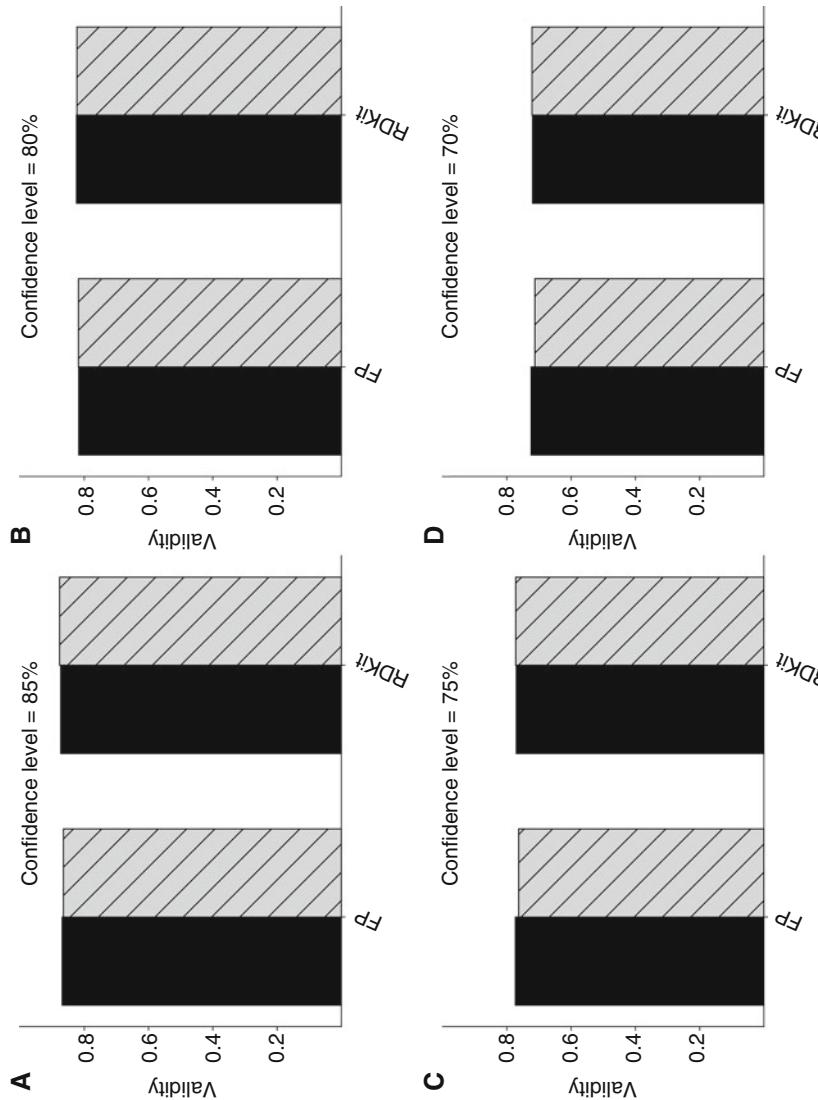


Fig. 3 Validity of the models on *D. magna* for the toxic class at different confidence levels using fingerprints and structural descriptors. Training set in black and test set in striped gray

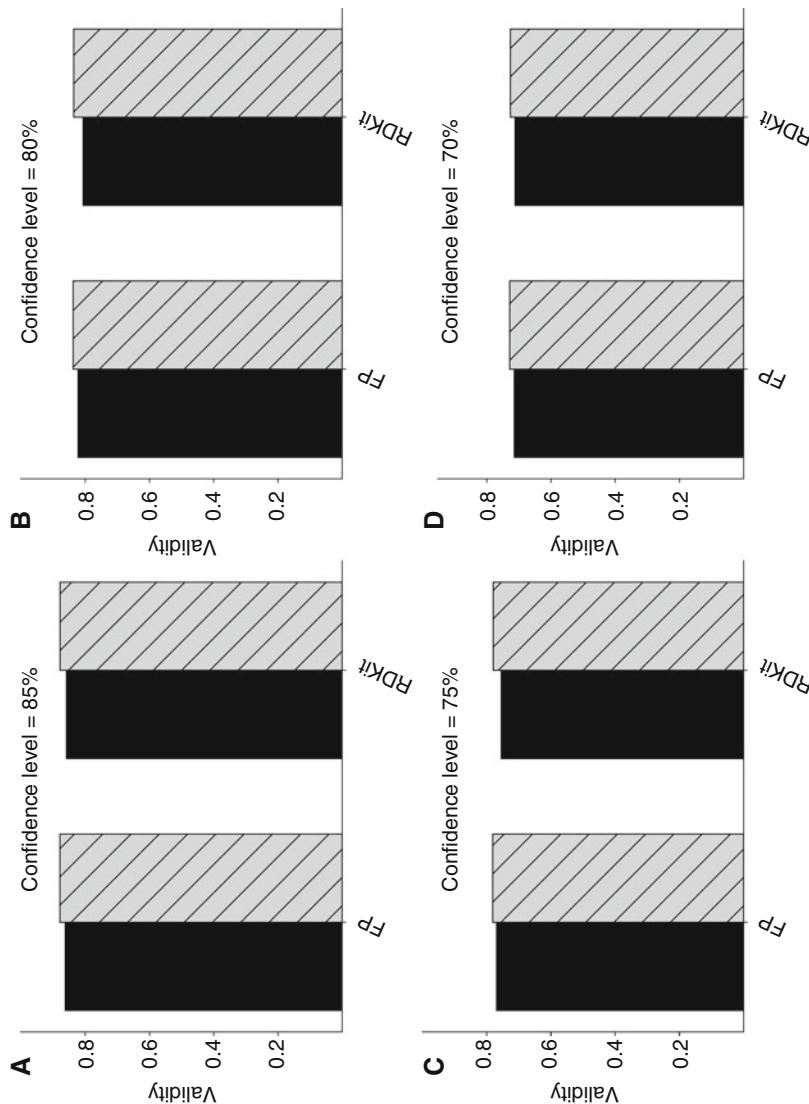


Fig. 4 Validity of the models on *D. magna* for the nontoxic class at different confidence levels using fingerprints and structural descriptors. Training set in black and test set in striped gray

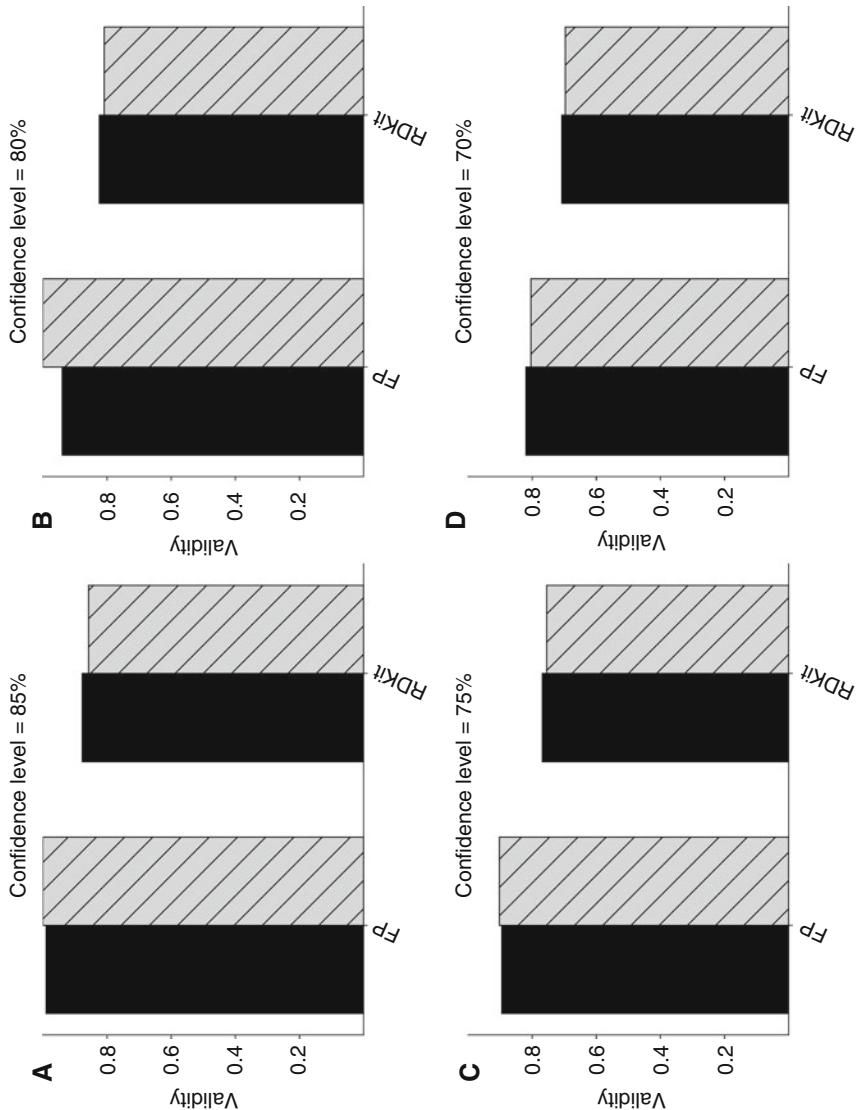


Fig. 5 Validity of the models on *P. subcapitata* for the toxic class at different confidence levels using fingerprints and structural descriptors. Training set in black and test set in striped gray

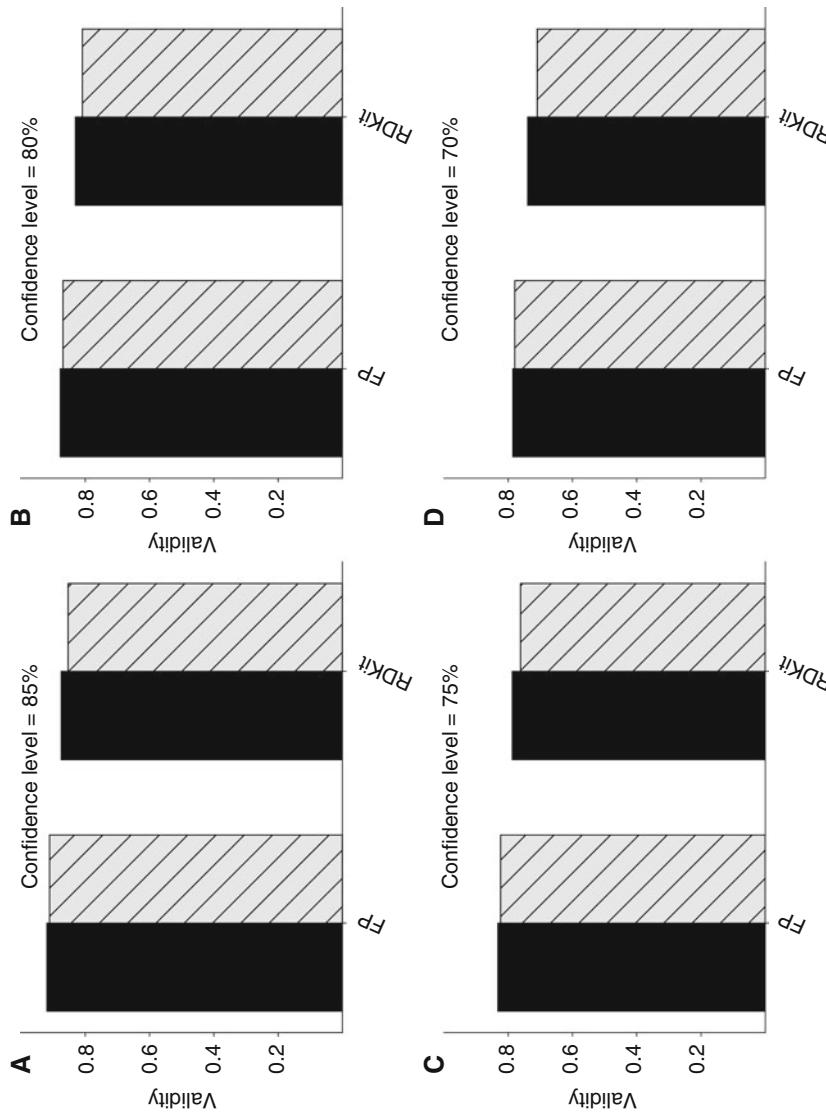


Fig. 6 Validity of the models on *P. subcapitata* for the nontoxic class at different confidence levels using fingerprints and structural descriptors. Training set in black and test set in striped gray

MCC for the training set and test set of 0.53 and 0.50, respectively, for *P. subcapitata* and 0.59 and 0.62, respectively, for *D. magna*, very similar to the results from Ding et al.

In line with the reported results by Ding et al., the outcomes for *P. subcapitata* were more challenging to predict and the efficiency of the models at high confidence levels were quite low. The close correspondence of the results illustrates an important point—a conformal predictor cannot be expected to increase the quality of the predictions as this is dependent on the underlying model. Instead, it is the utility of the predictions that is enhanced by the conformal process by assigning a well-defined confidence to the result.

Figure 7 shows how the efficiency is influenced by the confidence level of the predictor. Setting a high confidence level in many cases will increase the proportion of dual label (*both* class) predictions since the threshold for class inclusion becomes lower. This follows the intuitive logic that as the predictor is required to give more confident answers, fewer such single-label answers can be obtained given the same (constant) underlying model (predictor). The additional information provided by conformal prediction offers a new way of evaluating predictive models. Based on the efficacy, the user can weigh the desired confidence against the possible outputs at that level. For example, if 80% confidence is needed for the application at hand but models generated at that confidence level generate a very low efficacy, this shows that there is not sufficient information in the model to meet the set criteria. This might then be addressed by changing the underlying algorithms and data representations or, more likely, by the collection of additional or higher-quality data.

Applying conformal prediction to these dataset improves the utility of the predictions in several ways. Primarily, as is clear from the results presented above, the models do deliver a validity that corresponds to the set confidence level, thus greatly enhancing the usefulness of the models for decision-making as the expected error rate is clearly defined. Furthermore, the *both* and *empty* prediction generated by the models contain important information on what actions could be taken to improve the model. Acquiring experimental data for compounds with both labels will improve the definition of the decision boundary between the two classes, while data on the no label (*empty* class) compounds might increase the coverage of chemical space.

Another benefit is that using the validity and efficiency of the models, it is readily apparent that the fingerprint-based models for *P. subcapitata* are not suitable for deployment as they generate very inefficient models. This is a clear indication that in this representation, there is not sufficient information available in the data to distinguish the two classes with the desired confidence.

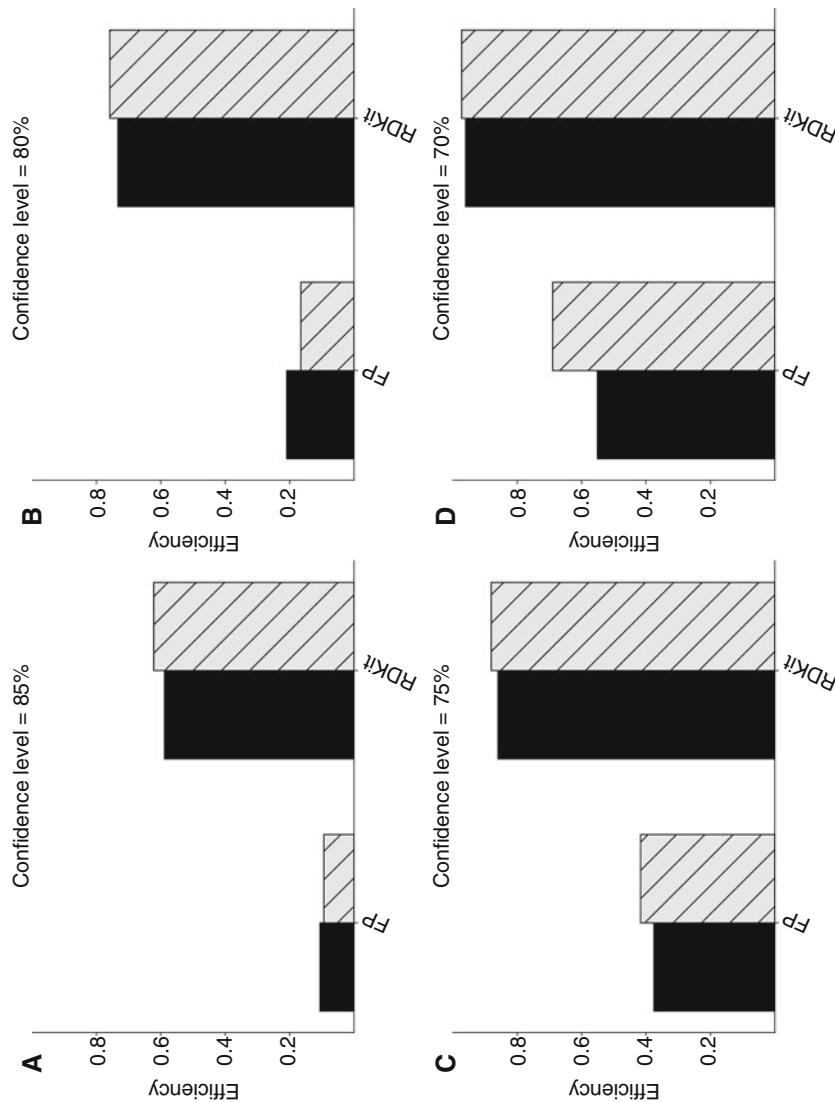


Fig. 7 Efficiency of the conformal predictor for *P. subcapitata* at different confidence levels using fingerprints and structural descriptors. Training set in black and test set in striped gray

4 Discussion

Despite being a useful method in many situations, there are certain limitations associated with conformal predictors. One obvious limitation, especially for smaller datasets, is the additional need for a calibration set [36]. The exception to this need for a separate calibration set is the RF algorithm where the out-of-bag compounds can be used as calibration set [25]. However, for most other algorithms employed for model building, this means that a certain percentage (20–30%) of the training set is set aside and cannot be used for building the model. Additionally, the need for a calibration set evokes questions on how should this calibration set be selected. Since the method operates under the exchangeability criteria, a random (stratified) selection of the calibration set that mirrors the training set seems to be a good and workable choice. Also, to cover descriptor space effectively, several randomly selected pairs of calibration and proper training sets can be used for predicting the outcome of the test set compounds. This latter approach is called aggregated conformal prediction and has been shown to perform well adding robustness to the outcome and especially for the minority class [26]. This is the approach used in this investigation.

As discussed briefly in the application section, the properties of a conformal prediction output can give additional information to the user. For example, perhaps the efficiency of the model, at the confidence level set by the user as needed for the decision at hand, is very low. The user can then examine the predictions to see how many compounds have been predicted as belonging to the *both* or to the *empty* class. If many compounds have been assigned to the *both* class, this means that the present model (classifier) cannot distinguish between classes, but the compounds are within the applicability domain of the model, and that the compound description, at present, lacks sufficient information. Thus, in order to improve the efficiency of the model new information needs to be incorporated into the compound description. If many compounds have been assigned to the *empty* (no) class, this means that the present model (classifier) cannot assign a class and that these compounds, in model space, are too dissimilar (out-of-domain) to the compounds in the calibration set. In this case (some of) these compounds should be tested and the model updated in order to expand the applicability domain of the model in order to increase efficiency for new predictions.

When dealing with prediction uncertainty in general, it is important to also understand the variations in the data that is used for model training and what the expected variation—and thus maximum performance of the prediction is likely to be [37].

We would also like to underline that although the usefulness of the predictions in many ways is enhanced by conformal prediction, in order to be meaningful, it still requires good-quality models trained on appropriate data [38, 39].

5 Conclusions and Outlook

Conformal prediction is a well-defined mathematical framework that generates predictions with defined confidence, as such conformal prediction is especially suitable for predictions in the area of safety where confidence is paramount. In contrast to many other techniques for confidence estimation, conformal prediction removes much of the ambiguity surrounding the definition of AD. The robust definition of conformal predictors and the relative ease of interpreting the confidence make them an ideal tool for enhancing the utility of predictions also in a regulatory setting.

In this chapter we have demonstrated how conformal prediction can enhance the utility of toxicity predictions by exemplifying the principles on two relevant datasets.

In our opinion, the low barrier to implement conformal prediction in combination with its clear utility for the field merits more people to consider this approach when developing predictive models for safety outcomes. A unified and robust confidence measure would be of great value for the field in general, and we envision that conformal prediction could be an important step toward achieving this.

Acknowledgments

The ARUK UCL Drug Discovery Institute is core funded by Alzheimer's Research UK (registered charity No. 1077089 and SC042474). The Francis Crick Institute receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust (FC001002).

References

1. Judson R, Richard A, David DJ, Houck K, Martin M, Kavlock R, Dellarco V, Henry T, Holderman T, Sayre P et al (2009) The toxicity data landscape for environmental chemicals. *Environ Health Perspect* 117(5):685–695
2. Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ Health Perspect* 111(10):1391–1401
3. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111(10):1361–1375
4. Jaworska J, Comber M, Auer C, Leeuwen C (2003) Summary of a workshop on regulatory

- acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Perspect* 111:1358
5. OECD: OECD principles for the validation, for regulatory purposes, of QSAR models. <http://www.oecd.org/Dataoecd/33/37/37849783.pdf>
 6. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena Pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 48 (9):1733–1746
 7. Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. *SAR QSAR Environ Res* 27 (11):865–881
 8. Williams RV, Amberg A, Brigo A, Coquin L, Giddings A, Glowienke S, Greene N, Jolly R, Kemper R, O’Leary-Steele C et al (2016) It’s difficult, but important, to make negative predictions. *Regul Toxicol Pharmacol* 76(Suppl C):79–86
 9. Bosnić Z, Kononenko I (2008) Comparison of approaches for estimating reliability of individual regression predictions. *Data Knowl Eng* 67 (3):504–516
 10. Lazic S, Edmunds N, Pollard C (2017) Predicting drug safety and communicating risk: benefits of a bayesian approach. *Toxicol Sci* 162:89–98
 11. Cortes-Ciriano I, van Westen GJP, Lenselink EB, Murrell DS, Bender A, Malliaivin T (2014) Proteochemometric modeling in a Bayesian framework. *J Cheminform* 6(1):35
 12. Wood DJ, Carlsson L, Eklund M, Norinder U, Stårling J (2013) QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. *J Comput Aided Mol Des* 27(3):203–219
 13. Aniceto N, Freitas AA, Bender A, Ghaforian T (2016) A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *J Cheminform* 8(1):69
 14. Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer, New York, pp 1–324
 15. Svensson F, Norinder U, Bender A (2017) Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol Res (Camb)* 6:73–80
 16. Forerryd A, Norinder U, Lindberg T, Lindstedt M (2018) Predicting skin sensitizers with confidence — using conformal prediction to determine applicability domain of GARD. *Toxicol Vitr* 48:179–187
 17. Cortés-Ciriano I, Bender A (2019) Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J Chem Inf Model* 59 (3):1269–1281
 18. Svensson F, Aniceto N, Norinder U, Cortes-Ciriano I, Spjuth O, Carlsson L, Bender A (2018) Conformal regression for quantitative structure-activity relationship modeling – quantifying prediction uncertainty. *J Chem Inf Model* 58(5):1132–1140
 19. Ding F, Wang Z, Yang X, Shi L, Liu J, Chen G (2019) Development of classification models for predicting chronic toxicity of chemicals to daphnia magna and Pseudokirchneriella subcapitata. *SAR QSAR Environ Res* 30(1):39–50
 20. Vovk V (2013) Conditional validity of inductive conformal predictors. *Mach Learn* 92 (2):349–376
 21. Chawla NV, Japkowicz N, Drive P (2004) Editorial : special issue on learning from imbalanced data sets. *ACM SIGKDD Explor News* 6(1):1–6
 22. Löfström T, Boström H, Linusson H, Johansson U (2015) Bias reduction through conditional conformal prediction. *Intell Data Anal* 19:1355–1375
 23. Norinder U, Boyer S (2017) Binary classification of imbalanced datasets using conformal prediction. *J Mol Graph Model* 72:256–265
 24. Papadopoulos H, Vovk V, Gammerman A (2011) Regression conformal prediction with nearest neighbours. *J Artif Intell Res* 40:815–840
 25. Johansson U, Boström H, Löfström T, Linusson H (2014) Regression conformal prediction with random forests. *Mach Learn* 97 (1):155–176
 26. Carlsson L, Eklund M, Norinder U (2014) Aggregated conformal prediction. In: Iliadis L, Maglogiannis I, Papadopoulos H, Sioutas S, Makris C (eds) Artificial intelligence applications and innovations: AIAI 2014 workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19–21, 2014. proceedings. Springer International Publishing, Berlin, pp 231–240
 27. Vovk V (2015) Cross-conformal predictors. *Ann Math Artif Intell* 74(1):9–28
 28. Sun J, Carlsson L, Ahlberg E, Norinder U, Engkvist O, Chen H (2017) Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *J Chem Inf Model* 57 (7):1591–1598

29. Norinder U, Carlsson L, Boyer S, Eklund M (2014) Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J Chem Inf Model* 54(6):1596–1603
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
31. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
32. IMI eTOX project standardizer. <https://pypi.python.org/pypi/standardiser>
33. MolVS standardizer. <https://pypi.python.org/pypi/MolVS>
34. RDKit: Open-Source Cheminformatics. <http://www.rdkit.org>
35. Linusson H, Norinder U, Boström H, Johansson U, Löfström T (2017) On the calibration of aggregated conformal predictors. In: Gammerman A, Vovk V, Luo Z, Papadopoulos H (eds) Conformal and probabilistic prediction and applications, 13–16 June 2017, vol 60. Machine Learning Research, Stockholm, pp 154–173
36. Johansson U, Ahlberg E, Boström H, Carlsson L, Linusson H, Sönströd C (2015) Handling small calibration sets in mondrian inductive conformal regressors. In: Gammerman A, Vovk V, Papadopoulos H (eds) Statistical learning and data sciences: third international symposium, SLDS 2015, Egham, UK, April 20–23, 2015, proceedings. Springer International Publishing, Cham, pp 271–280
37. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC50 data – a statistical analysis. *PLoS One* 8(4):e61007
38. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 50(7):1189–1204
39. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29(6–7):476–488



Chapter 13

Read-Across for Regulatory Ecotoxicology

Gulcin Tugcu, Serli Önlü, Ahmet Aydin, and Melek Türker Saçan

Abstract

Given the magnitude of chemicals that require ecotoxicity assessments for regulatory purposes, read-across allows for the filling in certain data requirements, such as endpoint estimation, screening and prioritization, and hazard identification, provided that they are justified and documented. In this chapter, we present a recompilation of recognized regulations and guidelines, as well as software and tools, used in grouping and read-across for ecotoxicology-related endpoints. Additionally, an exemplary read-across study for the bioconcentration factor prediction is included.

Key words Read-across, Ecotoxicology, Regulatory, Data gaps, REACH Regulation, Alternative methods, Similarity, Chemical analogue, Chemical category

Abbreviations

ADI	Applicability domain index
BCF	Bioconcentration factor
CAS	Chemical Abstracts Service
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
ECHA	European Chemicals Agency
EFSA	European Food Safety Authority
EU	European Union
FDA	Food and Drug Administration
HPV	High production volume
KNN	K-nearest neighbor
LOAEL	Lowest-observed-adverse-effect level
NOAEL	No-observed-adverse-effect level
NTM	Non-testing methods
OECD	Organisation for Economic Co-operation and Development
PBT/vPvB	Persistent, bioaccumulative, and toxic/very persistent and very bioaccumulative
QSAR	Quantitative structure-activity relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
US EPA	United States Environmental Protection Agency

1 Introduction

The environment is being continuously exposed to an immense amount of chemicals. At present, there are over 149 million unique chemical substances according to the Chemical Abstracts Service (CAS) Registry. Interestingly, only around 389,000 substances are inventoried globally under different regulatory frameworks, such as the high production volume (HPV) chemicals [1]. Notably, despite the plethora of the HPV chemicals, basic toxicity data remains as a requirement [2]. Furthermore, there is a constant regulatory gap due to the emergence of new chemicals. In this manner, regulation of chemicals, such as environmental hazard and risk assessment, is of utmost importance.

The European Union (EU) Regulation concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) [3] is one of the regulatory frameworks meant to ensure the protection of human health and the environment. The REACH Regulation requires industry to provide information on chemical substances regarding their intrinsic properties, manufacture, uses, exposures, possible hazards, and hazard management. Likewise, the industry is obliged to register chemical substances produced or imported in volumes of minimum 1 ton per year. Besides the protection of human health and the environment, the REACH Regulation also pioneers the global chemical management; several regional REACH adaptations, such as Korea [4], Turkey [5], and the United Kingdom [6] counterparts, have been coming into force worldwide.

The information on intrinsic properties of substances required by the REACH Regulation can be addressed by alternative data generation methods, such as grouping of substances and read-across [3]. Moreover, in order to avoid animal testing, it is a legal requirement to gather all the available information on intrinsic properties, such as physicochemical and various ecotoxicological properties, including the information derived by read-across.

1.1 Read-Across

Read-across is a non-testing strategy based on the fundamental principle that similar chemicals have similar chemical behaviors. Principally, read-across can be used for both qualitative and quantitative assessments of different physicochemical properties and various ecotoxicity endpoints. Read-across can be based on interpolation or extrapolation. There are two read-across approaches: (1) chemical analogue (analogue approach) and (2) chemical category (category approach). Chemical analogue approach can be considered for a limited chemical category, where trends in properties are not apparent [7], usually consisting of two chemicals: one source and one target chemical. In the usual practice, the endpoint information for a substance (source

chemical) is used to predict, i.e., “read-across,” the same endpoint information for another substance (target chemical), which is “similar” enough to the source chemical. The “similarity” is usually structural; however, it can be based on “similar” biological behaviors at the molecular level, such as mode/mechanism of action (MOA). Among the options to identify structural analogues, Tanimoto coefficient (also known as Jaccard similarity) is the most commonly used one. Tanimoto coefficient is an index of the similarity between two chemical structures represented as two vectors and is based on the topological descriptions of atoms and connecting bonds [8].

On the other hand, a chemical category is a group of chemicals with physicochemical, toxicological, and ecotoxicological properties that are likely to be similar or follow a regular pattern. The trend in a category of chemicals is often the presence, absence, or modulation of a particular effect for all members of the category, based on the presumption of a common MOA or adverse outcome pathway (AOP) [7, 9]. The similarities may be based on (1) a common functional group; (2) common constituents or chemical classes; (3) common precursors and/or the likelihood of common breakdown products via physical and biological processes, resulting in structural similarity; and (4) an incremental and constant change in the property across the category [10]. In order to identify if a query chemical structurally falls in a category or not (applicability domain of a category), the boundary conditions, such as the range of molecular weight, the logarithm of the octanol/water partition coefficient ($\log K_{ow}$ or $\log P$) values, water solubility, etc., should be defined.

1.2 Regulatory Read-Across, Documentation, and Guidance

The use of read-across is encouraged for specific regulatory data requirements, such as endpoint prediction, screening and prioritization, and ecotoxicity assessment. For instance, the REACH Regulation sets out the legislative framework for the use of grouping of substances and read-across for data gap filling as follows:

Application of the group concept requires that physico-chemical properties, human health effects and environmental effects or environmental fate may be predicted from data for reference substance(s) within the group by interpolation to other substances in the group (read-across approach).

Indeed, an analysis of the ecotoxicity data of 2887 substances submitted for REACH purposes indicated that for 10–15% of these substances, ecotoxicity values were derived based on read-across (10.8% for fish, 11.2% for aquatic invertebrates, and 14.6% for algae and aquatic plants) [11]. However, the entire process of read-across assessment including a rationale for analogue or category approach, expert judgment, and well-documented justification is required for regulatory acceptance. Transparency, reproducibility, data quality, and certainty are key for read-across in regulatory setting.

According to a recent report summarizing the current state of regulatory use of read-across in the EU, the United States, and Japan [12], despite some challenges, there have been significant efforts across the globe toward the standardization and good read-across practice [13]. Similarly, remarkable efforts have been made by various organizations for the development of most applicable regulatory and technical guidance, such as the Organisation for Economic Co-operation and Development (OECD) [9, 14], European Chemicals Agency (ECHA) [15–17], and European Centre for Ecotoxicology and Toxicology of Chemicals (ECE-TOC) [18]. Patlewicz et al. [19] surveyed the efforts for the read-across development as well as its scientific justification and documentation. This report provided a pivotal commentary perspective and a comprehensive review of the existing literature. They proposed a harmonized hybrid framework to help reconcile the common guiding principles and steps of the read-across process which should be helpful in expanding the scope and decision context of the existing frameworks.

1.3 Use of Read-Across for Industrial and Regulatory Purposes: A Literature Review

The read-across approach allows prediction for a specific endpoint of a chemical using experimental data available from reasonably similar compounds. Read-across is anticipated to be an important alternative for animal testing under the REACH Regulation [19].

Furthermore, the most frequently used Integrated Approaches to Testing and Assessment (IATA) is read-across. IATA comprises science-based approaches for chemical hazard characterization. IATA can include a combination of methods and can be informed by integrating results from one or many methodological approaches [e.g., (Q)SAR, read-across, *in chemico*, *in vitro*, *in vivo*]. The IATA Case Studies Project enabled the OECD to make a more reflective and dynamic analysis of the application of read-across for toxicological data gap filling [9, 20, 21].

As such, read-across can be applied for the prediction of any (eco)toxicological endpoint. The endpoint-specific modeling toxicity is usually related to human health endpoints, such as repeated-dose toxicity, skin sensitization, etc. There is a plenty of published reports in this field. However, read-across studies on environmental endpoints are scarce. A selection of the read-across efforts published in the literature is provided in the following paragraphs.

Wu et al. [22] described a framework that identifies potential analogues which can be used by a toxicologist for consideration in read-across. Their approach involves categorizing potential analogues based on various chemical and metabolic considerations, as well as modes of toxicity for categorizing potential analogues to their suitability, while explicitly detailing assumptions inherent in any read-across of toxicological data. Their result provided a framework to apply chemical, biochemical, and toxicological principles in a systematic manner to identify and evaluate factors that can

introduce uncertainty into structure-activity relationship (SAR) assessments while maximizing the appropriate use of all available data.

Vink et al. [23] illustrated some important aspects of applying toxicological read-across in human health risk assessment with a case study. They combined read-across as non-testing strategy with a tiered exposure assessment for the risk characterization of 1-methoxypropan-2-ol (PGME) as a representative for phase-in substances to be registered under REACH. PGME is currently used as a component of coatings and cleaners [24]. They selected three chemicals with an acceptable toxicological dataset (1) 2-propanol,1-ethoxy-(PGEE), (2) 2-propanol,1-propoxy (PGPE), and (3) propylene glycol n-butyl ether (PnB) which shared a similarity to PGME of at least 50%. PGEE was found to be the most related to PGME, based on both structure and physicochemical properties. Therefore, Vink et al. [23] used PGEE as the source chemical for read-across. They also used toxicological data available for PGPE and PnB to strengthen the outcome of the read-across. They provided data, which were comparable with experimental data available for target substance PGME, resulting in a realistic starting point for both qualitative and quantitative risk assessment.

Schüürmann et al. [25] used quantitative read-across for predicting the acute fish toxicity of organic compounds employing atom-centered fragments (ACFs) for evaluating chemical similarity.

Cronin et al. [26] provided information on chemical grouping, categories, and read-across to predict toxicity. Madden [27] provided information on data useful for category formation and read-across. Additionally, Cronin evaluated the categories and read-across for toxicity prediction allowing for regulatory acceptance [28] and expressed the state-of-the-art and future directions of category formation and read-across for toxicity prediction [29].

Rand-Weaver et al. [30] critically reviewed the evidence for read-across and found that few studies include plasma concentrations and MOA-based effects and highlighted the absence of documented evidence. They also attracted attention to a need for large-scale studies to generate robust data for testing the read-across hypothesis and to develop predictive models, the only feasible approach for protecting the environment.

Low et al. [31] represented a novel hybrid read-across method that is both predictive and interpretable by combining the simplicity and transparency of read-across methods with the benefits afforded by more sophisticated techniques such as ensemble modeling and instance-based learning while incorporating diverse data streams. They classified different types of endpoints (hepatotoxicity, hepatocarcinogenicity, mutagenicity, and acute lethality) using several biological data types (gene expression profiling and cytotoxicity screening).

Rorije et al. [32] provided read-across case studies for the estimation of the aquatic toxicity of five different fragrance substances and proposed a pragmatic approach for expressing uncertainty in read-across estimates. The estimated aquatic toxicity and their uncertainties are used to estimate freshwater compartment Predicted No Effect Concentrations (PNECs), which is directly used in risk characterization. The results of the musk fragrance read-across cases (musk xylene, musk ketone, and galaxolide) were compared to experimentally derived PNEC values.

Oomen et al. [33] used grouping and read-across approaches for risk assessment of nanomaterials. They proposed that these approaches facilitate for better use of available information on nanomaterials and are flexible enough to allow future adaptations related to scientific developments.

Schultz et al. [34] proposed a strategy for structuring and reporting a read-across prediction of toxicity and suggested a workflow for reporting a read-across prediction.

Benfenati et al. [35] integrated QSAR and read-across for environmental assessment addressing two main questions: (1) How do they solve the issue of the subjectivity in the evaluation of data and results, which may be particularly critical for read-across, but may have a role also for the QSAR assessment? (2) How do they take advantage of the results of both approaches to support each other? They developed a freely available program called ToxRead (www.toxgate.eu), which integrated with the output of QSAR, within a weight-of-evidence strategy, using the assessment of bioconcentration factors of chemicals.

Stanton and Kruszewski [36] quantified the benefits of using read-across and in silico techniques to fulfill hazard data requirements for chemical categories. They presented the actual benefits resulting from avoided testing through the use of read-across and in silico tools. They evaluated that the use of 100,000–150,000 test animals and the expenditures of \$50,000,000–\$70,000,000 were avoided when 261 noted substances are considered.

Zhu et al. [37] used omics data to establish biological similarity: Examples were given for in vitro stem cell models and short-term in vivo repeated-dose studies in rats used to support read-across and category formation. These preliminary biological data-driven read-across studies provided some new read-across approaches that can be used for chemical safety assessment.

Schultz et al. [38] used the category approach to read-across to predict repeated-dose toxicity for a variety of derivatives of 2-alkyl-1-alkanols as a case study. Specifically, the no-observed-adverse-effect levels (NOAELs) of 2-ethyl-1-hexanol and 2-propyl-1-heptanol, the source substances, can be read across with confidence to untested 2-alkyl-1-alkanols in the C5–C13 category based on the lowest-observed-adverse-effect level (LOAEL) of low systemic toxicity. These branched alcohols have metabolic pathways that have

significance to repeated-dose toxic potency. The chemical category is limited to the readily bioavailable analogues. Their findings revealed that the 90-day rat oral repeated-dose NOAEL values for the two source substances can be read across to fill the data gaps of the untested analogues in this category with uncertainty deemed equivalent to results from a TG 408 assessment.

Gajewicz [39] developed predictive read-across models based on real-life nanotoxicity data. The main practical difference of their algorithm compared to existing read-across approaches is the fact that it avoids the limitations associated with using only one descriptor or a maximum of two descriptors as indicated in the case of recently published nano-quantitative read-across algorithms. Gajewicz et al. [39] addressed the bottleneck for regulation of nanomaterials and proposed a quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. The proposed Nano-QRA approach is a simple and effective algorithm for filling data gaps in quantitative manner and provides an efficient tool to support the risk assessment of nanomaterials. Research priorities relevant to development of updating of nano-relevant regulations and guidelines covering read-across was reported by the EU NanoSafety Cluster [40].

Patlewicz et al. [41] reviewed in silico tools for grouping with a focus on the challenges involved in read-across development and its scientific justification and documentation.

Floris and Olla [42] summarized molecular similarity in computational toxicology within the context of the read-across approach. This book chapter reports an implementation of chemical similarity and the analysis of multiple combinations of binary fingerprints and similarity metrics in the context of the read-across technique. Their approach has been implemented in two open-source software tools for computational toxicology (CAESAR and VEGA).

Mellor et al. [8] considered various molecular fingerprints and similarity measures to be used to calculate molecular similarity. They investigated the value and concordance of the Tanimoto similarity values calculated using six widely used fingerprints within six toxicological datasets. Their results suggested that for read-across, generic fingerprint-derived similarities are likely to be most predictive after chemicals are placed into categories (or groups) and then similarity is calculated within those categories; ideally, the specific similarity measure should be appropriate to the chemistry and endpoint considered.

The focal step of acquiring acceptance of a read-across prediction is to identify and assess the uncertainties associated with it. Schultz et al. [43] assessed uncertainty in read-across: They identified and summarized the main sources of uncertainty that potentially impact the acceptance of a read-across argument. They also proposed a series of questions to evaluate toxicity predictions

based on knowledge gained from case studies, in order to progress further in this area.

Lombardo et al. [44] presented and discussed the recently published guideline from the European Food Safety Authority (EFSA) for integrating and weighting evidence for scientific assessment. They reported the criteria and application on the use of non-testing methods (NTM) within a weight-of-evidence strategy. They stated that NTM are a valuable resource for risk assessment of chemical substances and can be particularly useful when the information provided by different sources was integrated. Hence, this integration increases the confidence in the final result. They assessed bioconcentration factor (BCF) prediction using in silico models and demonstrated the suitability and effectiveness of in silico methods for risk assessment with this example and proposed a practical guide to end users to perform similar analyses on likely hazardous chemicals.

1.4 Available Software and Tools Used in Ecotoxicity-Related Read-Across Predictions

With the encouragement of the regulatory bodies, many computational tools have been developed for safety assessment in the last decades. Some of the prominent software and tools used in similarity and data search, grouping, metabolism prediction, and read-across predictions for the ecotoxicological endpoints and environmental fate properties are summarized below. These tools accept extensively used chemical identifiers such as chemicals' name, CAS registry number, EC number, and simplified molecular-input line-entry specification (SMILES) notation. Additionally, a structure drawing editor option may be available. Structural similarity, mechanism similarity, and structural alerts can be used to group chemicals into categories. Structural alerts may also be used as supporting tool for hazard and risk assessment [45]. Some of the QSAR methods, such as k-nearest neighbor (KNN) and hierarchical clustering, could be regarded as read-across prediction methods, since they group chemicals according to their structural similarities obtained from structural and/or physicochemical property descriptors and structural alerts [35].

OECD QSAR Toolbox (v. 4.3) [46] has been developed by LMC and supported by governments and industry. This stand-alone software performs profiling based on the identification of relevant structural features and potential mechanisms of a query chemical, grouping chemicals based on the structural and/or mechanistical features, and filling data gaps making use of available experimental data obtained from various databases. QSAR Toolbox supports making predictions and leaves the decision to the user.

US EPA TEST (v. 4.2.1) [47] performs mutagenicity and BCF endpoint predictions. The software includes four methods for BCF prediction, hierarchical clustering, single model, group contribution, and nearest neighbor, in addition to a consensus model. The dataset used in the BCF model consists of 676 chemicals compiled

from different sources [48]. Single model and group contribution method employ linear regression method. The others basically cluster the training set chemicals.

BCF/BAF (bioaccumulation factor) and biodegradation prediction models are available under EPISuite™ [49] of EPA. BCFBAF™, formerly called BCFWIN™, estimates fish BCF using two different methods. The first one is the regression-based model employing log P. The second one is the Arnot-Gobas method, which calculates BCF from mechanistic first principles. BCFBAF also incorporates prediction of apparent metabolism half-life in fish and estimates BCF and BAF for three trophic levels. Seven linear and nonlinear BIOWIN models estimate the probability of rapid aerobic and anaerobic biodegradation of organic compound in the presence of mixed populations of environmental microorganisms.

VEGA (v. 1.1.4) application has been developed by Istituto di Ricerche Farmacologiche Mario Negri (Laboratory of Environmental Chemistry and Toxicology) and KODE (www.vega-qsar.eu). Various human and environmental toxicological endpoints and physicochemical properties are predicted by VEGA software. Among the ecotoxicological endpoints, fish and *Daphnia magna* LC₅₀ and bee acute toxicity models are available. A toxicity classification for the query chemical based on the predicted LC₅₀ value can be made. Environmental endpoints include ready biodegradability model (IRFMN); kM/half-life model (Arnot/EPI-Suite); persistency models in sediment, soil, and water (IRFMN); and three BCF models (CAESAR, Meylan, and KNN/read-across). Prediction of the endpoint for the query chemical and the most similar structures available in the dataset with their experimental and predicted values are reported at the end of the modeling process. The overall reliability of each model is graded by applicability domain index (ADI) and reported.

ToxRead (v. 0.11) is another software developed by Istituto di Ricerche Farmacologiche Mario Negri, Politecnico di Milano, and KODE (<http://www.toxread.eu/>). The software assists users in making reproducible read-across evaluations as well as showing the similar chemicals, structural alerts, and relevant features in common between the chemicals. Data within the software are from VEGA platform. There are 860 experimental BCF values and log P values that are derived from KOWWIN model of EPI-Suite™ [35]. In the initial stage of the prediction, the number of chemicals similar to the target chemical is asked. Then, the most similar chemicals, in a number chosen by the user, are selected from the database based on the log P values.

Toxtree (v. 3.1) has been developed by Ideaconsult Ltd. [50]. The software implements several classification schemes using chemical structures, metabolic pathways, and descriptors calculated from chemical structures for predicting various endpoints.

Classifying the chemicals according to Cramer rules and extended Cramer rules and identifying structural alerts for various endpoints are available within this stand-alone software. For an ecotoxicological endpoint, START Biodegradability model makes estimations using a decision tree. At the end of the process, the query chemical is allocated into one of the three classes: Class 1 (easily biodegradable), Class 2 (persistent chemical), and Class 3 (unknown biodegradability).

Danish (Q)SAR Database [51], developed by the National Food Institute, Technical University of Denmark, has been in use since 2004. A collection of over 200 QSAR models from free and commercial sources for various physicochemical and environmental endpoints is available along with training data on the website. Substructure and similarity search as well as searching by model endpoint can be performed. Not ready biodegradability models of three commercial software and environmentally related endpoint models of EPISuite™ are available.

Toxmatch (v. 1.07) is also developed by Ideaconsult Ltd. The open-source and freely available software performs similarity evaluation of two chemicals. Highly similar chemicals can be used in source chemical selection and category formation to support the application of read-across. Three descriptors, which are relevant for BCF, are considered for similarity assessment. Then, the similarity metric Euclidean distance, calculated from effective diameter, maximum diameter, and log P, is used as the predictor. The dataset of experimental BCF values is the one used in the EPISuite's BCFBAF model. Read-across for BCF can be performed based on this dataset. The software allows the users to categorize the dataset into groups, such as a set of different ranges for a BCF value.

ChemIDplus (<https://chem.nlm.nih.gov/chemidplus/>) website managed by the National Institutes of Health (NIH) contains over 300,000 chemical structure records integrated with other TOXNET records. A search is available in various options, such as searching for similar compounds within a lower limit of similarity or for substructures.

ToxDelta [52] uses MCS (maximum common substructure) concept in a different manner. While the similarity of two compounds is compared and the MCS of two chemicals is found, the molecular fragments which are not in common are also listed. These fragments are categorized as active and inactive fragments and employed in, for example, the toxicity assessment.

CBRA software [31] evaluates the in vivo activity of chemicals using the chemical-biological read-across approach. The software employs calculated structural descriptors and experimentally measured in vitro bioactivity profiles as input. At the end of the process, it generates radial plots representing chemical and biological neighbors of the query chemical.

BioTransformer software consists of five modules, called “transformers,” predicting small molecule metabolism in mammals as well as in the soil and aquatic microbiota in the environmental compartments [53]. Chemicals having common metabolites can be used in read-across applications of RAAF scenarios 1, 3, or 5, where metabolites of the analogues are considered.

Analog Identification Methodology (AIM) (v1.01) tool (SRC, Inc., North Syracuse, NY) has been developed by SRC, Inc. for US EPA, in cooperation with EPA’s Risk Assessment Division. The software searches for analogues of the query chemical based on its chemical structure. At the end of the analysis, a list of potential analogues and corresponding toxicity data sources for the analogues is provided.

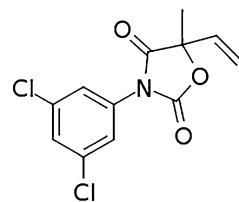
2 Case Study

Stepwise procedures for both analogue and category approaches are reported in the OECD guidance document [9]. These recommended approaches have common steps such as (1) inspecting the chemical for the possible chemical category/class, (2) identifying possible analogues, (3) gathering data for each chemical, (4) evaluating the data for adequacy, (5) constructing a data matrix, and (6) documentation. Here, we present a case study to predict BCF adopting analogue approach.

BCF is an important property describing fate and behavior of chemicals in the environment. The information on BCF can be used in PBT (persistent, bioaccumulative, and toxic) and vPvB (very persistent and very bioaccumulative) assessments, hazard classification, and chemical safety assessment, as well as in deciding whether long-term ecotoxicity testing might be necessary. Water solubility and log P are the key parameters estimating this property. Therefore, similar solubility and log P along with structural similarity are important in the read-across prediction of BCF.

Vinclozolin (3-(3,5-dichlorophenyl)-5-ethenyl-5-methyl-1,3-oxazolidine-2,4-dione) (Fig. 1) is a widely used fungicide with androgen receptor antagonism [54]. It is included in the OECD’s list of HPV chemicals. However, its experimental BCF value is not available in the literature.

As a first attempt of estimation of its BCF, “seven most similar options” were selected in ToxRead for potential source chemicals. Two of these analogue chemicals had estimated log P values, and their predictions were excluded from the list. These chemicals with their experimental BCF and corresponding log P values are given in Table 1. Given the fact that BCF values are highly correlated with log P, the analogues with close log P values to that of vinclozolin could be chosen as source chemicals. Propanil, iprodione, and linuron have log P values between 3 and 3.2 and experimental

**Fig. 1** 2D structure of vinclozolin
Table 1
Potential source chemicals obtained from the software used

Name	SMILES	Structure	Exp. Log P	Exp. BCF
Chlorpropham	CC(C)OC(=O)NC1=CC(=CC=C1)Cl		3.51	2.16 ^a
Propanil	CCC(=O)NC1=CC(=C(C=C1)Cl)Cl		3.07	2.05 ^a
Iprodione	CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl		3.00	1.85 ^a
Linuron	CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC		3.20	1.29 ^a
Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl		2.68	0.74 ^a
Oxadiazon	CC(C)OC1=C(C=C(C(=C1)N2C(=O)OC(=N2)C(C)(C)C)Cl)Cl		4.80	1.89 ^b

Experimental values: ^aToxRead ^bUS EPA TEST

BCF values between 1.29 and 2.05. Average log BCF of these three chemicals is 1.73.

US EPA TEST program was employed for the BCF prediction of the target chemical, vinclozolin. Four different predictions from

Table 2
Physicochemical properties^a of the potential analogues for vinclozolin

Name	CAS	Log P	pKa	Water solubility	MW g/mol	Melting point (°C)	Boiling point (°C)	Vapor pressure (Pa, 25 °C)
Vinclozolin (Target chemical)	50471-44-8	3.10	No ionizable atom found	2.6 mg/L (20 °C)	286.12	187.73	446.77	1.21E-005
Chlorpropham	101-21-3	3.51	12.89	89 mg/L (25 °C)	213.67	64.92	283.19	0.453
Propanil	709-98-8	3.07	13.90	152 mg/L (25 °C)	218.08	130.95	354.93	0.00376
Iprodione	36734-19-7	3.00	13.94	13.9 mg/L (25 °C)	330.17	233.57	544.90	2.01E-008
Linuron	330-55-2	3.20	11.94	75 mg/L (25 °C)	249.10	137.23	365.91	0.00162
Diuron	330-54-1	2.68	13.18	42 mg/L (25 °C)	233.10	126.39	353.86	0.00062
Oxadiazon	19666-30-9	4.80	No ionizable atom found	0.7 mg/L (24 °C)	345.23	178.65	431.28	4.5E-005

^aPhysicochemical properties are from EPISuite™; pKa predictions are made using MarvinSketch (v. 17.29.0) (ChemAxon, 2017)

available models (hierarchical clustering, 1.47; single model, 2.16; group contribution, 1.63; and FDA, 2.04) and a consensus prediction value (1.83) were obtained. The only chemical in the database with structural similarity greater than 50% (0.56) was oxadiazon. Close values of the predicted and experimental values (1.92 and 1.89, respectively) for oxadiazon support the reliability of the model.

Three models of VEGA software were used for the log BCF prediction. The same chemicals appeared in these models as potential source chemicals. However, their similarity metrics are different. CAESAR and Meylan models resulted in ADIs of 0.85 and predictions of 1.37 and 1.71, respectively. Although KNN/read-across prediction appears to show low reliability with ADI of 0.70, all predictions, varying from 1.37 to 1.71, are in agreement indicating vinclozolin is non-bioaccumulative. Carbonyl residue alert was found in all chemicals under study in CAESAR model. This alert is found for a large number of non-bioaccumulative chemicals, even when log P value is very high. Presence of >C=O polar group increases hydrophilicity, providing lower values of BCF.

The analogue chemicals assessed by the software packages have similarities to the target chemical ranging between 0.56 and 0.84. The target and the source chemicals have similar physicochemical properties (Table 2), leading to similar environmental behavior. Additionally, close pKa values indicate that the analogue chemicals are not expected to hydrolyze. The experimental values within the prediction software packages are adequate, and the predictions obtained are also in agreement. Apparently, estimated log BCF of vinclozolin is less than 3.3 based on the available information from BCF calculations, and this chemical could be categorized as not bioaccumulative [55].

3 Conclusions/Future Prospects

Toxicity assessment of chemicals to human and the environment is required for regulatory reasons. Justified and well-documented read-across enables to provide information on endpoint estimation, screening and prioritization, and risk characterization. In this chapter, we presented the regulatory read-across, documentation, and guidance together with the use of read-across for industrial and regulatory purposes. We also summarized free software packages and websites to be used for read-across. As a case study, we provided bioconcentration factor (BCF), an important property describing fate and behavior of chemicals in the environment, information from read-across by employing analogue approach. Given the popularity and continuous efforts toward the development, justification, and documentation, the future seems to be promising for grouping, category formation, and read-across.

References

- CAS Registry. <http://support.cas.org/content/chemical-substances>. Accessed Apr 2019
- Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, Dellarco V, Henry T, Holderman T, Sayre P (2008) The toxicity data landscape for environmental chemicals. Environ Health Perspect 117(5):685–695
- European Commission, Regulation No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). Official Journal of the European Union, L 396/1–849
- http://elaw.klri.re.kr/kor_service/lawView.do?hseq=31605&lang=ENG. Accessed Apr 2019
- Çevre ve Şehircilik Bakanlığı, Kimyasalların Kaydı, Değerlendirilmesi, İzni ve Kısıtlanması Hakkında Yönetmelik (KKDİK). Türkiye Cumhuriyeti Resmi Gazete, 23 Haziran 2017, Sayı: 30105 (Mükerrer), 464–855
- https://www.legislation.gov.uk/ksi/2019/758/pdfs/ksi_20190758_en.pdf?title=reach. Accessed Apr 2019
- Guidance Document for using the OECD (Q) SAR application toolbox to develop chemical categories guidance on grouping of chemicals, OECD series on testing and assessment No. 102, 2009
- Mellor C, Robinson RM, Benigni R, Ebbrell D, Enoch S, Firman J, Madden J, Pawar G, Yang C, Cronin M (2019) Molecular fingerprint-derived similarity measures for toxicological read-across: recommendations for

- optimal use. *Regul Toxicol Pharmacol* 101:121–134
9. Organisation for Economic Cooperation and Development (OECD) (2014) Guidance on grouping of chemicals, 2nd edn, No. 194, series on testing & assessment. ENV/JM/MONO(2014)4, OECD, Paris
 10. https://www.epa.gov/sites/production/files/2014-10/documents/ncp_chemical_categories_august_2010_version_0.pdf. Accessed Apr 2019
 11. Sobanska MA, Cesnaitis R, Sobanski T, Versonnen B, Bonnomet V, Tarazona JV, De Coen W (2014) Analysis of the ecotoxicity data submitted within the framework of the REACH Regulation. Part 1. General overview and data availability for the first registration deadline. *Sci Total Environ* 470:1225–1232
 12. Chesnut M, Yamada T, Adams T, Knight D, Kleinstreuer N, Kass G, Luechtefeld T, Hartung T, Maertens A (2018) Regulatory acceptance of read-across. *ALTEX* 35 (3):413–419
 13. Ball N, Cronin MT, Shen J, Blackburn K, Booth ED, Bouhifd M, Donley E, Egnash L, Hastings C, Juberg DR (2016) t4 report: toward good read-across practice (GRAP) guidance. *ALTEX* 33(2):149
 14. Guidance on Grouping of Chemicals (2007) OECD series on testing and assessment No. 80, Organisation for Economic Co-operation and Development, Paris
 15. ECHA, Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.6: QSARs and grouping of chemicals
 16. Read-Across Assessment Framework (RAAF) (2015) ECHA-15-R-07-EN
 17. Read-Across Assessment Framework (RAAF) (2017) ECHA-17-R-01-EN
 18. ECETOC (2012) Technical Report 116 Category Approaches, Read-Across, (Q)SAR
 19. Patlewicz G, Cronin MT, Helman G, Lambert JC, Lizarraga LE, Shah I (2018) Navigating through the minefield of read-across frameworks: a commentary perspective. *Comput Toxicol* 6:39
 20. Organisation for Economic Cooperation and Development (OECD), Report on Considerations from Case Studies on Integrated Approaches for Testing and Assessment (IATA), First Review Cycle (2015) Case studies on grouping methods as a part of IATA, No. 250, series on testing & assessment. ENV/JM/MONO(2016)
 21. Organisation for Economic Cooperation and Development (OECD) (2018) Report on considerations from case studies on integrated approaches for testing and assessment (IATA), third review cycle (2017), case studies on grouping methods as a part of IATA, No. 289, series on testing & assessment. ENV/JM/MONO(2018) 25, OECD, Paris.
 22. Wu S, Blackburn K, Amburgey J, Jaworska J, Federle T (2010) A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments. *Regul Toxicol Pharmacol* 56(1):67–81
 23. Vink S, Mikkers J, Bouwman T, Marquart H, Kroese E (2010) Use of read-across and tiered exposure assessment in risk assessment under REACH—A case study on a phase-in substance. *Regul Toxicol Pharmacol* 58(1):64–71
 24. SIDS Initial Assessment Report for SIAM 17 (2003) Propylene glycol ethers, Arona
 25. Schüürmann G, Ebert R-U, Kühne R (2011) Quantitative read-across for predicting the acute fish toxicity of organic compounds. *Environ Sci Technol* 45(10):4616–4622
 26. Cronin MTD, Madden JC, Roberts DR, Enoch SJ (2013) Chemical toxicity prediction: category formation and read-across, vol 17. Royal Society of Chemistry, Cambridge
 27. Madden JC (2013) Sources of chemical information, toxicity data and assessment of their quality. In: Cronin MTD , Madden JC, Enoch SJ, Roberts DW (eds) Chemical toxicity prediction: category formation and read-across. Royal Society of Chemistry, Cambridge, pp 98–126
 28. Cronin MTD (2013) Evaluation of categories and read-across for toxicity prediction allowing for regulatory acceptance. In: Chemical toxicity prediction: category formation and read-across. Royal Society of Chemistry, Cambridge, pp 155–167
 29. Cronin MTD (2013) The state of the art and future directions of category formation and read-across for toxicity prediction. In: Chemical toxicity prediction: category formation and read-across. Royal Society of Chemistry, Cambridge, pp 168–179
 30. Rand-Weaver M, Margiotta-Casaluci L, Patel A, Panter GH, Owen SF, Sumpter JP (2013) The read-across hypothesis and environmental risk assessment of pharmaceuticals. *Environ Sci Technol* 47(20):11384–11395
 31. Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A (2013) Integrative chemical–biological read-across approach for chemical hazard classification. *Chem Res Toxicol* 26(8):1199–1208

32. Rorije E, Aldenberg T, Peijnenburg W (2013) Read-across estimates of aquatic toxicity for selected fragrances. *Altern Lab Anim* 41:77–90
33. Oomen A, Bleeker E, Bos P, van Broekhuizen F, Gottardo S, Groenewold M, Hristozov D, Hund-Rinke K, Irfan M-A, Marcomini A (2015) Grouping and read-across approaches for risk assessment of nanomaterials. *Int J Environ Res Public Health* 12 (10):13415–13434
34. Schultz T, Amcoff P, Berggren E, Gautier F, Klaric M, Knight D, Mahony C, Schwarz M, White A, Cronin M (2015) A strategy for structuring and reporting a read-across prediction of toxicity. *Regul Toxicol Pharmacol* 72 (3):586–601
35. Benfenati E, Roncaglioni A, Petoumenou M, Cappelli C, Gini G (2015) Integrating QSAR and read-across for environmental assessment. *SAR QSAR Environ Res* 26(7–9):605–618
36. Stanton K, Kruszewski FH (2016) Quantifying the benefits of using read-across and in silico techniques to fulfill hazard data requirements for chemical categories. *Regul Toxicol Pharmacol* 81:250–259
37. Zhu H, Bouhifd M, Kleinstreuer N, Kroese ED, Liu Z, Luechtefeld T, Pamies D, Shen J, Strauss V, Wu S (2016) t4 report: supporting read-across using biological data. *ALTEX* 33 (2):167
38. Schultz TW, Przybylak KR, Richarz A-N, Mellor CL, Bradbury SP, Cronin MT (2017) Read-across of 90-day rat oral repeated-dose toxicity: a case study for selected 2-alkyl-1-alkanols. *Comput Toxicol* 2:28–38
39. Gajewicz A (2017) Development of valuable predictive read-across models based on “real-life”(sparse) nanotoxicity data. *Environ Sci Nano* 4(6):1389–1403
40. Stone V, Önlü S, Bergamaschi E, Carlander D, Costa A, Engelmann W, Ghanem A, Hartl S, Hristozov D, Scott-Fordsmand JJ (2017) Research priorities relevant to development or updating of nano-relevant regulations and guidelines
41. Patlewicz G, Helman G, Pradeep P, Shah I (2017) Navigating through the minefield of read-across tools: a review of in silico tools for grouping. *Comput Toxicol* 3:1–18
42. Floris M, Olla S (2018) Molecular similarity in computational toxicology. In: Computational toxicology. Springer, pp 171–179
43. Schultz TW, Richarz A-N, Cronin MT (2019) Assessing uncertainty in read-across: questions to evaluate toxicity predictions based on knowledge gained from case studies. *Comput Toxicol* 9:1–11
44. Lombardo A, Raitano G, Gadaleta D, Benfenati E (2018) Criteria and application on the use of nontesting methods within a weight of evidence strategy. In: Computational toxicology. Springer, New York, pp 199–218
45. Valsecchi C, Grisoni F, Consonni V, Ballabio D (2019) Structural alerts for the identification of bioaccumulative compounds. *Integr Environ Assess Manag* 15(1):19–28
46. QSAR Toolbox. <http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm>. Accessed Apr 2019
47. US EPA TEST. <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>. Accessed Apr 2019
48. US EPA TEST manual. <https://www.epa.gov/sites/production/files/2016-05/documents/600r16058.pdf>. Accessed Apr 2019
49. EPISuite. <https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>. Accessed Apr 2019
50. Patlewicz G, Jeliazkova N, Safford R, Worth A, Aleksiev B (2008) An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* 19(5–6):495–524
51. Danish (Q)SAR Database, Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark. <http://qsar.food.dtu.dk>. Accessed Apr 2019
52. Golbamaki A, Franchi A, Manganelli S, Manganaro A, Gini G (2017) ToxDelta: a new program to assess how dissimilarity affects the effect of chemical substances. *Drug Des* 6 (153):2169.–0138.1000153
53. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform* 11(1):2
54. Pohanish RP (2014) Sittig’s handbook of pesticides and agricultural chemicals. William Andrew Publishing, Norwich, New York
55. ECHA Guidance on Information Requirements and Chemical Safety Assessment Chapter R.11: PBT/vPvB assessment Version 3.0 June 2017



Chapter 14

Methodological Protocol for Assessing the Environmental Footprint by Means of Ecotoxicological Tools: Wastewater Treatment Plants as an Example Case

Roberta Pedrazzani, Pietro Baroni, Donatella Feretti, Giovanna Mazzoleni, Nathalie Steimberg, Chiara Urani, Gaia Viola, Ilaria Zerbini, Emanuele Ziliani, and Giorgio Bertanza

Abstract

The ecotoxicological tools reveal to be profitably employable within the assessment of the so-called environmental footprint, which is commonly based on the results of a chemical monitoring. Due to the heterogeneity of biological endpoints and the possibility to explore several exposure frames, as well as to consider higher levels of organization (from cells to organisms and mesocosms), the definition of a protocol is desirable.

Key words Activated sludge, Baseline toxicity, Bioassays, Endocrine disruption, Environmental footprint, Genetic toxicity, Protocol, Modes of action, Multi-tiered approach, Wastewater

1 Introduction

Wastewaters are complex mixtures, in which a multitude of pollutants, emitted from different anthropogenic sources, is concentrated. Conventional wastewater treatment plants (WWTPs) guarantee, in some cases, only a partial removal of these pollutants [1]; on the other hand, raw and treated wastewaters may reach the surface waters through sewer overflows and WWTP effluents, respectively. Therefore, in order to define the optimal system for wastewater management (collection and transport systems, overflow regulating devices, treatment technologies), a deeper investigation of the impact of sewage (raw and treated) on the ecosystem is required [2].

For this purpose, two different approaches may be adopted: the life cycle approach, based on chemical input data, which quantify the impacts on each environmental compartment (including toxicity toward freshwater ecosystem), and procedures based only on

biological assays and ecotoxicological tools. The latter represents the basic tool, from which the following approaches set in: the Biological Assays and Diseases (BAD) [3, 4], the Whole Effluent Toxicity (WET) [5–7], the Direct Toxicity Assessment (DTA) [8], and the Whole Effluent Toxicity [9, 10, 11]. Each procedure is characterized by intrinsic limitations. Our approach tries to combine these methodologies with the aim of overcoming their shortcomings and lacks and integrating their strength points.

Indeed, the detection and quantification of all the substances and their biotic and abiotic transformation products potentially present in an environmental matrix cannot be a workable solution [12–14]. The single components of the mixture might be unknown or present at concentrations below the limits of detection [15, 16]. Most of all, chemical analyses do not allow to predict the impacts of a mixture on an organism or on the whole ecosystem, due to possible additive/subtractive/synergistic/antagonistic effects [17, 18]. Only biological assays enable to take into account these complex interactions and quantify the overall impact of wastewater as a whole [19].

The assessment of toxicity entails setting the boundaries of the biological investigation, so as to eliminate the risk of haziness, possibly arising from the results of the investigation. Actually, toxicological tools should be chosen and employed with a careful attention to the information they can provide; the definition of the criteria underpinning both the selection of the biological assays and the interpretation of the results for the environmental impact assessment is still a matter of debate [20]. Furthermore, the results of the bioassays cannot be directly used for the environmental footprint assessment by means of standardized methodologies based on life cycle assessment (LCA) criteria.

For these reasons, an integrated strategy based on the combination of chemical and biological analyses should be adopted for better investigating the impact of effluents [21, 22]. This work presents a protocol, which includes consecutive stages: from setting criteria for sample collection and preparation to the definition of a battery of multi-tiered tests, of the procedures of experimental data elaboration, and, finally, to the assessment of the environmental footprint and, consequently, the proposal of a role in the decision processes (Fig. 1). In particular, this protocol was applied to three real-scale WWTPs located in Northern Italy, receiving domestic and industrial wastewaters and equipped with conventional and innovative (Membrane Biological Reactor, MBR) treatment technologies.

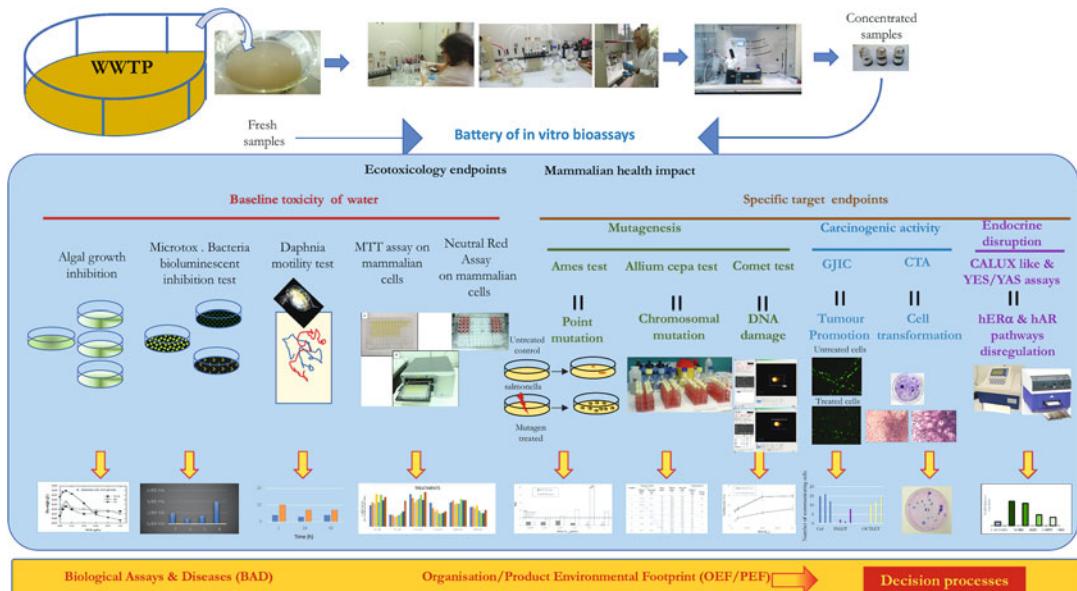


Fig. 1 Scheme of the protocol for assessing the environmental footprint by means of ecotoxicological tools: from the choice of bioassays to the participation at decision processes

2 Sampling

The acquisition of samples represents a crucial step in the evaluation of the plant performance and hence of the characteristics of its effluent. A first bond, which cannot be avoidable, is the observance of sludge age and hydraulic retention time. A second bond lies in the wide variability of (trace) pollutant content throughout a day and a week: influent and effluent single grab samples are not representative at all of the actual trends (see, *inter alia*, [23–25]). A third bond relates to the need to consider a plant working in a steady state (unless specific conditions have to be monitored, as in case of the operation starting). Finally, significant wastewater changes depending on seasonal activities (due, for instance, to industrial processes and tourism flows) entail ad hoc monitoring campaigns.

Our protocol requires identifying a minimum set of sampling points: the influent, after the mechanical pre-treatments, the final effluent, and the excess sludge. It was applied on conventional activated sludge treatment plants and membrane bioreactor plants.

Samples are always taken over a 2-week period, by means of a refrigerated auto-sampler, equipped with Teflon pipes and dark glass containers. In our case, daily aliquots of each 24-hour composite flow-proportional samples were mixed (either raw or processed, as described in paragraph 3), in order to obtain a cumulative sample, representative for the whole period. Thus, we overcome the constraints inherent to the continuous variability of flow rate and quality of the target streams.

3 Sample Preparation

The first operation consists in separating liquid and solid phases (filtration through glass fiber filters, to retain particles larger than 1.6 µm). Liquid samples were then acidified with H₂SO₄ until pH=4.0 and filtered (at a maximum flow rate of 10 mL/min) through trifunctional silica tC18 cartridges (10 g Sep-Pak Plus C18 Environmental Cartridges, Waters Chromatography), following a slightly modified version of the US EPA 525.2 method [26]. The cartridges were previously activated by adding ethyl acetate, acetone, methanol, and distilled water (40 mL each). Cartridge elution was performed with ethyl acetate, acetone and methanol (40 mL each). Eluates were then treated in a rotating vacuum evaporator and submitted to drying under nitrogen gas at ambient temperature. Residues were dissolved in dimethyl sulfoxide, to obtain approximately a 20,000-fold concentration. Solid phases retained by filters, as well as excess sludge samples, were submitted to soxhlet extraction for 6 h (solution of acetone/n-hexane 1:1), dehydration by anhydrous Na₂SO₄ addition, drying under gentle N₂ stream at ambient temperature and dissolution in sterile DMSO.

4 Choice of Bioassays

As widely underlined in literature, the use of the ecotoxicological tools requires a first clear definition of the goal, i.e., the question to answer. While assessing the environmental impact of an effluent, it is necessary to consider the standardized tests, which are commonly provided for by national and international regulations. In this case, one should carefully consider the results beyond the mere compliance with set values and use them for a more complex overall evaluation, as described in paragraph 6. Afterward, it should be advisable to include both prokaryotes and eukaryotes, as well as unicellular and multicellular, animal and vegetal organisms. Another criterion might lead to include an assay for each trophic level (producer, consumer, and decomposer) together with specific tests aimed at assessing toxicity for humans. Finally, diverse modes of actions should be investigated; in our protocol, we propose the endocrine disruption (selected because of the wide literature about wastewater estrogenic compounds, identified from the 1970s), the genetic toxicity, and the carcinogenic activity (in order to apply the models described in paragraphs 6 and 6.2, which include specifically the neoplastic diseases). The list might be increased by adding the research of specific biomarkers (i.e., of the oxidative stress) and the application of new tools for assessing the contribution of epigenetic regulation to toxicity induction. Table 1 displays the tests

Table 1

List of the bioassays included in the proposed protocols: mode of toxic actions, definition, and relative objects of measurement

Mode of toxic action	Bioassay	Measured by
Baseline toxicity	Algal growth inhibition test	Determination of the growth inhibition of the unicellular green alga <i>Raphidocelis subcapitata</i>
	Bioluminescence inhibition test	Reduction of the natural bioluminescence of marine bacteria <i>Aliivibrio fischeri</i> (formerly <i>Vibrio fischeri</i>)
	Acute toxicity of water flea	Determination of mobility inhibition of the freshwater cladoceran <i>Daphnia magna</i>
	Neutral red uptake assay	Detection of viable cells via the uptake of the dye neutral red
	MTT reduction assay	Formation of formazan salts due to mitochondrial enzymes
Endocrine disruption	ERE-tk_Luc_MCF-7	Luciferase activity quantification in human breast cancer cell line
	YES/YAS	Chromogenic substrate quantification after incubation with recombinant <i>Saccharomyces cerevisiae</i>
Genetic toxicity	Ames test	Point mutations in bacterium <i>Salmonella typhimurium</i>
	<i>Allium cepa</i> test	Chromosomal mutation in root cells
	Comet test	DNA damage in human leukocytes
Carcinogenicity	In vitro cell transformation assay (CTA)	Number of malignant foci or transformed colonies
	Tumor promotion	Gap junction-mediated intercellular communication

included in the proposed protocol, together with the joined modes of action and the phenomena/indicators measured.

Finally, it is worth noting the remarkable significance of “to avoid/reduce behaviors” (also by performing laboratory activities), which may adversely affect the organism welfare and the environment. Therefore, the choice of bioassays should be based also on the 3Rs (Replacement, Reduction, and Refinement) and Green Toxicology principles [27, 28].

5 Bioassays: Materials and Methods

5.1 Baseline Toxicity

This section includes both standardized and non-standardized protocols. In any case, while choosing a technique, its repeatability, reproducibility, sensitivity, and accuracy should be taken into account. Even more important, before planning a toxicological

survey, one should select the tests leading to clear and exploitable answers: the main risk, otherwise, is the collection of useless and unaccountable data.

5.1.1 Algal Growth Inhibition Test

The assay with unicellular green algae was performed according to the ISO standard [29]. Its execution is invaluable, because it involves photosynthetic organisms, and hence can capture possible interferences and impairments to the related mechanisms. Raw samples were submitted to this assay. EC₁₀ and EC₅₀ values were calculated.

5.1.2 Bioluminescence Inhibition Test

The assay with marine luminescent bacteria was performed according to the ISO standard [30]. Its strength resides in the established standardization, the execution rapidity, and the worldwide adoption by governmental and nongovernmental organizations. Raw samples were submitted to this assay. EC₁₀ and EC₅₀ values were calculated.

5.1.3 Acute Toxicity of Water Flea

The assay with Cladocera crustacean followed the ISO standard [31]. As for the luminescent bacteria assay, this is widely prescribed by international and national regulations, standardized and marketed. Raw samples were submitted to this assay. EC₁₀ and EC₅₀ values were calculated.

5.1.4 Neutral Red (NR) and MTT Assays

Basal cytotoxicity is often a prerequisite step for assessing toxic substance activity and mode of action in order to extrapolate their impact on human health. Moreover, determining the inhibitory concentration 50 (IC₅₀), for which 50% of cell population dies, allows working within sublethal doses for assessing other biological targets of substances.

Because substances can exhibit cytotoxicity, in our protocol, cytotoxicity assays were performed on a human breast tumor cell line (MCF-7) and on rat hepatic cell line (IAR203), being both the cell types influenced in vivo by various xenobiotics. The evaluation of cell viability is achieved through the NR and MTT tests, which explore the keeping of a dye in their cytoplasm/lysosomes and their metabolic activity, respectively (Fig. 2). In this case, both the MCF-7 and IAR203 cell lines were seeded at the density of 20,000 cells/cm² for both tests in 96-well plates. The effect of the doses 0, 0.001, 0.002.5, 0.005, 0.01, 0.025, 0.05, 0.1, and 0.25 liters equivalent was evaluated in triplicate, after 24 and 48 h of treatment.

MTT assay is based on the capability of metabolically active cells to reduce actively the water-soluble salt [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] into a insoluble formazan crystals [32]. This redox reaction depends on the mitochondrial respiration, thus reflecting the cellular energy capacity and, at the same time, indicating the cell viability (colorimetric technique). However, one must consider that other internal enzymes can

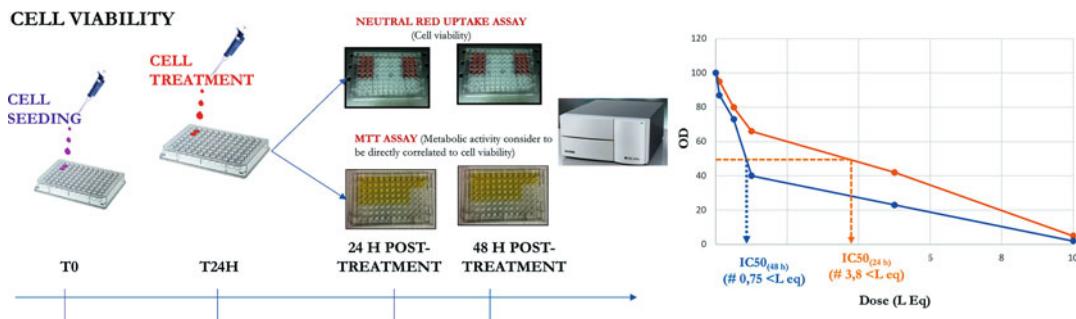


Fig. 2 Detection of cell viability: neutral red (NR) and MTT assays

react and reduce the MTT dye. In our case, culture medium was removed from the wells, replaced by a solution of MTT 0.2 mg/mL in Hanks'/Hepes buffer, and incubated for 2 h at 37 °C (5% CO₂ atmosphere). The MTT solution was then eliminated, and an isopropanol acid solution (0.4 M HCl in isopropanol) was added, to solubilize formazan crystals. Staining intensity was quantified by means of a SUNRISE spectrophotometric plate reader (Tecan, Italy) at the wavelength of 620 nm.

The neutral red uptake assay is so far the most used cytotoxic assay worldwide. It is based on the quantification of the 3-amino-7-dimethyl-2-methylphenazine hydrochloride (Neutral Red Dye) uptake by viable cells. They uptake actively the supravital dye that mainly accumulates in lysosomes, where it remains charged thanks to pH gradients and is consequently unable to exit. The absorbance reading at 540 nm, after release of dye under acidic conditions, reflects the cell membrane stability, hence the viability state [33]. In our case, NR solution was prepared the day before by diluting (1:80) in DMEM medium with 5% FBS and leaving it overnight at 37 °C. After centrifuging twice at 1,250 × g for 10 min at room temperature and cell exposure for 3 h at 37 °C (5% CO₂ atmosphere), fixation was performed by adding a solution of formalcalcium for 1 min; finally, cells were lysed with acetic acid/ethanol. Staining intensity was quantified, after 5 min of mixing, with a microplate reader at the wavelength of 540 nm.

The sample extracts dissolved in sterile DMSO were used.

5.2 Endocrine Disruption

Endocrine disruption can occur at different biological levels and target a wide range of receptors, affecting the regulation cascade processes. Literature has been recently focused on the quantification of hormone-like substances in environmental matrices and their possible effects on organisms and ecosystems. This protocol proposes the study of the estrogenic and androgenic activity, because the international regulations on water quality include some pollutants of whose estrogenicity is proven (see, for instance,

the case of bisphenol A, β -estradiol, and nonylphenol within the revision of the EU Drinking Water Directive [34]).

As for baseline toxicity, this protocol presents both standardized and non-standardized assays.

In all cases, the sample extracts dissolved in sterile DMSO were used.

5.2.1 YES/YAS

The assay with the genetically modified yeast was performed according to the ISO standard [35]: the employed organism was a recombinant strain of *Saccharomyces cerevisiae* stably transfected with the human estrogen receptor (hER α). In our case, we used the commercially available microplate assay XenoScreen YES (Xenometrix®, Switzerland), which follows exactly the ISO procedure. XenoScreen YAS (Xenometrix®, Switzerland) utilizes a recombinant strain of *Saccharomyces cerevisiae* stably transfected with the human androgen receptor (hAR). Both agonistic and antagonistic activities were evaluated.

5.2.2 ERE-tk_Luc_MCF-7

This peculiar cell line is of interest due to its estrogen sensitivity linked with the expression of the receptor α (ER α); besides, cells stably express the reporter gene luciferase under the control of the ERE (Estrogen Responsive Elements) sequences. In such a way, in presence of both the specific ER ligand (17β -estradiol) and estrogen-like chemicals, the classical estrogen signaling pathways are activated, and, after migration and dimerization of ERs, their binding to ERE sequences induces an activation of the luciferase reporter gene. The activity of luciferase is measured by quantifying the bioluminescence by means of a microplate luminometer. More in detail (Fig. 3), MCF-7 cells were seeded at 25,000 cells/cm 2 in 24-well plates with DMEM HG medium containing 10% FBS, 1X mix antibiotics-antimycotics. 24 h after seeding and culture medium removal (washing with phenol red deprived Hanks' buffer), cells were treated either with 17β -estradiol or with wastewater samples. Treatment was performed in DMEM HG culture medium (without phenol red that interferes with the endocrine-

ENDOCRINE DISRUPTION ON HUMAN MAMMARY CELLS

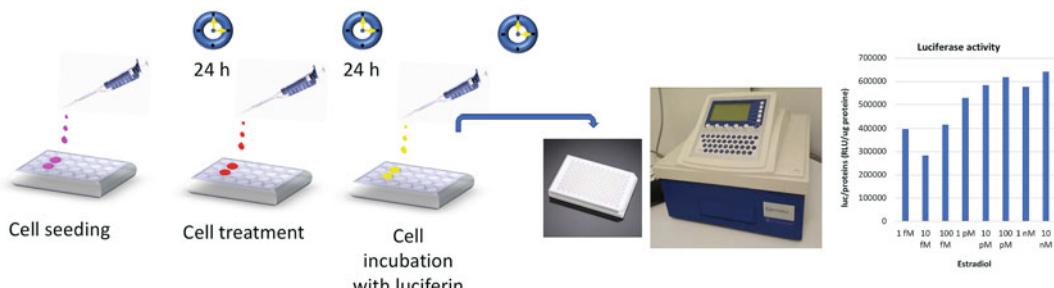


Fig. 3 Endocrine disruption on human mammary cells

disrupting activity quantification) with 10% charcoal stripped fetal bovine serum and 1X mix antibiotics-antimycotics. 17 β -Estradiol was suspended in 100% ethanol, whereas wastewater extracts were suspended in DMSO, as abovementioned. A standard curve was performed with 17 β -estradiol ranging from 1 fM to 10 nM. Before lysing cells, their morphology and survival were observed by microscopic observation. Fifty μ L of PLB 1X (Promega, Italy) was added to cell monolayers and let to act for 30 min in ice. A rubber policeman was used to harvest lysed cells. After a second wash with lysis buffer (PLB 1X), cell lysate was transferred into an Eppendorf tube and centrifuged at 10,000 rpm for 10 min at 4 °C. Supernatants were transferred in a new tube and left in room temperature for at least 30 min. 20 μ L of cell lysate was incubated with 100 μ L of luciferin (Promega, Italy) before measuring the relative luminescence. Media containing either DMSO or ethanol were used as negative controls. Protein content was quantified by the Bradford method (Biorad RC/DC assay, Biorad, Italy). Results were expressed as relative luminescence unit (RLU)/mg of proteins.

5.3 Genetic Toxicity

A battery of short-term mutagenicity tests revealing different genetic endpoints is proposed in order to detect possible phenomena of mutagenicity/genotoxicity induced by the exposure to the samples.

5.3.1 Ames Test

The *Salmonella*/microsome (or bacterial reverse mutation) test (Ames test) is the most widely validated mutagenicity test and is included in the Standard Methods for Examination of Water and Wastewater in 1998 as an official mutagenicity test for the aquatic environment [36]. By means of this test, it is possible to detect point mutations (base substitution and frameshift mutations) in *Salmonella typhimurium* strains [37]. *S. typhimurium* TA100 and TA98 strains were used to test concentrated wastewater in duplicate at increasing doses, corresponding to 0.001, 0.05, 0.1, 0.25, 0.5, 1, and 2 liters equivalent. The Ames test was performed with and without the metabolic activation (\pm S9), adding microsomal enzymes of rat liver to detect direct and indirect mutagens. Plates were incubated at 37 °C in darkness for 72 h; afterward revertant colonies were counted. 2-Nitrofluorene (10 μ g/plate) and sodium azide (10 μ g/plate) were used as positive controls for TA98 without S9 and TA100 without S9, respectively, and 2-aminofluorene (20 μ g/plate) for both strains with S9. DMSO and distilled water previously filtered through tC₁₈ silica cartridges were tested as negative controls. The results were considered positive in case two consecutive doses or the highest nontoxic dose caused a response at least twice that of the solvent control and at least two of these consecutive doses showed a dose-response relationship [36]. The results were expressed as mutagenicity ratio (MR), dividing the revertants/plate by the spontaneous mutation rate. In case the

response was positive, the results were also expressed as specific mutagenic activity (net revertants/liter) calculated by linear regression analysis of the dose-response curve. The sample extracts dissolved in sterile DMSO were used.

5.3.2 Single Cell Gel Electrophoresis Assay (SCGE)/Comet Test

The single cell gel electrophoresis (SCGE) assay or comet test is a simple method for measuring primary DNA damage in eukaryotic cells, as single-strand breaks (SSB) and double-strand breaks (DSB), excision repair sites, cross-links, and alkali-labile sites (ALS). By applying an electrical field, DNA fragments migrate toward the anode at a speed depending on its size and cause a comet image (Fig. 4). The assay was performed in alkaline conditions ($\text{pH} > 13$) using leukocytes from peripheral blood of a non-smoker donor [38]. The organic extracts of wastewater samples dissolved in DMSO were kept in contact with the leukocytes (1 h at 37°C with 5% CO_2) at increasing doses (0.001, 0.05, 0.1, 0.25, and 0.5 liters equivalent). Negative (DMSO) and positive controls (2 mM of ethylmethane sulfonate, EMS) were submitted to analyses. After incubation, about 5×10^5 cells were suspended in

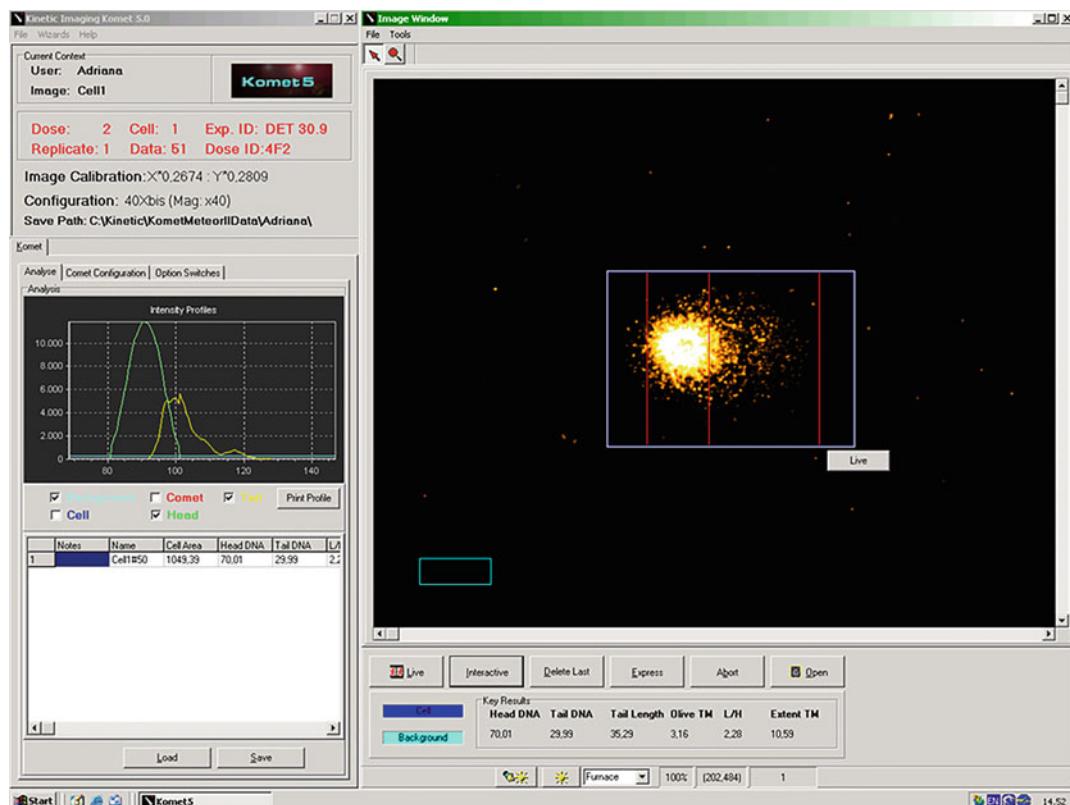


Fig. 4 Single cell gel electrophoresis assay (Comet test) damaged DNA generates the “tail” (Komet 5, Kinetic Imaging Ltd)

90 µL of 0.7% low melting agarose (LMA) and spread onto microscope slides pre-coated with 1% normal melting agarose (NMA). The cells were lysed overnight at 4 °C in a lysis solution (pH 10) containing 8 mM Tris–HCl, 2.5 M NaCl, 100 mM Na₂EDTA, 1% triton X-100, and 10% DMSO. After that, the slides were placed for 40 min in a horizontal gel electrophoresis tank filled with cold electrophoretic buffer (1 mM Na₂EDTA and 300 mM NaOH, pH > 13) to allow DNA unwinding. Electrophoresis was performed in the same buffer for 40 min at 25 V (1 V/cm) and 300 mA. After electrophoresis, the slides were neutralized with 0.4 M Tris–HCl (pH 7.5), stained with ethidium bromide (10 µg/mL) and analyzed using a fluorescence microscope (Olympus CX 41RF) equipped with a BP 515–560 nm excitation filter and an LP 580 nm barrier filter. The experiment was repeated twice. Fifty randomly selected cells per slide (two slides per sample) were analyzed. The extent of DNA migration was evaluated by means of a “visual score,” based on visual classification of DNA damage, and the comet parameter “tail intensity” (percentage of DNA migrated in the tail) was used as the measure of DNA damage, measured by an automatic imaging system (Komet 5, Kinetic Imaging Ltd). The results of Comet test were expressed as the mean ± standard deviation, and statistical significance was evaluated with one-way ANOVA followed by Dunnett’s multiple comparison test (comparing each exposure concentration to the negative control). The sample extracts dissolved in sterile DMSO were used.

5.3.3 Allium cepa Test

Two genotoxicity tests were carried out on raw samples using *Allium cepa* to detect chromosome aberrations (namely, bridges, buds, rings, polar slips, sticky, laggard, polyploidized and condensed nuclei, fragments, c-mitosis, and multipolar anaphases and metaphases, binucleated cells) and micronuclei [39, 40].

In a preliminary toxicity assay, 12 equal-sized young bulbs of onion ($\Phi \approx 2.5$ cm) were exposed for 76 h in darkness to undiluted and diluted water (1:2, 1:10, 1:20, 1:100, 1:200, and 1:100 dilution), replacing the sample solution every day. Root length was used to calculate the EC₅₀ value and to identify the concentration for the execution of *Allium cepa* genotoxicity assays, being the highest correspondent to the EC₅₀ value identified (the concentration causing a 50% reduction in root growth). Other macroscopic parameters (turgescence, consistency, color, and root tip shape change) were used as toxicity indexes [41], as displayed in Table 2, where plant roots growth inhibition test is included in the evaluation of freshwater ecotoxicity.

Chromosome aberrations (CA) and micronuclei (MN) tests were performed using six equal-sized young bulbs per sample;

after 72 h of pre-germination in saline solution (Rank solution), the bulbs were exposed to samples for 24 h [42]. Afterward, the roots were fixed in acetic acid and ethanol (1:3) for 24 h and lastly stored in 70% ethanol for the chromosome aberrations (CA) test [40]. In the micronuclei (MN) test, the bulbs, after exposure, were dipped in Rank solution for 44 h of recovery time (to cover two rounds of mitosis), fixed in acetic acid and ethanol (1:3) for 24 h, and lastly stored in 70% ethanol [39]. Rank saline solution (24 h exposure) and maleic hydrazide (10^{-2} M, 6-h exposure) were used as negative and positive controls, respectively. Five roots for each sample were considered for microscopic analysis, after Feulgen staining [39, 40]: 1000 cells/slide (5000 cells/sample) were scored for mitotic index (as a measure of cell division, hence sample toxicity), 200 in mitosis cells/slide (1000 cells/sample) for chromosomal aberrations, and 2000 in interphase cells/slide (10,000 cells/sample) for micronuclei frequency. Chi square test was performed for mitotic index and chromosomal aberration data analyses; the analysis of variance and Dunnett's multiple comparison test were performed to analyze the micronuclei frequency.

5.4 Carcinogenic Activity

Chemically induced carcinogenesis is assumed the result of multi-step events that can be recapitulated in initiation, promotion, transformation, and progression phases. Initiation is represented by multiple mutations and DNA damage events, commonly involving genes controlling key growth pathways; the promotion stage results in various epigenetic changes that enhance the growth of initiated cells. This cellular amplification progresses along the carcinogenesis process evolving in the progression step, in which the genetic instability results in the further and irreversible loss of growth control, gain in invasiveness, and metastatic properties. The complexity of carcinogenesis process is still discussed for trying to explain the spatiotemporal succession of deleterious events and the more specific classification of chemicals, their mode, and mechanism of action [43–46]. Plenty of in vitro tests have been proposed to assess mutagenic and carcinogenic properties of chemicals. Although the extrapolation to human risk assessment is difficult, they can provide major indications for further global evaluations of chemical/mixture toxicity.

5.4.1 Tumor Promotion

A test aimed at identifying chemicals/mixtures acting either by non-genotoxic carcinogenic induction (without DNA mutation) or by processes known to be important in carcinogenesis (down-/up-regulation of cell homeostasis: increased inflammation or cell proliferation, inhibition of cell differentiation, and death) is the in vitro study of the inhibition of gap junction intercellular communication (GJIC) [47, 48]. The predictivity of this test is about 70% [48]. To evaluate the intercellular communication mediated by gap junctions (GJIC), the scrape loading technique was used in

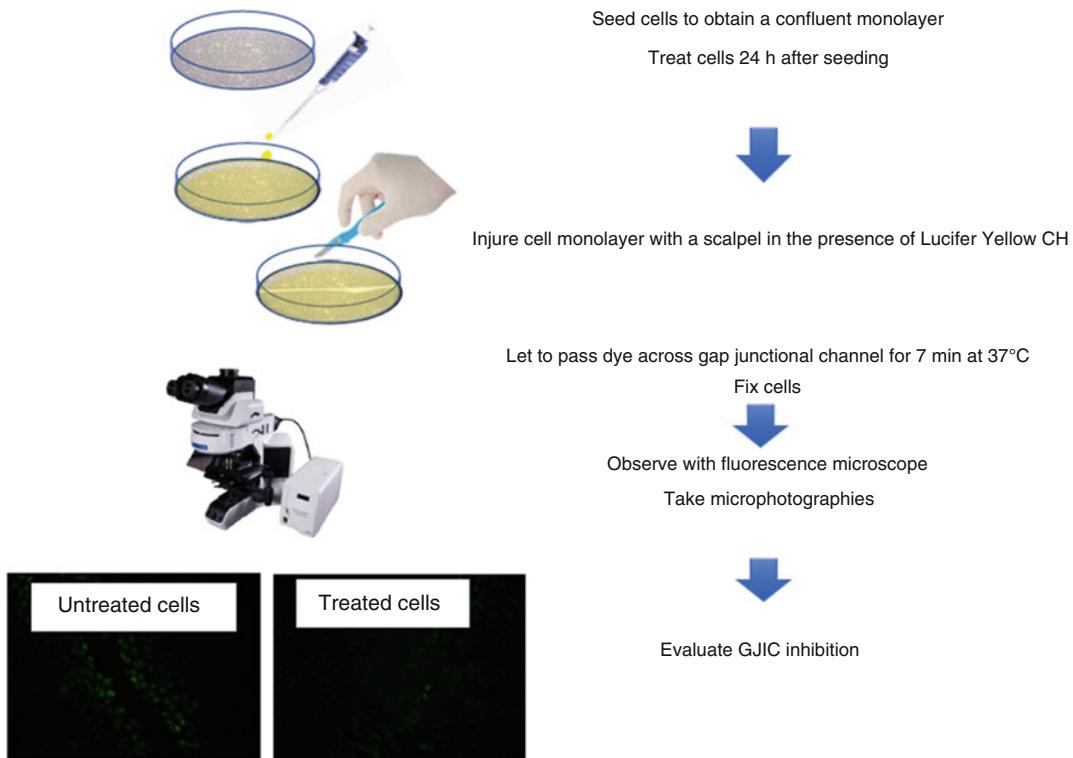


Fig. 5 Tumor promotion potential: gap junctional intercellular communication (GJIC) derived from [49]

subconfluent/confluent cells grown in monolayer [49] (Fig. 5). However, because samples are often scarce, we tried to miniaturize the method. Therefore, cells were into 24-wells on 12 mm round coverslips. Cell seeding density needs to be optimized for each cell type. Once a homogeneous cell monolayer was obtained, it was injured with a scalpel in the presence of the low molecular weight fluorescent dye, lucifer yellow CH (457.2 Dalton), which was incorporated by damaged cells all along the cut. The fluorescent tracer trapped inside the cytoplasm can spread to other adjacent cells only if they communicate via gap junction channels. In order to verify that GJIC are effectively responsible for the dye transfer, another dye (rhodamine dextran) was used, in concomitance, as the negative control; since its molecular weight is about 10,000 Dalton, thus it cannot pass throughout GJ channels (only molecules with a molecular weight lower than 1000 Dalton can pass through). As an example, we worked with a liver-derived cell line (the IAR203 cells). Cells were seeded at a density of $100,000/\text{cm}^2$ onto 24-well plates. After 24 h, the cells were treated for 6 h and then washed twice with PBS containing calcium and magnesium ions; finally, as abovementioned, cell monolayer was cut with in a lucifer yellow solution (0.5% in PBS). Cells were incubated for 7 min at 37 °C (5% CO₂ atmosphere; 97% relative humidity).

After washing twice with PBS and 5 min of fixation with formaldehyde at 4%, cells were washed again with PBS. The coverslips were mounted and rapidly observed under fluorescence microscopy. Images were acquired combining the microscope with the cellSens imaging software, to quantify the effect of substances/mixtures on the intercellular communication mediated by gap junctions. The sample extracts dissolved in sterile DMSO were used.

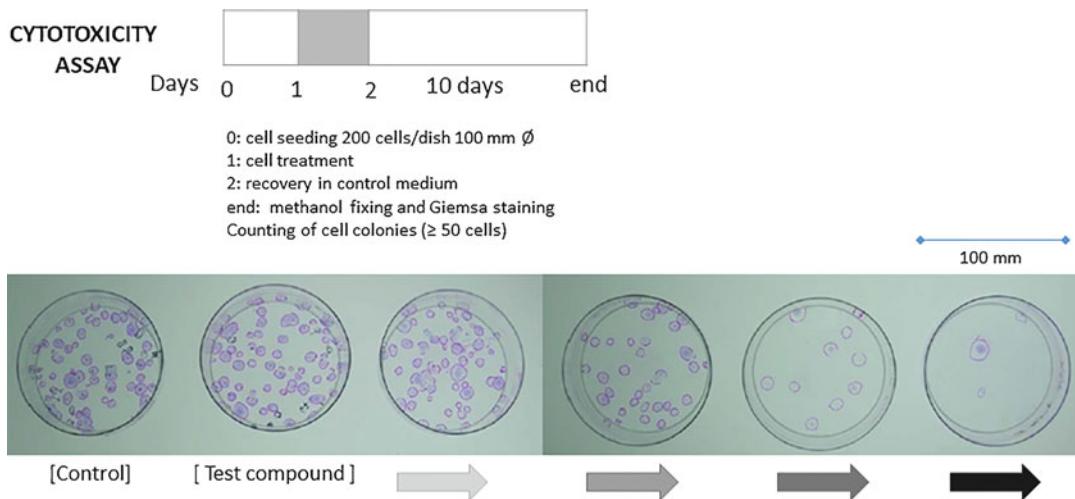
5.4.2 Cell Transformation Assays (CTA)

The cell transformation assays are in vitro tests mimicking some key stages of the in vivo carcinogenesis process and represent the most advanced in vitro tests for the prediction of human carcinogenicity induced by chemicals/mixtures [50]. The process of chemical-induced cell transformation in suitable cell models leads to phenotypic features, typical of tumorigenic cells. The in vitro transformed cells acquire the ability to induce tumors in susceptible animals, as a demonstration of their malignant potential [51].

Chemical carcinogens can be classified as either genotoxic or non-genotoxic, based on their mode of action. Genotoxic compounds (or their metabolites) are able to initiate the cells to carcinogenesis through a direct interaction with DNA, leading to structural and/or numerical chromosomal damages. Recommendations suggest the use of test batteries for genotoxicity evaluation before analyzing the carcinogenic potential. On the other hand, non-genotoxic carcinogens act via indirect or epigenetic mechanisms, at least initially, causing modifications to DNA structures and alterations of gene expression and signal transduction.

Thus, non-genotoxic carcinogens are compounds which exhibit negative results in genotoxicity tests but have the potential to induce cell transformation by means of a non-genotoxic mechanism. This suggests the importance of an integrated approach to the evaluation of the biological activity of environmental compounds/matrices.

Among the cell lines suggested by the Detailed Review Paper on Cell Transformation Assays [51], we selected the C3H10T1/2 clone 8 (C3H from here on) mouse embryonic fibroblasts (ATCC, CCL 226 lot. n. 58078542). These cells have a high sensitivity to carcinogenic compounds and a low spontaneous transformation rate. The expression of the neoplastic phenotype is visualized by means of transformed *foci* of high cell density and typical morphology. The *foci* of transformed cells are recognized under a microscope and classified by standard morphological features: multilayered growth, deep basophilic staining, random cell orientation at the edge of the *focus*, and invasiveness on the surrounding monolayer of normal cells. The morphological features of transformed cells of *foci* reflect the metabolic changes, genetic instability, altered growth control, and acquisition of immortalization, typical of the carcinogenesis process. Three types of *foci* are



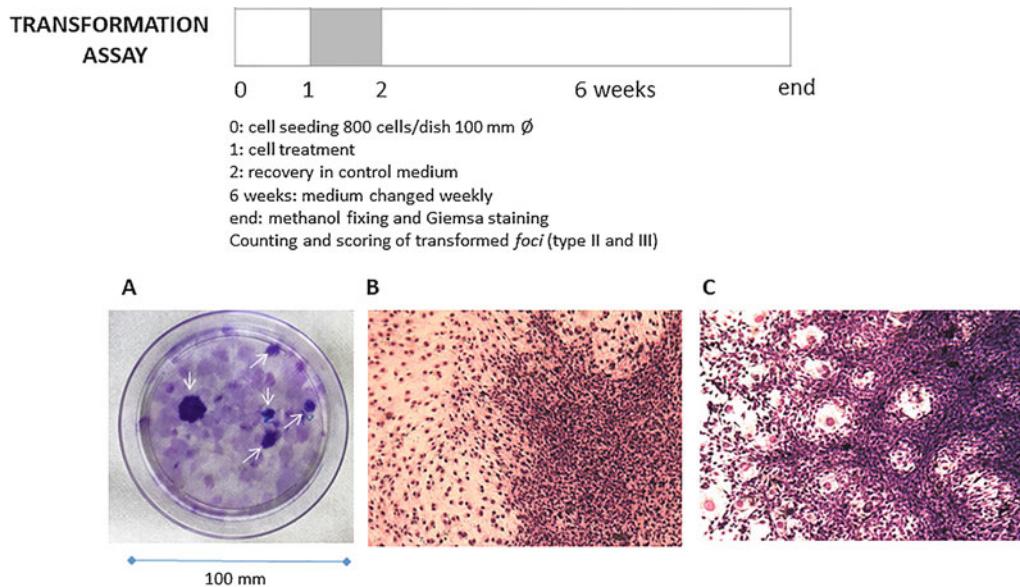
An example of results is shown: colonies of fixed and Giemsa stained cells are highlighted on the bottom of the dish. Darker arrows indicate increasing concentrations of test compound corresponding to a decrease in colonies number.

Fig. 6 Cytotoxicity assay

described: type I are not considered to be fully transformed as they do not produce tumors when injected into susceptible animal hosts, while type II and III are fully transformed and have undergone the malignant transformation.

The CTA is divided into two phases, according to the experimental protocol shown in Figs. 6 and 7. A preliminary cytotoxicity assay was performed with the aim of identifying a dose-response and of selecting non-cytotoxic concentrations of the chemical/matrix analyzed (Fig. 6). On day 0, cells were seeded (200 cells/dish, 5 dishes/sample) onto 100 mm diameter dishes. 24 h after seeding, complete control medium (Basal Medium Eagle containing 10% heat-inactivated fetal bovine serum, 1% glutamine, 0.5% HEPES 2 M, and 25 µg/mL gentamicin) was replaced with fresh medium (controls), with medium containing different dilutions of wastewaters or with medium containing all positive controls (e.g., 4 µg/m 3-methylcholanthrene). After 24 h of treatment, all the media were replaced with fresh control medium, and cells were allowed to grow. Seven to ten days after seeding, all the samples were fixed with methanol and stained with 10% Giemsa in distilled water. After removing the excess of staining solution by washing with distilled water, the samples were air-dried and ready for counting of the colonies under a stereomicroscope. Colonies are formed by at least 50 cells.

Next to the selection of non-cytotoxic concentrations of test chemical/matrix, the transformation assay was performed (Fig. 7). C3H cells were seeded on day 0 at a density of 800 cells/dish (100 mm diameter, 10 dishes/sample) and exposed 24 h after



A representative dish at the end of the assay in which *foci* (arrows) developed is shown (A); details of type II (B) and type III (C) *foci* are visualized under light microscopy.

Fig. 7 Cell transformation assay

seeding to test compounds for additional 24 h. After treatment, the cultures were fed weekly with complete medium until confluence was reached (around 2 weeks); then the serum concentration was reduced to 5% until the end of the assay. The transformation assay lasted 6 weeks, during which the cells exposed to the test compound(s)/matrix eventually underwent in vitro transformation, visualized by colonies of transformed cells, the *foci* [52, 53]. After 6 weeks, the cells were fixed (methanol) and stained (10% Giemsa in water), rinsed with water, and observed under a microscope for *foci* scoring and classification. *Foci* of type II and III (fully transformed) were scored and counted, and the number of transformed *foci* in treated samples was compared to those of negative and positive controls.

6 Elaboration Criteria of Experimental Results

Beyond their inherent meaning (such as the values of EC₅₀, IC₅₀, etc.), the results of bioassays can be employed as input data for the evaluation of the effects of a work, namely, a wastewater treatment plant, on the human health and the ecosystem. Our protocol proposes two ways: the first one, consisting in a standardized life cycle-based assessment, and the second one, an experimental approach, which correlates the emissions into different

environmental matrices with the probability of occurrence of human diseases. In the following, a brief description of the principles of both approaches is reported: the reader may refer to the cited literature for details.

6.1 Determination of Equivalent Concentration for LCA

The Organisation Environmental Footprint (OEF) and Product Environmental Footprint (PEF) protocols, described in the Recommendation 2013/179/EU [54], allow to quantify the possible impact on 15 environmental footprint (EF) impact categories generated in all life cycle stages of organizations or products.

The results of bioassays cannot be directly used in the environmental footprint assessment protocols based on LCA principles. Nevertheless, an innovative procedure, for integrating the results of bioassays in the OEF/PEF protocols, was recently proposed [55, 56] and applied to evaluate the impacts on freshwater and human toxicity (both cancer and non-cancer effects) of WWTP effluents.

This procedure is based on the conversion of the results of biological assays into equivalent concentrations of reference substances giving the same effects as those measured experimentally. In practice, different concentrations of proper reference chemicals are submitted, as described in [57], to the related bioassays for obtaining dose-response curves. The equivalent concentration is identified as the one exerting the same toxicity as that measured on the tested samples. Concentrations of reference substances, then, can be used as input data of the LCA-based models for estimating the contribution to different impact categories. Indeed, referring to the target organism and the specific endpoint, bioassays can be associated to impact categories, as shown in Table 2. When, for a given impact category, more bioassays are available, according to a conservative approach, the highest effect can be selected as the one representative of the impact of the analyzed sample.

Reference chemicals must be both sensitive for the specific endpoints and included in the list provided by the International Reference Life Cycle Data System (ILCD) for the related impact category. The evaluation of the potential toxicity in the OEF/PEF procedures is based on the characterization factors derived from the USEtox model [58]. The USEtox characterization factors for human toxicity (both carcinogenic and non-carcinogenic impacts) and for freshwater ecosystem toxicity are classified as “indicative” (or “interim”) when a high degree of uncertainty is still associated in the fate, exposure, or effects of the substance in the different environmental compartments, while they are labeled as “recommended” when the model is considered fully appropriate. For this reason, in the works cited above, the reference substances were chosen among the list of chemicals, whose characterization factors, for the selected impact categories, are recorded as

Table 2

Bioassays used in the environmental footprint assessment by means of OEF/PEF protocols and BAD approach

Bioassay	OEF/PEF impact categories	BAD approach
Algal growth inhibition test	FET	—
Acute toxicity of water flea	FET	—
Bioluminescence inhibition test	FET	—
Plant root growth inhibition test (<i>A. cepa</i>)	FET	—
Neutral red uptake assay	HT-NC	×
MTT reduction assay	HT-NC	×
ERE-tk_Luc_MCF-7	HT-NC	×
YES/YAS	HT-NC	×
Ames test	HT-C (?)	×
<i>Allium cepa</i> test	FET	—
Comet test	HT-C	×
In vitro cell transformation assay	HT-C	×
Tumor promotion	HT-C	×

“recommended,” thus minimizing the intrinsic uncertainty related to the USEtox model.

6.2 Direct Use of Results of Bioassays Within the BAD Approach

The BAD approach, extensively explained in [3, 4, 59], is a procedure that allows quantifying the impacts and the benefits on human health related to a WWTP, by considering the emissions into both water and air. A direct approach is used for the evaluation of the damage on human health due to air emissions; on the contrary, an indirect approach is proposed for water emissions. The conversion of both damages into an economic impact enables their comparison.

Regarding the air, both the so-called indirect (due to energy production) and direct emissions (N_2O from nitrogen removal processes and CH_4 leakage from biogas storage) are converted into an external cost, using values (expressed in €/kg_{emitted pollutant}) available in the literature [60, 61].

As for water emissions, first, bioassays able to highlight potential effects on human health must be selected, according to the list in Table 2. From the results of each bioassay, performed on influent and effluent samples, a percent reduction (if any) of the effect exerted toward a specific endpoint is calculated. The next step is the calculation of the corresponding reduction of the burden of diseases, associated to the performed bioassays, and the consequent

reduction of DALY (Disability Adjusted Life Year), which can be finally converted into an economic value through the gross domestic product of a country. For the application of this procedure, the World Health Organization database can be referred to ref. 62.

7 The Final Step: Facing the Decision Processes

Assessing the impact of effluents on ecosystem through a well-defined protocol is crucial to ensure that this important aspect plays an appropriate role in the context of decision processes concerning the design of new WWTPs or the update of existing ones. These processes typically involve a multiplicity of players (ranging from technical experts to political bodies and the general public) and, due to their complexity, belong to the field of Multi-Attribute Decision Making (MADM). In MADM problems the evaluation of each alternative involves multiple attributes, also called criteria, which may have different importance and may be not directly comparable due to their different nature (quantitative or qualitative) and non-reconcilable units of measurement. In particular, in addition to environmental aspects, the following main classes of criteria play a role in the assessment of a technological solution: technical, economical, administrative/normative, and sociocultural.

To properly manage this complexity and avoid the risk of oversimplifications and/or unbalanced evaluations, the adoption of suitable decision support systems providing a formal support to the assessment and overall aggregation of the abovementioned criteria is appropriate [63–65]. This in turn requires that the assessment of each criterion is carried out in a clear, accountable, and reproducible manner. Criteria failing to meet these requirements might risk playing a diminished role in the decision process, even in spite of their intrinsic importance, in favor of criteria whose assessment is traditionally regarded as more consolidated and reliable, e.g., economical evaluations. By defining a protocol for assessing the environmental footprint by means of ecotoxicological tools, we aim also at providing a contribution to reduce this risk and to ensure a proper consideration of these important assessments in the context of the overall decision processes.

Acknowledgments

We thank dr. Elisabetta Ceretti for her technical support in laboratory activity.

References

1. Krzeminski P, Tomei MC, Karaolia P, Langenhoff A, Almeida CMR, Felis E, Gritten F, Andersen HR, Fernandes T, Manaia CM, Rizzo L, Fatta-Kassinos D (2019) Performance of secondary wastewater treatment methods for the removal of contaminants of emerging concern implicated in crop uptake and antibiotic resistance spread: a review. *Sci Total Environ* 648:1052–1081. <https://doi.org/10.1016/j.scitotenv.2018.08.130>
2. Teodosiu C, Gilca AF, Barjoveanu G, Fiore S (2018) Emerging pollutants removal through advanced drinking water treatment: a review on processes and environmental performances assessment. *J Clean Prod Elsevier* 197:1210. <https://doi.org/10.1016/j.jclepro.2018.06.247>
3. Papa M, Alfonsín C, Moreira MT, Bertanza G (2016) Ranking wastewater treatment trains based on their impacts and benefits on human health: a “biological assay and disease” approach. *J Clean Prod* 113:311–317. <https://doi.org/10.1016/j.jclepro.2015.11.021>
4. Papa M, Pedrazzani R, Bertanza G (2013) How green are environmental technologies? A new approach for a global evaluation: the case of WWTP effluents ozonation. *Water Res* 47:3679–3687. <https://doi.org/10.1016/j.watres.2013.04.015>
5. Chapman PM (2000) Whole effluent toxicity TESTING—usefulness, level of protection, and risk assessment. *Environ Toxicol Chem* 19:3. [https://doi.org/10.1897/1551-5028\(2000\)019<0003:WETTUL>2.3.CO;2](https://doi.org/10.1897/1551-5028(2000)019<0003:WETTUL>2.3.CO;2)
6. Ra JS, Kim HK, Chang NI, Kim SD (2007) Whole effluent toxicity (WET) tests on wastewater treatment plants with *Daphnia magna* and *Selenastrum capricornutum*. *Environ Monit Assess* 129:107–113. <https://doi.org/10.1007/s10661-006-9431-2>
7. Hassan SHA, Van Ginkel SW, Hussein MAM, Abskharon R, Oh S-E (2016) Toxicity assessment using different bioassays and microbial biosensors. *Environ Int* 92–93:106–118. <https://doi.org/10.1016/j.envint.2016.03.003>
8. Gruiz K, Fekete-Kertész I, Kunghné-Nagy Z, Hajdu C, Feigl V, Vaszita E, Molnár M (2016) Direct toxicity assessment — methods, evaluation, interpretation. *Sci Total Environ* 563–564:803–812. <https://doi.org/10.1016/j.scitotenv.2016.01.007>
9. Norberg-King TJ, Embry MR, Belanger SE, Braunbeck T, Butler JD, Dorn PB, Farr B, Guiney PD, Hughes SA, Jeffries M, Journeel R, Léonard M, McMaster M, Oris JT, Ryder K, Segner H, Senac T, Van Der Kraak G, Whale G, Wilson P (2018) An international perspective on the tools and concepts for effluent toxicity assessments in the context of animal alternatives: reduction in vertebrate use. *Environ Toxicol Chem* 37:2745–2757. <https://doi.org/10.1002/etc.4259>
10. Gargosova HZ, Urminska B (2017) Assessment of the efficiency of wastewater treatment plant using ecotoxicity tests, vol 26, pp 56–62
11. Tonkes M, De Graaf PJF, Graansma J (1999) Assessment of complex industrial effluents in the Netherlands using a whole effluent toxicity (or wet) approach. *Water Sci Technol* 39:55–61. [https://doi.org/10.1016/S0273-1223\(99\)00253-X](https://doi.org/10.1016/S0273-1223(99)00253-X)
12. Välijalo P, Perkola N, Seiler TB, Sillanpää M, Kuckelkorn J, Mikola A, Hollert H, Schultz E (2016) Estrogenic activity in Finnish municipal wastewater effluents. *Water Res* 88:740–749. <https://doi.org/10.1016/j.watres.2015.10.056>
13. Escher BI, Bramaz N, Quayle P, Rutishauser S, Vermeirssen EL (2008) Monitoring of the eco-toxicological hazard potential by polar organic micropollutants in sewage treatment plants and surface waters using a mode-of-action based test battery. *J Environ Monit* 10:622–631. <https://doi.org/10.1039/b800951a>
14. Avberšek M, Žegura B, Filipič M, Heath E (2011) Integration of GC-MSD and ER-Calux® assay into a single protocol for determining steroid estrogens in environmental samples. *Sci Total Environ* 409:5069–5075. <https://doi.org/10.1016/j.scitotenv.2011.08.020>
15. Arlos MJ, Parker WJ, Bicudo JR, Law P, Marjan P, Andrews SA, Servos MR (2018) Multi-year prediction of estrogenicity in municipal wastewater effluents. *Sci Total Environ* 610–611:1103–1112. <https://doi.org/10.1016/j.scitotenv.2017.08.171>
16. Caldwell DJ, Mastrocco F, Anderson PD, Länge R, Sumpter JP (2012) Predicted-no-effect concentrations for the steroid estrogens estrone, 17 β -estradiol, estriol, and 17 α -ethynodiol. *Environ Toxicol Chem* 31:1396–1406. <https://doi.org/10.1002/etc.1825>
17. Escher BI, Aït-Aïssa S, Behnisch PA, Brack W, Brion F, Brouwer A, Buchinger S, Crawford SE, Du Pasquier D, Hamers T, Hettwer K, Hilscherová K, Hollert H, Kase R, Kienle C, Tindall AJ, Tuerk J, van der Oost R, Vermeirssen E, Neale PA (2018) Effect-based trigger values for in vitro and in vivo bioassays performed on surface water extracts supporting

- the environmental quality standards (EQS) of the European Water Framework Directive. *Sci Total Environ* 628–629:748–765. <https://doi.org/10.1016/j.scitotenv.2018.01.340>
18. Leusch FDL, Chapman HF, Korner W, Gooneratne SR, Tremblay LA (2005) Efficacy of an advanced sewage treatment plant in southeast Queensland, Australia, to remove estrogenic chemicals. *Environ Sci Technol* 39:5781–5786. <https://doi.org/10.1021/es0484303>
 19. Jarosová B, Bláha L, Giesy JP, Hilscherová K (2014) What level of estrogenic activity determined by in vitro assays in municipal waste waters can be considered as safe? *Environ Int* 64:98–109. <https://doi.org/10.1016/j.envint.2013.12.009>
 20. Pedrazzani R, Bertanza G, Brnardić I, Cetecioglu Z, Dries J, Dvarionienė J, García-Fernández AJ, Langenhoff A, Libralato G, Lofrano G, Skrbic B, Martínez-López E, Meriç S, Pavlović DM, Papa M, Schröder P, Tsagarakis KP, Vogelsang C (2019) Opinion paper about organic trace pollutants in wastewater: toxicity assessment in a European perspective. *Sci Total Environ* 651:3202–3221. <https://doi.org/10.1016/J.SCITOTENV.2018.10.027>
 21. Escher BI, Allinson M, Altenburger R, Bain PA, Balaguer P, Busch W, Crago J, Denslow ND, Dopp E, Hilscherova K, Humpage AR, Kumar A, Grimaldi M, Jayasinghe BS, Jarosova B, Jia A, Makarov S, Maruya KA, Medvedev A, Mehinto AC, Mendez JE, Poulsen A, Prochazka E, Richard J, Schifferli A, Schlenk D, Scholz S, Shiraishi F, Snyder S, Su G, Tang JYM, van der BB, van der LSC, Werner I, Westerheide SD, Wong CKC, Yang M, Yeung BHY, Zhang X, Leusch FDL (2014) Benchmarking organic micropollutants in wastewater, recycled water and drinking water with in vitro bioassays. *Environ Sci Technol* 48:1940–1956. <https://doi.org/10.1021/es403899t>
 22. Bertanza G, Pedrazzani R, Dal Grande M, Papa M, Zambarda V, Montani C, Steimberg N, Mazzoleni G, Di Lorenzo D (2011) Effect of biological and chemical oxidation on the removal of estrogenic compounds (NP and BPA) from wastewater: an integrated assessment procedure. *Water Res* 45:2473–2484. <https://doi.org/10.1016/j.watres.2011.01.026>
 23. Coes AL, Paretti NV, Foreman WT, Iverson JL, Alvarez DA (2014) Sampling trace organic compounds in water: a comparison of a continuous active sampler to continuous passive and discrete sampling methods. *Sci Total Environ* 473–474:731. <https://doi.org/10.1016/j.scitotenv.2013.12.082>
 24. Aymerich I, Acuña V, Ort C, Rodríguez-Roda I, Corominas L (2017) Fate of organic microcontaminants in wastewater treatment and river systems: an uncertainty assessment in view of sampling strategy, and compound consumption rate and degradability. *Water Res* 125:152. <https://doi.org/10.1016/j.watres.2017.08.011>
 25. Petrie B, Barden R, Kasprzyk-Hordern B (2014) A review on emerging contaminants in wastewaters and the environment: current knowledge, understudied areas and recommendations for future monitoring. *Water Res* 72:3. <https://doi.org/10.1016/j.watres.2014.08.053>
 26. Budde WL, JW Eichelberger TD, Behymer WL (1988). Method 525.2 determination of organic compounds in drinking water by liquid-solid extraction and capillary column gas chromatography/mass spectrometry revision 2.0Budde-Method 525.1 Revision
 27. Sambuy Y, Alloisio S, Bertanza G, Feretti D, Letasiova S, Mazzoleni G, Pedrazzani R, Caloni F (2018) Air, water and soil: which alternatives? Alternative models in environmental toxicology. *Altex* 35:254. <https://doi.org/10.14573/altex.1802121>
 28. Maertens A, Hartung T (2018) Green toxicology—know early about and avoid toxic product liabilities. *Toxicol Sci* 161:285. <https://doi.org/10.1093/toxsci/kfx243>
 29. ISO. (2012). ISO 8692:2012(en), Water quality — fresh water algal growth inhibition test with unicellular green algae
 30. International Organization for Standardization (2007) Water quality – Determination of the inhibitory effect of water samples on the light emission of *Vibrio fischeri* (Luminescent bacteria test) – Part 3: Method using freeze-dried bacteria. 11348–3. Geneva (CH)
 31. International Organization for Standardization (2012) Water quality — Determination of the inhibition of the mobility of *Daphnia magna* Straus (Cladocera, Crustacea) — acute toxicity test. 6341. Geneva (CH)
 32. Mosmann T (1983) Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J Immunol Methods* 65:55–63
 33. Repetto G, del Peso A, Zurita JL (2008) Neutral red uptake assay for the estimation of cell viability/cytotoxicity. *Nat Protoc* 3:1125–1131. <https://doi.org/10.1038/nprot.2008.75>

34. Laaninen T (2019) Revision of the drinking water directive
35. International Organization for Standardization (2018) Water quality – Determination of the estrogenic potential of water and waste water – Part 1: Yeast estrogen screen (*Saccharomyces cerevisiae*). 19040-1. Geneva (CH)
36. APHA, AWWA, WEF (2017) Standard methods for the examination of water and wastewater. E.W. Rice, R.B. Baird, A.D. Eaton, editors 23rd edn, Publisher: American Public Health Association, American Water Works Association, Water Environment Federation. Washington D.C. ISBN: 9780875532875
37. Maron DM, Ames BN (1983) Revised methods for the Salmonella mutagenicity test. *Mutat Res Mutagen Relat Subj* doi 113:173. [https://doi.org/10.1016/0165-1161\(83\)90010-9](https://doi.org/10.1016/0165-1161(83)90010-9)
38. Tice R R, Agurell E, Anderson D, Burlinson B, Hartmann A, Kobayashi H, Miyamae Y, Rojas E, Ryu J-C, Sasaki Y F (2000). Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing. *Environ Mol Mutagen* 35(3):206:221. [https://doi.org/10.1002/\(SICI\)1098-2280\(2000\)35:3<206::AID-EM8>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1098-2280(2000)35:3<206::AID-EM8>3.0.CO;2-J)
39. Ma TH, Xu Z, Xu C, McConnell H, Valtierra Rabago E, Adriana Arreola G, Zhang H (1995) The improved Allium/Vicia root tip micronucleus assay for clastogenicity of environmental pollutants. *Mutat Res Mutagen Relat Subj* 334:185–195. [https://doi.org/10.1016/0165-1161\(95\)90010-1](https://doi.org/10.1016/0165-1161(95)90010-1)
40. Cabaravdic M (2010) Induction of chromosome aberrations in the Allium cepa test system caused by the exposure of cells to benzo (a) pyrene. *Med Arh* 64:215–218
41. Fiskesjö G (1995) Allium test. In: *In vitro toxicity testing protocols*. Humana Press, Totowa, pp 119–127. <https://doi.org/10.1385/0-89603-282-5:119>
42. Rank J, Lopez L C, Nielsen M H, Moretton J (2002). Genotoxicity of maleic hydrazide, acridine and DEHP in Allium cepa root cells performed by two different laboratories. *Hereditas* 136(1):13–18. <https://doi.org/10.1034/j.1601-5223.2002.1360103.x>
43. Cohen SM, Boobis AR, Dellarco VL, Doe JE, Fenner-Crisp PA, Moretto A, Pastoor TP, Schoeny RS, Seed JG, Wolf DC (2019) Chemical carcinogenicity revisited 3: risk assessment of carcinogenic potential based on the current state of knowledge of carcinogenesis in humans. *Regul Toxicol Pharmacol* 103:100–105. <https://doi.org/10.1016/j.yrtph.2019.01.017>
44. Doe JE, Boobis AR, Dellarco V, Fenner-Crisp PA, Moretto A, Pastoor TP, Schoeny RS, Seed JG, Wolf DC (2019) Chemical carcinogenicity revisited 2: current knowledge of carcinogenesis shows that categorization as a carcinogen or non-carcinogen is not scientifically credible. *Regul Toxicol Pharmacol* 103:124–129. <https://doi.org/10.1016/j.yrtph.2019.01.024>
45. Wolf DC, Cohen SM, Boobis AR, Dellarco VL, Fenner-Crisp PA, Moretto A, Pastoor TP, Schoeny RS, Seed JG, Doe JE (2019) Chemical carcinogenicity revisited 1: a unified theory of carcinogenicity based on contemporary knowledge. *Regul Toxicol Pharmacol* 103:86–92. <https://doi.org/10.1016/j.YRTPH.2019.01.021>
46. Cohen SM, Arnold LL (2011) Chemical Carcinogenesis. *Toxicol Sci* 120:S76–S92. <https://doi.org/10.1093/toxsci/kfq365>
47. Ruch RJ, Trosko JE, Farber E (2001) Gap-junction communication in chemical carcinogenesis (multiple letters). *Drug Metab Rev* Taylor & Francis 33:117. <https://doi.org/10.1081/DMR-100000137>
48. Rosenkranz HS (2002) Exploring the relationship between the inhibition of gap junctional intercellular communication and other biological phenomena. *Carcinogenesis* 21:1007–1011. <https://doi.org/10.1093/carcin/21.5.1007>
49. El-Fouly MH, Trosko JE, Chang CC (1987) Scrape-loading and dye transfer. A rapid and simple technique to study gap junctional intercellular communication. *Exp Cell Res* 168:422. [https://doi.org/10.1016/0014-4827\(87\)90014-0](https://doi.org/10.1016/0014-4827(87)90014-0)
50. Vanparry P, Corvi R, Aardema MJ, Gribaldo L, Hayashi M, Hoffmann S, Schechtman L (2012) Application of in vitro cell transformation assays in regulatory toxicology for pharmaceuticals, chemicals, food products and cosmetics. *Mutat Res – Genet Toxicol Environ Mutagen* 744:111–116. <https://doi.org/10.1016/j.mrgentox.2012.02.001>
51. OECD (2007) Detailed review paper on cell transformation assays for detection of chemical carcinogens. OECD Series on Testing and Assessment (31)
52. Urani C, Stefanini FM, Bussinelli L, Melchiorreto P, Crosta GF (2009) Image analysis and automatic classification of transformed foci. *J Microsc* 234:269–279. <https://doi.org/10.1111/j.1365-2818.2009.03171.x>
53. Forcella M, Callegaro G, Melchiorreto P, Gribaldo L, Frattini M, Stefanini FM, Fusi P, Urani C (2016) Cadmium-transformed cells in the in vitro cell transformation assay reveal

- different proliferative behaviours and activated pathways. *Toxicol Vitr* 36:71–80. <https://doi.org/10.1016/j.tiv.2016.07.006>
54. European Commission (2013) Recommendation 2013/179/EU on the use of common methods to measure and communicate the life cycle environmental performance of products and organisations. *Off J Eur Union* 210. https://doi.org/10.3000/19770677.L_2013.124.eng
55. Pedrazzani R, Cavallotti I, Bollati E, Ferreri M, Bertanza G (2018) The role of bioassays in the evaluation of ecotoxicological aspects within the PEF/OEF protocols: the case of WWTPs. *Ecotoxicol Environ Saf* 147:742–748. <https://doi.org/10.1016/j.ecoenv.2017.09.031>
56. Pedrazzani R, Ziliani E, Cavallotti I, Bollati E, Ferreri M, Bertanza G Use of ecotoxicology tools within the environmental footprint evaluation protocols: the case of wastewater treatment plants. *Desalin Water Treat*. In press
57. Gruiz K, Meggyes T, Fenyvesi É (2015) Engineering tools for environmental risk Management: 2. Environmental toxicology. CRC Press
58. Rosenbaum RK, Bachmann TM, Gold LS, Huijbregts MAJ, Jolliet O, Jurasko R, Kochler A, Larsen HF, MacLeod M, Margni M, McKone TE, Payet J, Schuhmacher M, van de Meent D, Hauschild MZ (2008) USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater eco-toxicity in life cycle impact assessment. *Int J Life Cycle Assess* 13:532–546. <https://doi.org/10.1007/s11367-008-0038-4>
59. Papa M, Ceretti E, Viola GCV, Feretti D, Zerbini I, Mazzoleni G, Steinberg N, Pedrazzani R, Bertanza G (2016) The assessment of WWTP performance: towards a jigsaw puzzle evaluation? *Chemosphere* 145:291–300. <https://doi.org/10.1016/j.chemosphere.2015.11.054>
60. EEA (2011) Revealing the costs of air pollution from industrial facilities in EuropeEEA technical Report. <https://doi.org/10.2800/23502>
61. De Schryver AM, Brakkee KW, Goedkoop MJ, Huijbregts MAJ (2009) Characterization factors for global warming in life cycle assessment based on damages to humans and ecosystems. *Environ Sci Technol* 43:1689. <https://doi.org/10.1021/es800456m>
62. WHO (2013) WHO methods and data sources for global burden of disease estimates 2000–2011. *Glob Heal Estim Tech Pap WHO* 4:81
63. Bertanza G, Canato M, Laera G, Vaccari M, Svanström M, Heimersson S (2017) A comparison between two full-scale MBR and CAS municipal wastewater treatment plants: techno-economic-environmental assessment. *Environ Sci Pollut Res* 24:17383. <https://doi.org/10.1007/s11356-017-9409-3>
64. Bertanza G, Canato M, Laera G (2018) Towards energy self-sufficiency and integral material recovery in waste water treatment plants: assessment of upgrading options. *J Clean Prod* 170:1206. <https://doi.org/10.1016/j.jclepro.2017.09.228>
65. Bertanza G, Baroni P, Canato M (2016) Ranking sewage sludge management strategies by means of decision support systems: a case study. *Resour Conserv Recycl* 110:1. <https://doi.org/10.1016/j.resconrec.2016.03.011>

Part III

Case Studies and Literature Reports



Chapter 15

Development of Baseline Quantitative Structure-Activity Relationships (QSARs) for the Effects of Active Pharmaceutical Ingredients (APIs) to Aquatic Species

David J. Ebbrell, Mark T. D. Cronin, Claire M. Ellison, James W. Firman, and Judith C. Madden

Abstract

The aim of this work was to develop predictive approaches for acute and chronic toxicity in fish, *Daphnia*, and algae utilizing baseline toxicity models. Currently available public active pharmaceutical ingredient (API) ecotoxicity data were compared to published baseline toxicity QSARs and classification schemes for industrial chemicals. The results showed that methods of assessing ecotoxicity for industrial chemicals are not adequate for the assessment of APIs. To develop equivalent prediction methods for APIs, acute baseline toxicity QSARs for APIs based on hydrophobicity (as $\log P$) were constructed, and the lower limits of toxicity for the public API data were compared with published industrial baseline toxicity QSARs for fish, *Daphnia*, and algae. These baseline toxicity QSARs were subsequently compared to the available acute toxicity data from the iPiE database. Since 75% of APIs are ionizable, baseline toxicity QSARs were also constructed using $\log D$ at pH 7.0. For chronic toxicity baselines, uncensored NOEC and LOEC data from the iPiE database were plotted using either $\log P$ or $\log D$ at pH 7.0. An alternative methodology was used to develop chronic baseline toxicity QSARs which consisted of iteratively refining the line of best fit until approximately 90% of the values were above the baseline toxicity QSARs. These chronic baseline toxicity QSARs could subsequently be used to identify groups which exhibit toxicity in excess of the baseline (i.e., greater than $10\times$ the hydrophobicity-predicted toxicity).

Key words NOEC, LOEC, QSAR, Environmental Risk Assessment, Aquatic toxicity, Baseline toxicity, Excess toxicity

1 Environmental Risk Assessment for Active Pharmaceutical Ingredients

Since 2006, within the European Union (EU), an Environmental Risk Assessment (ERA) must accompany applications for marketing authorization of active pharmaceutical ingredients (APIs) [1]. The standard tests used to assess chronic and acute ecotoxicological effects of APIs are presented in Table 1 [2–6]. These tests relate to three trophic levels (fish, *Daphnia*, and algae) and aim to give a broad assessment of the most relevant environmental adverse

Table 1
Standard tests used to assess adverse effects of APIs on environmental species

Test guideline	Organism	Type	Endpoint(s)
OECD 201	Algae sp.	Acute and chronic	Inhibition of growth (EC_{50}), NOEC, LOEC
OECD 202	<i>Daphnia</i> sp.	Acute	Immobilization (EC_{50})
OECD 203	Fish	Acute	Lethality (LC_{50})
OECD 210	Fish (early life stage)	Chronic	Growth, survival, hatching (NOEC, LOEC)
OECD 211	<i>Daphnia magna</i>	Chronic	Reproduction, growth, immobilization (NOEC, LOEC)

effects (hence these three species were selected for model development within this study). Ecotoxicological data are used to calculate a Predicted No Effect Concentration (PNEC) for the most susceptible species [1]. The PNEC can be calculated using data from acute or chronic assessments; however chronic data are preferred as they are more representative of exposure scenario for wildlife. In most instances using available acute data, rather than commissioning chronic testing, is pragmatic as the acute assessments provide a conservative estimation of the PNEC (when appropriate correction factors are used). Vestel et al. [7] found that 85% of PNECs calculated using acute data were more protective than the corresponding chronic PNEC. However, problems arise when the mechanism of action that causes the lethality measured in acute assays differs to the mechanism of the sublethal effects observed in chronic test systems [8]. Thus acute data are only a suitable surrogate when the mechanism of action remains the same [9].

For many APIs, particularly those approved prior to 2006, the standard tests indicated in Table 1 have not been performed, and hence there is little understanding of their potential impact on environmental species. Many computational models, such as Quantitative Structure-Activity Relationships (QSARs), have been published and used to assess the environmental impacts of industrial compounds (e.g., [10–13]). The use of such models can help prioritize the APIs for which there are little or no ecotoxicity data. Creation of these models was made possible because of the quantity and quality of publicly available data for industrial compounds, especially for acute endpoints. Unfortunately, fewer data for APIs are available publically [14]; therefore only a small number of published models are available for these compounds [15, 16]. Although acute and chronic effects for industrial chemicals are well established, applying the same approach to APIs may

be problematic due to specific differences between APIs and industrial chemicals. For example, unlike many industrial chemicals, APIs are structurally complex and often contain multiple functional groups and are designed to be biologically active. Another important difference is in their potential for ionization. Approximately 75% of APIs contain an ionizable functionality, including single acidic or basic groups, or more complex multiprotic chemicals and zwitterions. One consequence of this is that although the logarithm of the octanol/water partition coefficient ($\log P$ also denoted as $\log K_{ow}$, indicating the relative distribution of neutral species) is considered to be an appropriate descriptor for the distribution of industrial compounds, for ionizable APIs alternative descriptors may be required. The logarithm of the distribution coefficient ($\log D$) which considers the relative distribution of both ionized and unionized species and liposome/water partition coefficient (a system more representative of biological membranes) have both been proposed. The greater quantity and quality of models developed to predict the ecotoxicity of industrial compounds are in part due to the amount of data available. For instance, toxicity values for industrial chemicals were generated in one, or a small number of, laboratory(ies) carried out using consistent methodology. Typically, the industrial data used are those generated by the Center for Lake Superior Environmental Studies at the University of Wisconsin, USA, which were summarized by Russom et al. [12]. In contrast, publically available data for APIs are limited with toxicity values being measured in multiple laboratories resulting in much greater variability.

In addition to the quality of the data used, development of QSARs for industrial compounds has also been assisted by the relative ease with which the main mechanisms of action of acute aquatic toxicity can be modelled. The majority of compounds cause toxicity through non-specific disruption of biological membranes as they diffuse into the organism; a mechanism known as narcosis. This mechanism of action is dependent on a compound's ability to move out of the aqueous environment and into or through the cell membranes. Hence it can be modelled using hydrophobic descriptors alone, the most common being $\log P$. Sanderson and Thomsen [17] showed a good correlation existed between $\log P$ and acute toxicity of APIs, suggesting narcosis as the most likely primary mechanism of toxicity for these compounds. It has been argued that descriptors more able to account for potential speciation of APIs (e.g., the liposome/water partition coefficient) allow for better models to be built [18–21]. The use of liposome/water partitioning or $\log P_{\text{has}}$ enabled acute toxicity values to be predicted for a large number of APIs. For example, Escher et al. [18] predicted toxicity values for approximately 90% of the APIs found in wastewater indicating the majority elicited acute toxicity effects via narcosis.

The other mechanisms of acute aquatic toxicity were summarized by Verhaar et al. [22] and also extended by Russom et al. [12] and Thomas et al. [23]. In addition to two mechanisms for narcosis (polar and non-polar), reactive and specifically acting mechanisms were also identified. The domain of each mechanism was defined by Verhaar et al. using simple 2D molecular substructures. When used in a decision tree approach, the Verhaar scheme can be used to classify compounds into one of the four classes: class 1, baseline toxicity and non-polar narcotics; class 2, polar narcotics; class 3, reactive compounds; and class 4, compounds acting via a specific, receptor-mediated mechanism. The simplicity of the scheme makes it amenable to encoding into software used in toxicity prediction; one example of which is Toxtree (more details are available via the website <http://toxtree.sourceforge.net>). The scheme was built using a training set containing a range of environmental pollutants with acute lethality concentrations in fish and has been shown to have good predictive capabilities in assigning mechanisms of action for other industrial compounds [24–26]. These assigned mechanisms enable local modelling of adverse effects that can be more reliable and easier to rationalize in a regulatory submission. However, the scheme does not yet sufficiently cover the structural domain of APIs, and in a recent study, 78% of APIs were found to be out of the domain for the classification scheme [7]. One reason for this may be that there are relatively few defined structures associated with allocation to class 4, whereas class 4 ideally should capture all specific receptor-mediated mechanisms, (i.e., interaction with (1) protein receptors, (2) enzymes, (3) ion channels, or (4) transporters [27]). Many APIs are designed to elicit activity via such interactions [18]. However, it has also been shown that the majority of APIs elicit acute lethality via a narcotic mechanism; therefore expansion of the structural domains of classes 1 and 2 may also improve the applicability of this scheme to APIs. Expanding Verhaar class 4 to cover all the specific mechanisms of APIs is challenging not only because there are a large number of mechanisms to consider but also because the number of compounds acting via any one specific mechanism is relatively small, leading to the development of less reliable local models. Another approach for identifying compounds that do not elicit toxicity via narcosis, and hence exhibit acute toxicity in excess of the baseline toxicity QSAR, is the use of toxic ratios (TRs) [21, 28, 29]. Toxic ratios are calculated from the difference between the adverse effect concentration predicted using a baseline toxicity QSAR and the measured concentration. Compounds exhibiting high TRs are likely to act via reactive or specific mechanisms (the latter being more probable for APIs). However, the use of TRs is dependent on the availability of reliable baseline toxicity QSARs (narcosis) for APIs. Thus, the aims of this work were to assess how measured acute aquatic toxicity data for APIs compare with values predicted using existing baseline toxicity QSARs (developed using data for

industrial compounds) and to derive more robust baseline toxicity QSARs, specific to APIs, that may be useful in identifying those exhibiting excess toxicity.

2 Data Collation

The data used throughout this study were collected and extracted from the database developed as part of the Intelligence-led Assessment of Pharmaceuticals in the Environment (iPiE) project. The database for this project was populated by contributions of data from industrial partners of the project. These data were generated by adhering to the relevant OECD test guidelines for acute and chronic ecotoxicity endpoints for fish, *Daphnia*, and algae (see Table 1). Initial work focused on publicly available data; data from the iPiE project were also used once they became available. Four primary data sources were used to collate publicly available acute and chronic aquatic toxicity data for APIs. These were the Mistrapharma database (www.mistrapharma.se) and publications by Sanderson and Thomsen [30], Vestel et al. [7], and Brausch et al. [31]. These contained structural identifiers (name, CAS, and SMILES) and toxicity data collated from over 150 original sources. The toxicity data related to lethality (LC₅₀), effect (EC₅₀), no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) for fish, algae, and invertebrates. The specific species tested were noted in the majority of cases (70% of records), but for others only the generic taxonomic group was provided (e.g., “fish”, *Daphnia*). A second API dataset was obtained from the iPiE database.

The data were combined into a single dataset and standardized by performing the following:

1. Removal of compounds which were not APIs (e.g., excipients and intermediates)
2. Removal of salts forms of APIs
3. Retrieval of molecular weights from Chemspider (www.chemspider.com) to enable the toxicity concentrations reported to be converted into mM units
4. Removal of records associated with censored toxicity values (i.e., recorded as > or <) or data otherwise unsuitable for modelling (e.g., ambiguous or missing units of measurement)

Initially, investigations focused on developing baseline toxicity QSARs for acute toxicity in three species (fish, *Daphnia*, and algae). Thus, NOEC, LOEC, and effect concentrations relating to anything other than lethality (or immobility in *Daphnia*) were removed from analysis of the acute data. The compiled datasets will henceforth be referred to as the “public API dataset” and the “iPiE API dataset.”

A dataset of acute toxicity values for industrial compounds was also compiled. This was to enable a comparison to be made between the baselines toxicity QSARs generated for the two different chemical classes (i.e., APIs versus industrial compounds). This also enabled a comparison of the chemical space occupied by the two broad classes of compounds to be made and an assessment of whether the behavior of APIs differs when compared to industrial compounds. The dataset of acute toxicity values for industrial compounds, as reported by Russom et al. [12], was used for this purpose. These data comprised LC₅₀ values for the fish species *Pimephales promelas* (fathead minnow) along with chemical names and generic modes of action (i.e., narcosis, reactive toxicity, etc.). These data will henceforth be referred to as the “industrial dataset.”

3 Results

3.1 Data Analysis

An analysis into the variability of the publicly available acute toxicity data was performed wherein both the median toxicity and the range of values (for APIs with more than one measured toxicity value) were examined. Figure 1 shows the results of the investigation into the variability of the publicly available data. Note that for some APIs, reported toxicity values span 3 orders of magnitude or more (see APIs ethinylestradiol and 17 β-estradiol in Fig. 1). Highly variable data lead to lower quality models being developed [32].

To enable a comparison of the chemical space of APIs and industrial compounds, ecotoxicity data relevant to both groups were collated from the literature. The sources for the public API data were publications by Brausch et al. [31], Sanderson and Thomsen [30] and Vestel et al. [7], and also the Mistrapharma database. The source of the industrial data was the publication by Russom et al. [12]. The collation criteria are fully described within Sect. 2. For comparative purposes, a summary of the collated data is presented in Tables 2 and 3 (for public APIs and industrial chemicals, respectively).

The plot of toxicity against hydrophobicity for fish from both the industrial and public API datasets is shown in Fig. 2. The APIs have a narrow range of both toxicity concentrations and log *P* values that generally fit well within that of the industrial compounds. However, the APIs do not show the same distinctive baseline toxicity QSAR that the industrial compounds do for the correlation of log (1/LC₅₀) with log *P*. This could in part be due to the more significant experimental variability in the collated data (multiple sources of experimental data with different fish species) or mechanisms other than narcosis dominating the observed acute toxicity mechanisms of APIs.

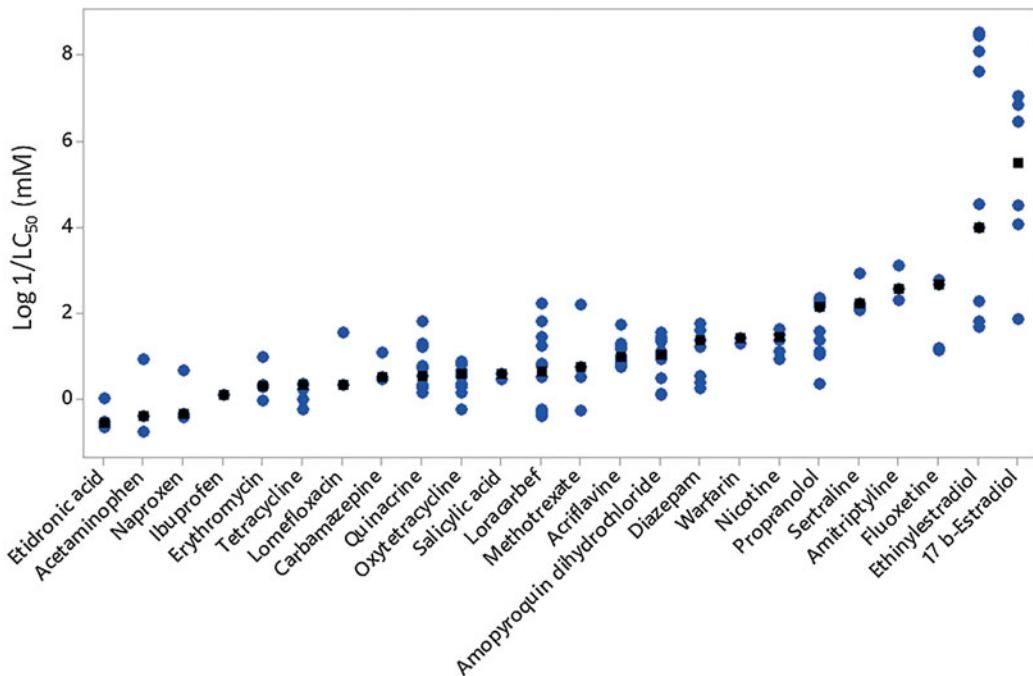


Fig. 1 The variability in public acute toxicity values in fish (blue circles indicate individual recorded values for each chemical; median values are indicated by black squares)

Table 2
Summary of the acute ecotoxicity data collated for public APIs

	Endpoints		
	Fish LC ₅₀	Daphnia EC ₅₀ or LC ₅₀ ^a	Algae EC ₅₀ or LC ₅₀
Number of compounds	152	234	166
Number of toxicity data points	272	644	388
Toxicity (log 1/LC ₅₀ or EC ₅₀) (mM) range	-1.27 — 4.34	-3.03 — 6.51	-2.59 — 5.71
Calculated log P range	-4.41 — 9.1	-9.36 — 9.1	-4.84 — 9.1

^aIn *Daphnia* assays immobilization can accurately be described as an EC₅₀ but may be reported as an LC₅₀; likewise for algae, the effect of growth inhibition may be reported as an LC₅₀

Table 3
Summary of the fish acute ecotoxicity data collated for industrial chemicals

	Fish LC ₅₀
Number of compounds	408
Number of toxicity data points	408
Toxicity (log 1/LC ₅₀) range	-2.96 — 6.38
Calculated log P range	-1.75 — 7.43

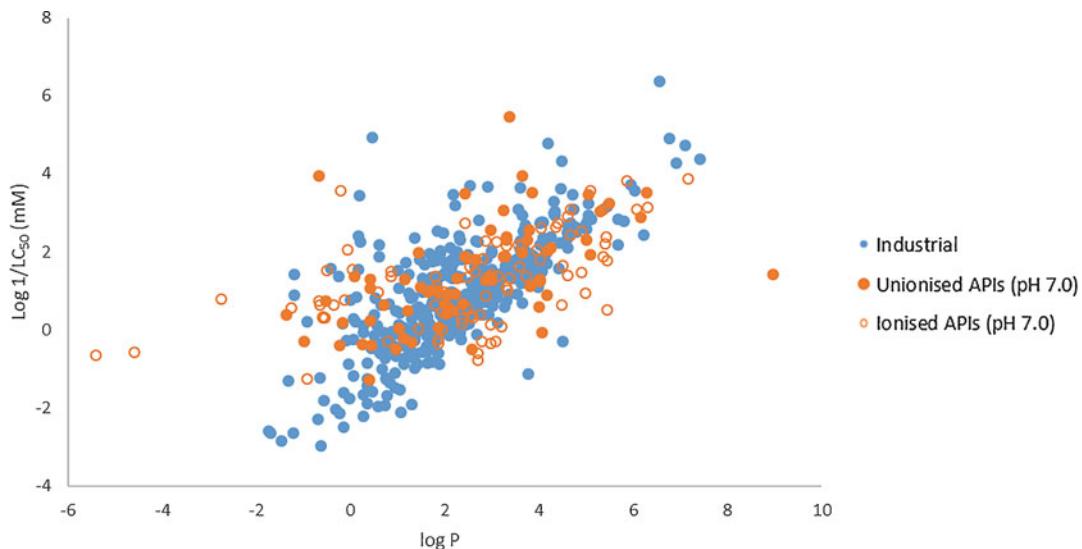


Fig. 2 Relationship between fish acute lethality data for APIs (unionized APIs at pH 7.0 are in bold orange circles, and ionized APIs at pH 7.0 are empty orange circles) and industrial compounds (blue circles) and $\log P$

3.2 Classification of Compounds Using the Verhaar Scheme and Comparison of the Chemical Space of Industrial Compounds Versus APIs

For the purposes of comparing the chemical space of industrial compounds to APIs, fish toxicity data selected from the “public API” dataset and the “industrial dataset” were used. To provide a basis for comparison, both the API and industrial datasets were analyzed using the Verhaar scheme as implemented in Toxtree version 2.6.13, supplemented by the use of the post-processing filters published by Ellison et al. [24], to classify the compounds according to putative mechanism/mode of action (reported in Table 4). This was done not only to enable a comparison between distribution of APIs and industrial compounds within each class but also to replicate the results of Vestel et al. [7] who showed that the majority of APIs are outside the domain of the scheme. The compounds were classified into one of the four Verhaar classes as described in Sect. 1, or into class 5. The chemical spaces were compared by plotting the inverse of the acute toxicity values ($\log (1/\text{LC}_{50}$ or EC_{50}) in mM concentrations) against hydrophobicity for all Verhaar classes. Hydrophobicity was represented by $\log P$ and was calculated using the ACD/Labs software, developed by Advanced Chemistry Development, Inc. [33]. The majority of APIs are out of the domain of the Verhaar scheme with 70% of the API dataset being classified as class 5 (out of domain), in agreement with the findings of Vestel et al. [7]. Conversely, 69% of the industrial compounds were classified into classes 1–4. This finding is, however, to be expected since information from industrial compounds was used to develop the structural rules of the Verhaar scheme. This suggests that the distinctive structural

Table 4

Number of compounds classified into each of the Verhaar classifications following the implementation of the Verhaar (modified) scheme in Toxtree and the Ellison et al. [24] extension rules

Verhaar class	Number of industrial compounds in class	Proportion of industrial compounds in class	Number of APIs in class	Proportion of APIs in class
1: Non-polar narcotics	110	0.27	7	0.06
2: Polar narcotics	50	0.12	8	0.06
3: Reactive compounds	89	0.22	17	0.13
4: Specifically acting compounds	33	0.08	6	0.05
5: Out of domain	126	0.31	89	0.70

moieties and chemical space of APIs were not covered by the training data and therefore the majority of APIs are outside of the domain of the Verhaar scheme.

Analysis by class of the APIs classified into Verhaar classes 1–5 (Fig. 3) reveals that most of the compounds which were assigned to classes 1, 3, and 4 did not show any significant differences in terms of the range of the toxicity and relationship to $\log P$ as compared to the industrial compounds. The plot for class 1 indicates that a baseline toxicity QSAR may be discernible for APIs acting via narcosis; however, few APIs fall into this category. For the APIs in class 5, there was no discernible relationship with $\log P$, and unlike the industrial compounds, it is more difficult to determine a significant baseline toxicity QSAR. It is clear that the Verhaar scheme needs to be expanded not only to cover the domain of APIs but also for the industrial compounds. These improvements could be achieved in two ways: extension of the structural characteristics of the narcosis domains (classes 1 and 2) to accommodate a larger proportion of the industrial compounds currently in class 5 and the expansion of the domain of class 4 to include more specifically acting APIs. However, before it is possible to classify compounds as either baseline or excess toxicants, it is imperative that representative baseline toxicity QSARs are developed for the toxicity observed for APIs. The collation of publicly available ecotoxicity data for APIs has resulted in the generation of a reasonably sized dataset, which can be used for modelling. However, while there are a reasonable number of publicly available ecotoxicity data for APIs, the variability within these data is significant.

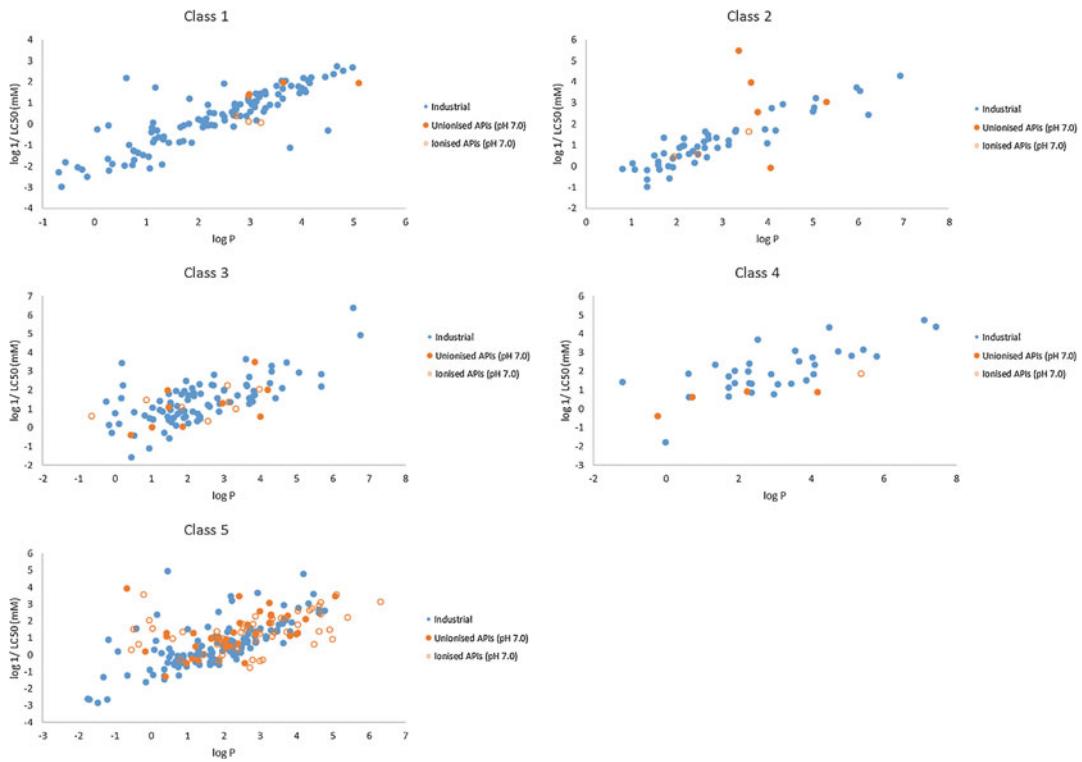


Fig. 3 Relationship between fish acute lethality data for APIs (unionized APIs at pH 7.0 are in bold orange circles, and ionized APIs at pH 7.0 are empty orange circles) and industrial compounds (blue) and $\log P$ for compounds classified to each of the five Verhaar classes

4 Discussion

4.1 Development of Baseline Toxicity QSARs Relevant to APIs

The next stage of the analysis involved developing baseline toxicity QSARs, relevant to APIs, using fish, algae, and *Daphnia* data. The baseline toxicity QSARs were constructed by comparing the lower limits of toxicity for the public API data with published baselines for each species. The published baseline toxicity QSARs had been developed using industrial data, the details for which are provided in Table 5; the models were converted to represent $\log(1/\text{LC}_{50}$ or EC_{50}) in mM concentrations where necessary. From this analysis upper and lower limits (cutoff points) for the correlation of toxicity with $\log P$ were empirically derived. The experimental reliability of $\log P$ values greater than 5.5 is questionable because of issues with poor solubility; the lower limits represent the concentration at which the minimum level toxicity is observed. The relationship between $\log P$ and acute toxicity between the observed cutoff points was then determined through visual comparison of the public API data with the industrial models. This resulted in linear P -based models being produced which were applicable

Table 5
Published baseline toxicity QSARs constructed using industrial compounds

Taxonomic group	Model ID	Model	Reference
Fish	F ₁	Log (1/LC ₅₀) (mM) = 0.94 log P — 1.83	Austin et al. [13]
	F ₂ ^a	Log (1/LC ₅₀) (mM) = 0.87 log P — 1.87	Könemann [11]
Algae	A ₁	Log (1/EC ₅₀) (mM) = 0.97 log P — 1.95	Hsieh et al. [34]
	A ₂	Log (1/EC ₅₀) (mM) = 0.90 log P — 1.40	Tsai and Chen [35]
<i>Daphnia</i>	D ₁ ^a	Log (1/EC ₅₀) (mM) = 0.82 log P + 1.58	Zhang et al. [36]
	D ₂	Log (1/EC ₅₀) (mM) = 0.79 log P — 1.24	ECOSAR

^aThese models have been altered in the report to represent log(1/LC₅₀ or EC₅₀) in mM concentrations

between two defined cutoff points for log P values (this method was only applied for the development of acute baseline toxicity QSARs).

Previous work on building baseline toxicity QSARs has focussed on collating data for a range of compounds likely to cause toxicity through non-polar narcosis (e.g., aliphatic alcohols) and adding data from other chemical classes (that act via the same mechanism), until a robust model is created using a significant number of compounds. This approach is problematic for APIs as their structures are more complicated than simple industrial compounds. Therefore, it is difficult to classify them as potential non-polar narcotics based on their structure. Even classification approaches such as the Verhaar scheme fail in this task because of the structural differences between APIs and simpler industrial compounds. As such, a more pragmatic approach was required to build API relevant baseline toxicity QSARs. To this end, QSARs of published industrial baseline toxicity QSARs were plotted along with API data to examine how the two datasets compared (Figs. 4, 5, and 6). By visual inspection it was then possible to overlay a baseline toxicity QSARs that was in keeping with other models and then to describe this line using a linear equation.

For all species, existing baseline toxicity QSARs fit within the lower segment of toxicity values but do not describe the lower limit of the observed toxicities. This is possibly due to the inherent differences between industrial chemicals and APIs (e.g., the majority of APIs may be ionized at pH 7). For this reason the manually fitted models described in Fig. 7 are a better description of the true lower limit of the observed toxicity. In addition, the lower plateau created by an absolute minimum observable toxicity and the high solubility cutoff are clearly visible. The precise descriptions of the models shown in Fig. 7 are described in Table 6.

The models shown in Fig. 7 are distinct from those developed previously for industrial chemicals not only in the approach used to

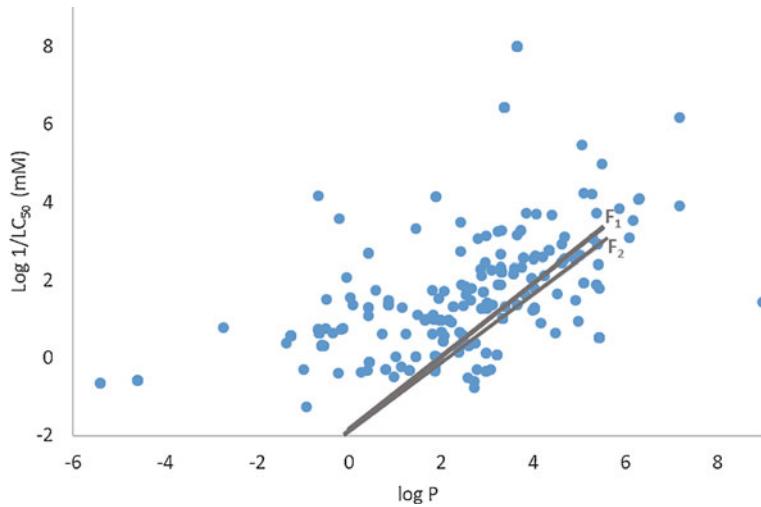


Fig. 4 Comparison of acute API toxicity to fish with published baseline toxicity QSARs (F₁-F₂ described in Table 2)

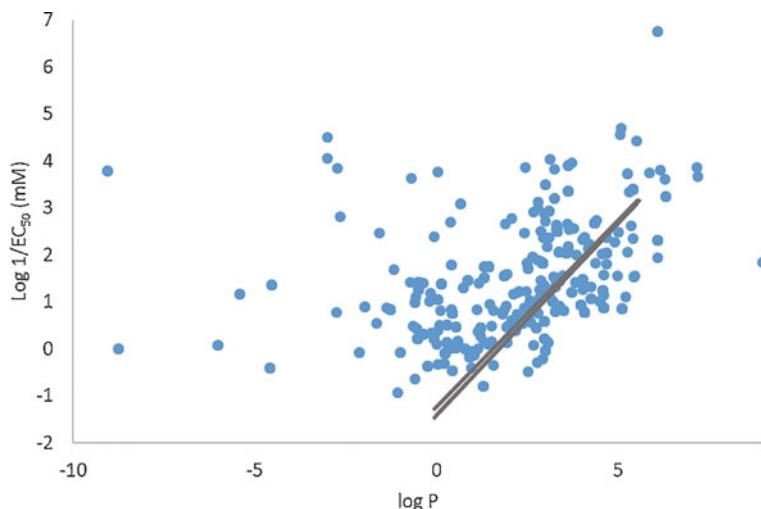


Fig. 5 Comparison of acute API toxicity to *Daphnia* with published baseline toxicity QSARs (D₁ (upper line); D₂ (lower line) as described in Table 2

build them but also as to which structures can be classified as exhibiting toxicity through non-polar narcosis. Industrial compounds are often put into the category of “baseline toxicant” based on the presence or absence of explicit structural features. Unfortunately, it is not as easy to structurally classify APIs as they often contain a variety of functional moieties. Therefore different methods are required to identify compounds where the observed toxicity can be explained and modelled through the phenomenon of non-polar narcosis.

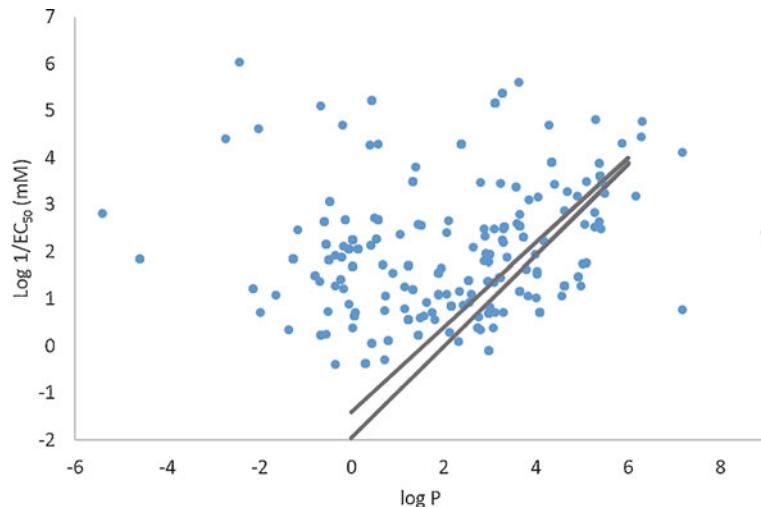


Fig. 6 Comparison of acute API toxicity to algae with published baseline toxicity QSARs (A_1 (lower line); A_2 (upper line) described in Table 2)

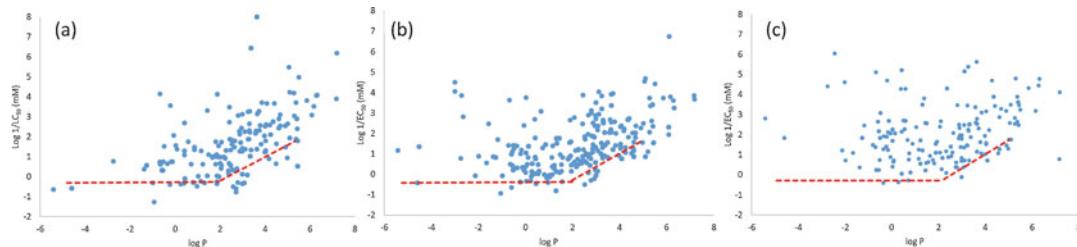


Fig. 7 Acute baseline toxicity QSARs derived herein for (a) fish (b) *Daphnia*, and (c) algae

Table 6
Summary of the acute ecotoxicity data collated available APIs from the iPiE project

	Endpoints		
	Fish LC ₅₀	<i>Daphnia</i> EC ₅₀ or LC ₅₀ ^a	Algae EC ₅₀ or LC ₅₀
Number of compounds	37	55	79
Number of toxicity data points	70	79	91
Toxicity (log 1/LC ₅₀ or EC ₅₀) range	0.20 – 4.80	-0.83 — 5.32	-0.17 — 5.22
Calculated log P range	-5.41 — 6.28	-9.04 — 8.98	-4.78 — 5.85

^aIn *Daphnia* assays immobilization can accurately be described as an EC₅₀ but may be reported as an LC₅₀; likewise for algae, the effect of growth inhibition may be reported as an LC₅₀

The models presented herein are intended to represent the minimum observed ecotoxicity of APIs. Thus they can provide an initial indication of the lowest concentration of concern. This may, or may not, reflect the observed experimental outcomes for any

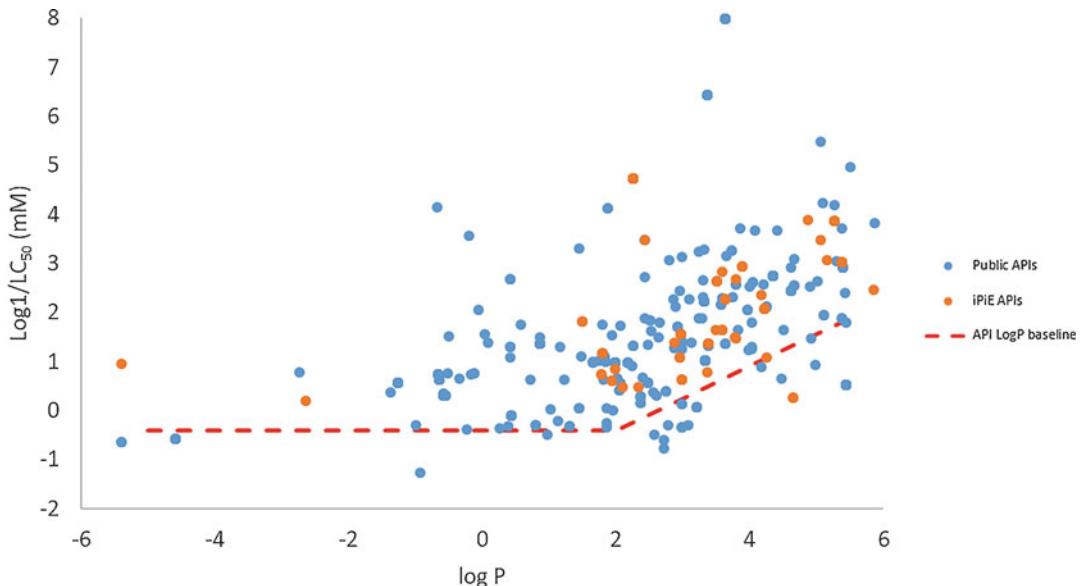


Fig. 8 Baseline toxicity QSAR for fish acute toxicity versus $\log P$ showing the distribution of publicly available data and data from the iPiE project

particular compound but does provide a starting point. The minimum level of toxicity for APIs appears to be less than that of the widely modelled industrial compounds, which raises the interesting question of why the industrial baseline toxicity QSARs do not represent the absolute minimum level of toxicity. However, the baseline toxicity QSARs developed here do allow for the identification of groups that exhibit toxicity in excess of the baseline toxicity QSARs. Further analysis is required to model the patterns observed between a chemical's structural properties and distance of the observed toxicity from the baseline toxicity QSARs.

Initially models for acute toxicity were built using the public data that were available from the outset of the project. When data from the iPiE project became available, these were used to validate the models developed previously using the public data. Comparing the data extracted from toxicological studies within the iPiE project with the three baseline toxicity QSARs developed using the publicly available data shows that the acute toxicity of the chemicals from the iPiE dataset falls within a similar range to the public data (Figs. 8, 9, and 10 for fish, *Daphnia*, and algae, respectively). The acute data within the iPiE dataset are summarized in Table 6.

Given that 75% of APIs are ionizable, it is possible that $\log D$ may be more suitable to model lipophilicity. With this in mind, acute baseline toxicity QSARs using $\log D$ were established adopting an identical method used for the $\log P$ baseline toxicity QSARs. Acute baseline toxicity QSARs are shown in Figs. 11, 12, and 13 for all three species (for fish, *Daphnia*, and algae, respectively). All

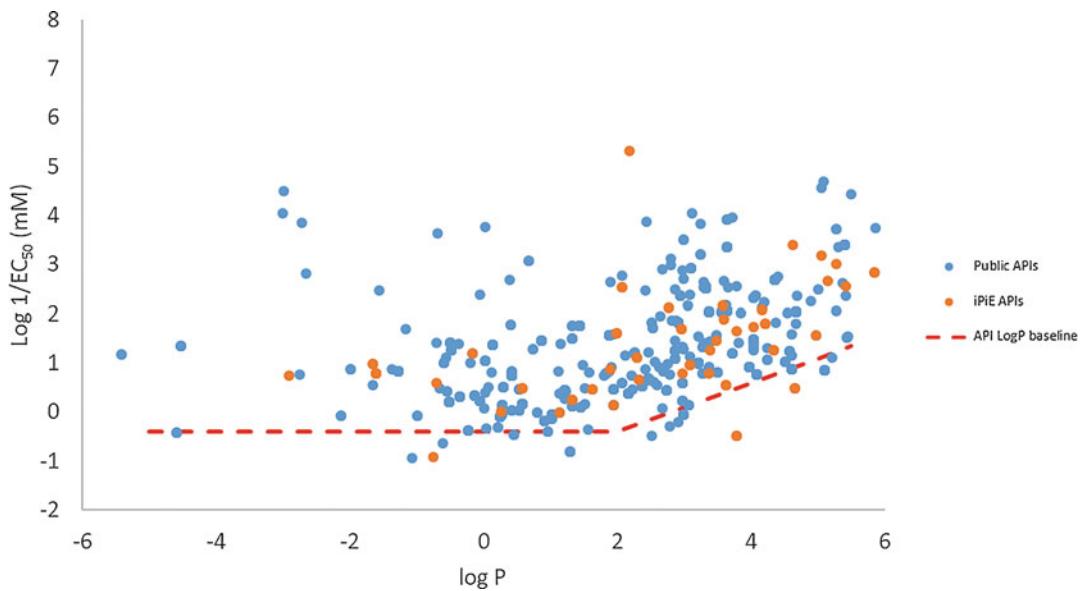


Fig. 9 Baseline toxicity QSAR for *Daphnia* acute toxicity versus log P showing the distribution of publicly available data and data from the iPiE project

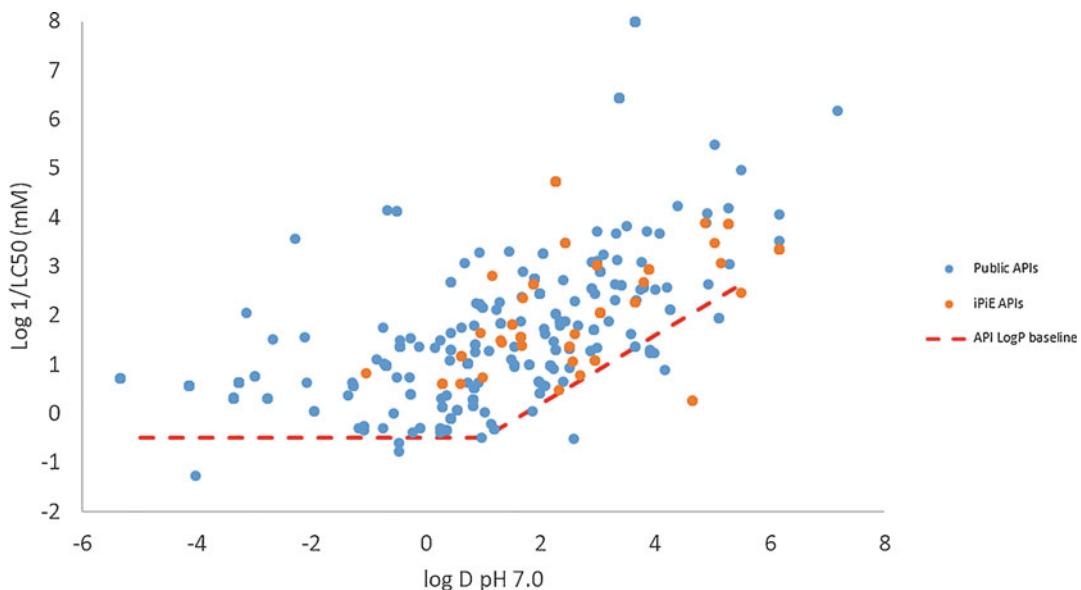


Fig. 10 Baseline toxicity QSAR for acute algal toxicity versus log P showing the distribution of publicly available data and data from the iPiE project

models for log D are shown in Table 12. Overall the log P and log D baseline toxicity QSARs are quite similar. However, the point at which toxicity increases linearly typically occurs at 1.0 log unit for the log D baseline toxicity QSARs in comparison to 2.0 log units for the log P baseline toxicity QSARs (compare Tables 11 and 12).

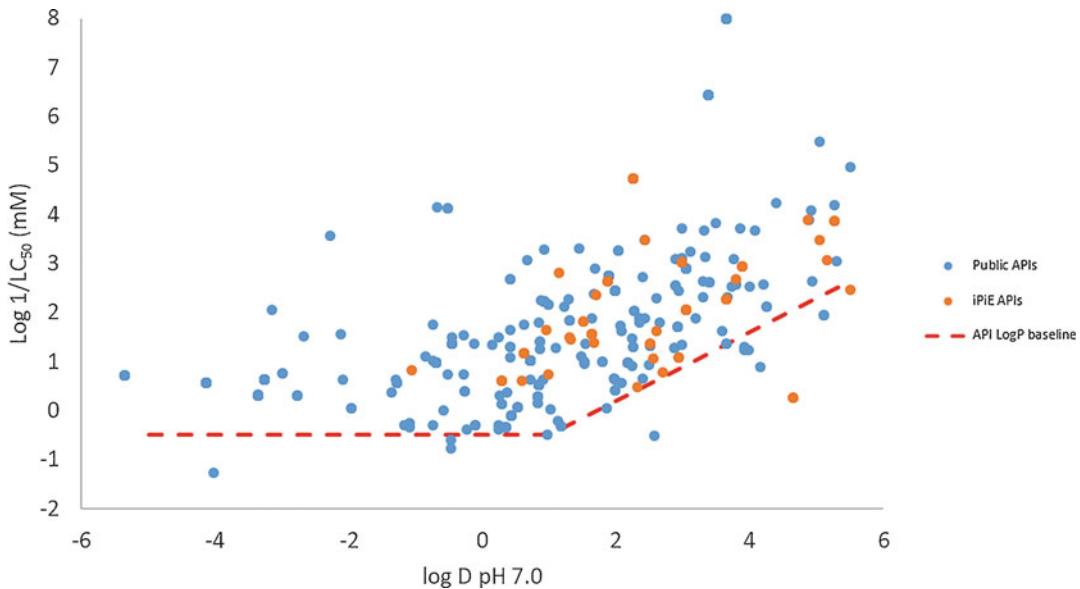


Fig. 11 Baseline toxicity QSAR for fish acute toxicity versus $\log D$ with public API and iPiE API data

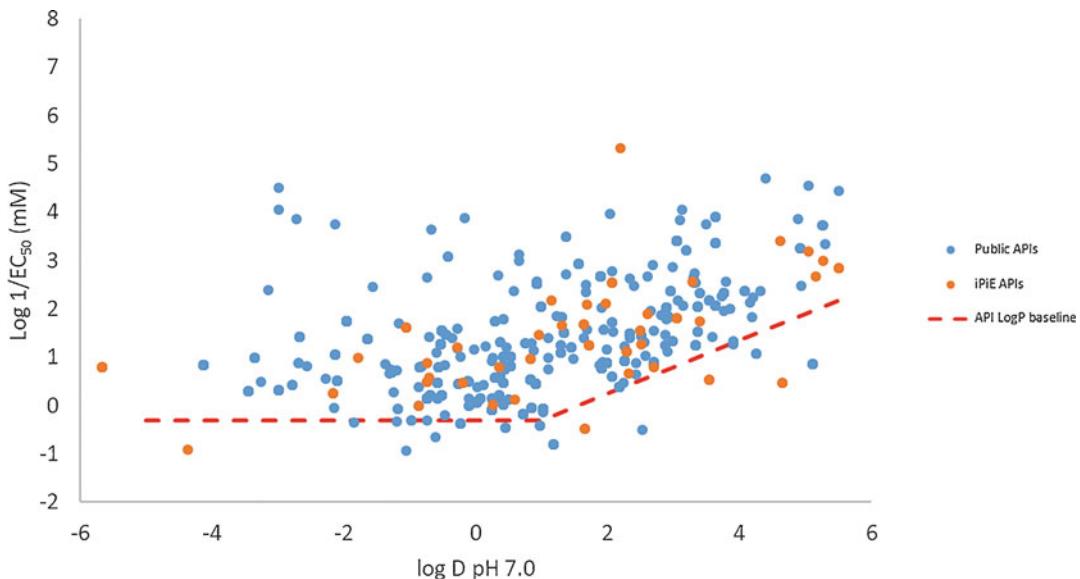


Fig. 12 Baseline toxicity QSAR for *Daphnia* acute toxicity versus $\log D$ with public API and iPiE API data

Although there have been numerous studies that have successfully derived acute baseline toxicity QSARs, the number of studies focusing on the development of chronic baseline toxicity QSARs is limited. In our approach we adopted an alternative methodology to that used for the development of the acute baseline toxicity QSARs in order to develop chronic baseline toxicity QSARs using iPiE data. The analysis was carried out using the NOEC and LOEC

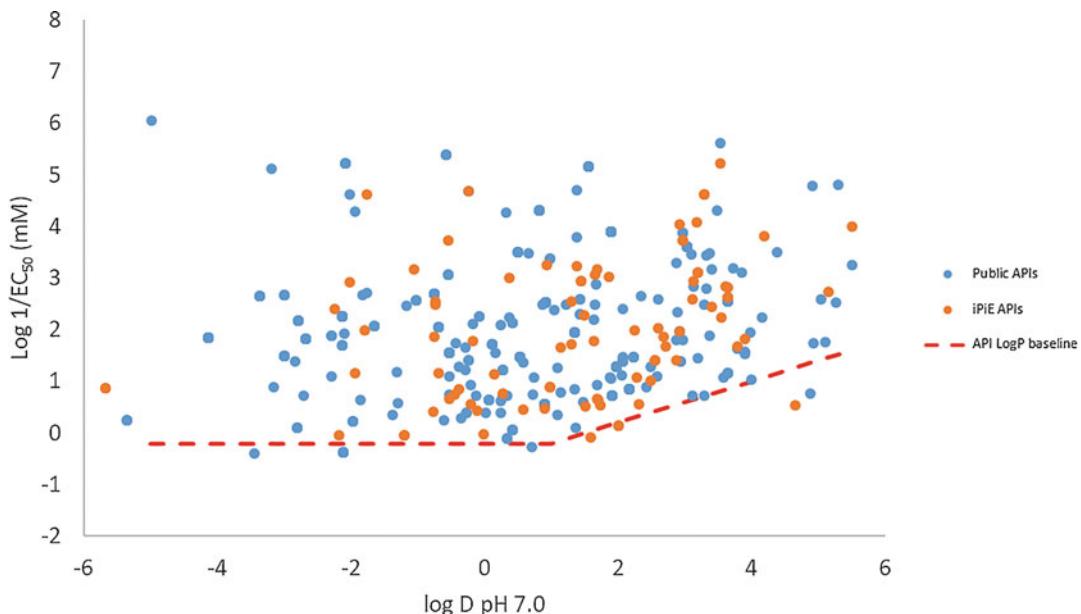


Fig. 13 Baseline toxicity QSAR for algae acute toxicity versus $\log D$ with public API and iPiE API data

values collated using analogous criteria to the collation of the acute toxicity values (refer to Sect. 2). The chronic data available via the iPiE project are summarized in Table 7.

Before developing chronic baseline toxicity QSARs, the iPiE NOEC data were analyzed to assess their usability. NOEC values were determined from a range of test concentrations, and the NOEC is equal to highest test concentration where there was no statistically significant effect on the recorded observation. However, in some cases this was equal to the reported maximum test concentration. This tells us little about the true chronic toxicity of the API as the concentration which represents the highest concentration where toxicity does not occur can lie anywhere beyond this point. By looking at the study designs, it was possible to identify which of the reported NOEC values were identical to the maximum test concentration for that study. This was 50% of the values for fish, 30% of the values for *Daphnia*, and 25% of the values for algae (Figs. 14, 15, and 16 for fish, *Daphnia*, and algae against $\log D$ at pH 7.0, respectively, with all values shown in Table 8). It is worth noting that some of the reported NOEC values were greater than the maximum tested concentration or were not equal to any of the reported test concentration (29 values for fish, 10 for *Daphnia*, and 35 for algae; these values were removed from the analysis). Interestingly, most of the values where the NOEC was equal to the maximum test concentration lie within the lower toxicity range where the baseline toxicity QSAR would likely fit. As such using these values would have a significant effect on the nature of the baseline toxicity QSARs for minimum toxicity, and therefore these

Table 7
Summary of the chronic ecotoxicity data collated for APIs within the iPiE project

	Endpoints					
	Fish		<i>Daphnia</i>		Algae	
	NOEC	LOEC	NOEC	LOEC	NOEC	LOEC
Number of compounds	102	59	103	76	132	55
Number of toxicity data points	335	131	257	143	317	104

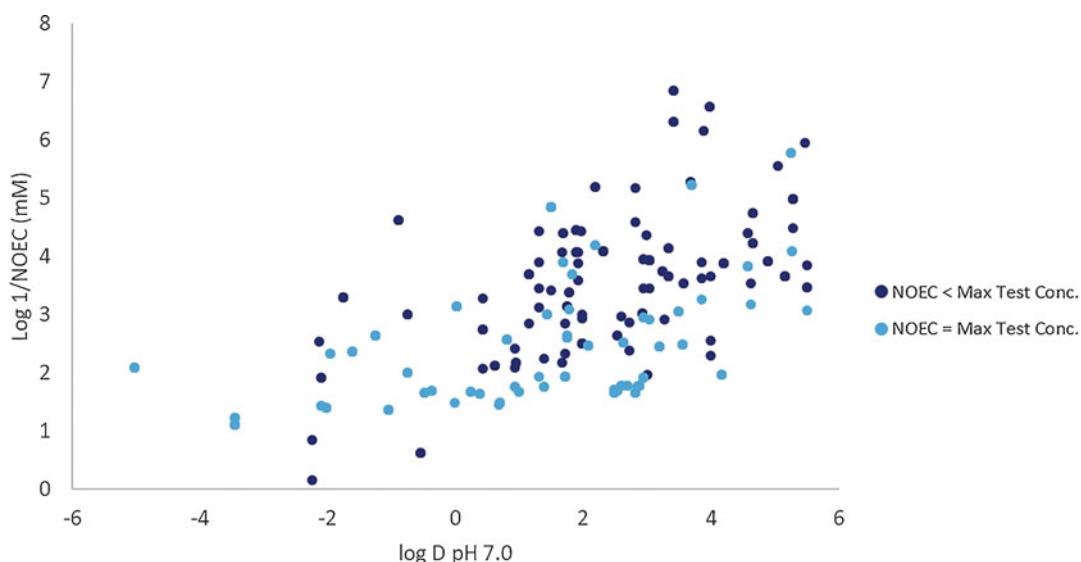


Fig. 14 Analysis of reported fish NOEC values compared to tested concentration ranges against $\log D$ at pH 7.0

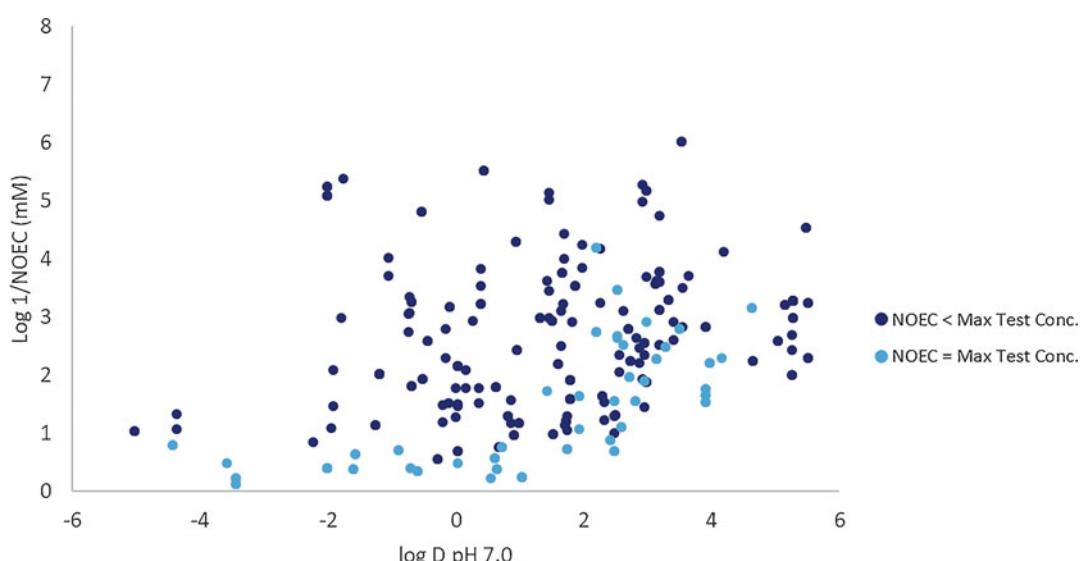


Fig. 15 Analysis of reported *Daphnia* NOEC values compared to tested concentration ranges against $\log D$ at pH 7.0

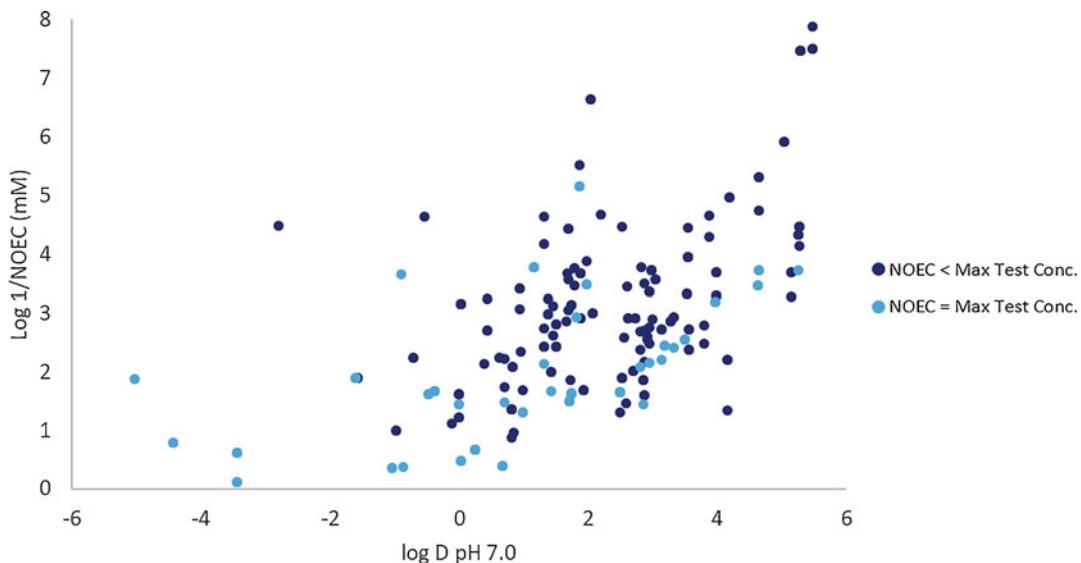


Fig. 16 Analysis of reported algae NOEC values compared to tested concentration ranges against $\log D$ at pH 7.0

Table 8
Relationship between reported NOEC values and test concentration ranges for all three species

	<i>Fish</i>		<i>Daphnia</i>		<i>Algae</i>	
	Number of values	Number of APIs	Number of values	Number of APIs	Number of values	Number of APIs
Number of NOECs = maximum test concentration	167	55	77	36	79	39
Number of NOECs < maximum test concentration	139	59	170	79	203	91

values were removed from all subsequent analyses. It was possible to carry out this analysis on the iPiE data due to the plethora of information that came with the toxicity values. However, carrying out a similar analysis on the available public data may prove more challenging as not all of the same information is available. As such the chronic toxicity baseline toxicity QSARs were developed using iPiE data alone.

The remaining NOEC values were used to developed chronic NOEC baseline toxicity QSARs using $\log P$ and $\log D$ at pH 7.0. These baseline toxicity QSARs were developed using an alternative method to the acute baseline toxicity QSARs. The following method was adopted to ensure that the baseline toxicity QSARs represented the minimum level of toxicity for APIs using a less subjective approach (see Fig. 17):

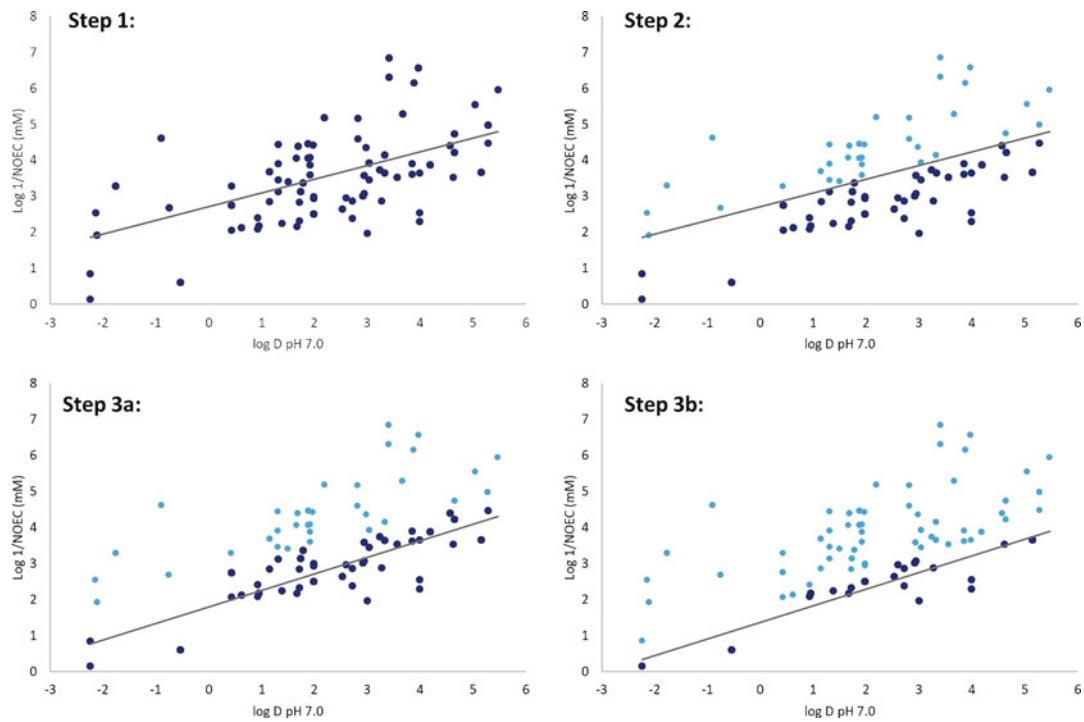


Fig. 17 The steps used to develop the chronic baseline toxicity QSARs for fish NOEC values, the values used to generate the lines are shown in dark blue, and values omitted at each iteration are shown as light blue (steps described in the text)

Table 9

The percentage of values which were above the developed line for each iteration

	Species (% of values above the baseline)					
	Fish		<i>Daphnia</i>		Algae	
	$\log P$	$\log D$	$\log P$	$\log D$	$\log P$	$\log D$
Iteration 1	45	44	42	42	47	44
Iteration 2	74	74	72	75	73	74
Iteration 3	88	90	86	88	85	85

1. A line of best fit was developed for the data.
2. All values that fall above the line of best fit were removed.
3. Steps 1 and 2 were repeated until approximately 90% of the data were above the baseline toxicity QSARs; this was the third iteration for all species (see Table 9).

The resulting NOEC baseline toxicity QSAR using $\log P$ can be found in Fig. 18 and Table 11, and the $\log D$ baseline toxicity

Table 10

The number of original LOEC values for each species and the number of additional LOEC values generated from the analysis

	Species		
	Fish	Daphnia	Algae
Number of original LOEC values ^a	107	119	77
Number of additional LOEC values ^a	15	25	54

^aNote these values may not match up with the number of NOECs for each species for the reasons discussed in the text

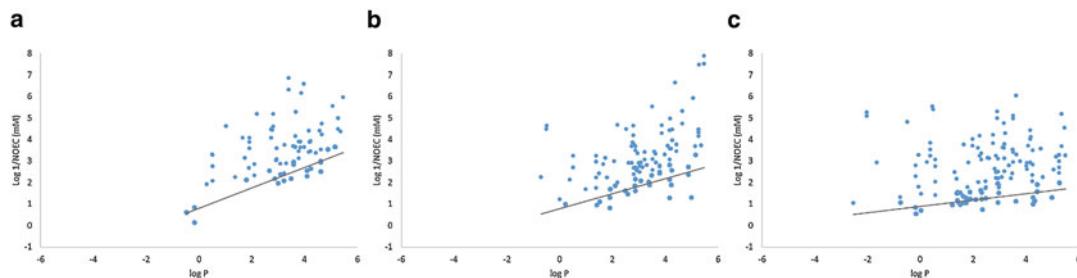


Fig. 18 Chronic NOEC baseline toxicity QSARs developed for (a) fish, (b) *Daphnia*, and (c) algae using $\log P$

Table 11

Derived acute and chronic baseline toxicity QSARs showing the correlation of toxicity with $\log P$ for fish, *Daphnia*, and algae as depicted in Figs. 8, 9, 10, 18, and 20

Fish	
Acute toxicity	If $\log P < 2$ then $\log(1/\text{LC}_{50})\text{mM} = -0.4$; If $2 < \log P < 5.5$ then $\log(1/\text{LC}_{50})\text{mM} = 0.45 \log P - 1.3$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.47 \log P + 0.81$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.48 \log P + 0.45$
<i>Daphnia</i>	
Acute toxicity	If $\log P < 2$ then $\log(1/\text{EC}_{50})\text{mM} = -0.6$; If $2 < \log P < 5.5$ then $\log(1/\text{EC}_{50})\text{mM} = 0.6 \log P - 1.8$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.35 \log P + 0.93$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.34 \log P + 0.43$
Algae	
Acute toxicity	If $\log P < 2$ then $\log(1/\text{EC}_{50})\text{mM} = -0.5$; If $2 < \log P < 5.5$ then $\log(1/\text{EC}_{50})\text{mM} = 0.60 \log P - 1.7$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.47 \log P + 0.81$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.12 \log P + 0.56$

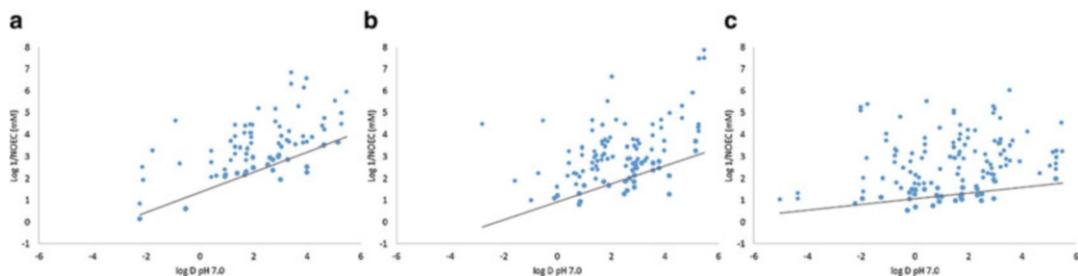


Fig. 19 Chronic NOEC baseline toxicity QSARs developed for (a) fish, (b) *Daphnia*, and (c) algae using $\log D$ at pH 7.0

Table 12

Derived acute and chronic toxicity models against $\log D$ for fish, algae, and *Daphnia* as depicted in Figs. 11, 12, 13, 19, and 21

Fish	
Acute toxicity	If $\log D < 1$ then $\log(1/\text{LC}_{50})\text{mM} = -0.5$; If $1 < \log D < 5.5$ then $\log(1/\text{LC}_{50})\text{mM} = 0.7 \log D - 1.2$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.46 \log D + 1.40$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.52 \log D + 0.77$
<i>Daphnia</i>	
Acute toxicity	If $\log D < 1$ then $\log(1/\text{EC}_{50})\text{mM} = -0.3$; If $1 < \log D < 5.5$ then $\log(1/\text{EC}_{50})\text{mM} = 0.55 \log D - 0.85$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.41 \log D + 0.93$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.35 \log D + 0.75$
Algae	
Acute toxicity	If $\log D < 1$ then $\log(1/\text{LC}_{50})\text{mM} = 0.1$; If $1 < \log D < 5.5$ then $\log(1/\text{EC}_{50})\text{mM} = 0.45 \log D - 0.35$
Chronic toxicity (NOEC)	$\log 1/\text{NOEC} (\text{mM}) = 0.13 \log D + 1.1$
Chronic toxicity (LOEC)	$\log 1/\text{LOEC} (\text{mM}) = 0.17 \log D + 0.67$

QSARs can be found in Fig. 19 and Table 12. The NOEC toxicity values for fish and *Daphnia* are more influenced by $\log P$ and $\log D$ than in algae. Additionally, in comparison to the acute toxicity data, the range of $\log P$ and $\log D$ for the remaining compounds is smaller resulting in a smaller range of applicability for the chronic baseline toxicity QSARs.

In terms of toxicity, the use of NOEC baseline toxicity QSARs may be trivial as this is a baseline of no toxicity and therefore provides limited information about the concentration at which the chronic toxicity of an API occurs for that specific endpoint effect. One potential solution to this is to develop baseline toxicity

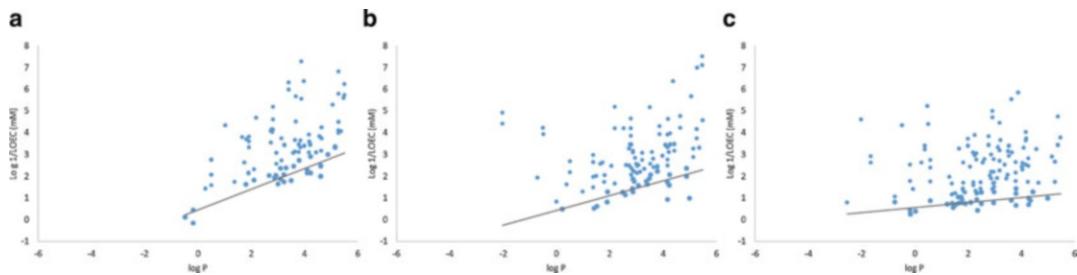


Fig. 20 Chronic LOEC baseline toxicity QSARs developed for (a) fish, (b) *Daphnia*, and (c) algae using $\log P$ (refer to Table 11 for model details)

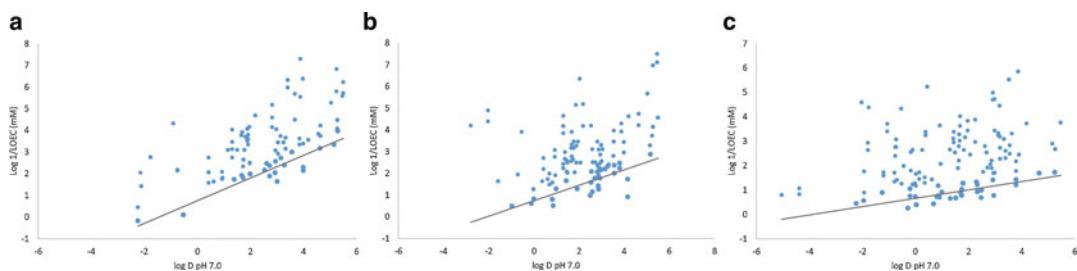


Fig. 21 Chronic LOEC baseline toxicity QSARs developed for (a) fish, (b) *Daphnia*, and (c) algae using $\log D$ at pH 7.0 (refer to Table 12 for model details)

QSARs based on the LOEC observation for an API. However, compared to the available NOEC data from the iPiE database, the number of available LOEC values is substantially lower. In addition to identifying the number of usable NOECs within the data set, it was possible to infer some additional LOEC values based on the reported NOEC value and test concentration ranges. This was achieved by taking the next test concentration up from the reported NOEC value as, in theory, the test concentration that is one increment higher than the reported NOEC should be the concentration that elicited the endpoint effect of interest (and thus should be the LOEC for that chemical and endpoint effect). The number of additional LOEC values generated is reported in Table 10. It was not possible to infer an additional LOEC for all NOEC values as some of the reported NOEC values did not match up to any of the reported concentration ranges, or the reported concentration units were inconsistent or incomprehensible. Baseline toxicity QSARs were developed using the total number of LOEC values for all three species using the same method for the NOEC baseline toxicity QSARs (Figs. 20 and 21 for baseline toxicity QSARs using $\log P$ and $\log D$, respectively). The equations for all the chronic and LOEC baseline toxicity QSARs using $\log P$ and $\log D$ are summarized in Tables 11 and 12, respectively.

5 Conclusions

Acute and chronic baseline toxicity QSARs for APIs have been developed based on public data and data from the iPiE project. The models developed here (see Tables 11 and 12) are to be integrated into the iPiE system software iPiEsys. Publicly available data for APIs are highly variable due to tests being performed using different species and in different laboratories. This makes the development of high quality, robust QSAR models challenging. However, the work described herein has demonstrated that it is possible to extract the lowest level of observed toxicity from the data and thus empirically derive baseline toxicity QSARs for the minimum levels of acute and chronic toxicity. Issues with data quality and usability were identified for chronic toxicity such as the use of NOEC values that are equal to the maximum test concentration in the study designs. While it was possible to identify specific groups of APIs likely to show toxicity above the baseline toxicity QSARs and those predicted to show a baseline level of toxicity, more work is required to successfully categorize all APIs.

Acknowledgments

The financial contribution of the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution through the European Union Innovative Medicines Initiative (IMI) iPiE Project (Grant Agreement no. 115735) is gratefully acknowledged.

References

- European Medicines Agency (EMEA, 2006): Guideline on the environmental risk assessment of medicinal products for human use. Doc. Ref. EMEA/CHMP/SWP/4447/00 corr 2
- OECD (2018) Test no. 201: freshwater alga and cyanobacteria, growth inhibition test. Available online at http://www.oecd-ilibrary.org/environment/test-no-201-alga-growth-inhibition-test_9789264069923-en;jsessionid=3he2xatcu4u0i.x-oecd-live-03
- OECD (2004) Test No. 202: *Daphnia sp.* Acute immobilisation test. Available online at http://www.oecd-ilibrary.org/environment/test-no-202-daphnia-sp-acute-immobilisation-test_9789264069947-en. Accessed 17 Oct 2017
- OECD (1992) Test No. 203: fish, acute toxicity test. Available online at http://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en. Accessed 17 Oct 2017
- OECD (2018) Test no. 210: fish, early-life stage toxicity test, growth inhibition test. Available online at http://www.oecd-ilibrary.org/environment/test-no-210-fish-early-life-stage-toxicity-test_9789264203785-en;jsessionid=3he2xatcu4u0i.x-oecd-live-03
- OECD (2018) Test no. 211: *Daphnia magna* reproduction test. Available online at http://www.oecd-ilibrary.org/environment/test-no-211-daphnia-magna-reproduction-test_9789264185203-en;jsessionid=3he2xatcu4u0i.x-oecd-live-03
- Vestel J, Caldwell DJ, Constantine L, D'Arco VJ, Davidson T, Dolan DG et al (2016) Use of acute and chronic ecotoxicity data in

- environmental risk assessment of pharmaceuticals. Environ Toxicol Chem 35:1201–1212
8. ECETOC Technical Report No. 120: activity-based relationships for aquatic ecotoxicology data: use of the activity approach to strengthen MoA predictions (2013)
 9. Crane M, Watts C, Boucard T (2006) Chronic aquatic environmental risks from exposure to human pharmaceuticals. Sci Total Environ 367:23–41
 10. Cronin MTD, Dearden JC, Dobbs AJ (1991) QSAR studies of comparative toxicity in aquatic organisms. Sci Total Environ 109:431–439
 11. Könemann H (1981) Quantitative structure-activity-relationships in fish toxicity studies. 1. Relationship for 50 industrial pollutants. Toxicology 19:209–221
 12. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA (1997) Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). Environ Toxicol Chem 16:948–967
 13. Austin T, Denoyelle M, Chaudry A, Stradling S, Eadsforth C (2015) European chemicals agency dossier submissions as an experimental data source: refinement of a fish toxicity model for predicting acute LC50 values. Environ Toxicol Chem 34:369–378
 14. Webb SF (2004) A data-based perspective on the environmental risk assessment of human pharmaceuticals I - collation of available ecotoxicity data. In: Kummerer K (ed) Pharmaceuticals in the environment: Sources, fate, effects and risks, 2nd edn. Springer, Berlin, pp 317–342
 15. Kar S, Roy K (2010) First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals. Chemosphere 81:738–747
 16. Tugcu G, Turker Sacan M, Vracko M, Novic M, Minovski N (2012) QSTR modelling of the acute toxicity of pharmaceuticals to fish. SAR QSAR Environ Res 23:297–310
 17. Sanderson H, Thomsen M (2007) Ecotoxicological quantitative structure-activity relationships for pharmaceuticals. Bull Environ Contam Toxicol 79:331–335
 18. Escher BI, Baumgartner R, Koller M, Treyer K, Lienert J, McArdell CS (2011) Environmental toxicology and risk assessment of pharmaceuticals from hospital wastewater. Water Res 45:75–92
 19. Escher BI, Bramaz N, Mueller JF, Quayle P, Rutishauser S, Vermeirissen ELM (2008) Toxic equivalent concentrations (TEQs) for baseline toxicity and specific modes of action as a tool to improve interpretation of ecotoxicity testing of environmental samples. J Environ Monit 10:612–621
 20. Escher BI, Eggen RIL, Schreiber U, Schreiber Z, Vye E, Wisner B et al (2002) Baseline toxicity (narcosis) of organic chemicals determined by *in vitro* membrane potential measurements in energy-transducing membranes. Environ Sci Technol 36:1971–1979
 21. Escher BI, Hermens JLM (2002) Modes of action in ecotoxicology: their role in body burdens, species sensitivity, QSARs, and mixture effects. Environ Sci Technol 36 (20):4201–4217
 22. Verhaar HJM, van Leeuwen CJ, Hermens JLM (1992) Classifying environmental-pollutants. 1. Structure-activity-relationships for prediction of aquatic toxicity. Chemosphere 25:471–491
 23. Thomas P, Dawick J, Lampi M, Lemaire P, Presow S, van Egmond R et al (2015) Application of the activity framework for assessing aquatic ecotoxicology data for organic chemicals. Environ Sci Technol 49:12289–12296
 24. Ellison CM, Madden JC, Cronin MTD, Enoch SJ (2015) Investigation of the Verhaar scheme for predicting acute aquatic toxicity: improving predictions obtained from Toxtree ver. 2.6. Chemosphere 139:146–154
 25. Ellison CM, Piechota P, Madden JC, Enoch SJ, Cronin MT (2016) Adverse outcome pathway (AOP) informed modeling of aquatic toxicology: QSARs, read-across, and interspecies verification of modes of action. Environ Sci Technol 50:3995–4007
 26. Enoch SJ, Hewitt M, Cronin MTD, Azam S, Madden JC (2008) Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree. Chemosphere 73:243–248
 27. ECETOC Technical Report No. 102: Intelligent testing strategies in ecotoxicology: mode of action approach for specifically acting chemicals (2007)
 28. He J, Fu L, Wang Y, Li JJ, Wang XH, Su LM et al (2014) Investigation on baseline toxicity to rats based on aliphatic compounds and comparison with toxicity to fish: effect of exposure routes on toxicity. Regul Toxicol Pharmacol 70:98–106
 29. Su LM, Liu X, Wang Y, Li JJ, Wang XH, Sheng LX et al (2014) The discrimination of excess toxicity from baseline effect: effect of bioconcentration. Sci Total Environ 484:137–145
 30. Sanderson H, Thomsen M (2009) Comparative analysis of pharmaceuticals versus

- industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicol Lett* 187:84–93
31. Brausch JM, Connors KA, Brooks BW, Rand GM (2012) Human pharmaceuticals in the aquatic environment: a review of recent toxicological studies and considerations for toxicity testing. In: Whitacre DM (ed) *Reviews of environmental contamination and toxicology* 218. Springer, pp 1–99
32. Hrovat M, Segner H, Jeram S (2009) Variability of in vivo fish acute toxicity data. *Regulat Toxicol Pharmacol* 54:294–300
33. ACD/Structure Elucidator, version 15.01, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com (2015)
34. Hsieh SH, Hsu CH, Tsai DY, Chen CY (2006) Quantitative structure-activity relationships (QSAR) for toxicity of nonpolar narcotic chemicals to *Pseudokirchneriella subcapitata*. *Environ Toxicol Chem* 25:2920–2926
35. Tsai KP, Chen CY (2007) An algal toxicity database of organic toxicants derived by a closed-system technique. *Environ Toxicol Chem* 26:1931–1939
36. Zhang X, Qin W, He J, Wen Y, Su L, Sheng L et al (2013) Discrimination of excess toxicity from narcotic effect: comparison of toxicity of class-based organic chemicals to *Daphnia magna* and *Tetrahymena pyriformis*. *Chemosphere* 93:397–407



Chapter 16

Ecotoxicological QSARs of Personal Care Products and Biocides

Kabiruddin Khan, Hans Sanderson, and Kunal Roy

Abstract

The personal care products (PCPs) constitute various nonmedical products intended only for the application on the body surface and are not used to treat internal body problems like infections, etc. With a continuous change in culture and lifestyle in the society, the consumption of PCPs has increased several fold. In contrast, biocides are any chemical substances administered individually or in mixture with the intention of “destroying, deterring, rendering harmless, preventing the action of, or otherwise exerting a controlling effect on, any harmful organism by any means other than mere physical or mechanical action.” The exponential rises in domestic application of PCPs and biocides have rendered them to be potential causes of environmental pollution. Their continuous detection in river bodies mainly due to improper treatment and uncontrolled release via sewage treatment plants has proven to be a leading cause of harm to ecological species. Some of them have been proved to have potential to become contaminants of emerging concern (CEC). Insufficient ecotoxicological data of PCPs for their environmental behavior and ecotoxicity have rendered Scientific Committee on Consumer Safety (SCCS) administered by the Directorate-General for Health and Consumer Protection of the European Commission to release guidelines pertaining to safer use and risk associated with it. On the other hand, Biocidal Products Regulation (BPR) EU 528/2012 was enacted to improve functioning of the biocide market and to ensure a high level of protection of human and animal health and the environment. In silico tools such as quantitative structure-activity relationship (QSAR) and read-across can be employed using existing information to rapidly identify the potentially most toxic and hazardous toxic PCPs/biocides and prioritize the most environmentally hazardous ones. QSAR is widely used to obtain predictions of known/untested or not yet synthesized chemicals in order to prioritize them as various toxic classes of potential hazard causing ingredients. The present chapter enlists the information related to impact and occurrence of PCPs/biocides along with their persistence, environmental fate, risk assessment, and risk management. Additionally, a special emphasis is given on in silico tools such as QSAR which can be employed in prediction of environmental fate of personal care products and biocides mainly related to the ecotoxicity to aquatic species. Finally, a detailed report is prepared on endpoints, ecotoxicity databases, and expert systems frequently used for ecotoxicity predictions of personal care products and biocides with the aim to justify the development and implementation of in silico tools in early risk assessment and reduction of animal experimentation.

Key words Biocides, Ecotoxicity, CEC, In silico, PCPs, QSAR, Risk assessment, Risk management, Waste management

1 Introduction

1.1 Personal Care Products

Personal care products (PCPs) constitute a broad class of chemicals including perfumes, detergents, disinfectants, sunscreen, deodorants, soaps and sprays, etc. A great concern has been raised by the ecotoxicologists with the detection of higher PCP concentration in surface water, soil, and flowing water bodies. The major source of their accumulation could be attributed to human use, sewage treatment plants, and direct discharge into canals/river from industries, etc. Though the effects of these ingredients are yet to be found in detail, several reports of persistence and bioaccumulation to sufficient extent have been reported. The bioaccumulation characteristics of molecules render them acute, sublethal, or in many cases chronic effects on living organisms. The reported range of accumulation varies according to the nature of the body of the organism. Higher concentrations of these substances have been reported from sea water, sewage, and wastewater treatment plants, whereas in groundwater and drinking water, the reported concentrations were less. The reported concentrations of PCPs in surface water range from $\mu\text{g/L}$ to 0.01 ng/L. Sufficient amounts of synthetic musk are reported from sewage treatment plants, thus making them a potential candidate for “biopersistance.” Another major constituent of PCPs having potential to behave like strong persistent, bioaccumulative and toxic (PBT) candidates [1] include UV filters and stabilizers. These are reported in mussels, fishes, crustaceans, and dolphins with the concentration range of no less than 0.001 ng/L. The major reason for concern related to PCP toxicity is attributed to their bioaccumulative and endocrine disrupting nature [2]. A number of PCPs behave like endocrine disruptors which include phthalates (butyl benzyl phthalate, di-(2-ethylhexyl) phthalate), alkyl phenols (octyl and nonylphenol), parabens (methyl, ethyl, propyl, and butyl), dioxins and furans, bisphenols, polychlorinated biphenyls, etc. Polycyclic musks and UV filters are proved to have higher potential for bioaccumulation mainly due to their lipophilic nature, thus making them potential candidate for PBT-like substances. There are reports on adverse effect on reproduction of benthic organisms caused by UV filters. Synthetic musks have been shown to bioaccumulate mainly in aquatic organisms like mussels, fish, and mammals. Synthetic fragrances like tonalide and galaxolide are reported in water bodies and are found to induce oxidative and genetic damage in zebra fish. The endocrine disruption properties have been reported by UV filters, parabens, and polycyclic musks. Another major contaminant reported is triclosan, said to cause disturbance in metabolic pathways of *P. subcapitata*. It is also found that algal species is most sensitive toward triclosan and other disinfectants.

1.2 Biocides

Biocides are used either as a single compound or in a mixture “with the intention of destroying, deterring, rendering harmless, preventing the action of, or otherwise exerting a controlling effect on, any harmful organism by any means other than mere physical or mechanical action” (BPR, Regulation (EU) 528/2012) [3]. The BPR regulation ((EU) 528/2012)) [3] regulates the biocidal products in Europe. The regulation works with the intention of improving the functioning of the biocidal market, to keep a check on adverse effects of biocides on human and animal health along with the environmental effects; thus, it necessitates to get approval for a new product before being implemented for use. The pivotal requirement for biocide registrations includes safety, efficacy and toxicity, analytical procedures for detection and identification, and control of metabolites along with degradation products with emphasis to ecotoxicological studies. As per regulation, the products of biocides are divided into four categories which are further subdivided into 22 product types (the Annex V of the regulation). Group 1 contains disinfectants and algaecides (may or may not be used on humans directly), veterinary products, food or feed area, etc. Group 2 mainly consists of preservatives employed in product storage, wood decay, leather/rubber and preservatives for polymeric materials, etc. The 3rd group contains products like avicides, rodenticides, vermicides, and molluscicides, which are used to control invertebrates; insecticides, acaricides, and piscicides used to control arthropods; and finally attractants and repellants for controlling various other vertebrates. The last group consists of antifouling products along with taxidermist fluids and embalming agents. The major concern arising due to widespread use of biocides includes its increased demand, propensity with which they accumulate and contaminate environment, and ability to cause cancer and structural diversity with reference to functional group present in it [4, 5].

The gradual rise in the use of PCPs and biocides is expected to intensify many folds (expected to garner \$429.8 billion by 2022 [6, 7]). Such increase in consumption of PCPs and biocides would propel scientists for early risk assessment using a huge number of experimental data, extensive time, huge cost, and extensive animal testing for *in vivo* testing. Unfortunately, the available data on such compounds are very much limited. Currently available data is limited to certain species along with specific environmental conditions. In silico tools mainly quantitative structure–activity/toxicity relationship (QSAR/QSTR) can help in filling the data gap effectively. In QSAR, a small amount of experimental data can be employed to get predictions for a large dataset without having any experimental response values provided they fall within the domain of the model. Such models have been widely employed in hazard assessment of various organic groups of chemicals like PCPs, pharmaceuticals, biocides, endocrine disruptor chemicals and agrochemicals, etc.

[8–12]. The use of QSAR in early risk assessment is recommended by various regulatory agencies like European Centre for the Validation of Alternative Methods (ECVAM), United States Environmental Protection Agency (US EPA), the Agency for Toxic Substances and Disease Registry (ATSDR) and the European Union Commission’s Scientific Committee on Toxicity, Ecotoxicity, and Environment (CSTEE), etc. [10]. Nearly 535 PCPs were screened and prioritized by Gramatica et al. (2016) [13] based on their acute predicted response. Similar ranking of nearly 600 cosmetics employing QSAR technique was achieved by Khan et al. (2017) [14]. In recent years, a number of easily available online expert software have been developed for toxicity predictions of organic chemicals such as ECOSAR [15, 16]. Very limited number of QSAR studies have been performed with these categories of chemicals so far, and still a large number of compounds need to be prioritized based on their acute predicted responses. There is a need to pay a special attention from the ecotoxicity point of view on PCPs and biocides due to their increased consumption and accumulation in the environment. There is a significant lack of knowledge about the environmental fate of a huge number of PCPs/biocides and their metabolites. This necessitates the development of QSAR models to study ecotoxicological behavior, mainly environmental fate, persistence, and toxicity of PCPs and biocides.

2 A General Overview of PCPs and Biocides

2.1 Types of Different PCP and Biocidal Formulations

PCPs and biocides have become indispensable parts of our day-to-day life mainly in urban population. Since PCPs are down-the-drain use products, they are released into the environment; the detected quantities are found mainly in sewage and effluent treatment plants and in sludge. Adverse effects from cosmetic formulations have become a significant concern in the European Union (EU). The EU has assigned poison centers with the intentions to receive information related to compositions of hazardous mixtures present in cosmetics [17]. Some of the hazardous components reported include detergents, paints, adhesive, etc. The center helps in identification of probable cause of poisoning reported by the respective European country, consulted physician, professional users, and consumers. The inventory of cosmetic ingredient database gives the repository of formulations used in/as cosmetic formulation along with probable adverse effects in case of improper utilization [18]. As per the data published on the ECHA website (European Chemical Agency), the poison center receives 600,000 calls per annum which makes it about 1700 calls per day. Most of these cases are of accidental exposure mainly involving children, whereas fatalities due to poisoning are approximated at 400 per year. Similarly, BPR Regulation (EU) 528/2012 [3] keeps controls on biocidal

product in the European Union to ensure safety of humans, animals, articles, or materials in the context of biocidal toxicity. The European Chemicals Agency is again responsible for assessing hazard profile of new biocidal products under application for approval. Both PCPs and biocides (in many cases) are available in multicomponent mixtures. The chemical constituents get transformed as they pass from one compartment to another in the surrounding and eventually metabolized by the microflora. The chemical components present in PCPs and biocides constitute a very diverse chemical range thus putting a major challenge in ecotoxicological evaluation and risk assessment. The Organisation for Economic Co-operation and Development (OECD), US EPA, and International Organization for Standardization (ISO) provide standard protocols or tests following which the active chemicals or their metabolites can be evaluated for their acute toxicity mainly in algae, zooplankton, fish, and other invertebrates. The list of various formulations of the biocide/PCP category and their reported toxicities will be discussed here.

- *PCPs (cosmetics):* The list of ingredients employed in cosmetic formulations is very large, such as anti-dandruff (helps control dandruff), detangling agents (reduce or eliminate hair intertwining and help combing), depilatory (removes unwanted body hair), anticaking (allows free flow of solid particles and thus avoids agglomeration), anticorrosive (prevents corrosion), binding (provides cohesion), emulsifying (promotes the formation of intimate mixtures of nonmiscible liquids), bulking agents (reduce bulk density), chelating (reacts and forms complexes with metal ions could affect the stability and appearance of cosmetics), denaturant (renders cosmetics unpalatable, mostly ethyl alcohol containing products), anti-seborrheic (controls sebum production), buffering agent (stabilizes the pH), anti-static (reduces static electricity), film forming (produces a continuous film on the skin, hair, or nails), foaming agents (trap numerous small bubbles of air or other gas), foam booster (improves the quality of the foam produced by a system by increasing volume, texture, and stability), gel former, hair conditioner (leaves the hair easy to comb and makes them shiny, glossy, etc.), hair dyeing (colors hair), hair fixing (permits physical control of hairstyle), hair waving (sets hair in the style required), humectant (holds and retains the moisture), keratolytic (helps eliminate the dead cells of the stratum corneum), masking (reduces or inhibits the basic odor/taste of the product), moisturizing (increases the water content of the skin), nail conditioner, opacifying agent (reduces transparency or translucency of cosmetics), oral care (provides cosmetic effects to the oral cavity like cleansing, deodorizing, protecting), oxidizing agent (changes the chemical nature of another substance by

adding oxygen or removing hydrogen thus changing the chemical nature), pearlescent (imparts a nacreous appearance), UV absorber (protects the cosmetic product from the effects of UV light), UV filter (filters certain UV rays in order to protect the skin or the hair from harmful effects of these rays), and viscosity agents (increase or decrease the viscosity of cosmetics) [18].

Among the cosmetics, the most toxic products identified are dehydroacetate (a preservative) [14], surface active agents (such as quaternary ammonium compounds like benzylidemethyldodecylammonium chloride, decyltrimethylammonium bromide, didecyldimethylammonium chloride, hexadecyltrimethylammonium chloride, etc.) [19], UV filters (like Fluorescent Brightener 367) [13], polycyclic musks (like Musk 36A) [13], UV sun screeners (like UV-320) [13], synthetic musks (like Musk xylene and musk ketone) [20], synthetic fragrances (like galaxolide and tonalide) [13], parabens and phthalates, etc. [13, 17, 20, 21].

- *Biocides:* The biocidal products approved for use include insecticides, acaricides and products to control other arthropods, repellents and attractants, piscicides, fiber leather rubber and polymerized materials preservatives, slimicides, preservatives for products during storage, working or cutting fluid preservatives, wood preservatives, disinfectants and algaecides not intended for direct application to humans or animals, preservatives for liquid, construction material preservatives, veterinary hygiene, film preservatives, food and feed area, human hygiene, antifouling products, embalming and taxidermist fluids, rodenticides, and avicides, etc. [3].

The most toxic biocidal products reported include antifouling (Irgarol, tributyltins (TBTs), etc.) [3, 22]; anticorrosive agents (quaternary ammonium, phosphonium salts, isothiazoline, and heavy metal salts) [3]; film preservatives (Diuron, 2-octyl-2H-isothiazol-3-one (OIT), etc.); several insecticides (like Triflumuron and Phenothrin), acaricides (like D-Tetramethrin, Triflumuron), and piscicides (Rotenone); etc. [3].

2.2 Source of Release into the Environment

The understanding of toxicity to the environment caused by PCPs and biocides is incomplete without identifying the source and extent of release. Some of the most common sources will be discussed here.

- *Domestic disposal:* The household discharge contributes to a greater extent to ecotoxicity of PCPs, since PCPs as the final product are intended to be used for enhancement of personal appeal. Many of the cosmetic products used regularly are disposed in the surroundings; the major problem lies in absence of

any proper regulation for the disposal of cosmetic products. Since they are considered to be safe enough, they are generally dumped in the garbage with no regulation. In contrary, biocides are dealt with more caution while releasing in the environment. However, still many developing countries lack the regulation for the proper and safe release of biocides into the environment. The lack of regulations on disposal renders a continuous accumulation in landfills via dust bins or toilets thus resulting in terrestrial ecosystem potential risk [23].

- *Discharge through industries:* Industrial discharge accounts to a greater extent for ecotoxicity of PCPs and biocides; there is a significant amount of disposal at each stage of the manufacturing mainly in process quality control. Though the good manufacturing practices (GMP) are to be followed in manufacturing, a good number of cases have been reported for its violation mainly from the developing countries like India and China. Concentrations up to ng/L of UV stabilizers have been reported from the Kaveri, Vellar, and Thamirabarani rivers of Tamil Nadu, India [24]. The major source of river contamination in India is attributed mainly to direct industrial discharge from units situated in close proximity of the river [25]. Biocides are mainly used in food industry as disinfectants in the United Kingdom and are responsible for resistance developed against the bacterial species they are targeted [26].
- *Discharge through hospitals:* Though the release of PCPs through hospitals in the developing countries is limited, still it accounts for a large quantity since the discharge is continuous. Biocides, mainly disinfectants and antifouling agents, are widely used in hospitals and health-care facilities to control the microbial growth, wherein unspent or expired products are exposed to the environment on the daily basis, thus accumulating in surrounding [27, 28].
- *Competition for development of better products:* The competition among industries to provide more effective cosmetics has increased manifold with the progress of modern science. The sale of products mainly relies on product appeal, whereas products which fail to appeal customers end up into environment by the time of their expiry. Some of the good examples include competition in the market to deliver a better shampoo or oil brand. Not only the release of the products, this competition has also created a rift to manufacture more and more such products leading to an exponential rise in accumulation of such chemicals in the environment.
- *Other sources of release:* Other sources of release include products used on animals, plant, and inanimate objects with the intention to decontaminate them or for appeal purposes.

2.3 Risk Assessments of PCPs and Biocides

The authorization of any novel cosmetics and biocides needs to undergo several regulatory procedures and protocol in order to get approval for domestic or for commercial application for a limited period of time.

- **PCPs:** In the European Union, “Notes of Guidance for Testing of Cosmetic Ingredients and Their Safety Evaluation by the SCCS (Scientific Committee on Consumer Safety)” contains relevant information on the different aspects of testing and safety evaluation of cosmetic substances [29]. The guidance is designed to provide assistance to public authorities and to the cosmetic industry in order to improve harmonized compliance with the current cosmetic EU legislation. In 2009 legislative recast, the cosmetic Directive 76/768/EEC [30] was transformed into a regulation, and since 11 July 2013 onward, this Regulation (2009/1223/EC) was fully applicable. This legislation prohibits the marketing of finished products containing ingredients or combinations of ingredients that have been subject to animal testing after 2013, thus supporting the progress made with regard to the development and validation of alternative methods [31]. The safety evaluation or risk assessment goes through two different channels; the final product is examined by the commission as well as the industry for the consumer protection following strict written safety evaluation procedures. The risk assessment procedure is subdivided in the following four parts:
 - *Human health hazard assessment:* It is carried out in order to identify toxicological properties of the substance or its potential to damage human health. It is based on results of in vivo tests, in vitro tests, clinical studies, case reports, epidemiological studies, etc.
 - *Dose-response assessment:* Here, the relationship between the exposure and the toxic response is evaluated. The concentrations such as no adverse effects observed (NOAEL), lowest dose at which an adverse effect is observed (LOAEL), and dose without any effect observed (NOEL) play a crucial role in dose-response assessment.
 - *Exposure assessment:* Here, the amount and frequency of human exposure are determined (specific groups at potential risk, e.g., pregnant women, children, etc.).
 - *Risk characterization:* Margin of exposure (MoE) is mostly calculated for oral toxicity studies and in some cases from a dermal toxicity study. Equation 1 is used for the oral toxicity study.

$$\text{MoE} = \text{NOAEL}_{\text{sys}} / \text{SED} \quad (1)$$

where $\text{NOAEL}_{\text{sys}}$ is the dose descriptor for the systemic exposure and SED represents the Systemic Exposure Dose.

- *Biocides:* An approval is necessary for the new active constituents before authorization of a biocidal product can be granted. The active substances are first assessed by an evaluating Member State competent authority, and the results of these evaluations are forwarded to the ECHA's Biocidal Products Committee, which prepares an opinion within 270 days. The opinion serves as the basis for the decision on approval which is adopted by the European Commission. The approval of an active substance is granted for a defined number of years, not exceeding 10 years and is renewable. It also includes exclusion and substitution criteria for evaluation of active constituents. The products meeting the exclusion criteria will eventually not be approved for the use. The exclusion criteria include a variety of substances such as carcinogens; mutagens and reprotoxic substances; endocrine disruptors; persistent, bioaccumulative, and toxic (PBT) substances; and very persistent and very bioaccumulative (vPvB) substances. Possible derogations are foreseen for the products needed on the grounds of public health or of public interest when no alternatives are available. However, for the derogatory products, the approval is only granted for 5 years. The other group includes active substances which need substitution owing to their toxicological features directly affecting public health or the environment with an objective to replace or phase out with more suitable alternatives over time. The intrinsic hazardous properties in combination with the use will be considered as a candidate for substitution only if the active substances have at least one quality of exclusion criteria. Some of the adverse properties of molecules making it a potential candidate for substitution include respiratory sensitization, higher toxicity threshold than already approved product for same use, a potent PBT substance, high-risk chemical for humans as well as environment, and product having significant amount of non-active isomers or impurities.

2.4 Risk Management of PCPs and Biocides

Since there is a lack of protocol on disposal of these products (mainly in the developing countries), risk management becomes the sole criterion to put a limit on the release of PCPs/biocides into the environment. The international community defines risk management as “the process of identifying, evaluating, selecting, and implementing actions to reduce risk to human health and to ecosystems.” The goal of risk management is “scientifically sound, cost-effective, integrated actions that reduce or prevent risks while taking into account social, cultural, ethical, political, and legal considerations” [32–34]. Here, we will discuss some approaches

which can help reduce release of PCPs and biocides in the environment.

- *Market Surveillance*

For the ease of management of adverse effects, the European countries have set up the Platform of European Market Surveillance Authorities for Cosmetics (PEMSAC) in order to ensure a coherent approach to consumer product issues [29, 30, 35]. The PEMSAC members meet twice a year to discuss about the market surveillance analytical methods. The PEMSAC aims to facilitate following operations:

1. Coordinating activities
2. Exchanging information
3. Developing and implementing joint projects
4. Exchanging expertise and best practices in cosmetics market surveillance

On the other hand, for biocides, the European Chemical Agency (ECHA) under the Committee for Risk Assessment (RAC) has setup market surveillance program in order to check for the compliance of laid rules, and it is undertaken as part of member state existing monitoring programs.

- *Awareness and Training*

One of the major crucial steps in controlling the effects of PCPs/biocides toward the environment is by making people aware about the consequences of exposure and providing them with sufficient training in order to fight with consequences or to avoid them [36]. The knowledge about disposal of various products included in PCPs/biocides is the first step in the way of reducing the input of those hazards into the ecosystem. This knowledge can be made available to users, stakeholders, and community by awareness programs or by notifying on the product labels. The industry needs to be very much responsible and alert as they are the major sources of potential hazards of these products, and most of the ingredients (or APIs) are released into the environment without adequate waste treatment. Additionally, each raw material in the process of product development should have a safety data sheet (MSDS) intended to provide employees and emergency personnel with information on processes for handling such products safely; the information could include physical data, toxicity and health first aid, hazards, storage, reactivity, protective equipment, disposal, and spill-handling procedures. The related people who are involved in risk management should possess information mainly about the product flows from sources of industries, households, and retail shops.

3 Application of QSAR in Ecotoxicological Analysis

So far we have discussed about the products, release, regulatory requirements, and risk assessment of PCPs and biocides. Here, we will discuss about the already developed successful QSAR models for ecotoxicity of PCPs and biocides. QSAR works on the principle of correlating structural features (theoretical descriptors) with the studied response (toxicity or activity). In the current chapter, we have tried to explore the all possible modeling work done on ecotoxicity of PCPs and biocides. To achieve this, a list of all publications (to our knowledge) related to ecotoxicity of PCPs and biocides was obtained through SCOPUS search engine [37]. The obtained papers were thoroughly analyzed in order to obtain the following points, namely, (1) source of data collection, (2) method employed, (3) software used, (4) and objective of the study. By following this approach, we have been able to report a number of data sources of cosmetics and biocides including databases, literature, etc. We have also encountered a list of various software tools which are used in QSAR modeling which will be discussed below.

3.1 Application of QSAR in Ecotoxicity of PCPs

The screened papers were segregated into two separate groups: the first group contains the QSAR models developed employing a mixture of different products (e.g., a mixture of two or more different types of PCPs/biocides called as the global approach), while the second group consists of papers dealing with selective class of QSAR models developed by taking one particular class of chemicals like group of surfactants (local QSAR approach).

- *Global QSAR Models on PCPs*

1. Papa et al. [38] developed five QSTR models by using 1105 heterogeneous organic compounds including pharmaceuticals and personal care products for the half-life predictions in humans. Two-dimensional descriptors generated from PaDEL-descriptor software were employed in model development, whereas the models were validated using metrics like R^2 , Q^2 (LOO/LMO) for internal validation, while the test set was evaluated using the Q^2_{F1} metric [39]. Additionally, to investigate the stability of the models in prediction or error of prediction, the residual mean squared errors (RMSE) were also calculated. The developed models were applied to predict the potential *in vivo* half-life of nearly 1300 PCPs within the applicability domain of the model. The results obtained in this work were utilized to assess PBT nature of the chemicals and to hypothesize chemicals of highest concern [38]. The modeling study was performed using QSARINS software.

2. Gramatica et al. (2018) [40] in an article demonstrated the use of combining principal component analysis (PCA) with predictive toxicological approach in order to provide an insight into screening, ranking, and predicting PBT nature of PCPs along with pharmaceuticals. The data were collected from literatures, whereas PCA along with PBT model implemented in QSARINS were used for screening and ranking of employed dataset.
3. The cytotoxicity of personal care products along with pharmaceuticals on the rainbow trout (*O. mykiss*) species and on RTL-W1 liver cell line was analyzed using 2D QSAR approach [41]. The models were developed using genetic algorithm followed by ordinary least squares approach, and the descriptors were calculated from SPARTAN [42] and Dragon software [43]. The model validation was achieved by R^2 , Q^2 , and R^2_{ext} ; additionally, Golbraikh and Tropsha criteria were checked in order to supplement the finding of the results. The data for QSAR modeling were collected from the literature, and model development was achieved using QSARINS software [44].
4. Gramatica and colleagues [13] developed highly robust models to predict acute toxicity of PCPs in three key organisms, namely, algae, crustacean, and fish. The models were developed following the strict OECD principles for the validation of QSARs. The models were developed based on ordinary least squares approach, while the descriptors were implemented through PaDEL-Descriptor software. The QSAR models were applied to predict acute toxicity for over 500 PCPs without experimental data; a trend of acute aquatic toxicity was highlighted by PCA allowing the ranking of inherently more toxic compounds, using a multi-criteria decision making approach for prioritization purposes. Finally, a QSAR model for the prediction of this aquatic toxicity index (ATI) was proposed to be applicable for the a priori chemical design of non-environmentally hazardous PCPs. The model predicted toxicity was compared with the ECOSAR [15] tool for the error analysis. Following the in silico QSAR approach, a total of 66 chemicals related to PCP ingredients (mainly UV filters and phthalates) were selected for inclusion into the final priority list for further more definitive evaluation, focusing on their necessary experimental tests. In this way, cost, time, and animal sacrifice can be reduced. The dataset was collected mainly from the ECOTOX database (available at <https://cfpub.epa.gov/ecotox/>), and some compounds were taken from other literature. The process of model development, ranking, and prioritization was accomplished using the QSARINS tool [44].

5. A collection of 534 heterogeneous personal care products was screened for their PBT properties by applying different tools, namely, the Insubria PBT Index, QSAR model included in the QSARINS software, and the US-EPA PBT Profiler by Cassani and Gramatica 2015 [45]. Finally, a priority list of the potentially most hazardous PCPs was proposed as identified by the QSAR model. This study also showed that the PBT Index could be a valid tool to evaluate more environmentally sustainable chemicals immediately from the molecular structures, thus avoiding unnecessary synthesis and expensive tests.
6. de Garcia and colleagues [46] classified a mixture of PCPs and pharmaceuticals based on their ecotoxicity values obtained by bioluminescence and respirometry assays along with predictions obtained by US EPA ecological structure–activity relationship (ECOSAR). The data classification was achieved by Globally Harmonized System of Classification and Labelling of Chemicals. Finally, as per the European Medicines Agency, the real risk of impact of these compounds in wastewater treatment plants (WWTPs) and in the aquatic environment was predicted. The global order of the species' sensitivity to the PPCPs considered was as follows: *A. fischeri* > algae > crustaceans > fish > biomass of WWTP. The compound used in this study was procured from Sigma-Aldrich and Fluka Chemicals (with purity $\geq 95\%$).
7. Toropova and Toropov 2018 [47] demonstrated the application of CORAL software in ecotoxicological analysis of organic pollutants containing personal care products employing the QSAR technique. The index of ideality of correlation (IIC) was suggested as a criterion of predictive potential of QSAR. The data implemented in this work were taken from literature.
8. An extensive scaling and prioritization study on a large dataset of 15,145 organic pollutants (containing PCPs and other ingredients) was performed by Edwin John Matthews in 2019 [48]. For the in silico predictions of chemical disposition (CD) (intestinal absorption, membrane permeability, distribution, sequestration, toxicokinetics parameters) and chemical toxicity (CT) (genetic, carcinogenicity, developmental, teratology), Percepta [49] and ADMET Predictor [30] software employing QSAR models were used. The results of the study demonstrated that chemicals with different purposes and colorants with different chemical classes had markedly different profiles of toxicological activities. The methodology was proposed to be capable of evaluating smaller sets of chemicals on a

routine basis to identify and detect potential CD and CT signals (or emerging chemical hazards) in food and cosmetic ingredients. The data implemented in this work were taken from several database like CenterWatch [50], Drugs@FDA [51], Good Scents Company database [52], CFSAN Thesaurus [53], FDA EAFUS List [54], Health Canada [55], European Food Safety Authority (EFSA) [56], GRAS Notice Inventory [57], Fragrance Products Information Network (FPINVA) [58], Environmental Protection Agency (EPA) [59], Registry of Toxic Effects of Chemical Substances (RTECS) database [30], FPIN_LP: Common Fragrance Chemicals in Laundry Products and Cleaners [60], GIVAUDAN and IFF: fragrance manufactures [61], International Cosmetic Ingredient Dictionary (ICID) and Handbooks [62], permanent and semipermanent hair dyes Internet database [63], CERES [64], Color of Art Pigment Database [65], ink dye stuff database [66], and Internet stains database [67].

9. Monika Batke and colleagues [68] categorized a large database with complex endpoints of toxicity using the read-across technique. The conceptual approach of read-across technique works on the principle of structural similarity with shared mechanism of action. They combined two databases on repeated dose toxicity, RepDose database [69], and ELINCS database [70] to form a common database for the identification of categories. For categorization of chemicals, the predictive clustering tree (PCT) approach [68] was adopted based on structural and on toxicological information to detect groups of chemicals with similar toxic profiles and pathways/mechanisms of toxicity. The read-cross was implemented to fill the data gaps, as many of the structures in the two databases lack experimental toxicity. Finally, they proposed improvements for a follow-up approach, such as incorporation of metabolic information and more detailed mechanistic information. The clustering was performed with a clustering tool provided as a free web service (accessible at <http://mlc-reach.informatik.uni-mainz.de>) [71].
10. An extensive in silico analysis of potential chemical-induced eye injury through irritation and corrosion caused by industrial, household, and cosmetic ingredient chemicals was performed by Verma and Matthews 2015 [72] by artificial neural network (ANN) c-QSAR (classification QSAR) approach [72]. They developed 21ANN c-QSAR models to predict eye irritation using the ADMET Predictor™ program [72] using a diverse data set of 2928 chemicals. The developed models could be used to fill the data gaps for

the safety assessment of cosmetic ingredient chemicals. The data for modeling was collected from literatures and several databases like material safety data sheets (MSDSs), FDA-approved drugs for ophthalmology (available at www.medilexicon.com/drugs-list/eyes.php), FDA-approved cosmetic colors in the eye area (available at www.fda.gov/forindustry/coloradditives/coloradditiveinventories/ucm115641.htm), Household Products Database (available at <http://hpdb.nlm.nih.gov>), Hazardous Substances Data Bank (available at <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>), and pharmaceutical drugs [73] from the Elsevier PharmaPen-dium database containing the preclinical, clinical, and post-marketed eye irritant adverse effects (available at www.pharmapendium.com).

11. Hisaki et al. (2015) [74] reported models for prediction of maximum no observed effect level (NOEL) for repeated-dose, developmental, and reproductive toxicities using 421 chemicals for repeated-dose toxicity, 315 for reproductive toxicity, and 156 for developmental toxicity collected from Japan Existing Chemical Database (JECDB) [74]. The artificial neural networks (ANN) were constructed to predict NOEL values, while descriptors were selected based on molecular orbital (MO) calculations. Model validations were achieved using root mean square (RMS) errors after tenfold cross-validation (0.529 for repeated dose, 0.508 for reproductive, and 0.558 for developmental toxicity). Commercially available TOPKAT software [75] was implemented in this work [74].
12. A group of 558 cosmetic ingredients randomly selected were analyzed for their mutagenicity, carcinogenicity, developmental toxicity, and skin sensitization by Plosnik et al. (2015) [76]. For the said analysis, models embedded in the CAESAR programs were utilized. The investigated data set was compiled from Inventory CosIng database (available at <http://ec.europa.eu/consumers/cosmetics/cosing/>). The CAESAR program provides experimental values for some of the compounds, whereas the rest of them were predicted correctly. For mutagenicity, the experimental data were known for 66 compounds, while for 6 of them, the predictions were wrong. For other three properties, only the limited number of experimental data was reported; however, the predictions were correct [76].
13. An extensive QSAR analysis on 596 cosmetics was carried out by Khan and Roy in 2017 [14] in order to prioritize the molecules of concern. They developed validated partial least squares QSAR models for three distinct species, namely,

P. subcapitata, *P. promelas*, and *D. magna*, following stringent OECD protocols for QSAR validation (a defined endpoint, unambiguous algorithm, sufficient statistics to indicate goodness of fit and predictivity, AD analysis, and mechanistic interpretation). The data for modeling was obtained from the ECOTOX database and published literatures, and the validation of the models was done using stringent validation criteria [77]. The study suggested a linear positive correlation of logP, size, and presence of sulfur with cosmetic toxicity against the studied endpoints. They have also compared their results with those of ECO-SAR, an online expert for toxicity prediction. Finally, they have prioritized molecules of concern following a scaling technique as described in Eq. 2.

$$\Upsilon_{\text{scaled}} = (\Upsilon_{\text{predicted}} - \Upsilon_{\text{obs_min}}) / (\Upsilon_{\text{obs_max}} - \Upsilon_{\text{obs_min}}) \quad (2)$$

- *Local QSAR Models on PCPs*
 - *Surfactants*

1. Surfactants, one of the major ingredients of various cosmetic products, are believed to possess several stabilizing factors for the final formulation. Nica and colleagues [19] analyzed five surfactants of quaternary ammonium compounds (QACs) category, namely, benzylidemethyl-dodecylammonium chloride, decyltrimethylammonium bromide, didecyldimethylammonium chloride, hexadecyltrimethylammonium chloride, and tetradecyltrimethylammonium bromide. They have successfully demonstrated the mode of action of these surfactants either alone or in a mixture form on *A. Fischeri* through the application of the QSAR models. They have also found out that only hexadecyltrimethyl ammonium chloride behaved as a polar narcotic, with a low reactivity toward the bacterial cell membrane. All the statistical analyses and the Ix (100*I%) values, with the related confidence intervals at 95%, were obtained using R® software [78, 79].

- *UV filters*

1. Jentzsch et al. [80] performed an in silico prediction to study transformation of ethylhexyl methoxycinnamate (EHMC) (a class of UV filters) and its environmental fate along with transformation products (TPs). Several software packages were used in that study like CASE Ultra V.1.5.2.0, MetaPC V.1.8.1 (both MultiCASE Inc.) [81]; the combined statistical and rule-based OASIS

Catalogic software V.5.11.6 TB from the Laboratory of Mathematical Chemistry, University Burgas, Bulgaria [5]; and the statistical QSAR Leadslope software (Version 3.2.6-1) [80]. Additional study was also performed to check for biodegradability of EHMC and their transformation products. The performed methodology was proved helpful in identifying potent TPs, and the application of multiple model prediction in getting good results was justified.

2. The toxicological effects of benzophenone (BP)-type UV filters on *D. magna* were determined by Liu et al. [82] using comparative molecular field analysis (CoMFA) and density functional theory (DFT). The major focus of this study was to understand the underlying mechanism of toxicity. Additionally, sensitivity of benzophenone against *Daphnia magna* and *Dugesia japonica* was studied employing interspecies correlation technique (QSTTR) [83]. The results demonstrated that the mechanism underlying the toxicity of BPs to *P. phosphoreum* is primarily related to their electronic properties, and the mechanism of toxicity to *D. magna* is hydrophobicity. It was also found that *D. magna* was more sensitive than *P. phosphoreum* to most of the BPs, with the exceptions of the polyhydric BPs [82].

– *Fragrances*

1. Papa et al. [84] proposed predictive MLR-QSAR models developed from fragrance data (79 compounds) against three toxicological endpoints, namely, oral LD₅₀, inhibition of NADH-oxidase (EC₅₀), and the effect on mitochondrial membrane potential (EC₅₀) in mouse. Theoretical molecular descriptors were calculated by using DRAGON software [43], and the best QSAR models were developed according to the principles defined by the Organization for Economic Co-operation and Development. The predictions obtained for all of the 79 compounds for these two endpoints have almost 80% correlation, which means that the same chemicals induce both inhibition of mitochondrial NADH-oxidase and depolarization of mitochondrial membrane. The actual fragrance dataset consisted of 39 compounds collected from the literature, while other fragrance materials of interest, such as nitro-musks, macrocyclic musks, as well as other terpenes and cinnamic acid derivatives, were included for predictive purposes in the study to give a total of 79 compounds [84].

– *Preservatives*

1. The dehydroacetic acid (DHA), a widely used preservative, was analyzed for its potential toxicities caused by its by-product using in silico QSAR approach following the bioassay conducted in vitro on *A. fischeri*. The result of the study stated that personal care products containing DHA must be protected from direct sunlight to prevent photodegradation [85]. The predicted toxicity values of by-products on *A. fischeri* were obtained from Toxicity Estimation Software Tool (TEST) [85].
2. A report on the development of physiologically based pharmacokinetic (PBPK) models for parabens mainly for methyl-, propyl-, and butylparaben on male and female Sprague-Dawley rats was reported [86]. The QSAR model was coupled with quantitative in vitro to in vivo extrapolation (IVIVE) study for hydrolysis in portals of entry including the intestine, skin, as well as liver. The finding of the models provided a very good agreement with the published time-course data in the blood and urine from controlled dosing studies in rat and human and demonstrates the potential value of quantitative IVIVE in expanding the use of human biomonitoring data in safety assessment.

– *Antioxidants*

1. The degradation products of α-tocopherol due to UV-visible rays in a cosmetic emulsion were predicted using Toxicity Estimation Software Tool (TEST) developed by the US Environmental Protection Agency. Toxicity evaluations were based on the physical characteristics of a chemical structure (molecular descriptors) [87].

– *Skin sensitizers*

1. A quantitative in silico QSAR model for predicting skin sensitization was reported using k-nearest neighbors approach with an in-house dataset of 1096 murine local lymph node (LLNA) studies [88]. The model predicts the Globally Harmonized System of Classification and Labelling of Chemicals skin sensitization category of compounds well, predicting 64% of chemicals in an external test set within the correct category. Of the remaining chemicals in the dataset, 25% were overpredicted, and 11% were underpredicted. Derek Nexus 5.0.2, an expert knowledge-based system for toxicity predictions (Lhasa), was used to make an in silico assessment of the skin sensitizing potential of the chemicals in the internal and external validation datasets.

3.2 Application of QSAR in Ecotoxicity of Biocides

The number of literatures showing application of QSAR in analyzing biocide toxicity is limited till date; very few people have tried to implement QSAR in ecotoxicity of biocides. In recent years, an increased attention is paid in studying environmental impact of biocides because of their cidal nature against living organisms.

- *Global QSAR models on biocides*

1. The very first comprehensive and promising application of QSAR on biocides is demonstrated by Khan et al. (2019) [89]. They had compiled biocide toxicity data on *Daphnia* and fish from various databases such as the OECD QSAR Toolbox v. 4.2 (available at www.qsartoolbox.org), Pesticide Properties Database (PPDB) database (available at <https://sitem.herts.ac.uk/aeru/ppdb/>), Office of Pesticide Programs (OPP) Pesticides Ecotoxicity Database (available at <http://www.ipmcenters.org/ecotox/>), European Food Safety Authority (EFSA) (<http://www.efsa.europa.eu/>) database, ECOTOX (available at <https://cfpub.epa.gov/ecotox/>) database, and the AMBIT (available at <http://cefic-lri.org/toolbox/ambit/>) database. They proposed robust QSAR models for biocides employing 133 data for immobilization on *Daphnia* and 88 data for mortality against fish following strict OECD (Organization for Economic Cooperation and Development) guidelines for QSAR validation. The findings of the paper stressed on linear dependency of toxicity of biocide on lipophilicity while an inverse dependence on polarity. The results of QSAR study also suggest that presence of nitrogen atoms increases fish toxicity and presence of sulfur/phosphate/thiophosphate moiety enhances *Daphnia* toxicity [89].
2. Rauert et al. (2014) [90] demonstrated the use of QSAR and read-across coupled with different regulatory criteria for PBT chemicals in the process of PBT/vPvB identification and substitution by alternatives. The results of the study show that the different mandatory measures imposed by the various regulations along with performed methodology should be considered together on harmonized PBT/vPvB identification in order to ensure that the truly problematic substances are identified.
3. Scholz et al. [91] attempted to prove the importance of alternative integrated testing strategies (ITSs) in early risk assessment of plant protection products, pharmaceuticals, biocides, feed additives, and effluents which can provide the operational means to combine the different promising alternative methods in a powerful and predictive approach that allows significant reduction of animal testing suitable from the ethical point of view.

4. The toxicity of biocides was estimated using a correlation statistic using QSAR approach [92], and the biocides were classified in different categories. The result of the study focuses on the importance of methinic group in controlling biocidal activity and concludes that QSAR can be used in the modulation of biocidal toxicity using simple molecular indices or descriptors.
5. The physiological modes of action of six biocides were analyzed based on predictive QSAR approach by Neuwoehner et al. (2008) [93]. The observed pattern of inhibition of reproduction and cell volume growth along with cell division were found to be closely related to the physiological modes of action of reference chemicals with well-known modes of toxic action in synchronous green algae. The proposed scheme of the paper can be used for initial screening and priority setting of biocides.
6. An ecotoxicological analysis of 26 dithiocarbamates (DCs) and related compounds on guppies (*Poecilia reticulata*), water fleas (*Daphnia magna*), green algae (*Chlorella pyrenoidosa*), and bacteria (*Photobacterium phosphoreum*) was performed by Van Leeuwen et al. [94]. The variations in toxicity of biocides were found to be directly correlated with n-octanol/water partition coefficient (nearly 100% explained variation) as explained by results of QSAR. In conclusion, DCs were classified as broad-spectrum biocides having cytotoxic properties against studied species.

- *Local QSAR models on biocides*
 - *Antifouling agents*

1. A set of 71 tributyltins (TBT) as antifouling compounds was analyzed for their PBT behavior using QSAR prediction programs such as BIOWIN™ (a biodegradation probability program), KOWWIN™ (log octanol-water partition coefficient calculation program) [95], and ECOSAR™ (Ecological Structure Activity Relationship Program) [15] by Cui et al. (2014) [22]. Their method highlights the importance of freely available toxicity prediction tools as mentioned above for the estimation of biodegradation of toxic compounds. A rapidly biodegradable chemical is said to be a suitable candidate for the antifouling agent as this can mitigate predicted ecological effects of compounds. They found out that 31 out of 71 compounds were rapidly biodegradable hence suited as a safe alternative for antifouling agent. Among the different class of chemicals taken into consideration, natural products were relatively more biodegradable when compared to synthetic one. The study suggests

low molecular weight (<400) natural product, and their analogues yield “green” antifoulants.

– *Antimicrobial agents*

1. The environmental fate and effect of 2,8-dichlorodibenzo-p-dioxin and photodegradation of triclosan (TCS), i.e., 2-hydroxy-8-chlorodibenzodioxin, was studied by Yuval et al. (2017) [96]. The QSAR results hinted at the non-biodegradable nature of studied compounds along with their transformation products, thus making them potential chemicals for environmental pollutants.

4 Prioritized Molecule Among PCPs and Biocides Using QSAR Approach

Finally, we list here a prioritized list of PCPs and biocides having potential to behave as potent environmental pollutants estimated using only QSAR and other computational (mainly predicting) approaches without involvement of experimental procedures. Tables 1 and 2 list the proposed molecules, uses along with the references from where they have been taken.

5 Conclusion

In conclusion, the current chapter highlights what we know about PCP and biocide- derived environmental toxicity and potential risk reported in the literature for various environmental compartments. The continuous report of accumulation of these chemicals causes concern among policy makers. Another major reason for concern related to PCPs and biocides is attributed to their explicit mode of action which differentiates them from other organic chemicals. We already saw above how extensively the demand of cosmetics and biocides is increasing due to modernization and globalization of the developed and developing countries (see Fig. 1) [6, 7]. The cosmetic market is considered to be booming because of its beautifying products such as lipstick, eye shadows/liners/highlighters, mascara, foundation cream, and concealer along with some medically beneficial products such as anti-dandruff agents, anti-hair fall agents, etc. In contrast, biocides find their application in cosmetics, medicines, household products, food products, domestic pest control, disinfectants, and industrial care products.

The environmental concern starts when these contaminants enter into different compartments of environment such as soil water and sediments either in an original form or in the form of metabolites thus affecting natural flora and fauna. Looking at the complexity of bioassay of these chemicals (mainly due to cost of

Table 1
List of PCPs prioritized/reported employing QSAR and other in silico models

No.	Name	Use	Reference
1	2-(20-Hydroxy-30,5'-di-tert-butylphenyl)benzotriazole	UV filter/sunscreen	[38]
2	2,4-Di-tert-butyl-6-(5-chloro-2Hbenzotriazol-2-yl)phenol	UV filter/sunscreen	[38]
3	Benzyl dimethyl dodecyl ammonium chloride	Surface active agents	[19]
4	Didecyl dimethyl ammonium chloride	Surface active agents	[19]
5	Tetradecyl trimethyl ammonium bromide	Surface active agents	[19]
6	Triclocarban	Antimicrobial agent	[13]
7	Fluorescent brightener 367	UV filter	[13]
8	Phenethyl Cinnamate	Fragrance	[13]
9	Diethylamino hydroxybenzoyl hexyl benzoate	UV filter	[13]
10	OCTRIZOLE	UV filter	[45]
11	2,4-Ditert-butyl-6-(5-chlorobenzotriazol-2-yl)phenol	UV filter	[45]
12	2-(Benzotriazol-2-yl)-4,6-bis(2-methylbutan-2-yl)phenol	UV filter	[45]
13	2-(Benzotriazol-2-yl)-6-butan-2-yl-4-tert-butylphenol	UV filter	[45]
14	2-(Benzotriazol-2-yl)-4,6-ditert-butylphenol	UV filter	[45]
15	Benzyl acetate	Fragrances	[84]
16	Cinnamyl acetate	Fragrances	[84]
17	γ -Methyl ionone	Fragrances	[84]
18	Hexyl salicylate	Fragrances	[84]
19	Linalool	Musk	[84]
20	Musk ketone	Fragrances	[84]
21	Phenethyl cinnamate	Fragrances	[84]
22	α -Amylcinnamyl alcohol	Fragrances	[84]
23	Celestolide, crysolide (ADBI)	Musk	[84]
24	DIMER-1	Antimicrobial agent	[85]
25	DIMER-2	Antimicrobial agent	[85]
26	PP2	Antimicrobial agent	[85]
27	α -Tocopherol PP1-6	Antioxidant	[87]
28	α -Tocopherol PP2	Antioxidant	[87]
29	α -Tocopherol PP3-2	Antioxidant	[87]
30	α -Tocopherol PP4-6	Antioxidant	[87]
31	α -Tocopherol PP4-10	Antioxidant	[87]

(continued)

Table 1
(continued)

No.	Name	Use	Reference
32	2-Methyl-2-phenyloxirane	Cosmetic ingredients	[76]
33	Ethyl benzene	Cosmetic ingredients	[76]
34	(1S,2R)-2-Amino-1-phenylpropan-1-ol	Cosmetic ingredients	[76]
35	Ethyl 1-methyl-4-phenylpiperidine-4-carboxylate	Cosmetic ingredients	[76]
36	Triclocarban	Antimicrobial	[13, 14]
37	Bis(2-ethylhexyl) benzene-1,2-dicarboxylate	Phthalate	[13, 14]
38	2-[4-(1,3-Benzoxazol-2-yl)naphthalen-1-yl]-1,3-enzoxazole	UV filter	[13, 14]
39	(Bis(methylcyclohexyl) phthalate)	Phthalate	[13, 14]
40	Dicyclohexyl benzene-1,2-dicarboxylate	Phthalate	[13, 14]
41	3,7-Dimethylocta-1,6-dien-3-yl (E)-3-phenylprop-2-enoate	Fragrance	[13, 14]
42	2-O-(2-Ethylhexyl) 1-O-hexyl benzene-1,2-dicarboxylate	Phthalate	[13, 14]
43	2-(2H-Benzothiazol-2-Yl)-6-(Dodecyl)-4-Methylphenol	UV filter	[13, 14]
44	2-Ethylhexyl (E)-3-(4-methoxyphenyl)prop-2-enoate	UV filter	[13, 14]
45	Hexyl 2-[4-(diethylamino)-2-hydroxybenzoyl]benzoate	UV filter	[13, 14]
46	1-O-Butyl 2-O-(2-ethylhexyl) benzene-1,2-dicarboxylate	Phthalate	[13, 14]
47	Bis(5-methylhexyl) benzene-1,2-dicarboxylate	Phthalate	[13, 14]
48	2-(Benzotriazol-2-yl)-4,6-bis(2-methylbutan-2-yl)phenol	UV filter	[13, 14]
49	Di hexyl benzene-1,2-dicarboxylate	Phthalate	[13, 14]
50	2-(Benzotriazol-2-yl)-4-(2,4,4-trimethylpentan-2-yl)phenol	UV filter	[13, 14]

experiments, time, and extensive labor involvement), the chemical laboratories fail to meet demands in detecting the ever-rising concentration of these emerging contaminants in the environment [8–12]. Looking at the scarcity of ecotoxicological data, the scientific researchers have come up with alternative methods of toxicity detection, and in silico methods such as QSAR are one of them. Using QSAR, one can utilize the available toxicity data of selected compounds in order to predict toxicity of untested or not yet synthesized chemicals in order to evaluate the risk of hazards posed by any chemical contaminants in general and PCPs and biocides in specific. In the last few decades, computational modeling of ecotoxicity predictions of diverse chemicals including PCPs and biocides has become a crucial step of ecotoxicological risk assessment. The experimental limitations such as cost, time, and animal sacrifice are believed to have been curbed to a greater extent using computational predictive approaches such as QSAR. The

Table 2
List of toxic biocides prioritized/reported employing QSAR and other in silico models

No.	Name	Use	Reference
1	Chlorothalonil	Antifouling products	[22]
2	Dichlofluanid	Antifouling products	[22]
3	Irgarol 1051	Antifouling products	[22]
4	TCMS pyridine	Antifouling products	[22]
5	TCMTB	Antifouling products	[22]
6	Diuron	Antifouling products	[22]
7	DCOIT	Antifouling products	[22]
8	Zinc pyrithione	Antifouling products	[22]
9	Copper pyrithione	Antifouling products	[22]
10	Zineb	Antifouling products	[22]
11	2,4-Di-tert-butyl-6-(dimorpholinomethyl)phenol	Antimicrobials	[92]
12	4,4'-(4-Isopropylphenyl)methylene)dimorpholine	Antimicrobials	[92]
13	4,4'-(Naphthalen-1-ylmethylene)dimorpholine	Antimicrobials	[92]
14	4,4'-(Pyren-1-ylmethylene)dimorpholine	Antimicrobials	[92]
15	3-Nitroaniline	Wood preservative	[93]
16	Irgarol® 1051	Antifouling products	[93]
17	Triclosan	Antimicrobials	[96]
18	Acrolein	Slimicides	[89]
19	Rotenone	Piscicides	[89]
20	Flufenoxuron	Wood preservatives	[89]
21	Cyhalothrin	Insecticides/acaricides	[89]
22	Deltamethrin	Insecticides/acaricides	[89]
23	Cypermethrin	Insecticides/acaricides	[89]
24	Acrinathrin	Insecticides	[89]
25	Malathion	Insecticides/acaricides	[89]
26	Chlorpyrifos	Insecticides/acaricides	[89]
27	Difethialone	Rodenticides	[89]
28	Hexaflumuron	Insecticides/acaricides	[89]
29	Fenbutatin oxide	Acaricide/miticide	[89]
30	Medetomidine	Antifouling products	[89]

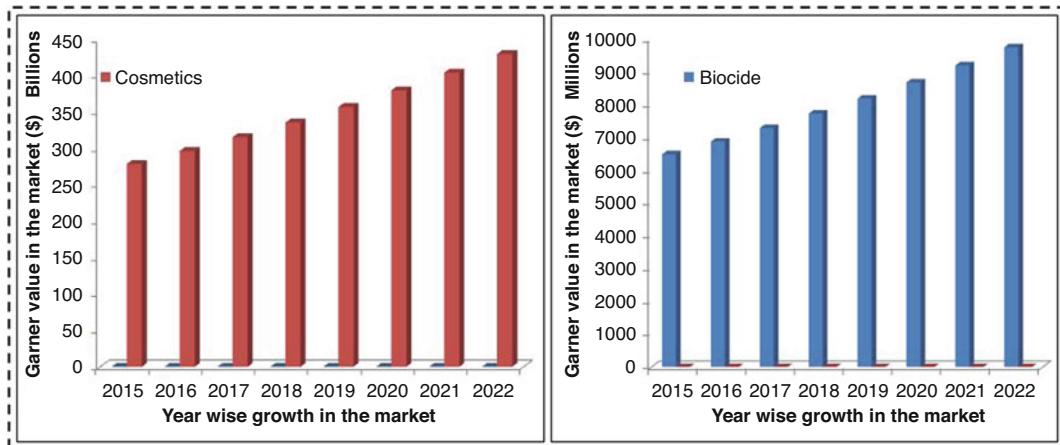


Fig. 1 The extent of rise in cosmetics and biocide markets value (in billions\$ for PCP and million\$ for biocides) in coming few year [6, 7]

described encouraging features of QSAR have made it compulsory in early risk assessment of environmental contaminants by various regulatory bodies such as ECVAM, US EPA, ATSDR, and CSTEE. Now, it is being used to predict the toxicity of new chemicals which are expected to become pollutants or toxicants after its use at present or in the future provided they withstand acceptability criteria as laid down and accepted by international community. The developed models on PCPs and biocides should be robust, statistically sound, and reliable [97] and should consist of diverse chemical datasets (large applicability domain) in order to have its application in ecotoxicity of various species.

The present chapter highlights the constituents, route, source, hazards, and potential risks associated with the exposure of PCPs and biocides to the environment. We have also seen the regulatory requirements or protocols as laid down by various regulatory bodies implemented in risk assessment of PCPs and biocides. The roles of the government authorities and different regulated policies regarding identification of risk of these products have also been discussed. Lastly, the chapter gives a detailed analysis of already applied QSAR models in ecotoxicological study of PCPs and biocides mainly focusing on the source of data collection, type of descriptor used, methodology used, and software employed which were among the few highlights. A list of chemicals of elevated concern to the environment obtained solely by QSAR and other predictive methods is presented at the end of the book chapter to depict some of their merits above experimental procedures. Looking at the number of models available on PCPs and biocides, it seems to be higher for PCPs compared to biocides, but the numbers are in general limited compared to other classes of industrial chemicals. Although it is definite that experimental

approaches can never be completely substituted with computational approaches, these approaches can be integrated with each other for better understanding. Thus, we can conclude that the need to develop more number of QSAR models is not only desirable but also confirms its nonpareil role in ecotoxicity prediction of PCPs and biocides.

Acknowledgments

KK thanks Indian Council of Medical Research, New Delhi for financial support in the form of a senior research fellowship.

References

1. De P, Roy K (2018) Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR QSAR Environ Res* 29:319–337
2. Kar S, SepAlveda MS, Roy K, Leszczynski J (2017) Endocrine-disrupting activity of per-and polyfluoroalkyl substances: exploring combined approaches of ligand and structure based modeling. *Chemosphere* 184:514–523
3. E. Union (2012) Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products. *Off J Eur Union L* 167:1–116
4. Devillers J, Mombelli E, Samsera R (2011) Structural alerts for estimating the carcinogenicity of pesticides and biocides. *SAR QSAR Environ Res* 22:89–106
5. Dich J, Zahm SH, Hanberg A, Adami H-O (1997) Pesticides and cancer. *Cancer Causes Control* 8:420–443
6. Available at <https://www.alliedmarketresearch.com/cosmetics-market> (2019)
7. Available at <https://www.alliedmarketresearch.com/biocides-market> (2019)
8. Khan K, Roy K, Benfenati E (2019) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. *J Hazard Mater* 369:707–718
9. Khan PM, Roy K, Benfenati E (2019) Chemometric modeling of Daphnia magna toxicity of agrochemicals. *Chemosphere* 224:470–479
10. Khan K, Benfenati E, Roy K (2019) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotox Environ Safe* 168:287–297
11. Hossain KA, Roy K (2018) Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in *Dugesia japonica* and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. *Ecotox Environ Safe* 166:92–101
12. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2019) Ecotoxicological modeling, ranking and prioritization of pharmaceuticals using QSTR and QSTTR approaches: application of 2D and fragment based descriptors. *Mol Inform*, 38, article 1800078, <https://doi.org/10.1002/minf.201800078>
13. Gramatica P, Cassani S, Sangion A (2016) Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives' design. *Green Chem* 18:4393–4406
14. Khan K, Roy K (2017) Ecotoxicological modelling of cosmetics for aquatic organisms: a QSTR approach. *SAR QSAR Environ Res* 28:567–594
15. Mayo-Bean K, Moran K, Meylan B, Ranslow P (2012) Methodology document for the ECOlogical Structure-Activity Relationship model (ECOSAR) class program. US-EPA, Washington DC
16. Khan K, Baderna D, Cappelli C, Toma C, Lombardo A, Roy K, Benfenati E (2019) Ecotoxicological QSAR modeling of organic compounds against fish: application of fragment based descriptors in feature analysis. *Aquat Toxicol* 212:162
17. ECHA (2019) <https://echa.europa.eu/-/position-centres-guidance>
18. European commission Cosmetic ingredient database (2019) Available at https://ec.europa.eu/growth/sectors/cosmetics/cosing_en. Access on March-May 2019
19. Di Nica V, Gallet J, Villa S, Mezzanotte V (2017) Toxicity of quaternary ammonium compounds (QACs) as single compounds and

- mixtures to aquatic non-target microorganisms: experimental data and predictive models. *Ecotox Environ Safe* 142:567–577
20. Yamagishi T, Miyazaki T, Horii S, Akiyama K (1983) Synthetic musk residues in biota and water from Tama River and Tokyo Bay (Japan). *Arch Environ Contam Toxicol* 12:83–89
 21. ICID, ICID International Cosmetic Ingredient Dictionary and Handbook (2008) 12th Edition and 2014 18th edition, published by The Cosmetic, Toiletry, and Fragrance Association, Washington, DC
 22. Cui Y, Teo S, Leong W, Chai C (2014) Searching for “environmentally-benign” antifouling biocides. *Int J Mol Sci* 15:9255–9284
 23. Ellis JB (2006) Pharmaceutical and personal care products (PPCPs) in urban receiving waters. *Environ Pollut* 144:184–189
 24. Vimalkumar K, Arun E, Krishna-Kumar S, Poopal RK, Nikhil NP, Subramanian A, Babu-Rajendran R (2018) Occurrence of triclocarban and benzotriazole ultraviolet stabilizers in water, sediment, and fish from Indian rivers. *Sci Total Environ* 625:1351–1360
 25. Govindarajalu K (2003) Industrial effluent and health status: a case study of Noyyal river basin. In: Proceedings of the third international conference on environment and health. Citeseer, Chennai, pp 15–17
 26. Holah J, Taylor J, Dawson D, Hall K (2002) Biocide use in the food industry and the disinfectant resistance of persistent strains of *Listeria monocytogenes* and *Escherichia coli*. *J Appl Microbiol* 92:111S–120S
 27. McLaughlin JK, Lipworth L, Tarone RE (2003) Suicide among women with cosmetic breast implants: a review of the epidemiologic evidence. *J Long-Term Eff Med Implants* 13:6
 28. Miller LG, Quan C, Shay A, Mostafaie K, Bharadwa K, Tan N, Matayoshi K, Cronin J, Tan J, Tagudar G (2007) A prospective investigation of outcomes after hospital discharge for endemic, community-acquired methicillin-resistant and-susceptible *Staphylococcus aureus* skin infection. *Clin Infect Dis* 44:483–492
 29. sccs (2019) https://ec.europa.eu/health/scientific_committees/consumer_safety/opinions_en
 30. ECSID, European commission Cosmetic ingredient database 2019 (2019) Available at https://ec.europa.eu/growth/sectors/cosmetics/cosing_en. Access on March-May 2019
 31. Roy K, Kar S (2016) In Silico models for eco-toxicity of pharmaceuticals, in: Springer, In Silico methods for predicting drug toxicity, pp 237–304
 32. Krewski D, Westphal M, Andersen ME, Paoli GM, Chiu WA, Al-Zoughool M, Croteau MC, Burgoon LD, Cote I (2014) A framework for the next generation of risk science. *Environ Health Perspect* 122:796–805
 33. Presidential/Congressional Commission on Risk Assessment Risk Management (1997) Risk assessment and risk management in regulatory decision-making. Final Report. Vol. 2.- Washington, DC:PCRARM. Available: <http://www.riskworld.com>. Accessed 4 Mar 2019
 34. Kar S, Roy K, Leszczynski J (2018) Impact of pharmaceuticals on the environment: risk assessment using QSAR modeling approach. In: Computational toxicology. Springer, NY, pp 395–443
 35. EFSA, European Food Safety Authority (EFSA) (2015) Website accessed in 2015. <https://www.efsa.europa.eu>
 36. Kummerer K (2007) Sustainable from the very beginning: rational design of molecules by life cycle engineering as an important approach for green pharmacy and green chemistry. *Green Chem* 9:899–907
 37. SCOPUS (2019) Available at <https://www.scopus.com/search/form.uri?display=basic>
 38. Papa E, Sangion A, Arnot JA, Gramatica P (2018) Development of human biotransformation QSARs and application for PBT assessment refinement. *Food Chem Toxicol* 112:535–543
 39. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 14:450–474
 40. Gramatica P, Papa E, Sangion A (2018) QSAR modeling of cumulative environmental end-points for the prioritization of hazardous chemicals. *Environ Sci Process Impacts* 20:38–47
 41. Önlü S, Saçan MT (2017) An in silico approach to cytotoxicity of pharmaceuticals and personal care products on the rainbow trout liver cell line RTL-W1. *Environ Toxicol Chem* 36:1162–1169
 42. Agarwal M, Frank MI (2019) Spartan: a software tool for parallelization bottleneck analysis, in: 2009 ICSE workshop on multicore software engineering. IEEE 2009:56–63
 43. Mauri A, Consonni V, Pavan M, Todeschini R, software D (2006) An easy approach to molecular descriptor calculations. *Match* 56:237–248

44. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34:2121–2132
45. Cassani S, Gramatica P (2015) Identification of potential PBT behavior of personal care products by structural approaches. *Sustainable Chem Pharm* 1:19–27
46. De García SAO, Pinto GP, García-Encina PA, Irusta-Mata R (2014) Ecotoxicity and environmental risk assessment of pharmaceuticals and personal care products in aquatic environments and wastewater treatment plants. *Ecotoxicology* 23:1517–1533
47. Toropova AP, Toropov AA (2018) Use of the index of ideality of correlation to improve models of eco-toxicity. *Environ Sci Pollut Res* 25:31771–31775
48. Matthews EJ (2019) In silico scaling and prioritization of chemical disposition and chemical toxicity of 15,145 organic chemicals. *Comput Toxicol* 9:100–132
49. Percepta, from Advanced Chemistry Development (ACD) Labs (2018) <http://www.acdlabs.com/products/percepta/>
50. Center Watch, website accessed in 2015. www.centerwatch.com/drug-information/fda-approvals/
51. Drugs@FDA, website accessed in 2015. <https://www.accessdata.fda.gov/scripts/cder/daf/>
52. The Good Scents Company (2018) <http://www.thegoodscentscompany.com>
53. CFSAN Thesaurus, accessed February 2013. <http://www.fda.gov/Food/FoodScienceResearch/ToolsMaterials/ucm181420.htm>
54. EAFUS List. Everything added to food in the United States. www.accessdata.fda.gov/scripts/fcn/fcnnavigation.cfm?rpt=eafuslisting
55. Health Canada, website accessed in 2015. <http://hc-sc.gc.ca/fn-an/securit/addit/list/11-preserv-conserv-eng.php>
56. European Food Safety Authority (EFSA), website accessed in 2015. <https://www.efsa.europa.eu>
57. FDA GRAS. GRAS notice inventory (2018) <https://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/NoticeInventory/default.htm>
58. Fragrance Products Information Network. <http://pw1.netcom.com/~bcb56/fpin.htm>, <http://www.fpinva.org/text/1a5d908-130.html> (web link no longer available)
59. Environmental Protection Agency (EPA) list of fragrance chemicals in household products (15/49 FPIN HOME PAGE OVERVIEW HEALTH FRAGRANCE MATERIALS. Accessed 23 Apr 2013
60. FPIN_LP: Common fragrance Chemicals in Laundry Products & cleaners, the FPINVA fragrances were compiled by Betty Bridges (RN, 08/2006, <http://www.fpinva.org/text/1a5d908-120.html>) from Aldrich's Flavors and Fragrances Catalog
61. GIVAUDAN & IFF fragrance manufacturers. <https://www.givaudan.com/>, <http://www.iff.com/>
62. ICID International Cosmetic Ingredient Dictionary and Handbook, 2008 12th Edition and 2014 18th Edition, published by The Cosmetic, Toiletry, and Fragrance Association, Washington, DC
63. Hair dyes. www.accord.asn.au
64. Arvidson KB, Chanderban R, Muldoon-Jacobs K, Mayer J, Ogungbesan A (2010) Regulatory use of computational toxicology tools and databases at the United States Food and Drug Administration's Office of Food Additive Safety. *Expert Opin Drug MetabToxicol* 6:793–796
65. Color of art database, the color of art pigment database, an artist reference. Accessed in 2013. <http://www.artoscreation.com/colorindes/index.html>
66. Ink Dystuffs. Accessed in 2012. http://www.trader-ina.com/Chemicals/Dyestuffs/Ink-Dyestuffs_3.html
67. Stainsfile Dye Index. Accessed in 2013. <http://stainsfile.info/StainsFile/dyes/dyes.htm>
68. Batke M, Gütlein M, Partosch F, Gundert-Remy U, Helma C, Kramer S, Maunz A, Seeland M, Bitsch A (2016) Innovative strategies to develop chemical categories using a combination of structural and toxicological properties. *Front Pharmacol* 7:321
69. Bitsch A, Jacobi S, Melber C, Wahnschaffe U, Simetska N, Mangelsdorf I (2006) REPDOSE: a database on repeated dose toxicity studies of commercial chemicals—a multifunctional tool. *Regul Toxicol Pharmacol* 46:202–210
70. Barabair F, Olsson H, Sokull-Klütgen B (2009) European List of notified chemical substances—ELINCS. JRC Scientific and Technical Reports, Brussels
71. A free web service tool. Accessible at <http://mlc-reach.informatik.uni-mainz.de>
72. Verma RP, Matthews EJ (2015) Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models

- (QSAR-21): part I: irritation potential. *Regul Toxicol Pharmacol* 71:318–330
73. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2017) Ecotoxicological assessment of pharmaceuticals using computational toxicology approaches: QSTR and interspecies QTTR modeling. In: Proceedings of MOL2-NET 2017, international conference on multidisciplinary sciences, 3rd edn. MDPI AG, p 1
 74. Hisaki T, née Kaneko MA, Yamaguchi M, Sasa H, Kouzuki H (2015) Development of QSAR models using artificial neural network analysis for risk assessment of repeated-dose, reproductive, and developmental toxicities of cosmetic ingredients. *J Toxicol Sci* 40:163–180
 75. Enslein K, Gombar VK (1997) TOPKAT 5.0 and modulation of toxicity. *Mutat Res-fund Mol M* 379:S14–S14
 76. Plošnik A, Zupan J, Vračko M (2015) Evaluation of toxic endpoints for a set of cosmetic ingredients with CAESAR models. *Chemosphere* 120:492–499
 77. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chromatogr Intell Lab Syst* 152:18–33
 78. Development Core Team R (2015) R: A language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. Available online at (<http://www.R-project.org/>)
 79. Ritz C, Streibig JC (2005) Bioassay analysis using R. *J Stat Softw* 12:1–22
 80. Jentzsch F, Olsson O, Westphal J, Reich M, Leder C, Kümmeler K (2016) Photodegradation of the UV filter ethylhexyl methoxycinnamate under ultraviolet light: identification and in silico assessment of photo-transformation products in the context of grey water reuse. *Sci Total Environ* 572:1092–1100
 81. Chakravarti SK, Saiakhov RD, Klopman G (2012) Optimizing predictive performance of CASE ultra expert system models using the applicability domains of individual toxicity alerts. *J Chem Inf Model* 52:2609–2618
 82. Liu H, Sun P, Liu H, Yang S, Wang L, Wang Z (2015) Acute toxicity of benzophenone-type UV filters for *Photobacterium phosphoreum* and *Daphnia magna*: QSAR analysis, interspecies relationship and integrated assessment. *Chemosphere* 135:182–188
 83. Kar S, Das RN, Roy K, Leszczynski J (2016) Can toxicity for different species be correlated?: the concept and emerging applications of interspecies quantitative structure-toxicity relationship (i-QSTR) modeling. *IJQSPR* 1:23–51
 84. Papa E, Luini M, Gramatica P (2009) Quantitative structure-activity relationship modelling of oral acute toxicity and cytotoxic activity of fragrance materials in rodents. *SAR QSAR Environ Res* 20:767–779
 85. De Vaugelade S, Nicol E, Vujovic S, Bourcier S, Pirnay S, Bouchonnet S (2018) Ultraviolet-visible phototransformation of dehydroacetic acid-structural characterization of photoproducts and global ecotoxicity. *Rapid Commun Mass Spectrom* 32:862–870
 86. Campbell JL, Yoon M, Clewell HJ (2015) A case study on quantitative in vitro to in vivo extrapolation for environmental esters: methyl-, propyl- and butylparaben. *Toxicology* 332:67–76
 87. De Vaugelade S, Nicol E, Vujovic S, Bourcier S, Pirnay S, Bouchonnet S (2017) UV-vis degradation of α -tocopherol in a model system and in a cosmetic emulsion-structural elucidation of photoproducts and toxicological consequences. *J Chromatogr A* 1517:126–133
 88. Canipa SJ, Chilton ML, Hemingway R, Macmillan DS, Myden A, Plante JP, Tennant RE, Vessey JD, Steger-Hartmann T, Gould J (2017) A quantitative in silico model for predicting skin sensitization using a nearest neighbours approach within expert-derived structure-activity alert spaces. *J Appl Toxicol* 37:985–995
 89. Khan K, Khan PM, Lavado G, Valsecchi C, Pasqualini J, Baderna D, Marzo M, Lombardo A, Roy K, Benfenati E (2019) QSAR modeling of *Daphnia magna* and fish toxicities of biocides using 2D descriptors. *Chemosphere* 229:8–17. <https://doi.org/10.1016/j.chemosphere.2019.04.204>
 90. Rauert C, Friesen A, Hermann G, Johncke U, Kehler A, Neumann M, Prutz I, Schonfeld J, Wiemann A, Willhaus K (2014) Proposal for a harmonised PBT identification across different regulatory frameworks. *Environ Sci Eur* 26:9
 91. Scholz S, Sela E, Blaha L, Braunbeck T, Galay-Burgos M, Garcia-Franco M, Guinea J, Kluver N, Schirmer K, Tanneberger K (2013) A European perspective on alternatives to animal testing for environmental hazard identification and risk assessment. *Regul Toxicol Pharmacol* 67:506–530
 92. Hernandez-Almirano R, Mena-Cervantes VY, Perez-Miranda S, Fernandez FJ, Flores-Sandoval CA, Barba V, Beltran HI, Zamudio-Rivera LS (2010) Molecular design and QSAR study of low acute toxicity biocides with 4, 4 \mathbb{A} \mathbb{E}^2 -dimorpholyl-methane core obtained by microwave-assisted synthesis. *Green Chem* 12:1036–1048
 93. Neuwoehner J, Junghans M, Koller M, Escher BI (2008) QSAR analysis and specific

- endpoints for classifying the physiological modes of action of biocides in synchronous green algae. *Aquat Toxicol* 90:8–18
94. Van Leeuwen CJ, Maas-Diepeveen JL, Niebeek G, Vergouw WHA, Griffioen PS, Luijken MW (1985) Aquatic toxicological aspects of dithiocarbamates and related compounds. I. Short-term toxicity tests. *Aquat Toxicol* 7:145–164
95. Meylan WM (2000) SRC KOWWIN Software SRC-LOGKOW Version 1.66, Syracuse Research Corporation, USA
96. Yuval A, Eran F, Janin W, Oliver O, Yael D (2017) Photodegradation of micropollutants using V-UV/UV-C processes; Triclosan as a model compound. *Sci Total Environ* 601:397–404
97. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* 3:11392–11406



Chapter 17

Computational Approaches to Evaluate Ecotoxicity of Biocides: Cases from the Project COMBASE

Sergi Gómez-Ganau, Marco Marzo, Rafael Gozalbes, and Emilio Benfenati

Abstract

The evaluation of the ecotoxicological profile of chemicals is of high relevance when a substance can have an impact on the environment, such as the case of biocides. Due to the high number of animal tests conducted each year for regulatory purposes and the ethical considerations that this entails, the requirement of alternative methods by companies and regulatory agencies is increasing. Within these, *in silico* tools are useful to minimize time, costs, and resources, and they can be applied as alternatives to traditional laboratory assays.

In this chapter, we present some computational models developed in the context of the EU LIFE+ project entitled “Computational tool for the assessment and substitution of biocidal active substances of ecotoxicological concern (COMBASE)” (<http://www.life-combase.com>). The main objective of the project was the development of a tool based on computational toxicology, integrating predictive models of the toxic effects associated with biocidal substances at different trophic levels. Here, different quantitative structure-activity relationship (QSAR) models for the estimation of ecotoxicity of biocides in microorganisms and fish are presented. First, an integrated model to predict the respiratory inhibition in activated sludge was developed, by combining sequentially a qualitative and a quantitative QSAR model. Previously to the development of the model, a set of 94 chemicals with known EC₅₀ values was selected to this study, based on their “biocide-like” structural features. Second, a model to predict LC₅₀ on rainbow trout was developed on a dataset made by collection data from OpenFoodTox database of the European Food Safety Authority (EFSA) and Pesticide Ecotoxicity Database of Office of Pesticide Programs (OPP) (<https://ecotox.ipmcenters.org/>).

Both models showed good performances and robustness and have been integrated in the VEGA last release (version 1.1.5; <https://www.vegahub.eu/>) as well as the specific COMBASE tool (<http://webtool.life-combase.com>).

Key words Biocides, QSAR, Biocidal Products Regulation (BPR), Activated sludge, Rainbow trout, VEGA, COMBASE

1 Introduction

Biocidal products such as disinfectants, wood preservatives, rodenticides, antifouling products, etc. are systematically used in our daily lives and in industry. During the last years, their concentration in the aquatic environments such as wastewater effluents, sludge, or

even drinking water has been rising due to their uncontrolled use and the inefficient wastewater treatment [1]. Increasing attention is being paid to the concerns related to the environmental occurrence and possible harmful impact of biocides which are becoming now a widely and well-recognized class of emerging environmental pollutants [1].

The authorization process and the placement of biocides on market are regulated by the Biocidal Products Regulation (BPR) 528/2012 [2]. Different tests for human toxicity, environmental safety, and control of residues and degradation products must be performed for biocidal products, including the risk assessment for the aquatic effects of biocides in different compartments and aquatic species [3].

The market of biocides in Europe is in expansion, and in silico methods have acquired a relevant role at the regulatory level in order to reduce time and costs of traditional laboratory assays. These computational approaches are increasingly being used for toxicity assessment and reduction of the need of in vitro or in vivo test. There are many situations where in silico methods have a key role in the hazard assessment of chemicals such as weight of evidence, assessment of degradation products and impurities in pharmaceutical products or plant protection products, or metabolite analysis [4].

The development of QSAR models represents one of the most widely used strategies within in silico methods. QSARs involve the construction of mathematical models derived from training sets of molecules with well-known chemical structures and biological/toxicological properties. The chemical structures of series of compounds are related by means of mathematical algorithms with physicochemical properties or biological/toxicological activities [5]. In the recent years, different QSAR models have been developed for predicting parameters for regulatory purposes [6]. The main reason for this increasing interest on QSARs is their acceptance at the regulatory level by different international instances, such as the European Chemicals Agency (ECHA; mainly regarding REACH norm), the European Food Safety Authority (EFSA), or the Food and Drug Administration (FDA) [6, 7]. QSAR models are validated by these organisms when they are developed following the standardized “Setubal rules” [8], and several well-known chemoinformatic programs such as VEGA [<https://www.vegahub.eu/>] include models completely acceptable and usable for regulatory processes.

The use of non-animal alternative test methods has also been foreseen in the Biocidal Products Regulation, Regulation (EU) 528/2012. The widespread use of non-testing methods would help SMEs in identifying, among others, the so-called metabolites of ecotoxicological concern while facilitating the registering of biocidal active substances within the BPR and reducing animal

testing. In this context, the EU LIFE + project entitled “Computational tool for the assessment and substitution of biocidal active substances of ecotoxicological concern (COMBASE)” has the objective to promote the sustainable use of biocidal active substances from a life cycle perspective. To this end, an online information system based on the combination of evidence-based decision support systems (EBDSS) and proven computational toxicology modeling approaches has been implemented.

The COMBASE project has been conducted by six partners from Spain and Italy: Istituto di Ricerche Farmacologiche Mario Negri (IRFMN; www.marionegri.it); INKOA Sistemas S.L. (www.inkoa.com); Instituto Tecnológico del Embalaje, Transporte y Logística (ITENE; www.itene.com); Xenobiotics S.L. (<https://xenobiotics.es>); Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA; www.inia.es); and ProtoQSAR S.L. (<https://protoqsar.com>). In this chapter we present the computational models developed to predict the ecotoxicity in aquatic environment at two trophic levels, microorganisms and fish.

2 Materials and Methods

2.1 Activated Sludge QSAR Models

2.1.1 Organism to Assess and Endpoint to Evaluate

The acute toxicity assay activated sludge, according to the OECD official Guideline 209 on “Activated sludge, Respiration Inhibition Test” [9], describes a method to determine the effects of a substance on microorganisms from activated sludge (largely bacteria) by measuring their respiration rate (carbon and/or ammonium oxidation) under defined conditions and in the presence of different concentrations of the test substance. With this test, it is possible to perform a rapid screening to assess the effects of chemicals on the microorganisms of the activated sludge.

Samples of activated sludge with test substance and without (blank controls) are incubated with synthetic sewage and the respiration rates are measured in an enclosed cell containing an oxygen electrode after a contact time of 3 h. The sensitivity of each batch of activated sludge is also tested with a suitable reference substance (i.e., 3,5-dichlorophenol). The test is typically used to determine the EC_x (e.g., EC₅₀) of the test substance and/or the non-observed effect concentration (NOEC).

2.1.2 Data Collection and Definition of a Biocide-Like Space

A COMBASE dataset has been built within the frame of the COMBASE project by collecting available information of existing biocides, both considering active substances and their metabolites. This database was implemented compiling toxicity data for biocide substances in organisms of the freshwater/marine and sewage treatment plant compartments. Several official and scientific databases were consulted, such as the Open Chemistry Database

Table 1
Physicochemical and structural features considered to differentiate biocides with respect to a set of standard chemicals

Physical properties	Molecular weight Total charge (sum of formal charges) Number of reactive groups
Atom counts	Number of atoms (including implicit atoms) Number of carbon atoms Number of hydrogen atoms (including implicit hydrogens) Number of heteroatoms Ratio between the number of carbon atoms and heteroatoms Number of heavy atoms Number of aromatic atoms Number of nitrogen atoms Number of oxygen atoms Number of hydrogen bond acceptors (number of nitrogen plus oxygen atoms) Number of hydrogen bond donors (number of OH and NH atoms) Number of fluorine atoms Number of chlorine atoms Number of bromine atoms Number of iodine atoms Number of halide atoms Number of phosphorous atoms Number of sulfur atoms Number of P and S Number of chiral centers Absence of atoms different from C, O, N, S, P, F, Cl, Br, I, Li, Na, K, Mg, Ca
Bond counts	Number of bonds (including implicit hydrogens) Number of single bonds (including implicit hydrogens) Number of double bonds Number of triple bonds Number of bonds between heavy atoms Number of rotatable single bonds Number of rigid bonds Number of aromatic bonds Number of rings Absence of $-(\text{CH}_2)_6\text{CH}_3$ chains
Adjacency and distance matrix descriptors	Diameter Radius Petitjean descriptor

(PubChem; <https://pubchem.ncbi.nlm.nih.gov/>), the ECOTOX-ecotoxicology knowledgebase (ECOTOX; <https://cfpub.epa.gov/ecotox/>), the Pesticide Properties Database (PPDB; <https://sitem.herts.ac.uk/aeru/ppdb/>), the TOXicology Data NETwork (TOXNET; <https://toxnet.nlm.nih.gov/>), or the Pesticide Action Network (PAN; <http://www.pesticideinfo.org/>). Data from

196 biocidal substances and 206 environmental metabolites were collected, and substances were categorized for their toxicity into four groups, considering values of L(E)C₅₀, according to EU Regulation (EC) No. 1272/2008. All of this information was made available for QSAR building purposes, but in the particular case of activated sludge, respiratory inhibition test, experimental EC₅₀ data after 3 h was only found for a very reduced number of chemicals, insufficient to build reliable QSARs.

Other datasets available within the OECD QSAR Toolbox v. 4.2. (www.qsartoolbox.org) were considered as potentially useful for our objective. Fifty-seven databases containing 84, 291 chemicals with almost 2.5 million measured data points with information for environmental fate and transport, physical chemical properties, ecotoxicology, and human health hazards are implemented in QSAR Toolbox 4.3. The most significant for constructing the model were ECOTOX, Aquatic ECETOC, and Aquatic Japan MOE, but information from other datasets was also used. Nevertheless, we wanted to develop computational models specifically tailored for biocides, and those databases include any kind of chemicals. In consequence, it was decided to previously define a “biocide-like” chemical space composed of chemicals with a common set of structural features. Once these features were selected, their application to this global dataset could provide us a dataset that, even if not composed exclusively of biocides, is composed in some way of compounds with potential biocide activity.

With the aim of identifying a set of common biocide-like relevant features, a comparison of physical/structural parameters (Table 1) and cutoff values was carried out between the Physprop database (<https://www.srccinc.com>), a generalist set of around 6500 chemicals, and a biocide-specific dataset (the COMBASE dataset; <http://www.life-combase.com>). The Physprop dataset contains chemical structures, names, and physical properties for generic chemical compounds. On the other side, the COMBASE dataset has been built within the frame of the COMBASE project by collecting available information of the existing biocides, both considering active substances and their metabolites.

Biocide-like filters were defined as those features able to maximize the difference between a biocide-like compound and a generic chemical for both datasets. The different parameters (Table 1) to characterize the structures from both databases were calculated using different software: CDK [10], FAF-Drugs4 [11], and PaDEL descriptor [12].

One single dataset was built by merging all the information. The quality of the experimental data of this dataset was curated following a standard procedure [13]: compounds with a chemical structure not clearly defined were deleted, as well as inorganic compounds, metal complexes, salts containing organic polyatomic

counterions, mixtures, and substances of unknown or variable composition (UVCB). Also, duplicates and tautomers were checked. A further check of duplicates was done using the best tautomer to ensure that only one compound was present in the final dataset.

In case of multiple and different experimental values for the same compound, the variability was evaluated using the threshold established by the European Commission [14] as the ratio between the maximum value and the minimum experimental value (x/y) as follows:

- (a) If x/y was <3 , the geometric mean of the experimental data was kept.
- (b) If x/y was >3 , the compound was removed from the dataset.

In the remaining cases with different experimental values, the geometric mean was considered as the experimental value associated to the compound. After curation of the dataset, the biocide-like filters described in Subheading 2.1.1 were applied to the dataset, and finally 94 biocide-like compounds were selected to develop the models.

2.1.3 Development of the Model

An integrated model to predict the respiratory inhibition in activated sludge was arranged cascading a qualitative QSAR model and a quantitative QSAR model.

Qualitative QSAR Model for Activated Sludge

Starting from the dataset composed of 94 biocide-like chemical compounds, a binary model (toxic/nontoxic) was developed. A compound was considered toxic when the EC₅₀ for activated sludge respiratory inhibition test was <100 mg/L. Molecular descriptors were calculated by an in-house software module in which they are implemented as described by Todeschini et al. [15]. Constant variables, near-constant variables, and 0.95 pair-correlated variables were deleted. Once the variables were calculated, STATISTICA software [16] was used to carry out the model building. The whole dataset was randomly split into a training set (64%) and a validation set (36%), and the boosted tree method [16] was used for the classification model development. The algorithm for boosted trees evolved from the application of boosting methods to regression trees. The general idea is to compute a sequence of (very) simple trees, where each successive tree is built for the prediction residuals of the preceding tree. Thus, at each step of the boosting (boosting tree algorithm), a simple (best) partitioning of the data is determined, and the deviations of the observed values from the respective means (residuals for each partition) are computed. The next three-node tree is then fitted to those residuals, to find another partition that will further reduce the residual (error) variance for the data, given the preceding sequence of trees.

It can be shown that such “additive weighted expansions” of trees can eventually produce an excellent fit of the predicted values to the observed values, even if the specific nature of the relationships between the predictor variables and the dependent variable of interest is very complex (nonlinear in nature). Hence, the method of gradient boosting—fitting a weighted additive expansion of simple trees—represents a very general and powerful machine learning algorithm [16]. The variables in the model were selected using a sensitivity analysis. Sensitivity analysis in data mining and statistical model building/fitting generally refers to the assessment of the importance of predictors in the respective (fitted) models. In short, given a fitted model with certain model parameters for each predictor, what the effect would be of varying the parameters of the model (for each variable) on the overall model fit is studied. In Statistica Data Miner, sensitivity analysis is available via several options; the particular statistics and measures that are reported depend on the statistical or data mining method for which the sensitivity analysis is requested. In all CART (classification and regression tree) and boosted tree models, sensitivity and predictor importance is computed from the average importance of each predictor at each split point (split node) in the final tree model.

Quantitative QSAR Model

Once the qualitative model was developed, a quantitative model was developed using the compounds for which a precise value of EC₅₀ was known (all of them belonging to the group of “toxic” chemicals with EC₅₀ < 100 mg/L). Once the variables were calculated, the EC₅₀ was converted to LogEC₅₀, and STATISTICA software [16] was used to develop the model. The whole dataset was used to perform a multiple linear regression (MLR). The variables in the model were selected by using a forward stepwise analysis. In STATISTICA, the forward stepwise method employs a combination of the procedures used in the forward entry and backward removal methods. At step 1 the procedures for forward entry are performed. At any subsequent step where two or more effects have been selected for entry into the model, forward entry is performed if possible, and/or backward removal is performed if possible, until neither procedure can be performed. Stepping is also terminated if the maximum number of steps is reached.

When the model was developed, a leave-*one*-out cross validation (LOOCV) was carried out. Leave-one-out cross validation is a special case of k-fold cross validation with $k = n$, the number of observations. LOOCV consists of splitting the dataset randomly into n partitions. For each n -th iteration, $n-1$ partitions are used as the training set and the left-out sample is used as the test set. When the dataset is small, leave-one-out cross validation is appealing as the size of the training set is maximized [17].

2.2 Rainbow Trout QSAR Model

2.2.1 Data Collection

A model to predict LC₅₀ on rainbow trout was developed from a dataset made by collection data from OpenFoodTox database [18] of the European Food Safety Authority (EFSA) and Pesticide Ecotoxicity Database of the US EPA Office of Pesticides Programs (OPP) (<https://ecotox.ipmcenters.org/>). Two hundred and four values were collected from OpenFoodTox database and 521 from Pesticide Ecotoxicity Database. Filters to select data from the databases are the following:

- Substance type is pesticide or biocide.
- Species is rainbow trout.
- Fish age is less than 3 days.
- The guideline followed is OECD 203 [19].
- Active ingredient purity must be >80%.
- Exposure time is 96 h.
- Endpoint is LC₅₀.
- No qualifier is accepted.
- All measured units were converted in mg/L.

For compounds with multiple data, congruence of experimental data was evaluated; if the ratio between the highest and lowest value was greater than 3, the compound was discarded. For compounds with congruent multiple data, the geometric mean was calculated. After data collection, the dataset is composed of 393 compounds.

The model was developed using graphical program KNIME [20]. KNIME is a free JAVA program. The model was developed using descriptors calculated with the commercial software Dragon 7.0 [21] integrated in KNIME that permits to calculate over 5000 chemical descriptors.

2.2.2 Data Cleaning

Before developing a model, a data cleaning process was applied: 2 compounds were discarded because they were inorganics; 8 compounds were discarded because they were composed of a large number of carbons compared with the other dataset compounds, thus they were discarded after a box plot analysis based on the number of carbons in the molecule, and they were found as outliers; 7 compounds were discarded because in the PCA analysis made on 3481 Dragon descriptors, they were found as outliers; 47 compounds were discarded because the experimental value results were outliers in a box plot analysis.

2.2.3 Features Selection

After the data cleaning process, the final dataset was composed of 329 compounds. The final dataset was split in a training set (TS), a calibration set (CS), and a validation set (VS) (Fig. 1). VS corresponds to the 20% of the dataset, and the remaining compounds

were split in TS and CS with ratio 80/20%. Therefore, TS is composed of 212 compounds, CS 52 compounds, and VS 65 compounds. TS is used to develop the model, CS is used to select the best parameters to select descriptor and model parameters, and VS is used after model development to see the performance of the model on external data.

With Dragon, 2641 descriptors were calculated on the 329 compounds. All descriptors were normalized with “z score” methods. For descriptor selection, low variance filter was used to filter the descriptor that has variability lower than 0.1 (2591 descriptors collected). Then, high correlation filter was applied to filter descriptors that were correlated more than 0.6. After descriptor selection the model was developed using 335 descriptors.

2.2.4 Model Optimization

The model developed with tree ensemble (TE) statistic method has the following parameters: 9 levels, node size of 1, and 500 trees.

2.2.5 Applicability Domain

Tree ensemble methods provide a confidence value; therefore, to increase the performance of the models, the confidence value was used to improve the applicability domain criteria; thus, only predictions with a confidence higher than or equal to 0.65 were taken into account. To select these parameters, many confidence thresholds were tested on CS, and only results with a coverage higher than 0.6 were taken into account.

3 Results

3.1 Biocide-Like Chemical Space

Different cutoff values for the physicochemical and structural parameters listed in Table 1 were applied to both databases trying to identify a set of common biocide properties. Table 2 lists the properties and cutoff values selected and the number and percentage of compounds for both datasets that did not meet the filter criteria.

3.2 Qualitative QSAR Model for Activated Sludge

A qualitative QSAR model was developed starting from 94 biocide-like compounds by using boosted tree analysis and 8 descriptors selected by sensitivity analysis. The dataset was distributed in a training set (60 compounds) and validation set (34 compounds). The descriptors selected by the best model are shown in Table 3, and the statistics obtained with this model are presented in Fig. 2.

3.3 Quantitative QSAR Model for Activated Sludge

By using the 35 biocide-like compounds considered toxic ($EC_{50} < 100 \text{ mg/L}$) in the dataset, a quantitative QSAR model was developed by using the whole set and a MLR analysis (Table 4). The equation obtained for the quantitative model was:

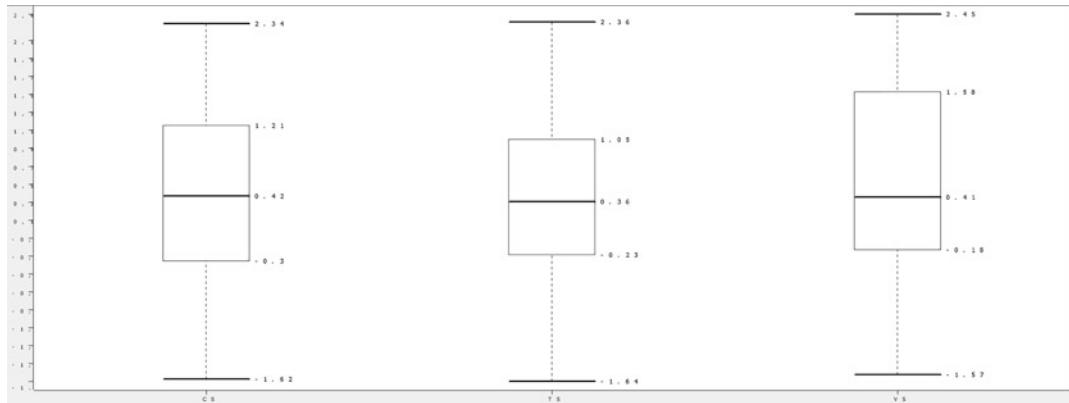


Fig. 1 Box plot of the LC₅₀ distribution in the three sets: TS (training set), CS (calibration set), and VS (validation set)

Table 2

Features and cutoff values maximizing the differences between biocides and generic chemicals

Biocide-like filter	Biocide dataset	Generic organic chemical dataset
Sum of P and S ≤ 2	2 (0.78%)	135 (2.03%)
Number of heteroatoms ≥ 1	2 (0.78%)	263 (3.95%)
Number of rigid bonds ≥ 1	14 (5.45%)	1831 (27.53%)
Number of P ≤ 0	1 (0.39%)	275 (4.13%)
Number of Cl ≤ 3	0 (0.00%)	262 (3.94%)
Total	19 (7.39%)	2766 (42.47%)

$$\begin{aligned} \text{LogEC}_{50} = & 2.28 + 0.05 \text{MinHBint2} - 0.00017 \text{ATSC7v} \\ & + 0.005 \text{VE3_DzZ} + 12.16 \text{AATSC4e} \\ & - 0.13 \text{BCUTp} - 11 \end{aligned}$$

Figure 3 shows visually the adjustment between predicted and experimental values, and Table 5 includes the statistical parameters of the MLR equation.

Furthermore, a leave-one-out cross validation and a fivefold cross validation were carried out. The statistics for the analysis were a Q² _{loo} of 0.69 and a Q² _{5fold-CV} of 0.69.

3.4 Rainbow Trout QSAR Model

The model performances are shown in Table 6.

The performance of the model without confidence value is good; TS has R² of 0.96 because TE must overfit the TS in order to have a good performance (Fig. 4), while CS and VS have comparable performance (R² 0.41 and 0.42; see Figs. 5 and 6, respectively). The performance of the model using the confidence value

Table 3
Descriptors selected by the binary QSAR model

MaxHother	Maximum atom-type H E-state: H on aaCH, dCH2, or dsCH
MinwHBa	Minimum E-states for weak hydrogen bond acceptors
ETA_BetaP_ns_d	A measure of lone electrons entering into resonance relative to molecular size
Gats3c	Geary autocorrelation—lag 3/weighted by charges
MinsCH3	Minimum atom-type E-state: -CH3
ATSC4p	Centered Broto-Moreau autocorrelation—lag 4/weighted by polarizabilities
SpMax1_Bhm	Largest absolute eigenvalue of burden modified matrix—n 1/weighted by relative mass
GATS1i	Geary autocorrelation—lag 1/weighted by first ionization potential

Training set			
QSAR Predictions			
	Non-Toxic	Toxic	Total
Experimental Toxicity (EC₅₀ 3 hours)			
Non-toxic	36 (TN)	1 (FP)	97.29% (Sp)
Toxic	5 (FN)	18 (TP)	78.26% (Sn)
Total (%)	87.80%	94.73%	87.79% (Ac)

Validation set			
QSAR Predictions			
	Non-Toxic	Toxic	Total
Experimental Toxicity (EC₅₀ 3 hours)			
Non-toxic	17 (TN)	5 (FP)	77.27% (Sp)
Toxic	2 (FN)	10 (TP)	83.33% (Sn)
Total (%)	89.40%	66.66%	80.30% (Ac)

Fig. 2 Confusion matrix obtained with the binary QSAR model

threshold is increased. TS has R^2 of 0.99 with a coverage of 0.17 which is normal because the model already overfits. The performance of CS increases at R^2 0.55 on 62% of compounds (Fig. 7), and R^2 VS is 0.51 on 77% of compounds (Fig. 8), so the model prediction associated with the confidence value gives a good prediction for the LC₅₀ endpoint.

Table 4
Descriptors selected by the MLR-QSAR model

MinHBint	Minimum E-state descriptors of strength for potential hydrogen bonds of path length 2
ATSC7v	Centered Broto-Moreau autocorrelation—lag 7/weighted by van der Waals volumes
VE3_DzZ	Logarithmic coefficient sum of the last eigenvector from Barysz matrix/weighted by atomic number
AATSC4e	Average centered Broto-Moreau autocorrelation—lag 4/weighted by Sanderson electronegativities
BCUTp- 11	nhigh lowest polarizability weighted BCUTS

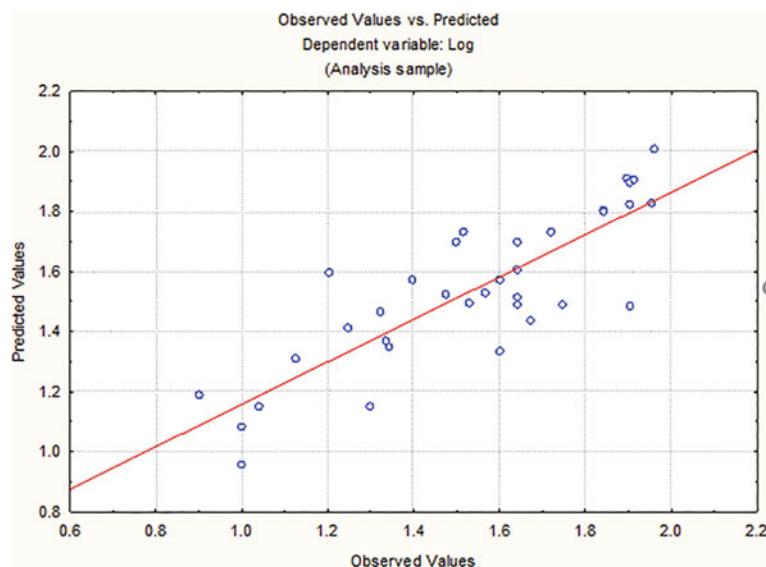


Fig. 3 Graphical representation of the adjustment between predicted and experimental values obtained by MLR

4 Discussion

The need for alternative methods in regulatory toxicology is stronger than ever, considering the legal requirements imposed by different international regulations (e.g., EU-REACH). The application of QSARs in this domain is experiencing an increasing interest worldwide, and in fact the use of such methods is stimulated by the mentioned regulations [6].

Biocidal products are commonly used to protect human beings and the environment against harmful organisms such as pests or bacteria. Nevertheless, biocides and their metabolites or

Table 5
Statistics obtained with the MLR-QSAR model

Quantitative model	Log EC ₅₀
<i>p</i>	0.000001
<i>F</i>	13.96
MS residual	0.03
Df	29
Df model	5
SS residual	0.92
MS	0.44
SS model	2.23
<i>R</i> ²	0.71

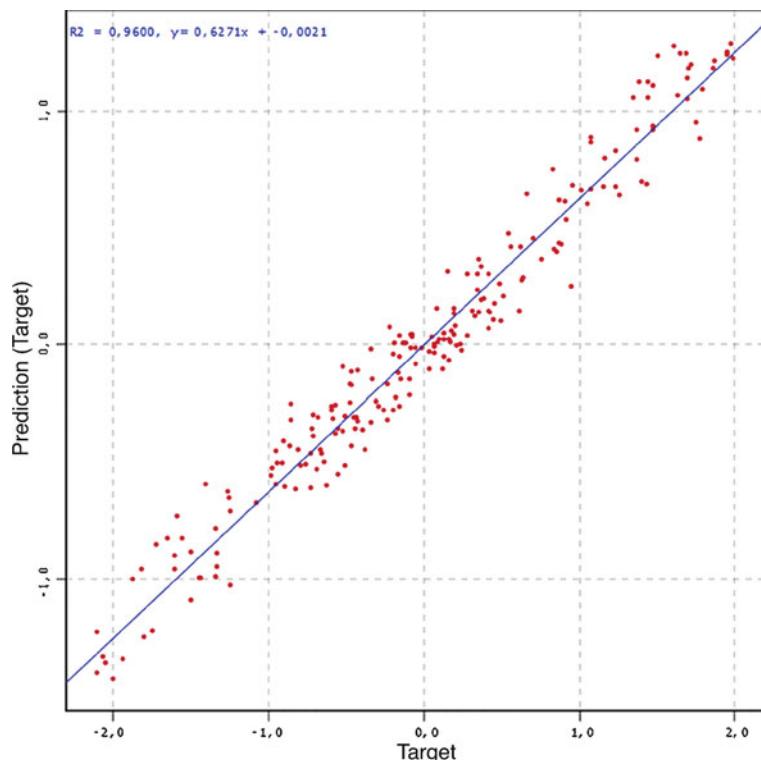


Fig. 4 Scatter plot of the TS prediction. Experimental value on *x*-axis and prediction on *y*-axis

transformation products can also have undesirable adverse effects. In the EU, the commercialization and use of biocidal products depend on the Regulation (EU) No. 528/2012 (commonly

Table 6
Tree ensemble performance to predict LC₅₀ in rainbow trout

	TS	CS	VS
R^2	0.96	0.41	0.42
Coverage	1	1	1

	TS C	CS C	VS C
R^2	0.99	0.55	0.51
Coverage	0.17	0.62	0.77

TS training set, CS calibration set, VS validation set
C means that the confidence value was taken into account

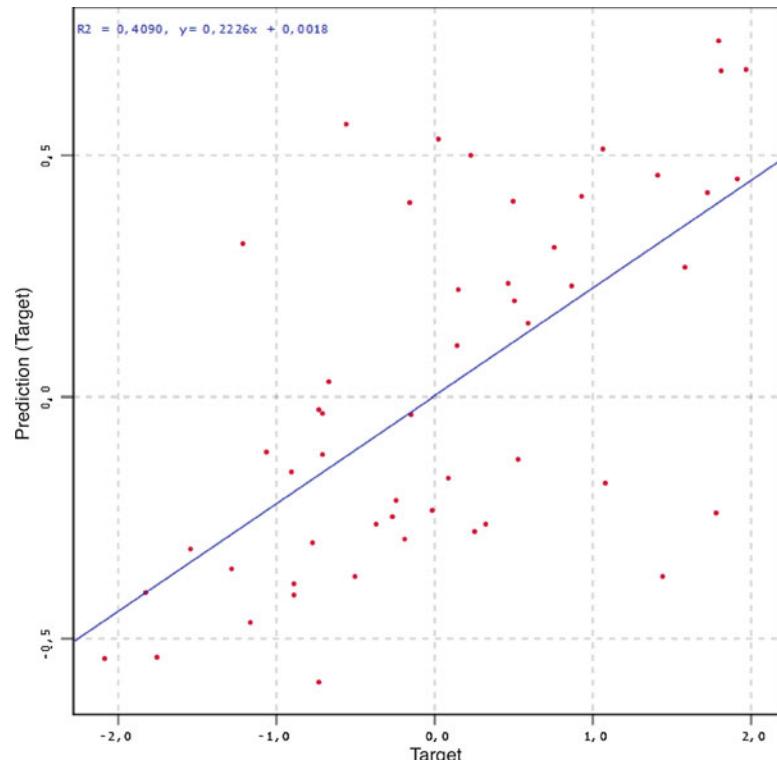


Fig. 5 Scatter plot of the CS prediction. Experimental value on x-axis and prediction on y-axis

known as the Biocidal Products Regulation (BPR)). The purpose of BPR is to harmonize the rules on the supply and use of biocidal products while ensuring a high level of protection of people and the environment. BPR establishes that biocidal products must be authorized before they can be made available on the market and

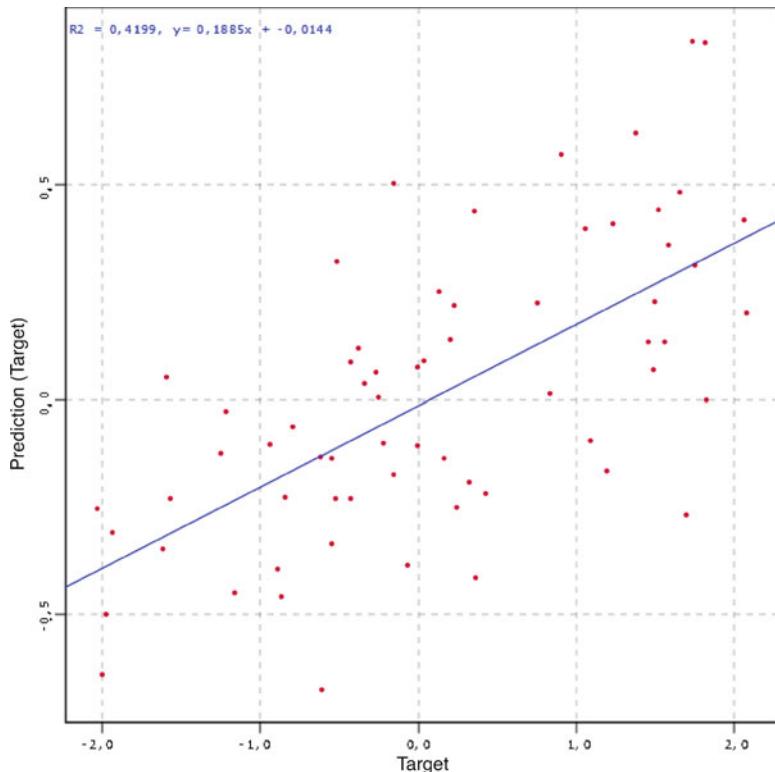


Fig. 6 Scatter plot of the VS prediction. Experimental value on x-axis and prediction on y-axis

used. Although BPR does not completely forbid animal testing, it tries to minimize it as much as possible, for example, by compiling companies to share data on tests on vertebrate animals and by expressly forbidding duplicating such tests. The use of alternative methods is also specifically foreseen in BPR, with special reference to computational methods such as QSARs.

In this context, the project COMBASE represents an opportunity to the increasing use of computational approaches. The models developed for the prediction of aquatic ecotoxicity are made available for free both by the well-known VEGA software and by the specific COMBASE tool, which is more focused on companies concerned by the BPR regulation.

First, an integrated model to predict the respiratory inhibition in activated sludge was developed, by combining sequentially a qualitative and a quantitative QSAR model. A set of 94 chemicals with known EC₅₀ values was selected to this study, based on their “biocide-like” structural features. Thirty-five of them were considered as toxic (EC₅₀ < 100 mg/L), and the other 59 were considered as nontoxic. The set of compounds was randomly distributed in a training set (60 structures) and a validation set (34 structures), and boosted tree analysis was performed yielding an accuracy of

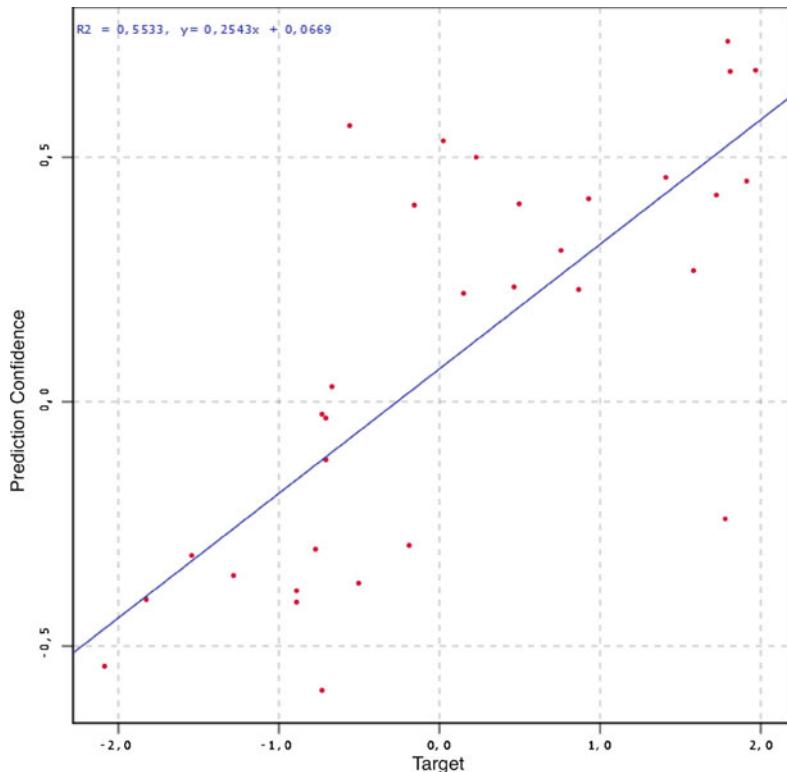


Fig. 7 Scatter plot of the CS prediction. Experimental value on x-axis and only prediction in the applicability domain on y-axis

adjustment between experimental and predicted EC₅₀ of 85% and 80% on the training and test sets, respectively. Consequently, a multiple linear regression (MLR) was carried out using the 35 compounds for which a precise EC₅₀ value was known. Statistics for the model reached a R^2 of 0.71, a $Q^{2\text{ loo}}$ of 0.69, and a $Q^{2\text{ 5fold-CV}}$ of 0.69. From a statistical point of view, probably the worst result observed was the percentage of compounds considered as false positives (i.e., chemicals predicted as toxic that were in fact non-toxic): 5 compounds from 15 (33.33%) in the validation set (Fig. 2). Nevertheless, it is important to consider that in this case, the result would follow the “precautionary principle.”

Second, a model to predict LC₅₀ on rainbow trout was developed on a dataset made by collection data from OpenFoodTox database [15] of the European Food Safety Authority (EFSA) and Pesticide Ecotoxicity Database of Office of Pesticides Programs (OPP) (<https://ecotox.ipmcenters.org/>). From these two databases, a dataset of 393 compounds with LC₅₀ values was extracted. The set of compounds was split into a training set (212) to develop the model, a calibration set (52) to select the best features and to set the best parameter model, and a validation set (65) to validate the model. TS and CS were used to perform the feature selection using

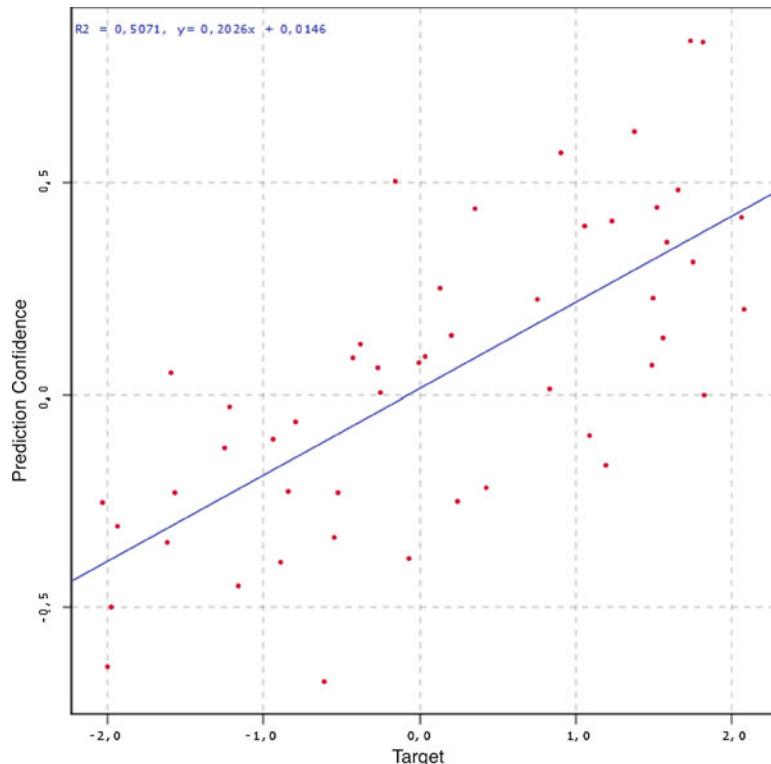


Fig. 8 Scatter plot of the VS prediction. Experimental value on *x*-axis and only prediction in the applicability domain on *y*-axis

low variance filter and high correlation filter methods. The set of features selected was used to develop the tree ensemble methods. The model prediction with confidence value gave satisfactory results; in fact in CS and TS, the R^2 was relatively 0.55 on the 62% of the set (applying the tool to select predictions based on higher confidence) and 0.51 on the 77% of the set.

Those statistics confirmed the robustness of both models (activated sludge and acute toxicity in fish) and their efficiency to estimate aquatic ecotoxicity. Both models are available for free in VEGA, including detailed reports and applicability domain coefficients, as well as in a simplest tool (COMBASE) more adapted to companies concerned by the BPR regulation and without specific personnel with experience or knowledge on chemoinformatics.

References

1. Gheorghe S, Stoica C et al (2019) Ecotoxicity of biocides (chemical disinfectants) – lethal and sublethal effects on non-target organisms. *Revista de Chimie (Bucharest)* 70(1):307–312
2. ECHA (2014) Transitional Guidance on Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products (Biocidal Products Regulation, the BPR). European Chemicals Agency, Helsinki, Finland 2014

3. Guidance on the Biocidal Products Regulation Volume IV Environment – Assessment and Evaluation (Parts B + C) Version 2.0, October 2017
4. Myatt GJ, Ahlberg E et al (2018) In silico toxicology protocols. *Regul Toxicol Pharmacol* 96:1–17
5. Gómez-Ganau S, De Julián-Ortiz JV, Gozalbes R (2018) Recent advances in computational approaches for designing potential anti-alzheimer's agents. Springer Book “Computational modeling of drugs against Alzheimer's disease”. Chapter 2, Pages 25–59 (Series: Neuro-methods, Kunal Roy (ed.), Vol. 132, ISBN 978-1-4939-7404-7)
6. Gozalbes R, de Julián Ortiz JV (2018) Applications of chemoinformatics in predictive toxicology for regulatory purposes, especially in the context of the EU REACH legislation. *Int J Quantitat Struct-Prop Relat* 3(1):1–24
7. Valerio LG Jr (2011) In silico toxicology models and databases as FDA critical path initiative toolkits. *Hum Genomics* 5(3):200–207. <https://doi.org/10.1186/1479-7364-5-3-200>
8. The Organisation for Economic Co-operation and Development (OECD) (2007) Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models. OECD Environment Health and Safety Publications. Retrieved from www.oecd.org/ehs/
9. Organization for Economic Cooperation and Development, Activated Sludge, Respiration Inhibition Test, OECD Chemicals Programme, Ecotoxicological Testing (1981)
10. Willighagen EL, Mayfield JW et al (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9 (1):33
11. Lagorce D, Sperandio O, Galons H, Miteva MA, Villoutreix BO (2008) FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* 9:396
12. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32(7):1466–1474
13. Cherkasov A, Muratov EN et al (2014 Jun 26) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57 (12):4977–5010
14. SANCO/10597/2003 –rev. 10.1 (2012)
15. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics. Wiley-VCH Verlag GmbH & Co. KGaA
16. StatSoft, Inc. (2007) STATISTICA (data analysis software system), version 8.0. <http://www.statsoft.com>
17. Cheng H, Garrick DJ, Fernando RL (2017) Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *J Anim Sci Biotechnol* 8:38. <https://doi.org/10.1186/s40104-017-0164-6>. eCollection 2017. PubMed PMID: 28469846; PubMed Central PMCID: PMC5414316
18. Dorne JL et al (2017) EFSA (European Food Safety Authority), 2017. OpenFoodTox: EFSA's open source toxicological database on chemical hazards in food and feed. *EFSA J* 15 (1):e15011. [3 pp.]. <https://doi.org/10.2903/j.efsa.2017.e15011>
19. OECD 203. OECD (1992) Test no. 203: fish, acute toxicity test, OECD guidelines for the testing of chemicals, section 2. OECD Publishing, Paris. <https://doi.org/10.1787/9789264069961-en>
20. Berthold MR, Cebron N, Dill F, Gabriel TR, Köller T, Meini T, Ohl P, Sieb C, Thiel K, Wiswedel B (2007) KNIME: The Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds.) Data Analysis, Machine Learning and Applications – Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V (GfKL 2007), Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, Germany, pp 319–326
21. Kode (2016) Kode srl, Dragon (software for molecular descriptor calculation) version 7.0.4. 2016, software available at: <https://chm.kode-solutions.net>



Chapter 18

QSAR Modeling of Dye Ecotoxicity

Simona Funar-Timofei and Gheorghe Ilia

Abstract

Dyes have a long history, beginning in ancient times, when natural plant and insect sources were used to create them. Although they are important components in our daily lives, various synthetic dyes have been found to be toxic to human, animal, and plant health, and to the environment. The complexity and costs of the experimental methods used to study dye toxicity and to treat dyeing wastewaters have guided researchers to study theoretical alternative (nonanimal) approaches, which are less expensive. Quantitative structure–activity/property relationship (QSAR/QSPR) techniques can be used for complementary knowledge of in vivo effects in both animals and humans, and for avoiding expensive experimental effluent treatment processes. In this chapter, QSAR/QSPR methods employed in the estimation of dye ecotoxicity are presented. QSAR models reported in the literature for acute toxicity, allergenicity, mutagenicity, carcinogenicity, degradation products, animal and plant toxicity, abiotic degradation and decoloration, bioelimination and bioreduction, and adsorption removal of dyes are analyzed and discussed. QSAR/QSPR techniques combined with virtual screening approaches constitute a powerful tool in the design of dyes with improved ecotoxicological properties.

Key words Dyes, Ecotoxicity, QSAR, QSPR, Toxicity, Allergenicity, Mutagenicity, Carcinogenicity, Ecology, Biodegradation

1 Introduction

Dyes have been known of for a long period of time, and they play an important part in our daily lives. They are substances that impart color to the substrate by temporarily destroying any crystal structure of the colored substance. Dyes adhere to surfaces by physical adsorption, by mechanical retention, by covalent bond formation, or as complexes with salts or metals. Until the middle of the nineteenth century, natural plants and insects were the sources used for obtaining dyes, but then there was a rapid turn to synthetic manufacturing processes [1]. In 1856, Perkin obtained the first synthetic dye of technical significance, called mauveine. Along with his father and brother, he founded the first factory to produce synthetic dyes. This was the start of the dye and pigment industry. Fuchsine was synthesized in 1856, and production began in 1859.

In 1862, Griess discovered and created diazo compounds and azo dye chemistry. Alizarin, an anthraquinone dye, was synthesized in 1868, the first sulfur dye in 1873, and indigo in 1878 [2].

Dyes can be classified by chemical composition or by application. On the basis of their chemical composition, dyes can be divided into azo, nitroso, nitro, diarylmethane, triarylmethane, anthraquinoid, xanthene, cyanine, acridine, quinone-imine, thiazole, and phthalocyanine dyes. On the basis of their application, dyes can be divided into acid, basic, reactive, direct, disperse, vat, mordant, azoic, and sulfur dyes. However, there is no systematic nomenclature for dyes. To resolve this problem, the Society of Dyers and Colourists and the American Association of Textile Chemists and Colorists have created a Color Index in which each dye has an individual Color Index number [3–6].

Nowadays, there is growing awareness of the damage caused to the environment by the use of dyes and chemicals, some of them being toxic or even carcinogenic. The textile industry is one of the most polluting industries because it requires chemicals. More than 20,000 different substances are used in the textile industry. A finite resource (e.g., water) is used at every step of textile processing, during which it is saturated with chemical additives. The process pollutes the environment by the heat of the effluent, its increased pH, and its content of dyes, defoamers, bleaches, detergents, optical brighteners, equalizers, and many other potentially harmful chemicals [7].

Azo dyes, which constitute a significant proportion of textile dyes and part of the dyes used for coloring, are discharged into the environment. Therefore, there is a need for development of non-genotoxic dyes and investment in research to find treatments for the water that is discharged [8]. Discharge of most azo dyes into the environment is undesirable because of their breakdown products (e.g., benzidine), which are toxic to aquatic life and mutagenic in humans [9].

The toxicity of azo dyes is well known. Some azo dyes can induce splenic sarcomas, bladder cancer, and hepatocarcinoma, or can cause a chromosomal aberration in mammalian cells. Malachite green has an effect on the immune and reproductive systems, and is a potential genotoxic and carcinogenic agent. CI Disperse Blue has been shown to cause frameshift mutation and base pair substitution in *Salmonella*. The genotoxic and cytotoxic effects of this dye on human cells have also been studied [10]. Azo dyes raise potential environmental concerns, considering their toxic, mutagenic, and carcinogenic effects [11, 12]. As the discharge of azo dyes into water bodies presents human and ecological risks, a few synthetic dyes have been tested to evaluate their potential toxicity. The results have shown that these dyes have toxic effects on a variety of organisms such as aquatic animals [13].

Dyes have high stability to light or temperature and are usually resistant to environmental degradation. This is the reason why they persist for a long time when they are discharged into the environment [14]. Toxicity of dyes is due to the direct action of the original compound or its intermediate metabolites such as naphthalene, benzidine, and other aromatic amines. Those compounds are by-products of cleavage of the azo bond by microorganisms and have been reported to be cytotoxic, genotoxic, carcinogenic, and mutagenic, increasing the incidence of bladder cancer. These by-products can be more dangerous than the dyes themselves, even at low concentrations [15, 16].

To minimize environmental damage and protect users and consumers against the toxicological impact of dyestuffs, the Ecological and Toxicological Association of the Dyestuffs Manufacturing Industry (ETAD) was set up in 1974. ETAD showed that from a total of approximately 4000 tested dyes, more than 90% showed median lethal dose (LD_{50}) values above 2×10^3 mg/kg, the most toxic being the groups of basic and direct diazo dyes [17, 18].

The quantitative structure–activity relationship (QSAR) is one of the major computational tools employed in environmental sciences. QSAR theoretical models relate a quantitative measure of a chemical structure to its physical property or biological activity, and are based on the principle that structurally similar chemicals are likely to have similar physicochemical and biological properties. QSARs correlate the activity of compounds with physicochemical properties and/or structural descriptors, and they can reduce or even replace the need for animal testing. These methods are especially applicable to acute ecotoxicity, but the greatest challenge is to predict chronic effects [19]. With respect to dyes, researchers and industrialists have focused their attention on the special properties of the different type of dyes. The most important properties are color, brightness, solubility, affinity, fastness, absorption maxima, mutagenicity, the diffusion constant, lipophilicity, bioelimination, and antimalarial activity. These characteristics can be assessed experimentally, using methods that require reagents and complex equipment, and are time consuming. The complexity of the experimental methods has guided researchers to investigate the relationships between the structure and properties or activities of dyes, using theoretical methods. These methods are accurate but are questionable for large molecules such as dyes. Quantitative structure–activity/property relationship (QSAR/QSPR) approaches can overcome these drawbacks because they can be extended to larger molecules included in large data sets, with quite good accuracy. These methods have been applied in pharmaceutical chemistry, environmental chemistry, and toxicology, to improve biological activities and physicochemical properties. Several QSAR studies of dye adsorption by cellulose fiber have been reported [20–31]. In this chapter, we present QSAR models described in the literature for the modeling of dye ecotoxicity.

2 A Short Introduction to QSAR Methodology

Since the beginning of the QSAR field in 1962 [32], thousands of QSAR and QSPR works have been reported, using a large range of end points and approaches [33]. Several published papers and books have presented guidelines on accurate approaches to the QSAR/QSPR process [19, 34–39]. QSAR modeling is based on the fundamental statement that the structure of a molecule (i.e., its geometric, steric, and electronic properties) contains the characteristics responsible for its physical, chemical, and biological properties [40]. By employing QSAR models, the biological activity of a new or untested chemical can be deduced from the molecular structure of “similar” compounds whose activities (properties) have already been experimentally assessed.

Toxicity-based QSAR models require effective knowledge of chemistry, toxicology, and statistics [41]. Knowledge of the toxicological and chemical information on which the QSAR model is based is essential for accurate prediction of toxicity and the quality of a QSAR model. Usually, three components are required for the creation of a new QSAR model [40]: (1) experimental biological activity or property data determined for a group of chemicals (i.e., the training set); (2) molecular structure and/or property data (i.e., the structural descriptors) for this group of chemicals; and (3) statistical approaches to find and validate the relationship between these two sets. Each method has advantages, disadvantages, and practical constraints [41].

The ideal QSAR model should (1) consider an adequate number of training molecules for sufficient structural diversity; (2) have large limits (of several orders of magnitude) of the dependent variable values in the case of regression models and an satisfactory distribution of molecules in each class for pattern recognition models; (3) be applicable for obtaining reliable predictions of new untested chemicals (thus, there is a need for validation and applicability domain tools); and (4) if possible, allow mechanistic information on the modeled end point to be obtained [40]. The mechanistic approach is essential for descriptive QSAR modeling but is less important for predictive QSAR modeling, which uses the statistical technique. The core of QSAR modeling lies in the statistical methods that are applied to relate the response (the dependent variable) to the molecular descriptors (the independent variables). A molecular descriptor is usually encoded into a useful number generated from a standardized experiment or a mathematical procedure. The central part of QSAR modeling resides in the statistical methods that are applied to relate the response to the molecular descriptors.

Several methods are used to derive mathematical models that relate the modeled end point to the chemical structure [36, 40]:

(1) supervised learning approaches such as multiple linear regression (MLR), discriminant analysis (DA; including linear DA, quadratic DA, and regularized DA), partial least squares (PLS), soft independent modeling of class analogy (SIMCA), factor analysis, canonical correlation analysis, principal component regression (PCR), classification and regression trees (CART), neural networks, adaptive least squares, genetic programming, and logistic regression; and (2) unsupervised learning approaches such as principal component analysis, cluster analysis, nonlinear mapping, k -nearest neighbors (KNN), correspondence analysis, Kohonen mapping, and self-linear learning machine organizing mapping (SOM).

The most common approaches used as multivariate methods for regression analysis applied in QSAR modeling are MLR, PCR, and PLS [40].

Pattern recognition methods such as DA, factor analysis, and cluster analysis can be used to complete or replace regression analysis in the development of QSARs for several toxic responses (e.g., nonpolar narcosis, polar narcosis, and uncoupling of oxidative phosphorylation) [42]. For each grouping identified by a pattern recognition technique, one can use multiple regression analysis to develop quantitative predictions of toxicity. This approach takes advantage of the best features of multiple regression analysis (e.g., generation of quantitative predictions and easy interpretation of the model) and reduces the effects of its limitations (e.g., overfitting caused by trying to have one model fit a large, complex data set). It also reduces some of the subjectivity inherent in assigning a substance to a particular class or group.

Lately, nonlinear models using neural networks, genetic algorithms, or hybrids of these two approaches have been used to develop more generalized QSAR models [42]. These generalized models are capable of handling a broad range of chemical structures and properties, functional groups, and modes of toxic action.

The most important value of a QSAR or QSPR is its predictivity—that is, how well it is able to predict end point values of compounds that are not used to develop the correlation (i.e., that are not in the training set) [33]. Two main methods are used to determine predictivity: internal cross-validation and external validation with a test set of compounds. It is generally accepted now that only QSARs and QSPRs that have been suitably externally validated can be considered reliable for both scientific and regulatory purposes [43].

Some guidelines for estimation of the validity of QSARs for regulatory purposes were proposed in 2002, as the “Setúbal principles”, at an international workshop in Setúbal, Portugal [33, 37]. Two years later, in 2004, they were modified by the Organisation for Economic Co-operation and Development (OECD) Work Program on QSARs as the *OECD Principles for*

the Validation, for Regulatory Purposes, of (Quantitative) Structure–Activity Relationships Models [44, 45]. The corresponding OECD guidelines for a valid QSAR/QSPR model are as follows: “To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; (5) a mechanistic interpretation, if possible” [46].

The applicability domain is a theoretical region in the physico-chemical space (the response and chemical structure space) for which a QSAR model should make predictions with given reliability [40]. This region is defined by the nature of the chemicals in the training set and can be characterized in various ways [47]. Not even a robust, significant, and validated QSAR model can be expected to reliably predict the modeled property for the entire universe of chemicals. In fact, only predictions for new chemicals falling within the model domain can be considered reliable, such as those for training chemicals and not extrapolations of the model.

QSAR models find broad applications for assessing potential impacts of chemicals, materials, and nanomaterials on human health and ecological systems [37]. For regulatory identification of potential health hazards, screening, and prioritization, predictive QSAR models of *in vivo* effects in both animals and humans are of great interest, implying specialized regulatory tools and databases, for model development and validation. Within the European Union (EU), the New Chemicals Policy of the European Commission (Registration, Evaluation, and Authorization and Restriction of Chemicals (REACH)) has proposed a new system for managing chemical information in a single regulatory framework.

In the field of toxicology, QSAR methods typically use toxicity end points from *in vitro* cell cultures or *in vivo* animal test systems, for which the mechanism of action is less well understood [37]. Other significant challenges are related to the chemical knowledge used for model building (i.e., the training set) and the chemical space to which models will be applied (i.e., the prediction space) and for what aim (mechanism elucidation, screening, prioritization, safety assessment, etc.).

In toxicity QSAR modeling, several conditions are needed for successful modeling [37]: (1) there are similar structures in the training set, considering a single target-mediated mechanism; (2) the toxicity dependent variable to be modeled is either non-target-specific or related to chemical reactivity principles; (3) the toxicity end point is connected to a well-defined molecular target or phenotype; and (4) toxicity data are available for a satisfactorily large number of diverse chemicals.

Despite the great promise of computational toxicology approaches, there continue to be areas of chemistry and chemical

risk assessment in which relevant test compounds are unavailable (such as in the early phases of chemical design or premanufacture review) or quantitative high-throughput screening test results are unattainable with current technologies (e.g., volatiles, reactives, insolubles, and metabolites) [37]. QSAR methods are being progressively more used in screening, testing prioritization, pollution prevention initiatives, green chemistry, hazard identification, and risk assessment.

To minimize the possible problems that might arise in QSAR model development and validation, several conditions need to be fulfilled [48]. The toxicological end points must be diverse, be reliable, be of high quality, and reflect well-defined and continuous data. The compound structural descriptors should be of high quality and reproducible. Whenever possible, a mechanistic interpretation of the QSAR would be an advantage. The statistical procedures that are used should be as accurate as possible and suitable for the end point being modeled, to allow for development of models that are as easily interpretable as possible. Validation of the developed QSAR model is important only within the descriptor space and the applicability domain.

3 Application of QSAR Models to Dye Ecotoxicity

3.1 QSAR Models for Dye Toxicity

3.1.1 QSAR Models for Acute Dye Toxicity

Acute toxicity refers to effects that take place within a short period of time after short-term exposure—for example, a single oral administration [49]. Generally, dyes have low acute oral toxicity. It has been noted that more than 80% of dyes have an LD₅₀ (rat, oral) value greater than 5000 mg/kg. Only 15 dyes (fewer than 1%) have LD₅₀ (rat, oral) values less than 250 mg/kg.

In a quantitative structure-toxicity relationship (QSTR) study of dye acute toxicity, the mouse intraperitoneal LD₅₀ values were related to the dye structural calculated parameters, using the PLS approach [50]. Twenty-eight dyes were modeled using molecular mechanics calculations and 821 zero-dimensional (0D), one-dimensional (1D), and two-dimensional (2D) descriptors were computed from the optimized structures, using Dragon software. An acceptable three-component PLS model with predictive power (checked according to the criteria proposed by Alexander et al. [43]) was obtained: $R^2 X(\text{Cum}) = 0.559$, $R^2 Y(\text{cum}) = 0.912$, $Q^2(\text{Cum}) = 0.514$). Five compounds were included in the test set. The dye toxicity values were increased by structural parameters, including the number of aliphatic secondary carbon atoms, the number of nonaromatic carbon atoms, the number of aliphatic tertiary carbon atoms, the number of positively charged nitrogen atoms, and the number of tertiary aromatic amines. The favorable features for lower toxicity were the increased numbers of secondary aromatic amines, number of donor atoms for hydrogen bonds and of the hydroxyl groups in the dye molecule, and dye hydrophilicity.

3.1.2 SAR Models for Dye Sensitization

Skin sensitization, also known as contact sensitivity or allergic contact dermatitis, is a T-lymphocyte-mediated delayed hypersensitivity reaction [51]. Certain textile (disperse, azo, and anthraquinone) dyes can cause contact dermatitis (e.g., “panty hose syndrome,” which is associated with the wearing of close-fitting athletic and fashion wear, such as velvet leggings). Oxidative hair dyes (especially those containing arylamines, e.g., *p*-phenylenediamine) have also been found to cause allergic skin reactions [51–54].

Cluster analysis was used in a QSAR model to identify the sensitization strength (expressed by experimental local lymph node assay data) of hair dyes, by means of their chemical structures [55]. Simplified molecular-input line-entry system (SMILES) notation was used to remove the duplicate structures. A cluster analysis was performed for grouping the dyes according to their chemical similarity, based on topological substructural molecular design (TOPS-MODE) descriptors [56]. The algorithm of *k*-means, with 10 clusters, was applied. Each cluster contained other “similar” chemicals. The TOPS-MODE QSAR model was used to estimate the likely sensitization potency in one of three bands: (1) strong/moderate sensitizers, (2) weak sensitizers, and (3) extremely weak sensitizers or nonsensitizers. Most (75%) of the 229 identified hair dyes were predicted to be strong/moderate sensitizers, 22% of the hair dyes were predicted to be weak sensitizers, and 3% were predicted to be extremely weak sensitizers or nonsensitizers. Eight of the most commonly used hair dyes were predicted to be strong/moderate sensitizers, including *p*-phenylenediamine, which is the most commonly used hair dye allergy marker in patch testing. These results are useful to improve the diagnostic work-up of hair dye allergy cases.

As part of the process of designing new hair dyes with desired properties, Williams et al. [57] performed a cheminformatics study to determine the skin sensitization potential of these compounds. Several physicochemical descriptors were computed for these compounds, using the KNIME Analytics platform [58]. The hierarchical clustering approach was then applied to obtain the chemical similarity of the studied compounds included in a created Hair Dye Substance Database [59]. A QSAR model was created using random forest and 2D descriptors computed for all of the chemical structures of the training set compounds (440 for the murine local lymph node assay used for the in vivo test for skin sensitization). The prediction performances of this model were sensitivity = 87.4% and specificity = 48.0%. The calculations using this model were conducted as an individual and consistent KNIME work flow. The Pred-Skin application was applied to each of the potential hair dyes for skin sensitization potential (human) predictions [60]. Most of them (269 hair dyes) were predicted to be sensitizers (and of these, 109 have been banned from use in hair dye products in the EU), 74 were considered to be sensitizers or to have some type of

sensitization potential, and 27 were not accepted as sensitizers; the sensitization potential of the remaining 12 substances was not reported. In the development of the QSAR model, the following steps were considered: (1) data curation/preparation/analysis (selection of compounds and descriptors), (2) model building, and (3) model validation/selection. A fivefold external cross-validation procedure was used. The best models were identified and selected according to acceptable threshold values of the correct classification rate (CCR, computed as the average of the sensitivity and specificity of the model) for the internal test sets (called an out-of-bag set in the random forest; *vide infra*). Then selected models were applied to the external set compounds to predict their experimental properties. In addition, 1000 rounds of Y-randomization were performed for each data set to ensure that the high accuracy of the models built with real data was not due to chance correlations.

3.1.3 QSAR Models for Dye Mutagenicity

Mutagenic dyes, which are very stable in the aquatic environment, have been found in several rivers [61]. Most carcinogens act through mutagenic mechanisms [62]. Azo dye carcinogenicity implies an intact azo linkage. In contrast, most mutagenicity studies involve azoreduction as a condition for activity. Most benzidine-based dyes are nonmutagenic unless some form of external activation system is used [63]. A few dyes (e.g., Direct Blue 15 (which is mutagenic for TA1538) and Direct Brown 31 (which is mutagenic for TA98)) have been reported to be mutagenic without exogenous activation. Under oxidative conditions, bacterial metabolism alone cannot generate mutagenic metabolites for the majority of these dyes. Many of these dyes become mutagenic through conversion via mammalian metabolism and can produce mutagenic metabolites after metabolism in a reductive and oxidative system.

Sushko et al. [64] used a QSAR model to predict the mutagenicity of dyes, using a training set of 4361 compounds, tested against the Ames mutagenicity test [57]. The same procedure described in Subheading 3.1.2 was used. The resulting models were selected on the basis of their prediction performances (sensitivity 79.5% and specificity 80.5%) and identified 30–60% of compounds with an accuracy of prediction similar to the interlaboratory accuracy of 90% in the Ames test.

A QSAR model for the mutagenic activity in the *S. typhimurium* TA98 bacterial strain with S9 activation of 43 aminoazobenzene dyes was studied using MLR and an artificial neural network (ANN) [65]. Geometric, electrostatic, quantum chemical, and hydrophobic descriptors were derived using the CODESSA software. The models were selected on the basis of the highest value of the squared regression coefficient and contained noncollinear descriptors, as determined by the pair correlation matrix. A five-descriptor MLR model was built, taking into account 85% of the

variation in the mutagenic activity. A better three-descriptor ANN model had a better result and accounted for over 90% of the variation in mutagenic activity. It was concluded that the ANN models predicted fewer false positives than the MLR models. For the same set of dyes, the comparative molecular field analysis (CoMFA) approach was used [66], with the most bioactive compound, as a template. Better statistical results were obtained in the CoMFA model including only the steric field. An electronegative and bulky group at the benzene ring and a small group attached to the aniline ring were considered to be favorable to reduce the mutagenic activity.

In another QSAR study on the same set of 43 dyes included in Garg et al. [65], three approaches were employed to study mutagenicity: the hologram quantitative structure-toxicity relationship (HQSTR), CoMFA, and comparative molecular similarity index analysis (CoMSIA) [67]. Similar results were obtained by these three methods. Fragment and donor-acceptor descriptors were included in the HQSTR model (a bulky group on the acceptor ring and a small group on the donor ring were found to be responsible for the mutagenicity decrease). In CoMFA and CoMSIA, steric effects (by bulky moieties on the acceptor ring and small groups ortho-attached to the terminal amine function) and electrostatic effects (by negative groups) were important in modeling the mutagenicity.

The computer automated structure evaluation (CASE) methodology was used to study the mutagenicity of phenylazoanilines [68] and 1-amino-2-naphthol derived azo dyes [69]. In this approach, the descriptors were selected automatically from a learning set composed of active and inactive molecules. The structural descriptors were activating (biophore) or inactivating (biophobe) single and continuous fragments. Once the training set was assimilated, CASE could be queried regarding the predicted activity of molecules of unknown activity. Thus, entry of an unknown chemical would result in the generation of all possible fragments ranging from 2 to 12 atoms accompanied by their hydrogens, and these would be compared with the previously identified biophores and biophobes. On the basis of the presence and/or absence of these descriptors, CASE predicted activity or lack thereof. In addition, CASE also used the descriptors to perform a multivariate regression analysis (QSAR) in which the activity was related to the biophores and biophobes (fragments having $\geq 90\%$ probability of being associated with mutagenicity were considered). Each of the biophores and biophobes was characterized by a likelihood of being associated with mutagenicity (which was derived from their distribution among mutagenic, marginally mutagenic, and nonmutagenic molecules in the database), with a confidence level. They were used to predict the probability of mutagenicity and

nonmutagenicity. It was found that the activity of the dye molecules was dependent upon an intact moiety that spanned the azo linkage; i.e., the azo bond must remain intact for mutagenicity. The study also addressed the effect of sulfonation on the activity of these azo dyes. It was revealed that sulfonation only at certain sites resulted in loss of mutagenicity.

MLR was used in a QSAR study of mutagenic activity in the TA98 + S9 system of 74 aminoazo dyes [70]. The dye structures were modeled using the semiempirical AM1 Hamiltonian. Structural (constitutional, topological, geometric, electrostatic, quantum chemical, and thermodynamic) parameters derived from the optimized structures were related to the mutagenic activity, using MLR and fuzzy logic with ANNs. In the last approach, the algorithm that was employed generated feed-forward network architecture for a given data set, and after generating fuzzy entropies at each node of the network, it switched to fuzzy decision making based on those entropies. Nodes and hidden layers were added as needed until the learning task was accomplished; in this study, the architecture was restricted to a single hidden layer. An MLR model including eight descriptors was chosen as the best model, with a regression correlation coefficient of 0.73. Better statistical results were obtained by the nonlinear model, with a correlation coefficient of 0.95 and a cross-validated correlation coefficient of 0.94.

Fuzzy logic methodology, combined with an ANN with a single hidden layer, was employed to learn and differentiate between mutagenic/carcinogenic and nonmutagenic/noncarcinogenic dyes [71]. Twenty-two azo dyes were optimized using the density functional theory (DFT) approach performed at the BP/DN** computational level. Several descriptors (including topological parameters) were calculated for these dyes. The set of dyes was split into 80% training and 20% test sets. The model generalization ability was checked using a fivefold cross-validation procedure. An algorithm for the creation and manipulation of fuzzy membership functions, which had previously been learned by a neural network from the data set under consideration, was designed and implemented. In this research, membership functions were used to calculate fuzzy entropies for measuring uncertainty and information. An 11-descriptor ANN model was found to describe the mutagenic activity of the dyes.

Two classification strategies using knowledge-based methods (the fragment-based model, which simply considers structural matching of rules sets addressing toxicity; and the joined mechanistic model, an expert system that takes into account a broad range of factors) and docking simulations were used to predict the mutagenicity of 354 azo dyes [72]. A training set of 321 compounds and a test set of 33 azo dyes were used. The classification models were evaluated using Cooper's parameters. The Matthews correlation

coefficient was also used for the quality of binary classification. The X-ray azoreductase structure cocrystallized with the azo dye Acid Red 88 was used for the docking calculations. It was concluded that the integration of multiple strategies and a weight-of-evidence approach might overcome the limitations inherent in single models.

3.1.4 QSAR Models for Dye Carcinogenicity

Azo dyes represent an important proportion of textile dyes and are known to be responsible for carcinogenicity (e.g., causing bladder cancer in humans; splenic sarcomas, hepatocarcinoma, and nuclear anomalies in experimental animals; and a chromosomal aberration in mammalian cells) [9]. The carcinogenicity of 44 organic colorants and benzidine-based dyes was evaluated by the International Agency for Research on Cancer (IARC) in a series of monographs [49]. Azo dyes are considered to be carcinogens if a carcinogenic aromatic amine is formed by reductive cleavage of one or more azo groups. It was found that 150 commercial azo dyes are susceptible to forming aromatic amines recognized to be animal carcinogens, of which 15 are considered to be relevant to the colorants industry.

Thirty-five azo dyes were examined as possible human carcinogens [73]. Azo dyes that were sulfonated on both sides of the azo bonds were considered to be noncarcinogenic in any species. Those that were half sulfonated and half nonsulfonated were sometimes carcinogenic in at least one test species. The inclusion of free alkylated or acetylated amine in the azo structures conferred the carcinogenicity property. In addition, all benzidine-containing and 3,3'-disubstituted benzidine-containing azo dyes were carcinogenic in at least one test species [74]. About 2000 azo dyes have been synthesized so far, and, of those, more than 500 are based on carcinogenic amines and more than 250 azo dyes are benzidine constituted [1].

Hair dyes, which contain one or several “primary intermediates” (e.g., *p*-phenylenediamines, *p*-aminophenols) and “couplers” (e.g., *m*-aminophenols, *m*-hydroxyphenols) are considered to be responsible for bladder cancer in humans because of the connection to aromatic amines [75].

Three publicly available QSAR models (OpenTox/Lazar, Toxtree, and OECD Toolbox) were tested and compared with respect to the carcinogenic potential of a set of color additives [76]. In this study, a data set of 44 color additives, which included approximately equal numbers of carcinogens and noncarcinogens, was employed. The carcinogenicity of these compounds was predicted with a reasonable degree of sensitivity (0.67–0.82). The highest degree of specificity (1.00) was obtained by the OpenTox/Lazar model. The other models overpredicted the carcinogenic potential of the compounds (Toxtree and OECD Toolbox gave specificity values of 0.47 and 0.25, respectively). By comparison, the bacterial

reverse mutation assay (Ames test) had sensitivity of 0.8 and specificity of 1.0 for predicting the carcinogenicity of compounds in this data set.

In a structure–activity relationship study, the photodynamic efficiency and the phototoxicity (expressed as the median inhibitory concentration (IC_{50}) in human epithelial type 2 (HEp-2) cells) against a carcinoma cell line of four xanthene dyes was related to structural dye features [77] in order to find photosensitizers for use in photodynamic therapy. The dye structures were optimized using DFT with use of B3LYP/6-31+G(d) as a basis set, followed by vibrational frequency analysis. The water medium was simulated using the integral equation formulation of the polarizable continuum model. Several reactivity parameters (the highest occupied molecular orbital (HOMO) and lowest occupied molecular orbital (LUMO) energies, the HOMO–LUMO energy gap, chemical hardness, electronic chemical potential, electrophilicity, area, volume, and dipole moment) were calculated for these optimized structures. In addition, the partition coefficient ($\log P$) values were determined spectrophotometrically, using the method of Poocher and Valenzeno [78], as a measure of the hydrophobic character of the dyes. It was concluded that Rose Bengal dye showed higher phototoxicity. Although the other dyes were less effective in killing cells under illumination, they had much lower intrinsic dark cytotoxicity.

Linear DA was used to test the carcinogenicity of 185 dyes by selecting and weighting important parameters [79]. The resulted discriminant score was related to substructural and other parameters with a positive/negative contribution to carcinogenicity, using the MLR approach. In addition, 42 dye structural parameters were calculated: 38 dichotomous and four connectivity indices. The parameters with positive coefficients were considered to be responsible for carcinogenicity, and those with a negative coefficient were considered to reduce it. The variables were ranked from the most to the least important with respect to their power to distinguish carcinogens from noncarcinogens, and then a resubstitution method for validation was applied. The compounds were classified as indeterminates when the probability of carcinogenicity was between $P = 0.3$ and $P = 0.7$; i.e., probabilities too close to chance (0.5) did not distinguish between positivity and negativity. These equations were considered to be useful for the prediction of the carcinogenic potential of untested compounds, rather than for the elucidation of mechanisms of carcinogenesis, with some limitations.

3.1.5 QSAR Models for Dye Metabolites (Aromatic Amines)

The toxicity of a dye is caused especially by its degradation products [15], obtained by the azo linkage breakdown by an enzyme (azoreductase) present in various microorganisms and in all tested

mammals, including humans [1]. Many of the resulted aromatic amines (e.g., benzidine) show a very high level of acute and chronic toxicity, and carcinogenicity [74, 80]. They can cause cancer of the genitourinary tract, pancreas, liver, gallbladder, bile duct, lung, large intestine, stomach, lymphopoiesis, and renal cells, as well as non-Hodgkin's lymphoma. They have been proved to be more dangerous than the parent compound [81]. For solvent-soluble dyes with nonpolar substituents, a solubilizing mechanism must occur in the organism before further degradation and excretion [82].

Computational models—which were directed on the stability of the nitrenium ion, anion formation energy, and hydrophobicity—and expert rule-based models were employed in a study of the mutagenicity of aromatic amines released from the cleavage products of 470 azo dyes used in clothing textiles [83]. At the first step, a modified in silico method was applied to predict Ames activities of primary aromatic amines by calculating the stability of the metabolically intermediate nitrenium ions [84]. A subset of primary aromatic amines was selected that (1) contained no electric charge in the formula, (2) had a molecular weight below 500 Da, (3) had no more than one stereo center, (4) had fewer than 10 rotatable bonds, (5) had only one aromatic amine functionality, and (6) did not contain aromatic nitro groups as they could exhibit Ames toxicity because of their nitro moiety. The cleavage products were downloaded as SMILES structures, salt-stripped, and neutralized, except those with a fixed formal charge (e.g., quaternary ammonium). In particular, carboxylic acids and basic nitrogen were drawn in their neutral forms. The most abundant protomer was selected on the basis of the primary aromatic amine structure. No reassessment of the major protomer form was performed on the nitrenium ions. The three-dimensional (3D) geometry of the structures was optimized using the MMFF94s force field, and the lowest energy conformation of the primary aromatic amine was used in further quantum mechanics calculations. The modeled cleavage products were allocated to one of the following priorities: priority 1 (P1) were potential mutagens to test with the highest priority, priority 2 (P2) were other potential mutagens to test, priority 3 (P3) were substances for which Ames test results could be found in a database or substances for which the prediction was borderline, priority 4 (P4) were substances that were predicted to be nonmutagens, and priority 5 (P5) were substances for which quantum mechanics calculation failed. Cleavage products were clustered according to the substituents found on the aryl ring in order to select substances that were representative of the structural diversity. Substances in the P1 group were selected from structures for which the ArNH^+ formation energy ($\Delta\Delta E_{\text{ArNH}^+}$) was lower than -15 kcal/mol , and that did not contain sulfonic acid, sulfonamide,

sulfonic ester, or 2-aminophenols. Substances with ArNH⁻ formation energy relative to PhNH₂ ($\Delta\Delta E_{ArNH^-}$) of <0 kcal/mol were chosen as being assigned Ames positive. On the basis of the result of this analysis and other criteria, the aromatic amines were assigned to different priority groups. Forty different aromatic amines were identified as potentially mutagenic, primarily in the Ames test, and these are probably released as cleavage products from approximately 180 parent azo dyes. From the 18 substances assigned to priority groups P1 or P2, only four substances (22%) were found to be mutagenic in the Ames screening test.

MLR and ANN QSAR models for the mutagenic activity TA98 + S9 system of 181 aromatic amine derivatives (having at least one amino group) were studied [85]. These compounds were energy optimized at the AM1 computational level. Geometric, electrostatic, quantum chemical, and hydrophobic descriptors were calculated and correlated with mutagenic activity, using MLR and fuzzy logic integrated with ANN approaches. In the last method, the architecture was restricted to a single hidden layer. The generalizability of the models was checked using a jackknife cross-validation procedure. An MLR model with five descriptors resulted, with a squared regression coefficient of 0.66, and the ANN models, which included 10 descriptors, had a correlation coefficient of 0.91.

In a review, Chung et al. presented QSAR models for monocyclic aromatic amine mutagenicity [86]. They concluded that among the calculated structural parameters of these compounds included in the QSAR models, the lowest unoccupied molecular orbital energy (E_{LUMO}), highest occupied molecular orbital energy (E_{HOMO}), and hydrophobicity were important in influencing the mutagenic activity.

The experimental mutagenic potencies of 95 aromatic amines toward a *Salmonella typhimurium* TA98 + S9 microsomal assay were studied by an MLR approach, using the Chebyshev polynomial expansion of the most significant descriptors and back-propagation neural networks [87]. Constitutional, topological, electrostatic, geometric, quantum chemical (derived from AM1 Hamiltonian calculations), and thermodynamic descriptors were calculated for these structures. MLR models in which the mutagenic activity was related to these descriptors were obtained using a forward descriptor selection. Then additional descriptors derived from nonlinear transformations obtained as the first five terms in the Chebyshev polynomial expansion were employed. The MLR models obtained in this way did not improve the model results. Hydrogen bonding, charge distribution, bond energy, and molecular conformations influenced the amine mutagenicity. The best six-parameter MLR model had worse statistical results than the nonlinear back-propagation neural networks model, which included six descriptors.

3.1.6 QSAR Models for Dye Toxicity to Animals and Plants

The complex formed by conjugation between malachite green dye and a lysozyme model protein was studied using computational methods (computer-aided molecular modeling) and experimental methods (steady-state and time-resolved fluorescence, and circular dichroism) [88].

Malachite green is a triphenylmethane dye that is broadly used in many industrial and aquacultural processes, and is associated with environmental and human health problems. The malachite green structure was modeled using the Tripos force field with Gasteiger–Hückel charges, with a gradient of 0.005 kcal mol⁻¹. The lysozyme–malachite green complexation was studied using docking calculations by the Surflex docking program, based on the crystal structure of lysozyme, which was downloaded from the Brookhaven Protein Data Bank (entry codes 6LYZ, resolution 2.0°A; <http://www.rcsb.org/pdb>). It was concluded that the principal forces in the lysozyme–malachite green complex were hydrophobic and π–π interactions, and that the polypeptide chain of lysozyme was partially destabilized upon complexation with malachite green. This information was considered to enhance the understanding of the toxicological action of malachite green in the human body.

Nelms et al. [89] studied the influence of hair dyes on mitochondrial dysfunction, using an *in silico* profiler. This study was based on oral repeat dose toxicity (no observed adverse effect level (NOAEL)) data for 94 hair dye chemicals, which were studied using a similarity analysis. Four categories of hair dyes were identified on the basis of key structural fragments, which were further used to develop a mechanistic hypothesis for the molecular initiating event for each category. Four structural alerts resulted, being related to the ability of aromatic chemicals to disrupt mitochondrial function because of their free radical chemistry, which assigned 56 of the 94 chemicals in the data set to a mechanism-based chemical category. This approach offered points of view for a key molecular initiating event that might be responsible for initiating an adverse outcome pathway paradigm, leading to chronic toxicity.

Decision rule QSAR models were developed to study the uptake of dyes into living cells and organisms [90]. Some of these models allow the prediction of which dyes are likely to enter cells and which dyes will be excluded. QSAR methods were employed in the study of dye intracellular accumulation, redistribution, loss from the cell, and metabolic modification. In these methods for each dye, numerical experimental and calculated structural parameters (e.g., the electric charge, acid dissociation constant (pK_a), solubility, conjugated bond number for the size of the aromatic system, ionic weight for ionic size, and $\log P$ for hydrophilicity/lipophilicity) were considered. Correlations between dye structural parameters and site(s) of their localization in a given cell structure were looked for. The mapping between a region in the parameter space representing a particular combination of physicochemical

properties and an uptake event or an intracellular location site was applied to derive a decision rule QSAR model for one mechanism of uptake, or for one mechanism of localization in a particular cell structure. The validity of such models was checked by exposing cells to previously unevaluated dyes that did (or did not) fall within the region of parameter space correlating with the uptake process or intracellular structure concerned. The prediction of the live cell staining by the dyes was checked using microscopic observations. The precise locations of the limiting boundaries in parameter space that corresponded to the uptake into particular organelles were explored in a similar way. Several ways of dye entry into living cells can be mentioned. For instance, passive diffusion through the plasma membrane is considered to be the simplest mode of dye entry, in which no cell physiological factors need to be considered. This process occurs when a dye can dissolve into the relatively fluid lipid bilayer but does not bind tightly either to lipid or protein membrane components. On the basis of this information, entry by passive diffusion can occur when a dye molecule falls into the following region of parameter space: $8 > \log P > 0$, amphiphilicity index (AI) < 8 , head group hydrophilicity (HGH) > -4 , head group size (HGS) < 400 , conjugated bond number (CBN) < 40 . An electric charge does not prevent passage through the plasma membrane. Ionized substituents do influence entry owing to their effects on the overall hydrophilic/hydrophobic character and on the amphiphilicity of dyes. Dyes that do not interact with cells are unlikely to be toxic, and interactions typically involve dye uptake of one kind or another. The potential hazards of dyes and of compounds metabolically derived from dyes can be determined by predicting the uptake, using QSAR models. Several dyes have been used to study their interactions with cells. The obtained QSAR models were quite limited to predict if the dyes exhibited "uptake" or "nonuptake" by a particular mechanism, or were "localized in" or "not localized in" a particular organelle. These predictions were therefore simplifications, to be regarded as indicative rather than regulatory. They had the advantage that they could use any set of dyes with localization data, even if the mechanism of the localization process was not known. Dyes that do not interact with cells are unlikely to be toxic, and interactions typically imply dye uptake of one kind or another. Consequently, QSAR models for predicting uptake can be used to provide an assessment of the potential hazards of dyes, and indeed of compounds metabolically derived from dyes. As already noted, though, many dyes do enter cells and so are potentially risky. Although the localization QSAR models say nothing directly concerning toxicity, the particular sites of dye localization within a cell may favor or limit subsequent toxic events.

The growth-inhibitory effect of 30 synthetic dyes on 22 strains of Gram-negative bacteria was studied by QSAR models [91]. Principal component analysis, a nonlinear mapping technique, and stepwise linear regression (for relating the dye strength (potency) and dye selectivity of the biological activity and their physicochemical parameter) were used. Stepwise regression analysis showed significant linear relationships between the strength (potency) and selectivity of the biological activity of dyes. The authors concluded that synthetic dyes showed marked biological activity toward both Gram-negative and Gram-positive bacteria. The strength and selectivity of the effect depended equally on the character of the test organisms and the chemical structure of the dyes. It was concluded that the strength of the effect depended on the type of dyes (anthracene, azobenzene, or trityl derivatives), and the hydrophobicity of dyes exerted a significant impact on the strength and selectivity of the biological effect.

3.2 QSAR Models for Dye Ecology

Not all of a dye is fixed on the fabric during the dyeing processes; a fraction of it remains unfixed to the fabric and is washed out [15, 92]. Approximately 10–15% of dyes are released into the environment during the dyeing process, making the effluent highly colored and aesthetically unpleasant [10]. These effluents are rich in dyes and chemicals, some of which are nonbiodegradable and carcinogenic, and pose a major threat to health and the environment.

Unutilized dyes and their metabolites produced during the production process need to be treated before discharge into the environment [93]. Several primary, secondary, and tertiary treatment processes such as flocculation, trickling filters, and electrodialysis have been used to treat these effluents [92]. However, these treatments have not been found to be effective against the removal of all dyes and chemicals used. The effluents contain not only a high concentration of dyes used in the industry but also the chemicals used at the various processing stages.

3.2.1 QSAR Models for Aquatic Toxicity of Dyes

The aquatic toxicity of 42 commercial dyes was analyzed using an in silico approach and ecological bioassays [14]. The chemical similarity (quantified by a similarity index) of these dyes was determined using istSimilarity v.1.0.5 software [94]. The list of the three most similar compounds for each dye was obtained. Dyes from the same chemical class were found to be generally similar. No correlation was found among dyes with the same color. Acute and short-term data were obtained for the water flea *Daphnia magna* and the microalga *Raphidocelis subcapitata*, according to their relative guidelines. In both cases, the assays were able to identify structures with potential ecotoxicity, but the algae were found to be more sensitive to dye toxicity, particularly if the effects on the biomass were considered.

Six QSAR modeling packages: Ecological Structure Activity Relationships (ECOSAR), Toxicity Prediction by Komputer Assisted Technology (TOPKAT), a probabilistic neural network (PNN), a computational neural network (CNN), the QSAR components of the Assessment Tools for the Evaluation of Risk (ASTER) system, and the Optimized Approach Based on Structural Indices Set (OASIS) system were compared for their ability to predict the toxic effects of several substances on biota, especially aquatic biota [42]. A data set of neutral organics, phenols, dinitro phenols, vinyl and allyl halides, esters, phosphate esters, aromatic amines, acrylates, hydrazines, imides, and others were used in the QSAR models to predict the 96-h median lethal concentration (LC_{50}) values in fathead minnows. For each QSAR modeling package, a linear regression analysis of the log of the measured toxicity versus the log of the predicted toxicity was performed. To derive a single measure of model performance, the packages were ranked (1 for the best performer, 6 for the worst performer) against each of seven performance statistics: the number of chemicals for which predictions were generated (except for comparisons where $n = \max$ or $n = \max - 1$ for all packages), mean absolute residual, mean squared residual, percentage of substances with differences between predicted and measured toxicity greater than a factor of 10, correlation coefficient, intercept, and slope. The mean overall rank was then calculated for each QSAR package. The best possible scores for the statistics used for the calculated mean rank would be 0 for the mean absolute residual and mean squared residual (no differences between predicted and measured toxicity values), 0 for the percentage of substances of substances with differences between predicted and measured toxicity greater than a factor of 10, 1 for the correlation coefficient, 0 for the intercept, and 1 for the slope (a perfect linear relationship between the log measured and log predicted toxicity would have an intercept of 0 and a slope of 1). The highest rank for the number of chemicals was given to the package that was able to generate predictions for the largest number of substances under consideration. The last statistic was not an indicator of model performance (in the statistical sense) but was indicative of model utility to users. PNN had the best overall model performance. TOPKAT had excellent model performances for substances within its optimum prediction space. Unfortunately, only 37% of the substances in the testing data set fell within the TOPKAT optimum prediction space, thus limiting its utility in programs that must screen large numbers of chemicals. No recommendations can be made from this analysis regarding the choice of a QSAR model for predictions of chronic toxicity or end points other than mortality.

The photoinduced acute toxicity of a series of anthraquinone dyes toward *D. magna* was studied using the time-dependent density functional theory (TD-DFT) approach [95]. The energy gap

between E_{LUMO} and E_{HOMO} was used to evaluate the photoinduced toxicity. After energy optimization, using the semiempirical PM3 Hamiltonian in MOPAC 2000 software, the optimized structures of the neutral molecules, radical anions, and radical cations of the anthraquinone dyes were optimized using single point and excited energy calculations at the B3LYP/6-31G(d,p) level of theory, with the Gaussian 03 program. The stationary points were checked using frequency calculations. The excited energies were calculated using TD-DFT. The solvent effects were taken into consideration by employing the self-consistent reaction field method with the integral equation of the polarized continuum model (water was used as the solvent). The energy gap between LUMO and HOMO was found to indicate the relative photoinduced toxicity of the dyes. TD-DFT calculations revealed that singlet oxygen and the superoxide anion could be generated through direct energy transfer or autoionization of the excited state of the dyes.

Newsome et al. [96] analyzed the aquatic toxicity of 200 dyes, using QSAR approaches. No QSAR correlations were found between the aquatic toxicity and their physicochemical properties in the case of charged (anionic, cationic, and amphoteric) dyes, for which the nearest analog SAR method was considered to be useful. They concluded that for neutral dyes, QSARs for other chemical classes, such as phenols and anilines, would be helpful to predict their aquatic toxicity. In addition, neutral dyes with molecular weights higher than 1000 daltons or a minimum cross-sectional diameter greater than 10 Å and with three or more acid groups in their structures would have reduced toxicity to fish and daphnids. Dinitro, phenols, and anthraquinones were considered to be toxic functional groups.

QSTR models were proposed for 206 phenols to model their toxicity against the ciliated protozoan *Tetrahymena pyriformis* [97]. MLR combined with a genetic algorithm and classification and regression tree modeling approach was used. The classification and regression tree models gave better results than the MLR models, with respect to the phenol toxicity mechanism of action and prediction.

The MLR approach was employed to study the influence of 96-h toxicity tests on *Chlorella vulgaris* algae for 67 phenols and aniline derivatives that can be used in environmental risk assessment [98]. Low-toxic-effect concentrations—the no-observed-effect concentration (NOEC) and the inhibitory concentration that resulted in 20% cell death (IC_{20})—were predicted using the MLR models. Satisfactory statistical results of the models with predictive power were obtained. Prediction of the IC_{20} was found to be more convenient than prediction of the NOEC, because the reported NOEC values were dependent on the concentrations tested. No mode of action of these compounds could be explained by this approach.

3.2.2 Environmental Fate and Exposure to Dyes

Most of the aromatic polycyclic hydrocarbons released into the environment are exposed to processes such as volatilization, chemical oxidation, bioaccumulation, and adsorption on soil particles [99]. The most important way in which they are eliminated is presently considered to be microbial transformation and degradation.

Several physical, chemical, and biological approaches (e.g., adsorption, coagulation–flocculation, reverse osmosis, oxidation, photodegradation, membrane filtration, and microbial degradation) can be applied for dye removal from wastewater [100]. Biological environmentally friendly methods are becoming increasingly capable and cost-effective in comparison with physicochemical dye removal approaches, which are expensive and have limited adaptability because of the waste products that are generated.

Biodegradation of dyes is an important topic, which can solve the problem of groundwater contamination with organic dyes released into the environment [101]. Lignolytic fungi or bacteria are microorganisms that can transform azo dyes into noncolored products or mineralize them. During the biodegradation process, the azo bond of the dyes is reduced by several bacteria (*Bacillus subtilis*, *Pseudomonas stutzeri*, *Streptomyces* (in aerobic conditions) or *Bacteroides*, *Eubacterium*, or *Clostridium* (in anaerobic conditions), and colorless amines are formed.

QSAR Models for Abiotic Degradation and Decoloration of Dyes

Photocatalysis can be used for dye color removal, mineralization, and toxicity reduction [102]. The efficiency of this method can be verified by measuring the toxicity (using a *Lactuca sativa* L. test) of the dye solution before and after photocatalysis. The apparent color removal rates obtained with the natural dye solution were first simply correlated with 2D calculated dye descriptors. These descriptors do not take into account the actual state of the dye molecule in the wastewater (hydrolyzed molecules, the presence of additives), which can modify the molecular structure. Full mineralization (or transformation into harmless by-products) is the last goal of degradation. It was concluded that some dye structural descriptors could be correlated with the apparent color removal rates at pH ranged between 5.8 and 6.9.

Four ecological water quality parameters—the molar absorption coefficients, photodegradation parameter quantum yields, biodegradability (expressed by the ratio between the 5-day biochemical oxygen demand (BOD_5) and the chemical oxygen demand (COD)) and the toxicity to *Vibrio fischeri* of samples prior to photodegradation and upon achievement of 95% decolorization—were employed in a QSAR study of the photodegradation of nine reactive triazine dyes [103]. The dye structures were modeled using the quantum chemical Austin model 1 (AM1) and modified neglect of diatomic overlap (MNDO) approaches. Quantum chemical dye structural descriptors and other parameters were calculated from these structures and used for correlation with the

ecological water quality parameters, using the variable-selection genetic algorithm combined with MLR methods. One- and two-variable MLR equations resulted. It was concluded that the initial toxicity of triazine dyes depends mostly on the polarizability and aromaticity of the particular dye molecule, while the toxicity of the dye solutions upon achieving 95% decolorization, when formed degradation by-products dominate, depends on the atomic masses, and the aromaticity of the parent dye molecules most probably influences the degradation pathway.

The MLR method was used in the study of decoloration and mineralization of 28 anionic water-soluble azo dyes under visible irradiation [104]. Heterogeneous photo-Fenton dye degradation was carried out using heterogeneous Fenton catalysts, based on amidoximated polyacrylonitrile fiber Fe complexes. Twenty-two dyes were included in the training set and six dyes in the test set. The dye decoloration percentage and the dye total organic carbon removal values were correlated with several dye structural descriptors: $NN=N$ (the number of azo linkages), and NAR (the number of aromatic rings), MW/S (the molecular weight divided by the number of sulfonate groups) and inorganic/organic value (I/O value; inorganic character divided by organic character). The descriptors MW/S and $NN=N$ were found to be the most important determining factors for dye degradation and mineralization. The increase in the values of these last two descriptors decreased the degradation percentage of total organic carbon (TOC) removal. Variation in the Fe content of the catalyst and the addition of sodium chloride did not influence the QSPR model equations.

A QSPR study of the discoloration rate of eight dyes degraded by a Mo–Zn–Al–O catalyst was performed using the PLS approach [105]. The dye structures were energy optimized by the DFT method at the B3LYP/6-31G(d,p) level. Twenty-six structural descriptors were derived from the minimum energy conformers and were related to the dye discoloration rate in catalytic wet air oxidation conditions. Two-component PLS models were obtained with good statistical results. It was concluded that three descriptors—the absolute hardness (η), the dipole moment (μ), and the most negative atomic net charges of the molecule (q^-)—influenced the discoloration rate of dyes by the Mo–Zn–Al–O catalyst.

Thirty-three organic compounds with diverse structures and applications were studied using QSAR models for the degradation of organic pollutants derived by an ozonation process under acidic conditions [106]. The removal ratio and kinetics (reaction rate constants) of these compounds (which also included dyes) were investigated using an ozonation process. These compounds were modeled using the DFT approach at the B3LYP/6-311G level. Several quantum chemical descriptors were calculated from the optimized structures. The compound reaction rate constants derived from the ozone degradation were correlated with these

descriptors, using a stepwise MLR approach. A test set of five compounds was used for external validation, and several criteria for internal and external validation were tested. Models with acceptable statistical results were obtained.

The dye decoloration process in a solution of four dyes was studied experimentally followed by theoretical modeling using a neural network approach [107]. The decoloration process was assessed by an ultraviolet hydrogen peroxide (UV/H₂O₂) process, from which the absorbance was estimated in each dye, as the maximum absorbance wavelength. A feed-forward neural network model was developed, based on hybrid variables, such as calculated structural dye parameters (e.g., the number of azo bonds and sulfonate groups) and process operational variables (such as temperature, initial pH, hydrogen peroxide volume, reactor operation time, and dye concentration). The relative importance of each input neuron in the output neuron was evaluated using the Garson method, which is based on the partition of the neural weights of the hidden and output layers of the neural network. The dye absorbance was considered to be an output variable. The Pearson correlation coefficient values were higher than 0.96 for the training, validation, and test sets, confirming good statistical results of the neural network models.

Degradation by oxidation of four acid dyes was studied using DFT calculations with use of the Coulomb-attenuating method (CAM)-B3LYP functional combined with the 6-31++G(d,p) basis set and the integral equation formalism-polarizable continuum model (IEF-PCM) solvation model in the presence of a water solvent [108]. The dye molecules were optimized and quantum chemical descriptors such as the local Fukui indices (calculated from radical attacks), hardness, dipole moment, and Gibbs solvation free enthalpy were calculated. Then the time-dependent functional density method TD-DFT approach was applied to determine the maximal wavelengths, the oscillator strengths, extinction coefficients, the energies of the excitations and the dipole moments of excitations. It was concluded that higher values of the maximal wavelengths, Fukui indices, and extinction coefficients decreased the molecular stability but increased the reactivity produced by radical attacks. Dye structures with low hardness values (low molecular stability), high wavelength values, and high oscillator strength values were the most susceptible to radical attacks.

QSAR Models for Bioelimination and Bioreduction of Dyes

The aerobic biodegradability of 25 sulfonated azo dyes was studied using DA [109]. The experimental biodegradability data contained four kinds of oxidation rates, including the following enzymes: horseradish peroxidase, peroxidase from *Streptomyces chromofuscus*, and two crude enzyme preparations from *Phanerochaete chrysosporium* (Mn peroxidase and ligninase). These data were treated as the category data, using the principal component approach. Therefore,

an indicator variable was calculated, with a value of 1 for dyes with fast biodegradation and 0 for dyes that are degraded slowly. The set of dyes was divided into two classes: eight dyes that were rapidly biodegradable and 17 dyes that were not rapidly biodegradable. This indicator variable was used as the dependent variable and was related using the MLR approach to other indicator variables, which expressed the biodegradability contribution of the substituents attached to a parent dye structure. Thus, several indicator variables were calculated from the dye structures, taking into account the presence/absence of functional groups, atoms, or fragments attached to a parent dye structure considered to express the biodegradability contribution of these dyes. In addition to this linear group contribution, an additional interaction model, accounting for the possible interactions between the substituent groups on the benzene ring, was developed, using the MLR approach. In this last model, the dependent variable was related to other indicator variables accounting for the interactions between the dye substituent groups. Linear group contribution models with interaction and noninteraction were obtained, respectively, using stepwise regression analysis. Two indicator variables in the interaction model and six descriptors in the noninteraction model were found to be significant for the substituent interaction. It was concluded that the most important structural moieties in controlling biodegradation were the interactions between the hydroxyl group in the para position relative to the azo linkage and its neighboring one- or two-electron-donating substituents of the methyl and/or methoxy group. Also, the sole contribution of the hydroxyl group in the para position was not important and could be omitted in the model.

Two QSAR models, for the fish bioconcentration factor and the octanol/water partition coefficient, were obtained for several compounds (including dyes) of environmental and toxicological interest, which were taken from diverse chemical classes [110]. Structural descriptors were calculated and were related to the aforementioned two dependent variables, using the MLR approach, combined with a genetic algorithm for variable selection. A training set of 290 compounds and a test set of 315 compounds were used in the model development of the bioconcentration factor. The compound polarizability, H-bonding, and chemical dimension were found to be important for the logarithm of the bioconcentration factor modeling. For the modeling of compound hydrophobicity (expressed as the logarithm of the octanol/water partition coefficient), 87 compounds (from which 31 were included in the test set) were used. All models were statistically validated internally (by cross-validation and bootstrap) and externally (by a priori splitting of the available data by a Kohonen map ANN in the training and prediction sets). The applicability domain was verified by the leverage method.

Quantitative structure–biodegradability relationships (QSBRs) were modeled for the biodegradability of 20 acid dyestuffs to find mechanistic explanations, using linear regression calculations [111]. The dye structures were modeled using the PM3 Hamiltonian, and quantum chemical descriptors (E_{HOMO} , E_{LUMO} , and the excited state energy (E_{ES})) and other descriptors were calculated. The dye biodegradability (obtained from a facultative aerobic process) was correlated with the calculated dye descriptors, using one- and two-descriptor linear regression models. The best equation included the molecular weight and E_{HOMO} , as descriptors, which were explained as nucleophilic reactivity and a molecular property.

A quantitative structure–property relationship study of the bioelimination of 103 anionic, water-soluble dyes was performed using the MLR approach [112]. The dye molecules were optimized using molecular mechanics (MM+) calculations (using the Fletcher–Reeves algorithm), and several dye descriptors were derived. The MLR results indicated that the dye bioelimination would be increased by larger molecular size/ionic charge ratios, containing many primary aromatic amines and unsulfonated naphthalene nuclei. The same effect was obtained with a small number of aliphatic alcohol groups.

In a bioremediation study, in silico docking calculations were performed for the study of degradation by laccase and azoreductase of *Aeromonas hydrophila* and *Lysinibacillus sphaericus* of six azo, anthraquinone, and phthalocyanine dyes (Reactive Red F3B, Remazol Red RGB, Joyfix Red RB, Joyfix Yellow MR, Remazol Blue RGB and Turquoise CL-5B) [113]. The color removal was determined using ultraviolet–visible light (UV-Vis) analysis and the biodegradation of these dyes, using *A. hydrophila* SK, was studied by gas chromatography–mass spectrometry (GC-MS) analysis; the BOD and COD removal efficiency was evaluated too. The dyes were docked to the binding sites of the oxidative enzyme laccase and the reductive enzyme azoreductase of *A. hydrophila* and *L. sphaericus* bacteria, using the FlexX docking approach. The docking results were analyzed on the basis of parameters such as stability, catalytic action, and selectivity for enzyme–dye interactions. The docking score of the enzyme–dye interaction was associated with the decolorization percentage. It was concluded that amino acids in the enzymes interacted with several dyes. Several types of dye–enzyme interactions were discussed.

A combined experimental and theoretical study of dye decolorization, using *A. hydrophila* SK16 and *L. sphaericus* SK13, was performed for five azo dyes [114]. Homology models were generated for laccase and azoreductase enzymes. The dye percentage decolorization was experimentally obtained by UV-Vis spectroscopy, high-performance liquid chromatography, Fourier transform infrared spectroscopy, and GC-MS. The final model was built using a target sequence alignment file, as well as the sequence of the

template along with its atomic coordinate file. The binding mode, molecular interaction with active site residues, and binding energy scores were investigated in the docking studies. The docking interactions and the amino acids at the binding site of the proteins interacting with the dyes were observed. The electrostatic interactions, hydrogen bonding, and hydrophobic interactions favored the bond formation between the amino acid residues of the enzymes and dyes. Experimental and theoretical results were compared. The docking score imitated the model of the in vitro decolorization percentage.

QSAR Models for Adsorption Removal of Dyes

A structure–activity relationship model for triarylmethane dye tracers was proposed [115]. Dyes are useful to measure groundwater flow velocity and to identify flow directions, hydraulic connections, and the pattern of water movement. The minimal sorption to soil materials was considered to be an optimal tracer feature. Using a sandy soil, the sorption properties of four dyes were determined experimentally, using the Langmuir isotherm. The resulted maximum adsorption capacity of the medium and the Langmuir coefficients were used as dependent variables. Several descriptors (including molecular connectivity indices) were calculated for these dyes and were correlated with the dependent variables, using a simple linear stepwise regression approach. An optimal triarylmethane water tracer was considered to include 4–6 SO₃ groups. It was concluded that the Langmuir coefficient values were dominated by the molecular size, branching pattern, and positions of substituents. The maximum adsorption might be more related to the interactions between the molecules and the soil medium surfaces than to the size or the shape of the dye molecules.

The mechanism of adsorption of 22 dyes onto activated carbon cloths was studied using the MLR approach [116]. The experimental initial kinetic coefficient (which is related to the initial adsorption rate weighted by the operating conditions) was related to the connectivity indices calculated for the dye molecules. It was concluded that the dye size may be the major structural feature influencing the adsorption rate. The saturation capacity (which is the mean of the steady-state adsorption capacities forming the plateau derived from dye rectangular adsorption isotherms) was chosen as the absorbability parameter and was correlated with the molecular connectivity indices. The adsorption capacities were influenced by structural details of the molecules.

The adsorption onto granulated activated carbon of 33 anthraquinone and azo dyes was studied experimentally and theoretically [117]. The dye molecules were optimized using the PM3 Hamiltonian, and several descriptors were further calculated. The maximum adsorption capacity of the adsorbent obtained from the experimental Langmuir isotherm was used as the dependent variable in the

QSPR models. The semiempirical PM3 method was used for dye structure optimization, and several descriptors were calculated from the minimum energy structures. The dye set was divided randomly into a training set including 25 dyes and a test set with eight dyes. MLR, support vector regression, and back-propagation neural network (using network architecture with three inputs, one hidden layer with three neurons, and one output) methods were used in the QSPR calculations. The neural network results were superior to the MLR results. In the MLR models, the descriptors were selected using a genetic algorithm. The MLR and support vector machine results were similar. The three-descriptor linear and nonlinear models were capable of accounting for more than 70% of the variation in the maximum adsorption capacity of the adsorbent.

4 Conclusions

The textile industry is one of the most polluting industries. Dyeing wastewaters contain nonbiodegradable dyes and substances, which can pose serious threats to the environment and to human, animal, and plant health. The complexity of the experimental methods used to resolve these problems has guided researchers to perform theoretical studies, which are less expensive and alternative (nonanimal) methods. In addition, several synthetic dyes used in daily life have been found to be toxic to human beings and to the environment. Quantitative structure–activity/property relationship (QSAR/QSPR) approaches are useful theoretical tools for avoiding expensive experimental effluent treatment processes and animal toxicity tests. Many supervised and unsupervised learning approaches, as well as virtual screening methods, have been reported in the literature on QSAR/QSPR studies of dye ecotoxicity. Safer dyes with improved properties that make them less toxic to human beings, animals, and the environment can be designed using ligand-based methods combined with structure-based methods.

Acknowledgements

This work was financially supported by Project No. 1.1/2018 of the “Coriolan Dragulescu” Institute of Chemistry of the Romanian Academy.

References

- Bafana A, Devi SS, Chakrabarti T (2011) Azo dyes: past, present and the future. Environ Rev 19:350–370
- Booth G, Zollinger H, McLaren K, Sharples WG, Westwell A (2000) Dyes, general survey. Ullmann’s encyclopedia of industrial chemistry, vol 11. Wiley-VCH, Weinheim, pp 675–729
- Zollinger H (2003) Color chemistry: synthesis, properties and applications of organic dyes

- and pigments, 3rd edn. Wiley-VCH, Weinheim
4. Garfield S (2002) Mauve: how one man invented a color that changed the world? Norton, New York
 5. Colour index online. Available from <http://www.colour-index.org/>. Accessed 15 Jan 2019
 6. Gregory P (1990) Classification of dyes by chemical structure. In: Waring DR, Hallas G (eds) The chemistry and application of dyes. Topics in Applied Chemistry. Springer, Boston
 7. Starovoitova D, Odido D (2014) Assessment of toxicity of textile dyes and chemicals via materials safety data sheets. Res Rev BioSci 9:241–248
 8. Chequer FMD, Dorta DJ, de Oliveira DP (2011) Azo dyes and their metabolites: does the discharge of the azo dye into water bodies represent human and ecological risks? In: Hauser P (ed) Advances in treating textile effluent. IntechOpen, Rijeka, pp 27–48
 9. Alabdulla WMS, Ali Albayati MB (2014) Biodegradation of azo dyes—a review. Int J Environ Eng Nat Resour 1:179–189
 10. Ratna, Padhi BS (2012) Pollution due to synthetic dyes toxicity & carcinogenicity studies and remediation. Int J Env Sci 3:940–945
 11. Sponza DT, Isik M (2005) Toxicity and intermediates of CI Direct Red 28 dye through sequential anaerobic/aerobic treatment. Process Biochem 40:2735–2744
 12. Nagel-Hassemer ME, Carvalho-Pinto CRS, Matias WG, Lapolli FR (2011) Removal of coloured compounds from textile industry effluents by UV/H₂O₂ advanced oxidation and toxicity evaluation. Environ Technol 32:1867–1874
 13. Young L, Yu J (1997) Ligninase-catalysed decolourization of synthetic dyes. Water Res 31:1187–1193
 14. Croce R, Cina F, Lombardo A, Crispenn G, Cappelli CI, Vian M, Maiorana S, Benfenati E, Baderna D (2017) Aquatic toxicity of several textile dye formulations: acute and chronic assays with *Daphnia magna* and *Raphidocelis subcapitata*. Ecotoxicol Environ Saf 144:79–87
 15. Ayadi I, Souissi Y, Jlassi I, Peixoto F, Mnif W (2016) Chemical synonyms, molecular structure and toxicological risk assessment of synthetic textile dyes: a critical review. J Develop Drugs 5:151
 16. Puvaneswari N, Muthukrishnan J, Gunasekaren P (2006) Toxicity assessment and microbial degradation of azo dyes. Indian J Exper Biol 44:618–626
 17. Anliker R (1979) Ecotoxicology of dye-stuffs—a joint effort by industry. Ecotox Environ Safety 3:59–74
 18. Golka K, Kopps S, Myslak ZW (2004) Carcinogenicity of azo colorants: influence of solubility and bioavailability—a review. Tox Lett 151:203–210
 19. Cronin MTD (2017) (Q)SARs to predict environmental toxicities: current status and future needs. Environ Sci: Proc Impacts 19:213–220
 20. Luan F, Xu X, Liu H, Soeiro D, Cordeiro MN (2013) Review of quantitative structure–activity/property relationship studies of dyes: recent advances and perspectives. Color Technol 129:173–186
 21. Polanski J, Gieleciak R, Wyszomirski M (2003) Comparative molecular surface analysis (CoMSA) for modeling dye–fiber affinities of the azo and anthraquinone dyes. J Chem Inf Comput Sci 43:1754–1762
 22. Polanski J, Gieleciak R, Wyszomirski M (2004) Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic mono-azo dyes. Dyes Pigments 62:61–76
 23. Timofei S, Kurunczi L, Suzuki T, Fabian WMF, Muresan S (1997) Multiple linear regression (MLR) and neural network (NN) calculations of some disazo dye adsorption on cellulose. Dyes Pigments 34:181–193
 24. Oprea TI, Kurunczi L, Timofei S (1997) Quantitative structure–activity relationship studies of disperse azo dyes. Toward the negation of the pharmacophore theory of dye–fiber interaction. Dyes Pigments 33:41–64
 25. Timofei S, Fabian WMF (1998) Comparative molecular field analysis (CoMFA) of heterocyclic monoazo dye–fibre affinities. J Chem Inf Comput Sci 38:1218–1222
 26. Timofei S, Schmidt W, Kurunczi L, Simon Z (2000) A review of QSAR for dye affinity for cellulose fibres. Dyes Pigments 47:5–16
 27. Timofei S, Schuurmann G (2002) Comparative molecular field analysis (CoMFA) of anionic azo dye–fibre affinities I: gas-phase molecular orbital descriptors. J Chem Inf Comput Sci 42:788–795
 28. Timofei S, Kurunczi L, Schmidt W, Simon Z (2002) Steric and electrostatic effects in dye–cellulose interactions by the MTD and CoMFA approaches. SAR & QSAR Environ Res 13:219–226
 29. Schuurmann G, Funar-Timofei S (2003) Multilinear regression and comparative

- molecular field analysis (CoMFA) of azo dye—fibre affinities II: inclusion of solution-phase molecular orbital descriptors. *J Chem Inf Comput Sci* 43:1502–1515
30. Kurunczi L, Funar-Timofei S, Bora A, Seclaman E (2007) Application of the MTD-PLS method to heterocyclic dye–cellulose interactions. *Int J Quantum Chem* 107:2057–2065
 31. Funar-Timofei S, Fabian WMF, Kurunczi L, Goodarzi M, Tahir AS, Vander Heyden Y (2012) Modelling heterocyclic azo dye affinities for cellulose fibers by computational approaches. *Dyes Pigments* 94:278–289
 32. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180
 33. Dearden JC, Cronin MTD, Kaiser KLE (2009) How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ Res* 20:241–266
 34. Walker JD, Dearden JC, Schultz TW et al (2003) QSARs for new practitioners. In: Walker JD (ed) QSARs for pollution prevention, toxicity screening, risk assessment, and web applications. SETAC, Pensacola
 35. Walker JD, Jaworska J, Comber MHI, Schultz TW, Dearden JC (2003) Guidelines for developing and using quantitative structure–activity relationships. *Environ Toxicol Chem* 22:1653–1665
 36. Livingstone DJ (2004) Building QSAR models: a practical guide. In: Cronin MTD, Livingstone DJ (eds) Predicting chemical toxicity and fate. CRC, Boca Raton
 37. Cherkasov A, Muratov EN, Fourches D et al (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010
 38. Schultz TW, Cronin MTD, Walker JD, Aptula AO (2003) Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J Molec Structure (Theochem)* 622:1–22
 39. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inf* 29:476–488
 40. Gramatica P (2013) On the development and validation of QSAR models. In: Reisfeld B, Mayeno A (eds) Computational toxicology, Methods in molecular biology (methods and protocols), vol 930. Humana, Totowa, pp 499–526
 41. Cronin MTD, Livingstone DJ (2004) Predicting chemical toxicity and fate. CRC, Boca Raton
 42. Moore DRJ, Breton RL, Macdonald DB (2003) A comparison of model performance for six quantitative structure–activity relationship packages that predict acute toxicity to fish. *Environ Toxicol Chem* 22:1799–1809
 43. Alexander DLJ, Tropsha A, Winkler DA (2015) Beware of R^2 : simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J Chem Inf Model* 55:1316–1322
 44. Environment Directorate, OECD [Organisation for Economic Co-operation and Development] (2004) The report from the Expert Group on (Quantitative) Structure–Activity Relationships [(Q)SARs] on the principles for the validation of (Q)SARs. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2004\)24&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2004)24&doclanguage=en). Accessed 10 Feb 2019
 45. OECD [Organisation for Economic Co-operation and Development] (2007) Guidance document on the validation of (quantitative) structure–activity relationships [(Q)SAR] models. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2007\)2&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en). Accessed 13 Feb 2019
 46. OECD [Organisation for Economic Co-operation and Development]. OECD principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models. <https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>. Accessed 27 Feb 2019
 47. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29
 48. Schultz TW, Cronin MTD (2003) Essential and desirable characteristics of ecotoxicity quantitative structure–activity relationships. *Environ Toxicol Chem* 22:599–607
 49. Clarke EA, Steinle D (1995) Health and environmental safety aspects of organic colorants. *Rev Prog Color* 25:1–5
 50. Funar-Timofei S, Kurunczi L, Vlaia V et al. (2008) Quantitative dye structure–toxicity relationships study by PLS. Paper presented at the 2nd European Computing Conference: New Aspects on Computers Research (ECC'08), Malta, 11–13 September 2008
 51. Uter W, Werfel T, White IR et al (2018) Contact allergy: a review of current problems from a clinical perspective. *Int J Environ Res Public Health* 15:1108. <https://doi.org/10.3390/ijerph15061108>

52. Rovira J, Domingo JL (2019) Human health risks due to exposure to inorganic and organic chemicals from textiles: a review. *Environ Res* 168:62–69
53. Nohynek GJ, Antignac E, Re T, Toutain H (2010) Safety assessment of personal care products/cosmetics and their ingredients. *Tox Appl Pharm* 243:239–259
54. Sosted H, Rustemeyer T, Goncalo M et al (2013) Contact allergy to common ingredients in hair dyes. *Contact Dermat* 69:32–39
55. Sosted H, Basketter DA, Estrada E et al (2004) Ranking of hair dye substances according to predicted sensitization potency: quantitative structure–activity relationships. *Contact Dermat* 51:241–254
56. Estrada E, Patlewicz G, Chamberlain M et al (2003) Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chem Res Toxicol* 16:1226–1235
57. Williams TN, Kuenemann MA, Den Driessche V et al (2018) Toward the rational design of sustainable hair dyes using cheminformatics approaches: step 1. Database development and analysis. *ACS Sustain Chem Eng* 6:2344–2352
58. Berthold MR, Cebron N, Dill F et al (2008) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications. Studies in classification, data analysis, and knowledge organization. Springer, Berlin
59. Alves VM, Muratov E, Fourches D et al (2015) Predicting chemically induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol Appl Pharmacol* 284:262–272
60. Braga RC, Alves VM, Muratov EN et al (2017) Pred-skin: a fast and reliable web application to assess skin sensitization effect of chemicals. *J Chem Inf Model* 57:1013–1017
61. Edwards LC, Freeman HS, Claxton LD (2004) Developing azo and formazan dyes based on environmental considerations: *Salmonella* mutagenicity. *Mutation Res* 546:17–28
62. Levine WG (1991) Metabolism of azo dyes: implication for detoxication and activation. *Drug Metabol Rev* 23:253–309
63. Chung K-T, Chen S-C, Claxton LD (2006) Review of the *Salmonella typhimurium* mutagenicity of benzidine, benzidine analogues, and benzidine-based dyes. *Mutation Res* 612:58–76
64. Sushko I, Novotarskyi S, Korner R et al (2010) Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. *J Chem Inf Model* 50:2094–2111
65. Garg A, Bhat KL, Bock CW (2002) Mutagenicity of aminoazobenzene dyes and related structures: a QSAR/QPAR investigation. *Dyes Pigments* 55:35–52
66. Pasha FA, Dal Nam K, Cho SJ (2007) CoMFA based quantitative structure toxicity relationship of azo dyes. *Molec Cell Toxicol* 3:145–149
67. Pasha F, Muddassar M, Chung HW et al (2008) Hologram and 3D-quantitative structure toxicity relationship studies of azo dyes. *J Mol Model* 14:293–302
68. Rosenkranz HS, Klopman G (1989) Structural basis of the mutagenicity of phenylazo-aniline dyes. *Mutation Res* 221:217–234
69. Rosenkranz HS, Klopman G (1990) Structural basis of the mutagenicity of 1-amino-2-naphthol-based azo dyes. *Mutagenesis* 5:137–146
70. Sztandera L GA, Hayik S et al (2003) Mutagenicity of aminoazo dyes and their reductive-cleavage metabolites: a QSAR/QPAR investigation. *Dyes Pigments* 59:117–133
71. Sztandera L, Trachtman M, Bock C et al (2003) Soft computing in the design of non-toxic chemicals. *J Chem Inf Comput Sci* 43:189–198
72. Gadaleta D, Porta N, Vrontaki E et al (2017) Integrating computational methods to predict mutagenicity of aromatic azo compounds. *J Environ Sci Health C* 35:239–257
73. Brown MA, De Vito SC (1993) Predicting azo dye toxicity. *Critical Rev Environ Sci Tech* 23:249–324
74. Chung K-T (2016) Azo dyes and human health: a review. *J Environ Sci Health C* 34:233–261
75. Bolt HM, Golka K (2007) The debate on carcinogenicity of permanent hair dyes: new insights. *Critical Rev Toxicol* 37:521–536
76. Brown R, White S, Goode J et al. (2013) Use of QSAR modeling to predict the carcinogenicity of color additives. In: Paper presented at the ASME 2013 Conference on Frontiers in Medical Devices: Applications of Computer Modeling and Simulation (FMD 2013), Washington DC, 11–13 September 2013
77. Buck STG, Bettanin F, Orestes E et al (2017) Photodynamic efficiency of xanthene dyes and their phototoxicity against a carcinoma cell

- line: a computational and experimental study. J Chem:Article ID 7365263, 9 pages. <https://doi.org/10.1155/2017/7365263>
78. Pooler JP, Valenzeno DP (1979) Physico-chemical determinants of the sensitizing effectiveness for photooxidation of nerve membranes by fluorescein derivatives. Photochem Photobiol 30:491–498
 79. Enslein K, Borgstedt HH (1989) A QSAR model carcinogenicity: for the estimation of example application to an azo-dye. Toxicology Lett 49:107–121
 80. Osugi ME, Umbuzeiro GA, De Castro FJ, Zanoni MV (2006) Photoelectrocatalytic oxidation of remazol turquoise blue and toxicological assessment of its oxidation products. J Hazard Mater 137:871–877
 81. Wong PK, Yuen PY (1996) Decolorization and biodegradation of methyl red by *Klebsiella pneumoniae* RS-13. Water Res 30:1736–1744
 82. Hunger K (2005) Toxicology and toxicological testing of colorants. Rev Prog Color 35:46–89
 83. Brüschweiler BJ, Merlot C (2017) Azo dyes in clothing textiles can be cleaved into a series of mutagenic aromatic amines which are not regulated yet. Regul Toxicol Pharmacol 88:214–226
 84. Bentzien J, Hickey ER, Kemper RA et al (2010) An in silico method for predicting Ames activities of primary aromatic amines by calculating the stabilities of nitrenium ions. J Chem Inf Model 50:274–297
 85. Bhat KL, Hayik S, Sztandera L, Bock CW (2005) Mutagenicity of aromatic and heteroaromatic amines and related compounds: a QSAR investigation. QSAR Comb Sci 24:831–843
 86. Chung K-T, Kirkovsky L, Kirkovsky A, Purcell WP (1997) Review of mutagenicity of monocyclic aromatic amines: quantitative structure–activity relationships. Mutat Res 387:1–16
 87. Karelson M, Sild S, Maran U (2000) Non-linear QSAR treatment of genotoxicity. Mol Simulat 24:229–242
 88. Ding F, Li X-N, Diao J-X et al (2012) Potential toxicity and affinity of triphenylmethane dye malachite green to lysozyme. Ecotoxicol Environ Saf 78:41–49
 89. Nelms MD, Ates G, Madden JC et al (2015) Proposal of an in silico profiler for categorisation of repeat dose toxicity data of hair dyes. Arch Toxicol 89:733–741
 90. Horobin RW (2014) Where do dyes go inside living cells? Predicting uptake, intracellular localisation, and accumulation using QSAR models. Color Technol 130:155–173
 91. Oros G, Cserhati T, Forgacs E (2003) Separation of the strength and selectivity of the microbiological effect of synthetic dyes by spectral mapping technique. Chemosphere 52:185–193
 92. Hassan MA, El Nemr A (2017) Health and environmental impacts of dyes: mini review. Am J Environ Sci 1:64–67
 93. Xie X, Liu N, Yang F et al (2018) Comparative study of antiestrogenic activity of two dyes after Fenton oxidation and biological degradation. Ecotoxicol Environ Saf 164:416–424
 94. Floris M, Manganaro A, Nicolotti O et al (2014) A generalizable definition of chemical similarity for read-across. J Cheminform 6:39
 95. Wang Y, Chen J, Ge L et al (2009) Experimental and theoretical studies on the photo-induced acute toxicity of a series of anthraquinone derivatives towards the water flea (*Daphnia magna*). Dyes Pigments 83:276–280
 96. Newsome LD, Nabholz JV, Kim A (1996) Designing aquatically safer chemicals. In: DeVito S (ed) Designing safer chemicals, ACS symposium series, chapter 9. ACS, Washington, DC
 97. Abbasifar F, Zare-Shahabadi V (2017) In silico prediction of toxicity of phenols to *Tetrahymena pyriformis* by using genetic algorithm and decision tree-based modeling approach. Chemosphere 172:249–259
 98. Tugcu G, Sacan MT (2018) A multipronged QSAR approach to predict algal low-toxic-effect concentrations of substituted phenols and anilines. J Hazard Mater 344:893–901
 99. Rojo-Nieto E, Perales-Vargas-Machuca JA (2012) Microbial degradation of PAHs: organisms and environmental compartments. In: Singh SN (ed) Microbial degradation of xenobiotics. Springer, Berlin
 100. Mondal PK, Chauhan B (2012) Microbial degradation of dye-containing wastewater. In: Singh SN (ed) Microbial degradation of xenobiotics. Springer, Berlin
 101. Weglarz-Tomczak E, Gorecki L (2012) Azo dyes—biological activity and synthetic strategy. Chemik 66:1298–1307
 102. Byberg R, Cobb J, Diez Martin L et al (2013) Comparison of photocatalytic degradation of dyes in relation to their structure. Environ Sci Pollut Res 20:3570–3581
 103. Kusic H, Koprivanac N, Bozic AL (2013) Environmental aspects on the photodegradation of reactive triazine dyes in aqueous

- media. *J Photochem Photobiol A Chem* 252:131–144
104. Li B, Dong Y, Ding Z (2013) Heterogeneous Fenton degradation of azo dyes catalyzed by modified polyacrylonitrile fiber Fe complexes: QSPR (quantitative structure property relationship) study. *J Environ Sci (China)* 25:1469–1476
105. Xu Y, X-Y C, Li Y et al (2016) Quantitative structure–property relationship (QSPR) study for the degradation of dye wastewater by Mo–Zn–Al–O catalyst. *J Mol Liq* 215:461–466
106. Zhu H, Guo W, Shen Z et al (2015) QSAR models for degradation of organic pollutants in ozonation process under acidic condition. *Chemosphere* 119:65–71
107. Guimaraes OLC, Silva MB (2007) Hybrid neural model for decoloration by UV/H₂O₂ involving process variables and structural parameters characteristics to azo dyes. *Chem Eng Prog* 46:45–51
108. Elhorri AM, Belaid KD, Zouaoui–Rabah M, Chadli R (2018) Theoretical study of the azo dyes dissociation by advanced oxidation using Fukui indices. DFT calculations. *Comput Theor Chem* 1130:98–106
109. Suzuki T, Timofei S, Kurunczi L et al (2001) Correlation of aerobic biodegradability of sulfonated azo dyes with the chemical structure. *Chemosphere* 45:1–9
110. Papa E, Dearden J, Gramatica P (2007) Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. *Chemosphere* 67:351–358
111. Y L, D-L X (2007) Quantitative structure–activity relationship study on the biodegradation of acid dyestuffs. *J Environ Sci* 19:800–804
112. Greaves A, Churchley J, Hutchings M et al (2001) A chemometric approach to understanding the bioclimination of anionic, water-soluble dyes by a biomass using empirical and semi-empirical molecular descriptors. *Water Res* 35:1225–1239
113. Srinivasan S, Sadasivam SK, Gunalan S et al (2019) Application of docking and active site analysis for enzyme linked biodegradation of textile dyes. *Environ Pollut* 248:599–608
114. Srinivasan S, Shanmugam G, Surwase SV et al (2017) In silico analysis of bacterial systems for textile azo dye decolorization and affirmation with wetlab studies. *Clean (Weinh)* 45:1600734. <https://doi.org/10.1002/cle.201600734>
115. Mon J, Flury M, Harsh JB (2006) A quantitative structure–activity relationships (QSAR) analysis of triarylmethane dye tracers. *J Hydrol* 316:84–97
116. Metivier-Pignon H, Faur C, Cloirec PL (2007) Adsorption of dyes onto activated carbon cloth: using QSPRs as tools to approach adsorption mechanisms. *Chemosphere* 66:887–893
117. Orucu E, Tugcu G, Sacan MT (2014) Molecular structure–adsorption study on current textile dyes. *SAR QSAR Environ Res* 25:983–998



Chapter 19

Ecotoxicological QSARs of Mixtures

Pathan Mohsin Khan, Supratik Kar, and Kunal Roy

Abstract

In this era of advanced industrialization, all the living beings and environment are exposed to multicomponent mixtures of different classes of chemicals such as organics, pesticides, heavy metals, and pharmaceuticals which may cause direct or indirect hazards to humans, wildlife, aquatic systems, and ecosystems. The regulatory authorities have mostly relied on the single chemical risk assessment, instead of considering the impact of complex chemical mixtures. It is also well known that toxicity data for the individual components is available for a fraction of all existing chemicals in environment. The condition is much worse as there is minimal toxicity data for complex multicomponent chemical mixtures, and the nature of toxicity of a mixture (synergism and/or antagonism) will be entirely different from the toxicity of the single chemicals. A number of regulatory authorities have proposed several methodologies and guidance for the evaluation of hazardous effects of multicomponent chemical mixtures. However, a standard, significant, and reliable approach for evaluation of toxicity of chemical mixtures and their management across diverse monitoring sectors is lacking. In the present chapter, we have illustrated the basic concepts of mixture toxicity assessment, such as concentration addition, independent action, and interaction (synergism and/or antagonism), as well as focused on the computational approaches, such as quantitative structure-activity relationship (QSAR), which is already proven as an efficient alternative method for toxicity prediction of chemicals by regulatory authorities for decision making. Subsequently, we have also provided a brief detail on several ongoing research projects in the European Union (EU), funded by the current European Research and Innovation Programme Horizon 2020 or the Seventh Framework Programme for mixture toxicity prediction. The present chapter also explains the importance of evaluation of chemical mixture toxicity and essential steps in basic QSAR modelling in the context of mixtures. Additionally, we have reported the successful application of QSAR in the prediction of mixture toxicity of different classes of chemicals such as pharmaceuticals, pesticides, metals, and organic industrial chemicals.

Key words Component-based approach, EuroMix, Human risk assessment (HRA), EUToxRisk, Generalized concentration addition (GCA) models, Interactions, Mixture toxicity assessment, QSAR of mixtures

1 Introduction

The entire world of living organisms is continuously being exposed to a huge number of chemicals and different combinations of them, i.e., mixtures in ever-changing dose/concentration, for different periods via food or feed, drinking water, polluted air, consumer

products, material, and goods. The possible combinations of chemical mixtures are increasing exponentially due to the release of a massive amount of chemical wastes into the environment via advanced industrialization and overuse or misuse and improper discharge of pharmaceuticals, agrochemicals, and metals in the form of metal oxide nanomaterials, etc. The occurrence of chemical mixtures in the environment without documentation about their individual component identity, concentration, and dangerous effects on humans, as well as wildlife and aquatic life, makes it neither realistic nor beneficial to investigate each possible combination of the mixtures. However, existing regulations for human risk assessment (HRA) of chemicals predominantly consider the exposure and evaluation of toxicity of individual compounds, instead of focusing on the chemical mixtures. Focus on the risk assessment of exposure of chemical mixtures is given in rare cases, only if exposure to chemical mixtures is considered in a statutory framework. The ERA is frequently restricted to chemicals falling within the framework which often neglects co-exposure to chemicals that are enclosed by different sections of the legislation [1].

In the last decade, the European Commission (EC) reported the joint response of chemicals, i.e., mixtures [2]. The EC had stated significant concerns about the existing limitations for risk assessment of individual compounds and proposed advanced ways to make sure that hazardous effects associated with chemical mixtures are appropriately understood and examined. It stated that the EU laws fixed stringent limits for the quantity of specific chemicals permissible in food, water, air, and industrial products; however, the probable hazard responses of all these chemicals in a mixture are rarely assessed. Several regulatory authorities are in place to regulate the risk assessment of single as well as mixture of chemicals such as the US Environmental Protection Agency (US-EPA) [3], the Agency for Toxic Substances and Disease Registry (ATSDR) [4], the World Health Organization (WHO) [5], the nonfood Committees of the European Commission, and the European Food Safety Authority (EFSA) [6], Plant Protection Products (PPPs) (Regulation No 1107/2009, 283/2013 and 284/2013), and biocidal product regulation (Regulation No 528/2012), Water Framework Directive (Directive 2006/60/EC), all of which have made substantial progress toward establishing a practical framework which will be suitable for risk assessment of multiple component mixtures [7]. Although multiple approaches for risk assessment of the chemical mixture are continuously designed and applied by the researchers and regulators in particular cases, till now there are no general, reliable, accurately validated integrated methods across the different regulatory authorities. The most widely accepted framework for the mixture risk assessment was developed in a WHO/IPCS (International Programme on Chemical Safety) workshop [5]. This framework explained a standard method for the

estimation of hazardous effects of chemical mixtures, and it can be altered according to the requirements of users. However, its application is mostly hampered by huge data gaps on exposure as well as hazard effect information of chemicals. Interestingly, there are several EU research projects currently going on to fulfill the data gaps, to develop newer chemometric tools, and to design novel in vitro risk assessment methods in order to prioritize the mixture of concern. These researches are mainly funded by the present European Research and Innovation Programme Horizon 2020 or the Seventh Framework Programme.

Table 1 gives a short overview of the currently ongoing EU research projects on toxicity prediction of chemical mixtures [8].

Yang et al. [14] stated that “there is no such object as a single chemical exposure.” Therefore, most of the environmental pollutants commonly exist as mixtures, and the response of individual chemical components in a mixture may not resemble that examined from data of pure individual compounds. Interactions of different components among each other in a mixture can result in complex and significant modifications in the superficial characteristics of its pure component. The components of a mixture may act either increased (synergistic) or decreased (antagonistic) response in comparison with ideal (additive) behavior. Majority of chemicals occurring in the environment are at dose/concentrations far beneath than their individual median effective concentration 50% (EC₅₀), and sometimes it may be lower than their individual no observed effect concentration (NOEC), but they can still produce a harmful effect to the human and other living species spanning over different compartment of environment due to co-contamination with other chemicals in the mixture. The significance of the combined effect of co-contaminants has long been acknowledged by the regulating authorities [15], and there are several commonly used approaches to cover risk assessment of chemical mixtures [16–18]. The most widely used experimental approaches for the risk assessment of mixtures are based on the principal mechanisms of action, i.e., concentration addition (CA) and independent action (IA) method. The research area of risk assessment of the combined effect of mixtures is not new; rather a series of reviews were published to deal with diverse aspects of the combined effect of mixture pharmacodynamics [19], aquatic toxicology [20, 21], phytopharmacology [22], carcinogenicity [23], and environmental toxicology [24].

In this chapter, we have focused on the basic principles of mixture toxicity assessment and the essential steps in basic QSAR modelling as well as the application of QSAR in the prediction of mixture toxicity of pharmaceuticals, pesticides, metals, and organic chemicals.

Table 1
A short overview of currently ongoing EU research projects on toxicity prediction of chemical mixtures [8]

Project name	EDC-MixRisk [9]	HBM4EU [10]	Solutions [11]	EuroMix [12]	EUTOxRisk [13]
Titles	Interdisciplinary and integrated (epidemiological and experimental biological) approaches for risk assessment of mixtures of endocrine disruptive compounds on children	Coordinating and advancing human biomonitoring in Europe to provide evidence for chemical policymaking	Solutions for present and future emerging pollutants in land and water resources management	A tiered strategy for risk assessment of mixtures of multiple chemicals	An Integrated European “Flagship” Program Driving Mechanism-based Toxicity Testing and Risk Assessment for the twenty-first century
Duration	2015–2019	2017–2021	2013–2018	2015–2019	2016–2021
Webpage	http://edcmixrisk.ki.se/	https://www.hbm4eu.eu/	http://www.solutions-project.eu	http://www.euromixproject.eu/	http://www.eu-toxrisk.eu/
Aim	1. Analysis of EDC mixtures: Correlation with the hazardous health outcomes such as growth and metabolism, neurodevelopment, and sexual development in a defined population 2. Understanding of molecular mechanisms and pathways involved in adverse health issues 3. Design and development of a reliable, transparent, and stepwise outline by integrating	(a) Harmonizing the processes of human biomonitoring (HBM) across all the participated countries in the project, by providing the comparable human exposure data of chemicals and mixture of chemicals at EU level, to the policymakers (b) Identification of the exposure pathway of chemicals by connecting data of internal exposure to external exposure of chemicals (c) Obtained scientific	I. Development of reliable and efficient tools for risk assessment of chemical substance and mixtures II. Provides a better understanding of effect and exposure of organic chemical by assembling a complete sequence of integrated models and databases III. Validates the further significance of the new generation of tools in trans-European case studies IV. Provides a logical theoretical outline for	(a) To establish and state a novel, effective, significantly validated test methodology for risk assessment of chemical mixtures (b) Analyzing the sophisticated grouping methodology of chemicals for cumulative assessment (c) Establish ranking criteria for a chemical mixture depending on the hazard effect (using <i>in silico</i> tools) and exposure considerations (d) Exploration of	A. Repeated dose systemic toxicity of chemicals on vital target organs such as the lung, kidney, liver, and nervous system B. Developmental and reproductive hazards of chemicals C. The prime objective is to provide testing approaches to enable reliable, significant, robust, animal-free hazard and risk assessment of chemicals

<p>epidemiological and experimental research the for risk assessment of EDCs and their mixture effects</p> <p>(d) Adapting chemical risk assessment approaches to use HBM data to interpret the contribution of numerous exposure pathways to the total chemical body burden</p> <p>(e) Provides knowledge on the exposure pathways of chemicals for the design of targeted policy measures to reduce exposure and minimize the risk</p>	<p>evidence on the fundamental relations between human exposure to chemicals and hazards health effects</p> <p>(d) Adapting chemical risk assessment approaches to use HBM data to interpret the contribution of numerous exposure pathways to the total chemical body burden</p> <p>(e) Provides knowledge on the exposure pathways of chemicals for the design of targeted policy measures to reduce exposure and minimize the risk</p>	<p>the assessment, ranking, and decline of contaminants and mixtures thereof to protect European waters and to decrease environmental and human health hazards</p> <p>(e) Determine how to extrapolate the results of in vitro bioassay and in silico models to humans</p> <p>(f) To develop tools and models for toxicity assessment of chemical mixtures</p>
<p>Relevant EU REACH regulation legislation</p>	<p>Chemical regulations</p>	<p>Water Framework Directive; all other pieces of EU chemicals legislation with relevance for water pollution</p>
<p>Relevant EU strategies or activities</p>	<p>Community strategy for endocrine disruptors, chemical mixtures</p>	<p>Global food safety legislation, pesticides, contaminants, additives, water framework directive, REACH regulation</p>

2 Why Is the Assessment of the Toxicity of Chemical Mixtures Important?

Ecotoxicity assessment of individual chemical component is a myth as majority of chemicals in the environment are present in a complex mixture, which on exposure results in greater hazardous effects on living systems and ecosystem than the individual components. Hence, the estimation of single component toxicity against a particular organism and environmental compartments may not illustrate the real toxicity data in real-life scenario [25]. However, the estimation of toxicity of a chemical mixture is a much more composite problem than the estimation of individual chemical components, because the components in the mixture may affect the interaction pattern of each other which results in the significant changes in the final response output of each component. Each component of a chemical mixture may act by several modes such as additive action of biological endpoints or may act by an increasing response (synergistic) or by reducing (antagonistic) effects. Another major issue while modelling the toxicity of a chemical mixture is the experimental data of the existing components and their concentration in a particular mixture. In majority of the cases, it reveals that the concentration of individual components is far below than its median effective concentration 50% (EC_{50}), or even below of its individual, no observed effect concentration (NOEC), but still they can result in hazardous effects on human and the environment by interaction mechanism with other individual components present in the chemical mixture. Thus, the determination of toxicity of major components may not demonstrate the actual toxicity value of the final chemical mixture [26–29].

3 General Principles of the Mixture Toxicity Assessment

The toxicity from exposure to chemical mixtures can be assessed based on the two well-known and widely used fundamental approaches such as (1) whole-mixture approach and (2) component-based approach (as shown in Fig. 1) [8].

3.1 Whole-Mixture Approach

The whole-mixture effects can be determined by direct testing of the complete mixture itself, but it can also be estimated from data obtained from the mixture of similar chemical composition (the presence of similar individual components in same quantities). In this approach, a quantitative mixture-toxicity relationship can be directly assessed by employing the available experimental toxicity data of the whole mixture. The whole-mixture analysis can be achieved for intentional mixtures, such as direct exposure of the workers to pesticide formulations, as well as for indirect exposure of unintentional mixtures, for example, mixtures of organic chemicals

in river water, streams, etc. This method permits consideration of any unidentified chemicals present in the mixtures and any type of interactions among mixture components [30].

The major hurdle in the application of the whole-mixture approach is that the mixture is comprised of a large number of individual chemicals at different concentration ratio, and it is not practically feasible to perform the ecotoxicological analysis of each and every possible combination of chemicals in the whole mixture. Till date, the majority of analysis work published based on employing the whole-mixture approach has been mainly focused on environmental, dietary, or consumer product mixtures, while the whole sources in real life are much larger and more complex.

3.2 Component-Based Approach

The component-based approach depends on the effect of the individual components of the mixture to predict the joint outcome of the mixture. The selection of a suitable mathematical approach for the prediction of mixture toxicity primarily depends on whether the individual components of a mixture act by a similar mode of action or independent mode of action [31]. The ideal usage of component-based approach is mainly dependent on obtained information such as (1) information about the architecture of the mixture and the mechanism of action of each individual components of the mixture and (2) on the knowledge of different functional groups present in the chemical structure responsible for a similar or different activity or mechanism of action. All this type of information can be obtained from chemical structures and structure-activity relationships analysis (either qualitative or quantitative), toxicophore identification, structural alerts, and toxicological responses or effects [7].

The component-based approach for the prediction of ecotoxicity of chemical mixtures can further be categorized into three well-known and widely used fundamental concepts of actions [32], such as:

1. Similar action (dose/concentration addition)
2. Independent action
3. Interactions
4. Generalized concentration addition (GCA) models

3.2.1 General Overview

The effect of the complex mixtures of similarly acting individual chemicals can be determined based on the sum of the dose or concentration of individual components in the mixture, but in case of a complex mixture of independently acting chemicals, their toxicity effect can be evaluated based on the probability of response obtained from each individual chemicals or sum of their biological responses (addition). These concepts (independent action and dose/concentration addition) are based on the

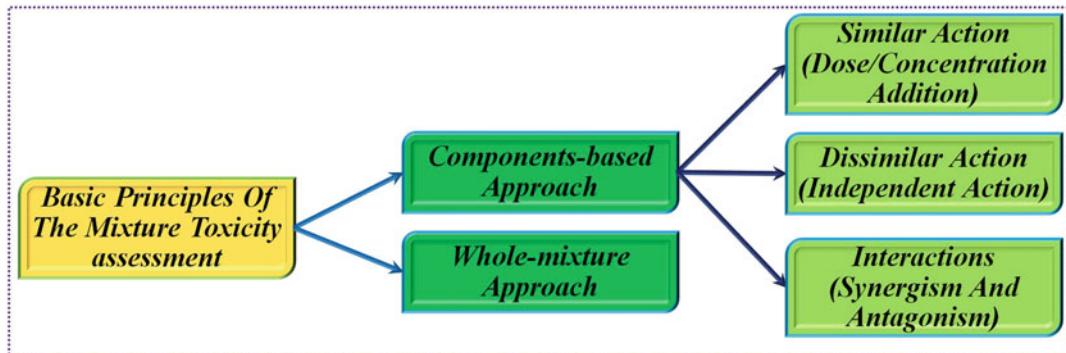


Fig. 1 Fundamental principles for toxicity assessment of chemical mixtures

hypothesis that individual chemical compounds in a complex mixture do not affect each other's toxicity, i.e., they do not interfere in the binding of each other with the biological target site, or they can act on the different targets, or they do not interact with each other at the biological target site. These types of chemical compounds can produce similar biological responses by a shared or analogous mechanism of action, or they act individually and may have dissimilar response endpoints and target organs. Both ideas have been proposed as default methods in regulatory risk assessment of chemical mixtures. However, it is rarely found that the chemical mixtures are composed of only similarly acting or only independently acting chemicals [7, 33, 34].

3.2.2 Similar Action (Dose/Concentration Addition)

The German pharmacologist Loewe in 1926 [35, 36] proposed the concept of concentration addition for the first time. The similar action or similar joint action is observed if each chemical component of the mixture acts on the same target site and possesses the same mode of action but exerts different biological response. The concept of concentration addition for a mixture of “ n ” number of individual components can be expressed mathematically as follows [37]:

$$C_{\text{eff}} = \sum (C_i / EC_i) \quad (1)$$

Here, C_{eff} is the overall effective concentration of a mixture which can be estimated by the sum of all the effective concentration of the compounds, C_i stands for the actual concentration of compound “ i ,” and EC_i stands for the concentration of the compound “ i ” at which the mortality is observed in 50% of the studied animals. Each fraction of C_i / EC_i termed as a “toxic unit,” provides information about the concentration or dose of an individual component in the mixture scaled for its relative potency [38]. If the summation of the toxic units equals 1 at a mixture concentration or dose which elicit the EC_i effect, then the mixture performs as per

the concentration addition (CA) concepts. Under such situations, each component of a complex mixture can be replaced by any other chemical component without altering the final toxicity of the mixture, as long as the summation of the toxic unit is constant. Such kind of alteration of the component usually retain the final response intake due to binding of compounds on the same biological target sites, i.e., individual components which have a similar mechanism of action and which are neither intercorrelated based on the physico-chemical property level nor with their toxicokinetic and toxicodynamic levels [39, 40].

3.2.3 Independent Action (IA)

This kind of effect is observed when chemicals act independently on different target sites such as molecular receptors, cells, and tissue of an exposed organism or with a probably different mode of action and do not influence the biological activity of each other [41–43]. As per these concepts, it is assumed that there are no physico-chemical as well as biological interactions among the components of the mixture and they do not influence or interfere with each other's uptake, distribution, and metabolism. The independent action of a binary mixture can be estimated by the joint probability of statistically independent events by using the following equations:

$$E(C_{\text{mix}}) = 1 - [(1 - E(C_1))(1 - E(C_2)) \dots \dots (1 - E(C_i))] \quad (2)$$

or

$$E(C_{\text{mix}}) = 1 - \prod_{i=1}^n (1 - E(C_i)) \quad (3)$$

Here, $E(C_{\text{mix}})$ stands for the combined effect of the mixture, and $E(C_i)$ is the effect of each single mixture component (i) present at the concentration (c_i). Effects are stated as parts of a maximum possible effect ($0\% \leq E \leq 100\%$) [32]. As per the above-stated equations, any chemical which exhibits $E(C_i)$ equal to zero does not contribute to the joint effect of the mixture. Accordingly, binary mixtures of independently acting chemical compounds pose no health concern, as long as the doses/concentrations of every single component of the mixture remain below their individual zero-effect levels (concentrations).

Estimation of the IA of the expected mixture effects requires the knowledge of individual component effects, which can be obtained from experimental data studied on the traditional ecotoxicological response endpoints (e.g., mortality, growth, reproduction), that each individual compound would show toxic effect if applied singly at a similar concentration at which it exists in the mixture. This analysis typically requires the concentration-response curves of all individual toxicants present in the mixture. On the other hand, the IA-predicted effect of a complex mixture $E(C_{\text{mix}})$ is always greater than the effect of individual constituent in the mixture, $E(C_i)$. This suggests that with an increasing number of

individual compounds in a mixture, lower outcomes have to be defined for each individual component in order to predict a convincing mixture effect. For example, suppose two individual compounds result in an individual effect of 29.3% at defined concentrations, their joint effect will be 50%. Conversely, if ten individual compounds are present in the mixture with 6.7% individual effects, they will produce the combined mixture effect of 50%. Therefore, large amounts of consistent ecotoxicological data which cover the region of low effects are required for the application of IA to multicomponent mixtures. Requisite number and helpful data are unavailable, and most importantly extensive experimental efforts are required for generating them. Therefore, the application of IA for the prediction of the common toxicities of multicomponent mixtures has so far been primarily limited to experimental mixture studies in which the necessary data were explicitly recorded. However, it should be noted that Posthuma and colleagues [44–48] reported the mixture toxicity assessment approach based on species sensitivity distributions (SSDs), which permits to compute the IA-expected species sensitivity distribution employing standard EC₅₀ and/or NOEC values, assuming that data for a sufficient number of *taxa* is at hand for each mixture component [49].

3.2.4 Interactions (Synergism and Antagonism)

Interactions explain the combined response of two or more individual chemicals of the mixture, which is either stronger (synergistic, potentiating, supra-additive) or weaker (antagonistic, inhibitive, subadditive, infra-additive) than the sum effects which can be estimated by dose/concentration addition or response addition. The interaction among chemicals includes all forms of combined effect other than the above-specified additive concept. Interactions might be as per the condition such as the relative intensity of dose/concentration of each component, the route(s), exposure duration and timing (including the biological persistence of the mixture components), and the biological target(s) site [32].

3.2.5 Generalized Concentration Addition (GCA) Models

The CA and IA are the most widely used approaches, but they are ineffective in some cases such as modelling of chemicals which have high potency but low efficacy. To overcome such problems, Howard and Webster have reported the generalized concentration addition (GCA) model [50]. The GCA acts based on the aggregate effect of a mixture by means of the efficacy and potency of the mixture's each chemical constituents. The simple way to explain the GCA model is illustrated in the following equation [28, 51]:

$$E = \frac{\text{max effect level}_A [A]/EC_{50A} + \text{max effect level}_B [B]/EC_{50B} + \dots}{1 + \frac{[A]}{EC_{50A}} + \frac{[B]}{EC_{50B}} + \dots} \quad (4)$$

Here, E denotes the effect of the chemical mixture at a particular concentration, whereas “max effect level A” is the maximal effect level of compound A , $[A]$ is the concentration of A in the mixture at an explicit mixture concentration, and EC_{50A} is the EC_{50} value of A and similar for chemical B , etc.

3.2.6 Realistic Confirmation on the Performance of CA and IA in Ecotoxicological Assessments of Chemical Mixtures

Globally several studies have been performed using the CA and IA approaches to predict the mixture toxicities, and the results were compared with the experimental toxicity in order to decide which concept is more suitable and predictive one. The evidence on the toxicity prediction of a large number of chemical mixtures such as heavy metals, endocrine disrupters, pharmaceuticals, agrochemicals, narcotics, and industrial organic chemicals using both concepts have been compiled, reviewed, and briefly illustrated.

- (a) Most of the reported experimental studies are applications of CA approach, whereas IA has been applied only in a limited number of investigated mixture studies. Even the comparison studies for the predictive performance of both approaches have been reported only in a small fraction of the literature.
- (b) As per the reported studies, the CA approach generally has high prediction power and is considered as the first and protective approach [52], while IA is only used to predict toxicities of virtually identical mixtures [53–55].
- (c) Majority of published toxicity assessment studies have been performed on a particular species of freshwater organisms, while chemical mixture toxicity studies on marine and terrestrial species as well as on a higher organism are still rare.
- (d) Based on the majority of reported investigations, the mixture modelling studies have only focused on the binary or tertiary mixtures [38, 52, 56–59]. Experimental analysis of more than two components or multicomponent chemical mixtures is inadequate. Moreover, chemical mixtures with known composition and concentration ratios do not imitate any real environmental condition, but these are developed with the purpose of investigating the theoretical mixture toxicity. For example, chemical mixtures of only similarly acting chemicals (e.g., ref. 60), and just dissimilarly acting chemicals (e.g., refs. 42, 43), and compounds which belong to the same class or of the same purpose (e.g., refs. 54, 59, 61) were used in different studies.
- (e) According to Belden et al. [56], for 88% of the selected pesticide mixtures, the prediction quality employing the CA hypothesis falls within a factor of 2 from the observed mixture toxicity, independent of the similar or dissimilar mode of action of the mixture components. It is also evident from the analysis by Cedergreen et al. [55] that there is a significant

deviation in predictions employing the CA approach (around 6% of the examined 158 data points). Although a good source of information on metal mixture toxicity was reported by Norwood and his co-worker [57], the majority of reported studies in the area of metal mixture toxicity would not provide enough knowledge or clear picture to determine the number of deviations between the observed and predicted mixture toxicity. The investigation of metal mixtures is further complicated because it is based on the fact that a number of metals are vital elements and every organism has a well-established active system for metal uptake, internal storage, and sequestration. Additionally, most of the metal interactions may happen at the level of bioavailability and absorption.

- (f) As we know that the chemical mixtures present in the environment may not be made up of only similar or dissimilarly acting chemical components, the major problem is how to apply the existing approaches for mixture prediction, and it has gained a lot of attention among the scientific community.

It should be noted that advanced CA- and IA-based methods have been put forward, based on SSDs [44, 62, 63], employing mechanistic modelling based on the dynamic energy budget (DEB) theory [64] or using tissue-residue approaches [65]. Despite their interesting characteristics, these approaches are presently not appropriate in several situations, particularly considering the exposure of industrial organic chemicals. In such cases, mostly the toxicity data of essential individual components of mixture for the requisite array of diverse species and taxa are not available (in the case of SSD approaches), and the knowledge on the relation between aqueous and internal body concentrations is not at hand (which is a prerequisite for the tissue-residue approach). Additionally, the accessible biological data (toxicokinetics and toxicodynamics) is also insufficient (in the situation of mechanistic models and tissue-residue strategies) [49].

4 Significance of Chemometric Approaches for Toxicity Assessment of Complex Chemical Mixtures

An enormous number of novel single chemicals are continuously being introduced and/or released, while several thousands already exist in the ecosystem with a lack of sufficient toxicity data against different living organisms and environment. However, the condition is even worse for the complex mixture toxicity data. The experimental toxicity evaluation of single as well as mixtures using animal models is time-consuming and costly, while the toxicity prediction employing computational approaches is relatively quick and easy. The *in silico* approaches predict the toxicity using several

learning algorithms (like multiple linear regression, partial least squares, artificial neural network, etc.) and computational expert systems [66]. Although it is essential to note that *in silico* approaches are not a complete alternative for *in vivo* and *in vitro* toxicity tests, these approaches can be used to complement experiments by reducing the number of animal testing, reduction in cost and time for toxicity assessment, and to predict the hazardous effects of novel chemicals prior to their development. Additionally, the advancement in the computer hardware and continuous development of novel learning algorithms lead to the use of computational approaches in diverse fields including the prediction of toxicity of chemical mixtures. The prime goals of computational methods (Fig. 2) used to predict the toxicity of chemical mixtures are the following [28]:

1. The intelligent application of computational approaches has reduced the number of animals sacrificed in the toxicity assessment.
2. Chemometric models using the existing chemical mixtures may be used to predict the toxicity of the untested or novel altered composition of chemical mixture against particular species or compartments if they fall within the applicability domain (AD) of the chemometric model.
3. Computational models are helpful for risk profiling of hazardous materials by the regulatory agencies.
4. The computational methods are significant and reliable tools to determine the magnitude of risk as well as help to plan how to reduce it.
5. Without any doubt, the computational tools are time and cost effective in comparison to the *in vitro* and *in vivo* toxicity experiments.
6. Helpful to fill the data gaps in ecotoxicity of chemical mixtures as a large portion of chemical mixtures have no toxicity data at all.

5 Quantitative Structure-Activity Relationship (QSAR) Modelling of Ecotoxicity of Mixtures

The quantitative structure-activity relationship (QSAR) approach is the most widely used chemometric technique that correlates the biological endpoints (property/activity/toxicity) of a molecule with its structural attributes [21, 22]. A conventional QSAR model is developed from the information obtained from a series of individual compounds (in the form of descriptors) and its response endpoint, while in case of QSAR modelling of mixtures, the descriptor information is obtained from two or more

components of the mixture [67]. Therefore, in the broad sense, a significant difference between the single molecule QSAR and QSAR of mixtures is related to the calculation of numerical descriptors and interpretation of the results, while the rest of modelling steps remain the same in both cases. Here, we have summarized the best practices of QSAR modelling which might be reasonably useful in modelling both individual compounds and mixtures [68].

In general, a QSAR model is developed by using one or more statistical tools to find out a reliable, robust, and significant correlation between the numerical features (descriptors) obtained from either a single compound or a mixture for a defined endpoint such as toxicity, biological activity, property, etc. [69–71]. As per the OECD (Organisation for Economic Co-operation and Development) guidelines, to develop a valid and acceptable model for regulatory assessment of chemical safety, one has to follow the five OECD principles, which are illustrated as (1) a defined endpoint; (2) an ambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness of fit, robustness, and predictivity; and (5) a mechanistic interpretation, if possible [72]. Any QSAR training exercises, including those for the mixtures, are executed in the consecutive steps starting from the data collection and data preparation, calculation of molecular descriptors, descriptor pooling or descriptor selection, model building (learning algorithms), validation of generated models based on the internal and external parameters, and finally the model exploitation (model interpretation). The performance and acceptability of QSAR models depend on how accurately each successive step has performed. Several common mistakes in the steps mentioned above result in the unacceptable QSAR models, and thus these should be avoided during QSAR model development as explained in several reviews [73–76]. Figure 3 depicts the schematic overview of QSAR modelling and its validation steps specific to mixtures.

Based on chemometric modelling algorithms, the QSAR models can be broadly classified into three categories: (1) regression-based QSAR approaches, (2) classification-based QSAR approaches, and (3) machine learning approaches [77, 78]. The regression-based QSAR approach includes multiple linear regression (MLR), principle component regression analysis (PCR), partial least squares (PLS), ridge regression, and genetic function approximation (GFA); classification-based QSAR approach includes linear discriminant analysis (LDA), and the machine learning approach includes artificial neural networks (ANN), Bayesian-regularized neural networks, support vector machine regression (SVM), decision tree, random forest (RF), naïve Bayesian classifier, and k-nearest neighbor method (k-NN). The QSAR model has been used to detect the essential chemical structural characteristics responsible for toxicity/activity/property to explore the possible mechanism behind the toxicity/activity/property of a

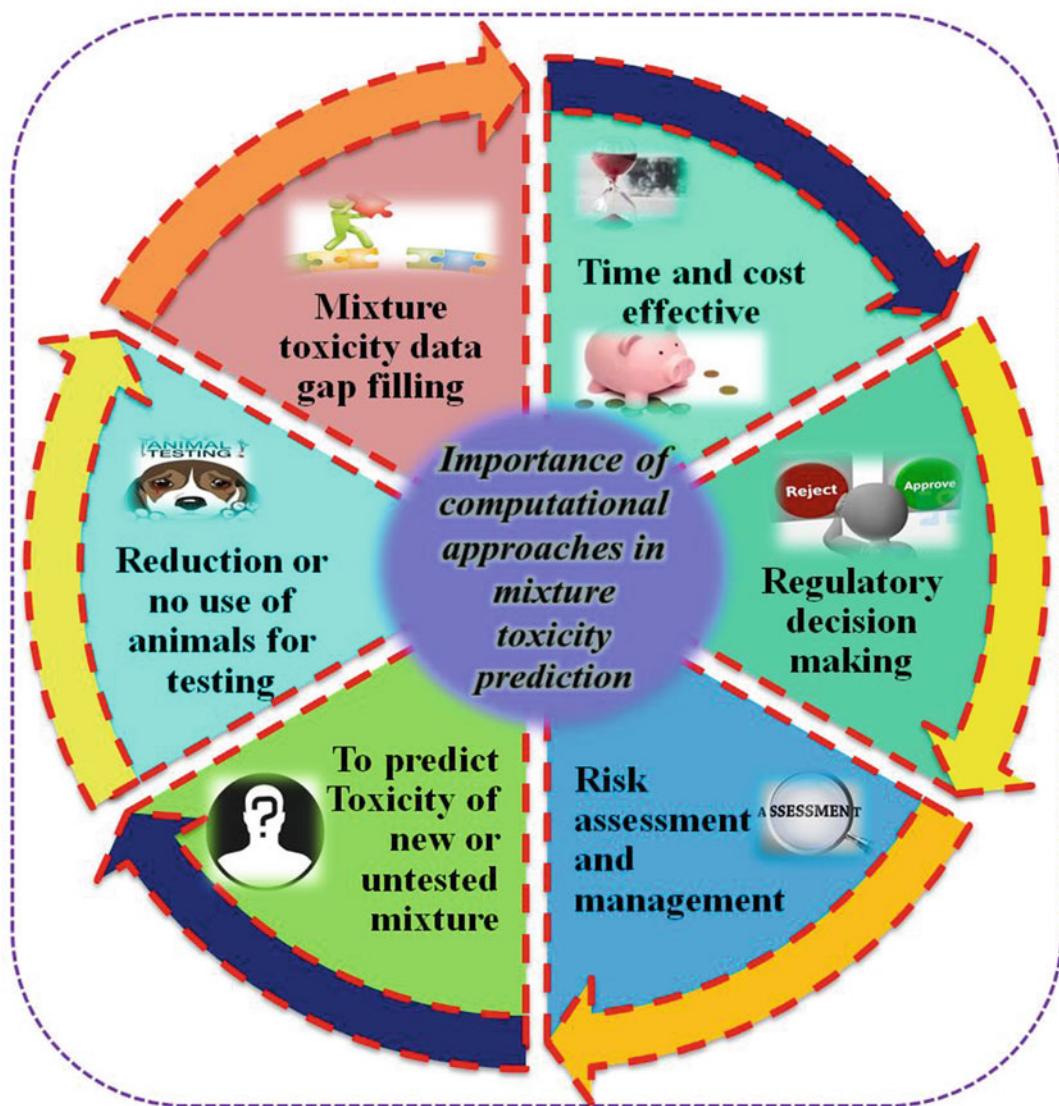


Fig. 2 Importance of computational approaches for a mixture's toxicity prediction

particular class of chemicals against a particular species. The following steps are very important for the generation of reliable/significant QSAR model of mixtures.

5.1 Data Collection and Data Preparation

The accuracy of the generated QSAR models is governed by the correctness of the input dataset. While preparing the dataset for QSAR modelling, care must be taken regarding the accuracy of the chemical structures and experimental response. The researcher should also be cautious about the use of data collected from several sources, because if the experimental protocols and conditions are different, then the data should not be clubbed into a single set. The collected data must be carefully checked to remove duplicates, salts,

etc. The best practices for data preparation before the modelling process have been described by Fourches et al. [75, 79, 80]. Unfortunately, in case of mixture dataset collection, the bigger problem is the quality of data obtained for QSAR modelling which is not clearly discussed in most of the reported studies [67].

5.2 Calculation of Molecular Descriptors

The prepared chemical structures are used to estimate the independent variables (descriptors) employing different software tools such as Padel-Descriptor [81], Dragon [82], SiMRS, and Alvadesc [83], etc. Descriptors are numerical quantities obtained from structures of single compounds in a quite straightforward process using software tools, and one can directly use the calculated descriptors as independent variables for QSAR modelling. However, the computation of descriptors for mixtures is very challenging. One way to calculate descriptors for mixtures is to estimate the values of variables from individual components, and then the obtained descriptor values from individual components are multiplied by their percentage ratio present in the mixture and finally added together to get the final descriptor value for mixtures. This is most commonly known as weighted descriptors approach and quite frequently used by researchers for mixture modelling [84, 85].

Descriptors of individual compound M : $D_1^M, D_2^M, \dots, D_n^M$.

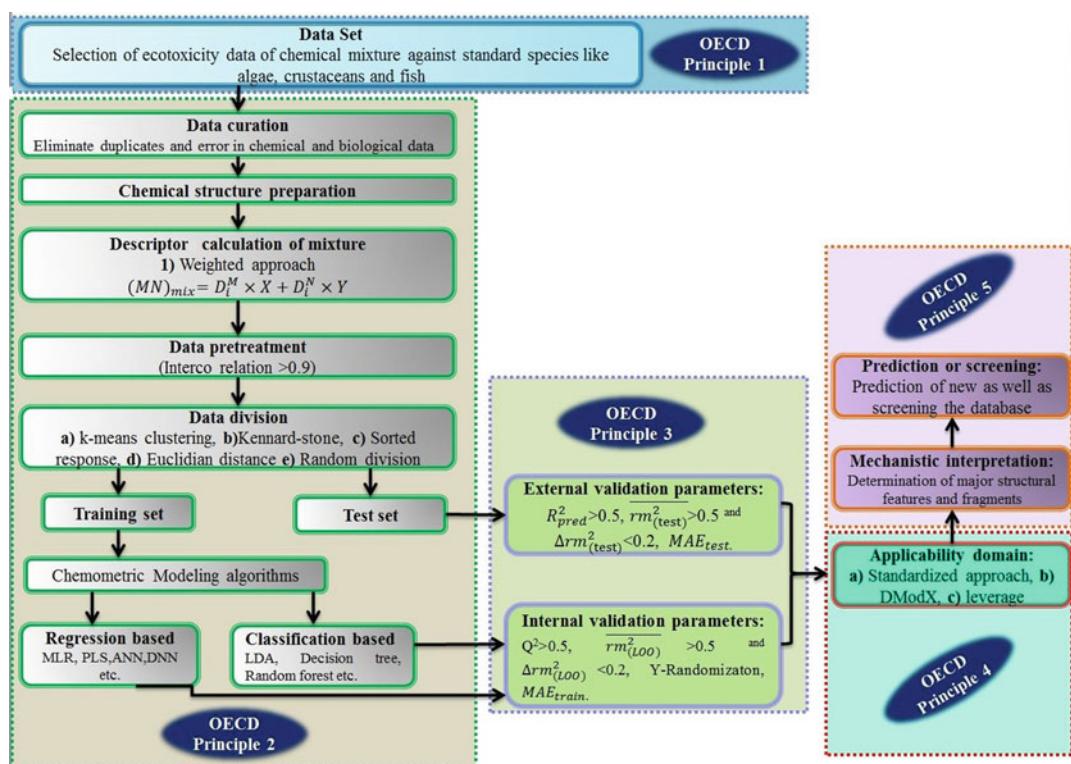


Fig. 3 A complete schematic overview of the development of a QSAR model

Descriptors of individual compound N : $D_1^N, D_2^N, \dots, D_n^N$

If compounds M and N are mixed in X/Y ratio, then i^{th} descriptor value of mixture M and N is estimated using the following equation:

$$(MN)_{\text{mix}} = D_i^M \times X + D_i^N \times Y \quad (5)$$

where D_i^M is i^{th} variable of individual compound M , D_i^N is i^{th} variable of individual compound N , and $X + Y$ is percentage ratio of each component in the mixture which is equal to 1% or 100%. Any error in the calculation of descriptors leads to an error in the final results. There are several categories of reported studies based on the type of descriptors used for modelling such as (1) descriptors based on the partition coefficient for mixtures, (2) integral (whole-molecule) additive descriptors (weighted sum of descriptors of individual components), (3) integral nonadditive descriptors of mixtures (mixture components are taken into account differently from the additive scheme), (4) fragment nonadditive descriptors (structural parts of different mixture components are simultaneously taken into account in the same descriptor), and (5) miscellaneous descriptors for QSAR modelling of mixtures [67].

5.3 Descriptor Selection

Descriptor selection is one of the most crucial steps in QSAR model development as it selects meaningful descriptors from a large pool of descriptor set. One cannot use the entire descriptor pool for modelling since it is computationally expensive and time-consuming. The principal aim of feature selection methods is the removal of redundant, noisy, or irrelevant descriptors while developing the QSAR models; in this way, the dimensionality of the input variable is reduced without loss of vital information. There are several methods available for descriptor selection such as stepwise forward selection (FS) and backward elimination (BE), genetic algorithm (GA), all possible subset selection and factor analysis, etc. For detailed information on several approaches of feature selection, please refer to a recently published review article by Khan and Roy [86]. The same feature selection methods as used for QSAR modelling of individual compounds can be used for QSAR modelling of mixtures.

5.4 Modelling Algorithms and Model Validation

QSAR is a statistical approach which quantitatively correlates the dependent variable (response endpoint) with a number of independent variables (descriptors). There are several modelling approaches such as PLS (partial least squares), MLR (multiple linear regression), principal component regression analysis, ridge regression, artificial neural network, etc. For detailed information on various methods of the model building, one can see the relevant literature [86]. The selection of appropriate modelling algorithm (linear and/or nonlinear) may affect the quality of the final predictions.

Finally, the validation of the developed models can be performed by considering several internal and external validation metrics, estimation of AD, and Y-randomization as per the OECD guidelines. The complete details of diverse validation parameters are available elsewhere [68, 87].

6 Application of QSAR in Ecotoxicity Prediction of Pharmaceutical Mixtures

Pharmaceuticals are continuously produced in large volumes and widely used as the therapeutic agents in human and veterinary medicines. Due to their extensive usage and heavy consumption, the occurrence of pharmaceuticals as a single component or as a mixture in the aquatic environment is of emerging concern [88]. Pharmaceuticals are frequently found in the form of complex mixtures (similar mode of action, synergistic form, additive form, etc.); hence different organisms in the environment are mostly exposed to the mixture of pharmaceuticals.

There are several studies which reported the presence of most toxic and concerning classes of pharmaceuticals such as antibiotics, antibacterial, analgesics, cardiovascular drugs, antidepressants, and antipsychotics as a single component and in a complex mixture in the environment. We have discussed below some of the recent applications of QSAR models in ecotoxicity prediction of the pharmaceutical mixture.

Białyk-Bielńska et al. [89] reported the mixture toxicity of most widely used six antimicrobial sulfonamides (SAs) and their two degradation products using both experimental and in silico (concentration addition approach) study (Fig. 4). They have utilized the toxicity data of sulfonamides toward two most sensitive organisms, viz., limnic green algae (*Scenedesmus vacuolatus*) and duck (*Lemna minor*). As per described studied, first they have evaluated individual toxicity of two transformation products (TPs) of SAs (sulfanilic acid (SNA) and sulfanilamide (SN)), and afterward, the authors have assessed the mixture toxicity of SAs among themselves (mixture 1) and with its TPs (mixture 2). The individual toxicity study of two TPs revealed that the observed EC₅₀ values for SN were 25.83 (21.69–29.94) mg/L and 5.09 (4.37–5.93) mg/L to *S. vacuolatus* and *L. minor*, respectively, while SNA has no observed toxicity against both studied organisms even at higher concentration range (up to 100 mg/L). On the other hand, the mixture toxicity evaluation suggests that both mixtures (mixture 1 and 2) show an effect less than the additive effect based upon their observed and predicted toxicity toward *S. vacuolatus* and *L. minor*.

The differences among chronic (24-hr exposure) and acute (15-min exposure) mixture toxicity were assessed employing the in silico approaches (docking-based receptor library of antibiotics and the receptor library-based QSAR model) by Zou et al.

[90]. They evaluated the toxicity of individual eight antibiotics, trimethoprim, and their binary mixture against *V. fischeri*. Subsequently, the authors developed the receptor library-based QSAR model using the individual chemical-receptors binding energy and the observed binding concentration of the individual compound (as shown in Eq. 6):

$$\begin{aligned} \log(EC_{50}) = & f_1 \left(E_{\text{binding}}^{A-\text{receptor}} \times \frac{C_a}{\sum C_i} \right) \\ & + f_2 \left(E_{\text{binding}}^{b-\text{receptor}} \times \frac{C_b}{\sum C_i} \right) \end{aligned} \quad (6)$$

where C_a and C_b are the observed concentrations of the individual compounds A and B in chemical mixtures, $\sum C_i$ the moral concentrations of the total chemicals in the mixture, and $E_{\text{binding}}^{A-\text{receptor}}$ the binding interaction of component A of chemical mixture. The study revealed that the risk quotients of antibiotic mixtures are only based on chronic mixture toxicity instead of other toxicity data like individual compound acute toxicity, acute chemical mixture toxicity, and individual compound chronic toxicity. In another similar study, Zou et al. [91] have reported the differences between the chronic (24 h) and acute (15 min) mixture toxicity of sulfonamides and their potentiators on the *Photobacterium phosphoreum* (Fig. 5). To model the toxicity, the authors calculated the toxic unit and EC_{50}^M response endpoint used to describe the acute and chronic toxicity of binary mixture by employing the following Eqs. 7 and 8, respectively:

$$TU = \frac{C_a}{EC_{50a}} + \frac{C_b}{EC_{50b}} \quad (7)$$

$$EC_{50}^M = \frac{C_a + C_b}{\frac{C_a}{EC_{50a}} + \frac{C_b}{EC_{50b}}} \quad (8)$$

where C_a and C_b are the observed concentrations of the component “a” and “b” in mixtures at median inhibition and EC_{50a} and EC_{50b} are the observed effective concentrations of the components “a” and “b.” Simultaneously, they have estimated that binding energy of chemical-receptors interaction using docking study and developed QSAR models for toxicity prediction of single as well as mixture of SAs using Eq. 6. From single compound toxicity, it was found that the pK_a played an essential role in the toxic effect of SAs because it helps the SAs to transport into the cell. Based on the obtained results, the authors have redeveloped QSAR models considering pK_a as one of the essential features which may improve the quality of prediction (Eq. 9):

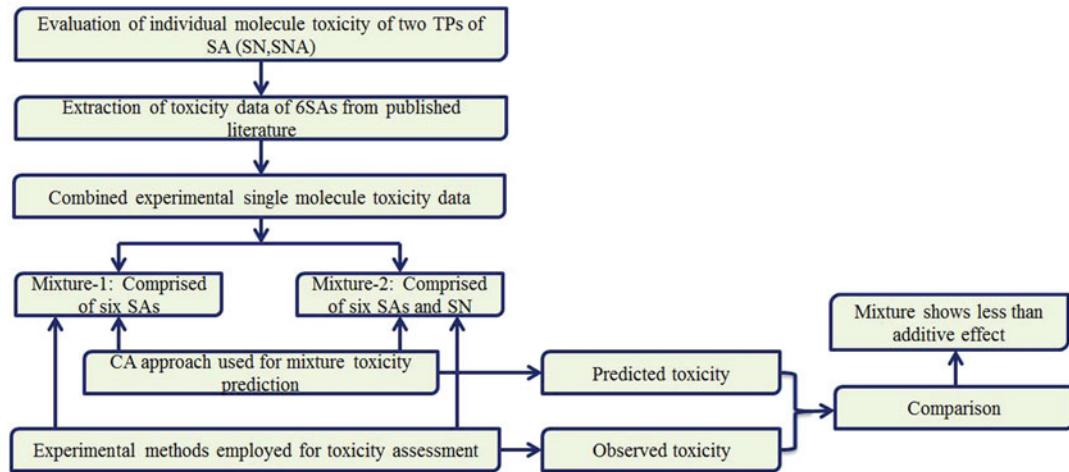


Fig. 4 Schematic overview of single chemical as well as mixture toxicity assessment of SA and its TPs using experimental and in silico approach

$$\begin{aligned} \text{Log}(EC_{50}) = & f_1 \times pK_a + f_2 \left(E_{\text{binding}}^{\text{A-receptor}} \times \frac{C_a}{\sum C_i} \right) \\ & + f_3 \left(E_{\text{binding}}^{\text{b-receptor}} \times \frac{C_b}{\sum C_i} \right) \end{aligned} \quad (9)$$

As per the QSAR analysis, it is revealed that the mechanism of acute and chronic mixture toxicity was based on the two points dissimilarity, i.e., (1) then receptor binding site of SAs is Luc (luciferase) in case of acute toxicity of mixture, while in case of chronic toxicity of mixture, it was Dhps (dihydrofolate reductase), and (2) variation in the actual concentration of binding between two different cases such as acute and chronic mixture toxicity. The study also suggested that the mixture of SAs may act antagonistically initially (acute mixture toxicity) while in chronic condition mixture act via the synergistic mechanism.

Long et al. [92] investigated the difference between the joint effect of sulfonamides and different antibiotics. They have employed the toxicity data of individual as well as a mixture of sulfonamides and several antibiotics toward the *E. coli*. As per the developed QSAR models, they have suggested that the difference of joint response between sulfonamides and various antibiotics was predominantly because of two aspects: (1) the target site (proteins, cell, tissue) of single chemicals and (2) the capability of antibiotics to interact with their target receptors, namely, the effective combined concentration. They have also introduced the concept of “effective concentration” for assessment of the mechanism of binary mixture toxicity more efficiently.

Wang et al. [93] assessed the combined toxicity of the eight quorum sensing inhibitors (QSIs) with three different classes of

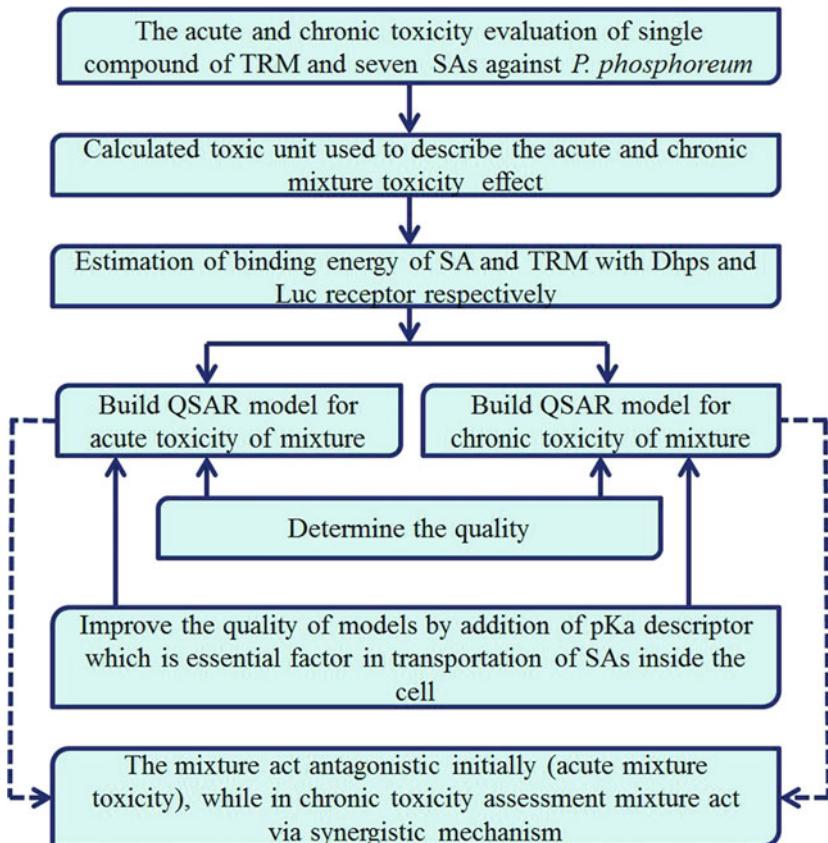


Fig. 5 Schematic representation of single chemical as well as mixture toxicity assessment of SAs (sulfonamides) and its potentiators (trimethoprim) (TRM) using QSAR method

predominantly used antibiotics, such as β -lactams, sulfonamides (SAs), and tetracyclines on *E. coli* (Table 2). The QSAR models for prediction of the combined toxicity were developed by employing the interaction energies of binding of receptor-ligand complex. The analysis showed that the SAs and QSI exhibit either additive or combined antagonistic response in the mixture toxicity test, although β -lactams and tetracyclines (TCs) displayed only antagonistic response with the QSI. The QSAR models proposed that the QSI in the mixtures showed more binding interaction with the target receptors than the antibiotics. In another study, Wang et al. [94] have reported the mixture toxicity of three different categories of most widely used antibiotics, i.e., sulfonamides (SAs), SA potentiators (SAPs), and TCs toward the three organisms, viz., *E. coli*, *V. fischeri*, and *B. subtilis*. The developed QSAR analysis proposed that the defined concentration ratio of each single component in a mixture might differ a lot from the designed concentration ratio; moreover, the TCs in the ternary mixtures changed the toxic ratio

of SAs and SAPs, which leads to the fluctuation in combined response of the ternary mixtures on different organism of study.

Cleuvers [95] reported the ecotoxicity of the nonsteroidal anti-inflammatory drugs (NSAIDs) such as diclofenac, ibuprofen, naproxen, and acetylsalicylic acid (ASA) (Table 2) against *Daphnia* and algae. The authors first evaluated the experimental single molecule toxicity against the target organism and subsequently employed the two reported QSAR models (Eqs. 10 and 11) [96, 97] to determine the mechanism of action of studied chemicals:

$$\text{Log}EC_{50}[\text{molL}^{-1}] = -0.95 \log K_{\text{ow}} - 1.32 \quad (10)$$

and

$$\text{Log}EC_{50}[\text{molL}^{-1}] = -1.00 \log K_{\text{ow}} - 1.23 \quad (11)$$

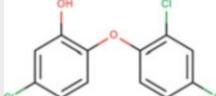
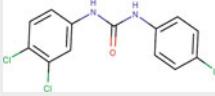
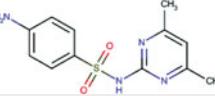
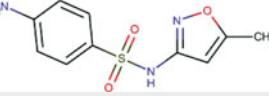
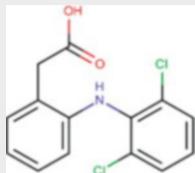
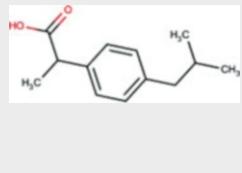
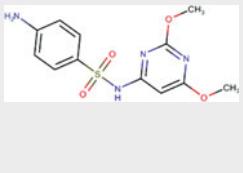
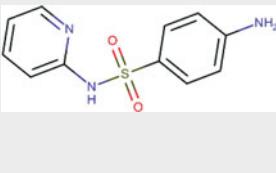
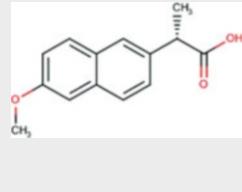
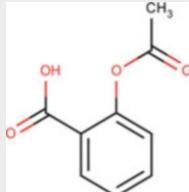
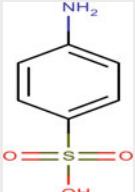
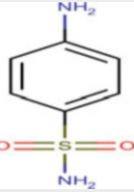
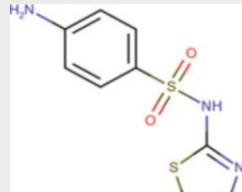
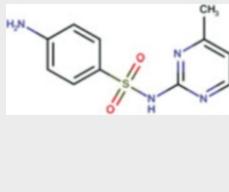
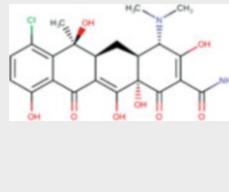
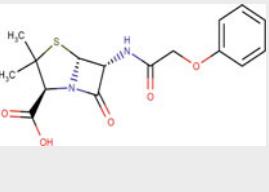
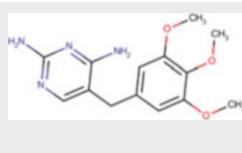
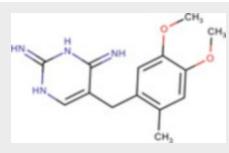
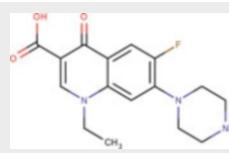
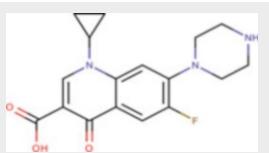
The QSAR analysis revealed that all compounds act by nonpolar narcosis suggesting that the increase in the n-octanol/water partitioning coefficient ($\log K_{\text{ow}}$) of the compounds will result in the increase of toxicity. On the other hand, they have also suggested that the combined toxicity of each component of the mixture can be precisely predicted by employing the concept of concentration addition. Escher et al. [98] reported the relative ecotoxicological risk assessment of beta-blockers as well as their metabolites employing a mode-of-action-based test battery and a QSAR method. The analysis shows that the beta-blockers followed the concept of concentration addition for mixture toxicity prediction. There are numerous other studies which reported the investigation of ecotoxicity of a pharmaceutical mixture [99–101]. Villa et al. [102] have reported the acute toxicity of antibacterial (triclocarban, triclosan, and methyl triclosan) compounds and their mixtures against the *V. fischeri*. The authors have estimated the individual chemical toxicity using an experimental approach, and the concentration-response data of every individual studied compound were fitted to Weibull function:

$$E = f(\alpha, \beta, c) = 1 - \exp(-\exp(\alpha + \beta * \log_{10}(c))) \quad (12)$$

Here, E stands for the concentration-inhibition ratio of an individual pesticide, while total concentration-inhibition ratio of a mixture, c , denotes the observed concentration of the pesticide, and α and β are the parameters to be computed by employing a nonlinear least squares (NLLS) technique. Afterward, they have used previously reported QSAR models to predict the nature of compounds, i.e., narcotics or polar narcotics against *V. fischeri*. The proposed QSAR model is completely based on the octanol/water partition coefficient. The QSAR model revealed that the triclosan and triclocarban (Table 2) act as polar narcotic compounds toward *V. fischeri*, whereas methyl triclosan acts as a narcotic. The mixture

Table 2

List of some of the most commonly observed pharmaceuticals as single components and/or in a complex mixture in the environment

			
Triclosan	Triclocarbon	Sulfadimidine	Sulfamethoxazole
			
Diclofenac	Ibuprofen	Sulfadimethoxine	Sulfapyridine
			
Naproxen	Acetylsalicylic acid	Sulfanilic acid	Sulfanilamide
			
Sulfathiazole	Sulfamerazine	Chlortetracycline	Penicillin V
			
Trimethoprim	Ormetoprim	Norfloxacin	Ciprofloxacin

toxicity of antibacterials was determined experimentally as well as predicted by using the most common approaches (CA and IA), and the results suggest that the observed mixture toxicity of antibacterial had no significant differences from those predicted by both CA and IA models.

7 Application of QSAR in Ecotoxicity Prediction of Mixtures of Agrochemicals

Agrochemicals (pesticides, insecticides, herbicides, and fertilizers) differ from the pharmaceuticals and other organic chemicals because they are intentionally designed or developed to elicit the toxic effect on one or more target organisms or pests. Although they are used for protection and enhancement of the yield of the crop, their toxicological effects are not limited to the targets for which they are applied. They will spread into the ecosystem via several physical ways and adversely affect other species such as human being, aquatic species, and wild animals. Humans belong to higher species than the target species for agrochemicals, so it is expected that they are unaffected by exposure of the least amount of these compounds [103]. However, a high dose of agrochemicals are toxic to humans and sometimes responsible for acute poisonings, but the exposure of agrochemicals in the mixtures even in low doses might show toxic actions [104, 105]. It is not surprising to discover a mixture of numerous agrochemicals in the surface water in agricultural areas [105, 106]. Liu et al. [107] have developed the dose-addition (DA) model to predict the joint ecotoxicity of the chemical mixture of herbicides which are coexisting with insecticides. They have selected five herbicides (simetryn, prometon, bromacil, Velpar, and diquat) and one organophosphate herbicide (dichlorvos) for toxicity assessment against the *Vibrio qinghaiensis*, employing the microplate toxicity test methodology. The dose-response data were fitted to a number of nonlinear functions and found that the dose-response curve of all the six pesticides was efficiently defined by the function known as Weibull function (Eq. 12). The generated Weibull models of the mixtures are significant and reliable with a statistical R-value of higher than 0.99 and an RMSE of lower than 0.020. The primary objective of the published report was to simplify whether the DA model can estimate the risk of the mixture which is composed of herbicides and insecticides with the likely different modes of actions but with the similar toxicity endpoints. The study suggests the DA model can significantly predict the combined toxicity of mixture. Gutowski et al. [108] investigated the toxicological hazards of S-metolachlor, its commercial product Mercantor Gold, and their photoproducts employing the two different approaches such as a water-sediment test and QSAR approach. They have applied three different QSAR models for ecotoxicity prediction of SM (S-metolachlor) and bio-TPs (biotransformation product). The applied QSAR revealed that the observed bio-TPs might be highly toxic than its parent compounds. Thus, it is highly recommended that toxicity of parent as well as its bio-TPs should be further evaluated and care should be taken for a detailed risk assessment of the chemical.

8 Application of QSAR in Ecotoxicity Prediction of Heavy Metals and Their Mixtures

Heavy metals are the most prevalent contaminants of major concern because they are nondegradable, and thus, they persist in the environment for a longer time [109]. Although they are naturally occurring elements, the majority of atmospheric pollutions occur due to several activities of humans such as mining and smelting processes, advanced industrial manufacturing, and heavy usage of metals for agriculture purpose [110]. The release of metals into the environment ultimately results in the harmful effects on humans, animals, and other biotic organisms on exposure.

Metal contaminants are hardly found alone in the environment [111]; cadmium, zinc, cobalt, and nickel are most commonly used in several industrial applications/processes, and thus they are often found as a mixture. It is more essential to determine their combined effect instead of considering a single element for risk assessment. The experimental toxicity of metals was determined by considering the several microbial parameters such as growth rate, biomass determination, inhibition of bioluminescence, and enzyme activity. The toxicity assessment of metal mixtures is done mostly based on bacterial bioluminescence. However, the QSAR approach is also employed for the toxicity assessment of metal mixtures. For example, Li et al. [112] have reported single as well as joint toxicity of cadmium (Cd) and nine substituted phenol against the *P. phosphoreum* (Fig. 6). Initially, the authors performed experimental toxicity assessment of individual substituted phenol and Cd which suggested that the individual chemical toxicity occurs due to different substituted groups on phenols ($\text{OH} > \text{NH}_2 > -\text{CH}_3\text{O} > \text{NO}_2$) along with the position of the substituted group (para>ortho>meta). Further, they carried out joint toxicity of Cd and substituted phenols employing the toxic unit and additive index as response endpoint. The final QSAR model was obtained using stepwise linear regression method by considering the joint toxicity as response endpoint and descriptors estimated from substituted phenols using several computational programs such as ClogP, MOPAC-PM3, CS Chem3D Ultra, Micro QSAR, and Molecular Modeling Pro. The models are comprised two descriptors which are the logarithm of *n*-octanol/water partition coefficient ($\log P$) reflecting the lipophilicity of compounds and the heat of formation (ΔH_f) contributing toward the stability of molecules with statistical quality of R^2 0.855, 0.878 and 0.780 at low, medium, and high concentration of cadmium, respectively. They concluded that the substituted phenols show similar combined effects at different concentrations of Cd.

Similarly, Su et al. [113] have reported the individual as well as combined toxicity of copper and 11 nitroaromatic compounds against the *P. phosphoreum* (Fig. 7). First, they have experimentally

determined the joint toxicity of a binary mixture of every single nitroaromatic compound and Cu, followed by calculated molecular descriptors using several computational tools, and built a QSAR model to predict the combined toxicity of a binary mixture of each individual nitroaromatic compound and Cu. The reported QSAR models obtained employing the stepwise linear regression method comprised of two different descriptors with statistical quality of R^2 0.828, 0.727, and 0.732 at low, medium, and high concentration of Cu, respectively. The QSAR analysis showed that the toxicity of nitroaromatic compounds at low concentration of Cu increases when Connolly solvent-excluded volume (CSEV) of nitroaromatic compounds decreases, while it increases when the dipolarity/polarizability (S) increases. In contrast, the toxicity is directly proportional to the Connolly accessible area (CAA) at medium and high Cu concentrations, i.e., increment in CAA value of compounds leads to higher toxicity. In simple word, at low concentration of Cu, the binary joint effects of Cu and nitroaromatic chemicals are a simple addition. On the other hand at medium and high concentration, the joint effect of Cu and nitroaromatics are antagonistic.

9 Application of QSAR in Ecotoxicity Prediction of Organic Chemical Mixtures

All the living organisms are exposed to several types of organic chemicals for a different period of time. Over time, the organic compounds have been released or have continued to be formed naturally, for example, the release of organic substances by fires and volcanoes in the environment [114]. After the evolution, and mainly from the last two centuries, due to the high rate of the industrial revolution, there has been an exponential increase in the number of organic chemicals in the environment [114]. The harmful effect of individual compounds on different organisms has been well explored, but toxicity assessment of a mixture of the organic chemical is still lacking. Numerous QSAR studies have reported the toxicity of diverse class of organic chemical mixtures such as aromatic amine and phenol mixtures [115], mixture of halogenated benzenes [116], combined effect of phenols and cadmium [112], mixture toxicity of alkanols [117], combined effect of nitrile and aldehydes [118], joint response of cyanogenic pollutants and aldehydes [119], collective toxic response of substituted phenols [120], combined response of nitrobenzene [121], mixture toxic effect of alkoxyethanol [122], poisonous effect of nitroaromatic and copper [113], mixture effect of chlorinated anilines and cadmium [123], toxicity of polynuclear aromatic hydrocarbon mixture [124], alkylphenols [125], perfluorinated carboxylic acid mixtures [126] perfluoroalkyl substance mixture [127], and mixture of halogenated chemicals [128] against different organisms including green algae, *Vibrio fischeri*, gram-positive and gram-negative

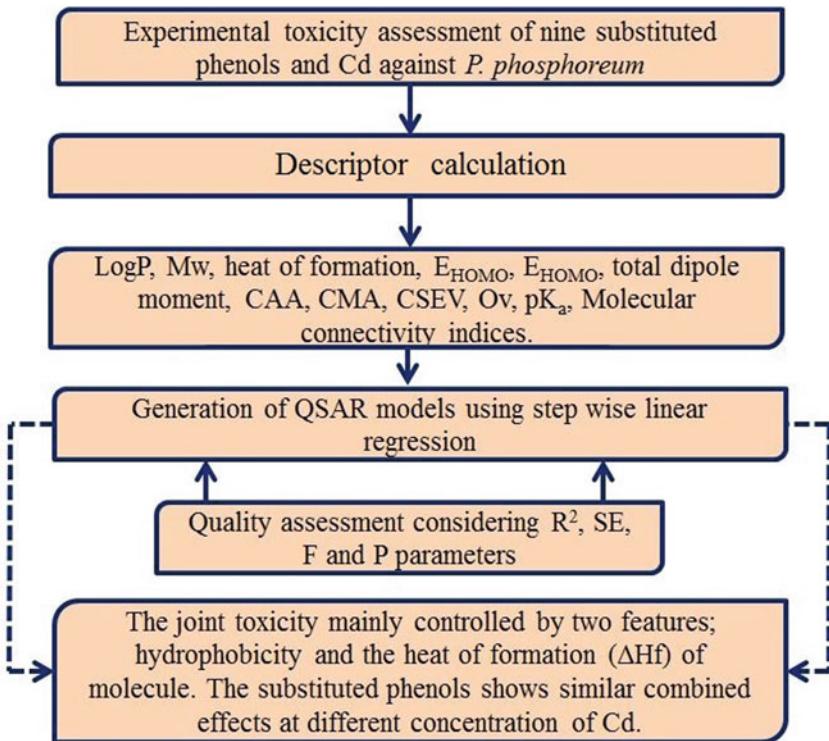


Fig. 6 Schematic workflow of single chemical as well as joint toxicity assessment of substituted phenols at different concentration of Cd (cadmium) using QSAR models

bacteria, *Dicrateria zhanjiangensis*, *Raphidocelis subcapitata*, amphibian (*Rana japonica*, *amphibian fibroblast cell line*), fish (zebrafish), etc. We have discussed below a few examples of the recent applications of QSAR models in ecotoxicity prediction of the organic chemical mixtures.

Lu et al. [115] have investigated the acute toxicity of 11 aromatic amines, 4 substituted phenols, and their 32 mixtures toward the river bacteria in natural waters and calculated the mixture toxicity index as response endpoint to describe the toxic effect using the following equations:

$$MTI = 1 - \log M / \log N \quad (13)$$

$$M = \sum TU_i = \sum C_i / IC_{50} \quad (14)$$

Here, C_i is the observed concentration of a single molecule existing in a mixture, IC_{50} is the median inhibition concentration of an individual component, and N is the number of individual components in the mixture. Subsequently, they have developed a QSAR model for mixture toxicity prediction using the dataset of 32 compounds with a squared correlation of $R^2 = 0.834$, comprising two molecular descriptors, namely, n-octanol/water partition coefficient ($\log P$) and the energy of the lowest unoccupied

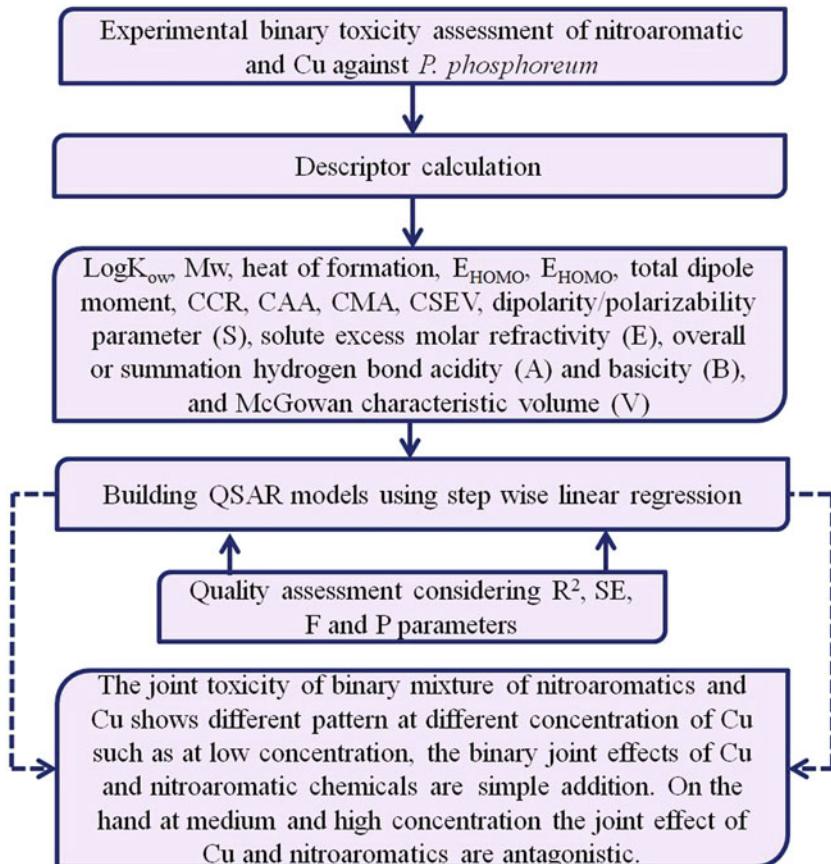


Fig. 7 Schematic workflow of combined toxicity assessment of binary mixture of nitroaromatics and Cu at different concentration of Cu using QSAR models

molecular orbital (E_{LUMO}) calculated using different software tools. They have also suggested that the developed model can be used successfully to predict the toxicity of any kind of mixtures such as binary, tertiary, or quaternary mixtures [115].

A QSAR model for prediction of combined toxicity of halogenated benzenes against *D. zhanjiangensis* was reported by Zeng et al. [116]. The QSAR model was obtained using a dataset of 49 compounds based on the simple octanol-water partition coefficient ($K_{ow,mix}$) descriptor with significant correlation coefficient ($R^2 = 0.879$) and least standard error ($SE = 0.124$). The $K_{ow,mix}$ descriptor values were estimated using the following equation:

$$K_{ow,mix} = \frac{w}{v} \times \frac{\sum_{i=1}^n \frac{Q_{water,i}}{1 + \frac{w}{v K_{Di}}}}{\sum_{i=1}^n Q_{water,i}^0 - \sum_{i=1}^n \frac{Q_{water,i}^0}{1 + \frac{w}{v K_{Di}}}} \quad (15)$$

where $K_{D,i}$ represents the partition coefficient of single component “ i ,” W stands for a total volume of solution, V denotes the volume of the only organic phase, $Q_{\text{water},i}^0$ is the initial concentration of component “ i ” in the aqueous phase, and “ n ” is the total number of individual compounds present in the chemical mixture. The value of W/V is 6.8×10^5 . They have concluded that the mixture toxicity primarily depends upon the partition coefficient of the mixture of halogenated benzenes which are considered as nonpolar narcotic chemicals, and it is known that the nonpolar narcotic chemicals affect the organism only by interaction with lipids of biomembranes [116].

Wang et al. [117] have reported the mixture toxicity of alkanols which are also narcotic compounds. They have measured the acute toxicity of 15 highly hydrophobic alkanols against *P. phosphoreum* by performing slight adjustment in octanol-water partition coefficient descriptors by considering the influence of volume effect and named it as equivalent octanol-water partition coefficient. Finally, the QSAR model of mixture toxicity was generated by employing the equivalent mixture octanol-water partition coefficient ($\log K_{\text{ow}}$). The QSAR model was developed using the simple linear regression technique with the statistical value of $R_{\text{adj}}^2 = 0.779$. As per mechanistic interpretation, the alkanols with higher lipophilicity can easily transport inside the cell via cell membrane (lipid bilayer) and produce a toxic effect to the target organism. It is also found that the alkanol toxicity increases with the increasing carbon chain length from methanol to dodecanol, and then the toxicity starts to fall down from tridecanol. This is because alkanol chemicals become too bulky for transport through the channel of biomembrane to inside the cell, which may result in decrease toxicity of alkanol. The authors concluded that developed linear model could be useful to predict mixture toxicity of new or untested mixture by using an equivalent mixture octanol-water partition coefficient.

Hoover et al. [127] have reported in vitro as well as in silico modelling approaches for toxicity estimation of the individual as well as a mixture of perfluoroalkyl substances using the amphibian fibroblast cell line. First, they have evaluated the cytotoxicity of PFAS as individual as well as in binary mixtures to amphibian fibroblast cell line. Second, the data obtained from in vitro studies were employed for in silico study, i.e., QSAR modelling of the individual as well as a chemical mixture. Among all the reported models, the best model comprised of the only single descriptor with a significant variance of 94% in training set and 90% predictive variance. Finally, the best model was used for prediction of 24 individual and 1380 binary mixtures [127]. As per the reported investigation, they have concluded that the combined effect of two very common PFAS (perfluorohexane sulfonate (PFHxS) and perfluorohexanoic acid (PFHxA)) was potentially higher than additive

effects, while shorter-chain PFAS may not lead to as many hazardous effects. Finally, they have suggested that the best QSAR model comprised vital structural features of the studied chemicals and might be useful as a query tool for toxicity prediction of novel and untested PEAS (individual as well as mixtures).

Kar et al. [128] have reported the QSAR models for toxicity prediction of halogenated chemicals (endocrine-disrupting pollutants) against zebrafish (*Danio rerio*) embryos using a dataset of nine compounds comprised of five single and four tertiary mixtures (Fig. 8). Four statistically significant QSAR models comprising of single descriptors were generated using GA-MLR technique for four different sets of division. The molecular descriptors for chemical mixture were calculated by using weighted descriptors approach as mentioned in Subheading 5.2. The final models were used for toxicity prediction of an external set of binary and tertiary mixtures [128]. The study revealed that the developed models for toxicity prediction of the halogenated chemical mixture were based on the concentration addition concept which signifies that similar MOA among all the studied compounds. They have also identified that compared to bromine, chlorine- and fluorine-based PFAS shows higher toxicity for the studied halogenated chemicals. The developed models served as a better tool for understanding the essential feature of studied chemicals and were used to predict the toxicity of new set of mixtures as well as single halogenated compounds.

Lin et al. [119] have developed a combined QSTR (quantitative structure-toxicity relationship) model using a dataset of 40 binary mixtures containing an aldehyde and a cyanogenic toxicant against the *P. phosphoreum*. The study revealed that the dataset was composed of mixtures of different observed effects (additive, synergistic, or antagonistic) from each other and the difference in the observed effect was based on the chemical-chemical interaction (formation of carbanion intermediate) between the cyanogenic toxicant and aldehydes. Further, chemical-chemical interaction analysis revealed that two essential features of single compound primarily contribute toward the combined effect of mixture. A QSAR model was developed by considering these two important features (Hammett constant used the charge of the carbon atom in the -CHO of aldehydes and C* is the charge of the carbon atom in the carbon chain of cyanogenic toxicants) with a correlation of $R^2 = 0.868$ and standard error (SE) 0.232. The external set of eight additional similar mixtures was used to determine the prediction quality of the built model, and it was found that the model significantly predicted toxicity with a correlation of $R^2 = 0.888$ and SE = 0.223 [119].

10 Future Perspective

The regulatory requirements and awareness for toxicity assessment of mixtures among several regulatory frameworks are scarce, even though several chemical compounds are subject to the provisions of more than one regulatory framework. At present, all the regulatory authorities, as well as toxicologists, have understood the importance and necessity of mixture toxicity evaluation instead of focusing on single chemical risk assessment. The ATSDR developed a strict direction for chemical mixtures which is equally similar to that in the US EPA guidance, although the ATSDR offers more weighting on physiologically based pharmacokinetic (PBPK) and pharmacodynamic (PBPD) modelling. Agencies like the National Institute of Environmental Health Sciences (NIEHS), National Toxicology Program (NTP), and National Institute for Occupational Safety and Health (NIOSH) initiated efforts to illustrate exposures, generate biomarkers, and assess environmentally relevant mixtures [129].

To improve the risk assessment of mixtures, there is a primary need for the design or development of novel databases with the toxicity information obtained from diverse experimental protocols using different species model at different time exposure. This toxicity information of mixtures might be useful for generation of computational models followed by expert systems. These types of expert systems might be helpful for mixtures risk assessment or, at the very least, for the estimation of dose-dependent interactive effects. In the present time, the databases cover only a small fraction of toxicity information of chemical mixture. Thus, a large number of efforts need to be employed to prepare an improved database in collaboration with an experimental and computational modeler.

Based on the present and future scenario of the mixture toxicity assessment study, we have illustrated a few points for future efforts:

- (a) The no-observed-adverse-effect level (NOAEL) dose should be considered for mixture risk assessment.
- (b) Till now there is no harmonized or universal approach for mixture risk assessment, and one needs to design or develop novel architecture for mixture toxicity assessment or modify old framework based on the type of mixture with an objective to solve the complex mixture issue because the traditional animal-based mixture risk assessment modelling practices are insufficient for such a complicated issue.
- (c) Combined efforts between experimental toxicologists and computational scientist are essential to resolve most of the problems and challenges associated with chemical mixture toxicity.

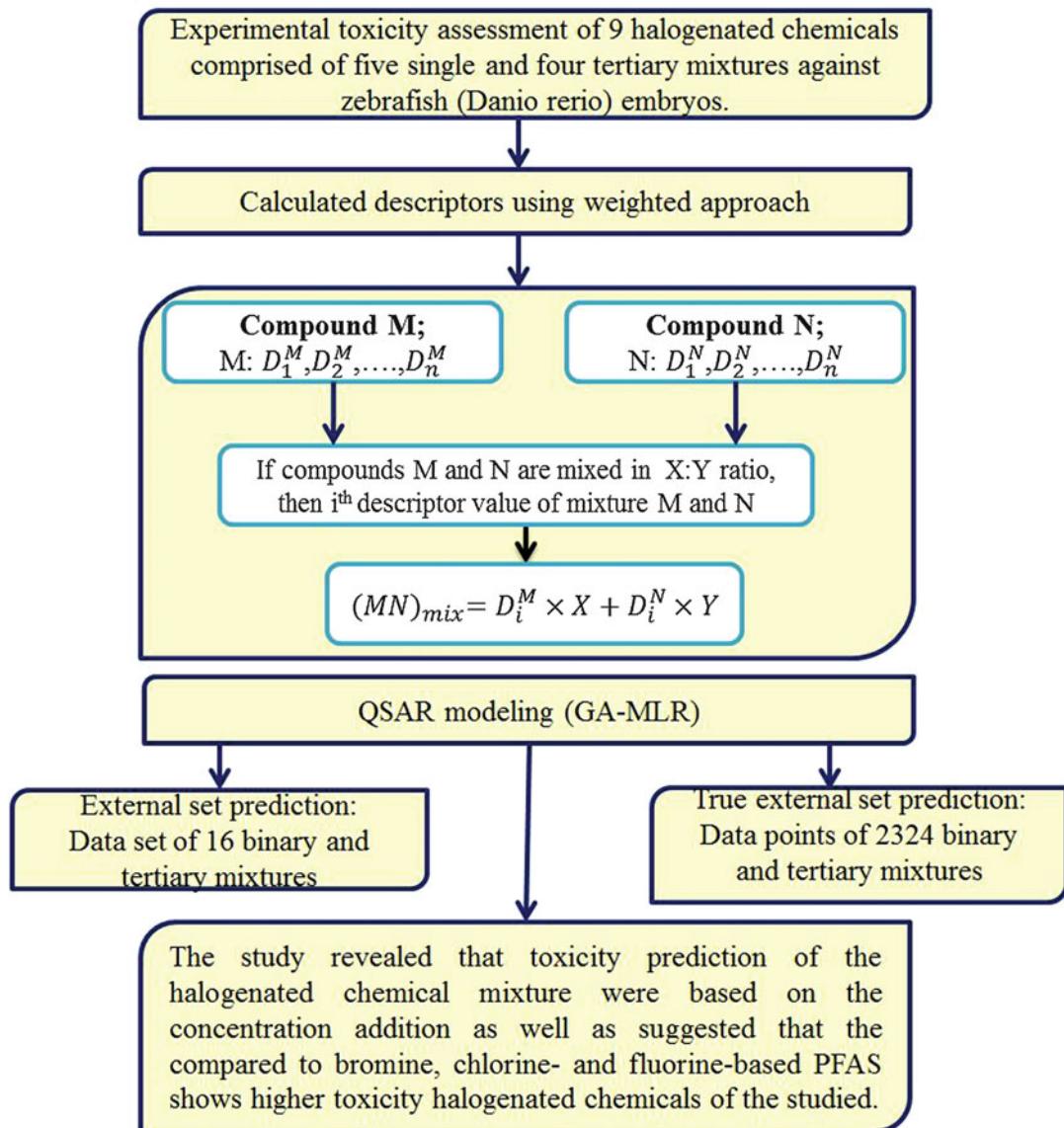


Fig. 8 Schematic diagram of mixture toxicity assessment of halogenated chemicals using QSAR approach

11 Conclusion

The risk assessment of chemical mixtures to the human and environmental has been of the major concern of toxicologist because the majority of chemicals in the environment exist in the form of a mixture. However, till now, the regulatory authorities and the majority of researchers have focused mainly on the risk assessment through toxicity investigation of individual chemicals, but the fact is that maximum chemicals exist as mixtures, generally at very low

levels or far below median effective concentration 50% (EC_{50}), while they can exhibit hazardous effects on human and environment by interaction mechanism with other individual components present in the chemical mixture. Therefore, the risk assessment may be misleading in such cases. The identification of each chemical present in the mixture as well as their percentage ratio is crucial before performing any toxicity quantification. It is also a well-known fact that the determination of the effects of different compositions of individual components in the mixture is quite a challenging task; therefore, the availability of mixture toxicity data is really scarce. In the present time, it is essential to determine the toxicity impact of chemical mixtures on the human and environmental health, and the chemometric QSAR tools are the alternative methods to assist in bridging the data gaps. In the current chapter, we have discussed different concepts of mixture toxicity modelling such as concentration addition, independent action, and interaction (synergism and antagonism), provided a short discussion on the ongoing projects in the EU for mixture risk assessment, and discussed the importance of the chemometric approach, i.e., QSAR in mixture toxicity prediction. We have also given an overview of essential steps involved in QSAR modelling, as well as cited successful applications of QSAR in toxicity assessment of different classes of chemical mixtures such as pharmaceuticals, pesticides, heavy metals, and organic chemical mixtures (Fig. 9). In our opinion, the precise information about the composition of a mixture with

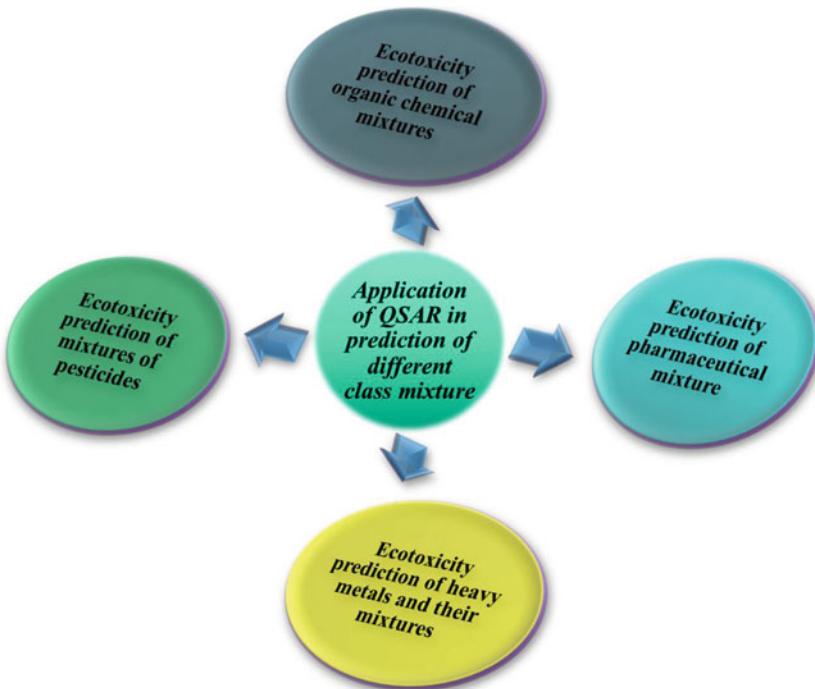


Fig. 9 Application of QSAR in the prediction of diverse classes of chemical mixtures

each individual component concentration and their mechanism of action to produce the toxicity is essential to develop the computational models for mixture toxicity assessment. Another main point to remember is that the mathematical models must be built by taking the physicochemical parameters related to the mechanism of action of chemicals into consideration, so that the model could be used for predictions for untested or unknown compounds/mixtures with a similar composition.

References

1. Evans RM, Martin OV, Faust M, Kortenkamp A (2016) Should the scope of human mixture risk assessment span legislative/regulatory silos for chemicals? *Sci Total Environ* 543:757–764
2. Commission, EE (2012) Communication from the Commission to the Council the combination effects of chemicals. *Chem Mix Com* 10
3. EPA, U (2007) Concepts, methods, and data sources for cumulative health risk assessment of multiple chemicals, exposures and effects: a resource document (final report). US Environmental Protection Agency, Washington, D.C.; EPA/600/R-06: 2007
4. Pohl HR, Mumtaz M, McClure PR, Colman J, Zaccaria K, Melia J, Ingerman L (2018) Framework for assessing health impacts of multiple chemicals and other stressors. CDC stack public health publication, Atlanta, USA
5. Meek M, Boobis AR, Crofton KM, Heinemeyer G, Van Raaij M, Vickers C (2011) Risk assessment of combined exposure to multiple chemicals: a WHO/IPCS framework. *Regul Toxicol Pharmacol* 60:S1–S14
6. EFSA, EFSA (2008) Opinion of the Scientific Panel on Plant Protection products and their Residues to evaluate the suitability of existing methodologies and, if appropriate, the identification of new approaches to assess cumulative and synergistic risks from pesticides to human health with a view to set MRLs for those pesticides in the frame of Regulation (EC) 396/2005. *EFSA J* 6:705
7. Scher S (2012) Opinion on the toxicity and assessment of chemical mixtures. Scientific Committees on Health and Environmental Risks (SCHER), Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR) European Commission, Brussels
8. Bopp SK, Barouki R, Brack W, Dalla Costa S, Dorne J-LC, Drakvik PE, Faust M, Karjalainen TK, Kephalopoulos S, van Klaveren J (2018) Current EU research activities on combined exposure to multiple chemicals. *Environ Int* 120:544–562
9. EDC-MixRisk (2019) <https://edcmixrisk.kise/>
10. HBM4EU (2019) <https://www.hbm4eu.eu/>
11. SOLUTIONS (2019) <https://www.solutions-project.eu/project/>
12. EuroMix (2019) <https://www.euromixproject.eu/>
13. EUToxRisk (2019) <http://www.eu-toxrisk.eu/>
14. Yang R, Thomas RS, Gustafson DL, Campain J, Benjamin SA, Verhaar H, Mumtaz MM (1998) Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling. *Environ Health Perspect* 106:1385–1393
15. Martens M, Mosselmans G, Fumero S, Jacobs G, Lafontaine A (1984) Some thoughts on a possible regulatory approach at EEC level on the classification and labeling of dangerous preparations. *Regul Toxicol Pharmacol* 4:145–156
16. Logan DT, Wilson HT (1995) An ecological risk assessment method for species exposed to contaminant mixtures. *Environ Toxicol Chem An Inter J* 14:351–359
17. europea, UeC (1996) Technical guidance document in support of commission directive 93/67/EEC on risk assessment for new notified substances and commission regulation (EC) N. 1488/94 on risk assessment for existing substances. Office for official publications of the European communities: 1996. European Commission, Brussels
18. OPP, EPA (2000) Proposed guidance on cumulative risk assessment of pesticide chemicals that have a common mechanism of Toxicity, US EPA, Washington DC
19. Greco WR, Bravo G, Parsons JC (1995) The search for synergy: a critical review from a response surface perspective. *Pharmacol Rev* 47:331–385

20. Altenburger R, Boedeker W, Faust M, Grimme LH (1993) Aquatic toxicology, analysis of combination effects. In: Handbook of hazardous materials. Academic Press, San Diego, pp 15–27
21. Kortenkamp A, Altenburger R (1998) Synergisms with mixtures of xenoestrogens: a reevaluation using the method of isoboles. *Sci Total Environ* 221:59–73
22. Bödeker W, Altenburger R, Faust M, Grimme L (1990) Methods for the assessment of mixtures of plant protection substances (pesticides): mathematical analysis of combination effects in phytopharmacology and ecotoxicology. *Nachr bl Dtsch Pflanzenschutzd* 42:70–78
23. Schmähl D (1980) Combination effects in chemical carcinogenesis. In: Further studies in the assessment of toxic actions. Springer, Berlin, pp 29–40
24. Monosson E (2004) Chemical mixtures: considering the evolution of toxicology and chemical assessment. *Environ Health Perspect* 113:383–390
25. Teuschler LK, Hertzberg RC (1995) Current and future risk assessment guidelines, policy, and methods development for chemical mixtures. *Toxicology* 105:137–144
26. Altenburger R, Backhaus T, Boedeker W, Faust M, Scholze M, Grimme LH (2000) Predictability of the toxicity of multiple chemical mixtures to *Vibrio fischeri*: mixtures composed of similarly acting chemicals. *Environ Toxicol Chem An Inter J* 19:2341–2347
27. Hayes AW (2007) Principles and methods of toxicology. Crc Press
28. Kar S, Leszczynski J (2019) Exploration of computational approaches to predict the toxicity of chemical mixtures. *Toxics* 7:15
29. Borzelleca JF (2001) The art, the science, and the seduction of toxicology: an evolutionary development. Principles and methods of toxicology
30. Brack W, Ait-Aissa S, Burgess RM, Busch W, Creusot N, Di Paolo C, Escher BI, Hewitt LM, Hilscherova K, Hollender J (2016) Effect-directed analysis supporting monitoring of aquatic environments—an in-depth overview. *Sci Total Environ* 544:1073–1118
31. Grotens JP, Feron VJ, Sühnel J (2001) Toxicology of simple and complex mixtures. *Trends Pharmacol Sci* 22:316–322
32. Scher S (2011) SCCS toxicity and assessment of chemical mixtures (pp 1–50), European commission, Brussels
33. Kortenkamp A, Backhaus T, Faust M (2007) State of the art report on mixture toxicity. Final Report to the European Commission under Contract Number 070307, The School of Pharmacy, University of London, London
34. Kienzler A, Berggren E, Bessems J, Bopp S, van der Linden, S, Worth A (2014) Assessment of mixtures-review of regulatory requirements and guidance. Joint Research Centre, Science and Policy Reports. European Commission, Luxembourg
35. Loewe S, Muischnek H (1926) Über kombinationswirkungen. *Naunyn Schmiedeberg's Arch Pharmacol* 114:313–326
36. Loewe S (1953) The problem of synergism and antagonism of combined drugs. *Arzneim Forsch* 3:285–290
37. Berenbaum MC (1985) The expected effect of a combination of agents: the general solution. *J Theor Biol* 114:413–431
38. Backhaus T, Sumpter J, Blanck H (2008) On the ecotoxicology of pharmaceutical mixtures. In: Pharmaceuticals in the environment. Springer, Berlin, Heidelberg, pp 257–276
39. Cedergreen N (2014) Quantifying synergy: a systematic review of mixture toxicity studies within environmental toxicology. *PLoS One* 9:e96580
40. Rodea-Palomares I, González-Pleiter M, Martín-Betancor K, Rosal R, Fernández-Piñas F (2015) Additivity and interactions in ecotoxicity of pollutant mixtures: some patterns, conclusions, and open questions. *Toxics* 3:342–369
41. Bliss C (1939) The toxicity of poisons applied jointly I. *Annu Appl Biol* 26:585–615
42. Faust M, Altenburger R, Backhaus T, Blanck H, Boedeker W, Gramatica P, Hamer V, Scholze M, Vighi M, Grimme L (2003) Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action. *Aquat Toxicol* 63:43–63
43. Backhaus T, Altenburger R, Boedeker W, Faust M, Scholze M, Grimme LH (2000) Predictability of the toxicity of a multiple mixture of dissimilarly acting chemicals to *Vibrio fischeri*. *Environ Toxicol Chem An Inter J* 19:2348–2356
44. De Zwart D, Posthuma L (2005) Complex mixture toxicity for single and multiple species: proposed methodologies. *Environ Toxic Chem An Inter J* 24:2665–2676
45. Suter G (2009) Extrapolation practice for ecotoxicological effect characterization of

- chemicals. *Integr Environ Assess Manag* 5:358
46. Posthuma L, Vijver M (2007) Exposure and ecological effects of toxic mixtures at field-relevant concentrations. Model validation and integration of the SSEO programme RIVM report, 860706002/2007
 47. Ragas AM, Teuschler LK, Posthuma L, Cowan CE (2011) Human and ecological risk assessment of chemical mixtures. In: *Mixture toxicity: linking approaches from ecological and human toxicology*. CRC-Press, New York, pp 157–212
 48. Jonker M, van Gestel CA, Kammenga JE, Laskowski R, Svendsen C (2016) *Mixture toxicity: linking approaches from ecological and human toxicology*. CRC press, New York
 49. Backhaus T, Faust M (2012) Predictive environmental risk assessment of chemical mixtures: a conceptual framework. *Environ Sci Technol* 46:2564–2573
 50. Howard GJ, Webster TF (2009) Generalized concentration addition: a method for examining mixtures containing partial agonists. *J Theor Biol* 259:469–477
 51. Hadrup N, Taxvig C, Pedersen M, Nellemann C, Hass U, Vinggaard AM (2013) Concentration addition, independent action and generalized concentration addition models for mixture effect prediction of sex hormone synthesis in vitro. *PLoS One* 8: e70490
 52. EC (2009) EC, State of the art report on mixture toxicity. European Union
 53. Faust M, Altenburger R, Boedeker W, Grimme L (1994) Algal toxicity of binary combinations of pesticides. *Bull Environ Contam Toxicol* 53:134–141
 54. Backhaus T, Faust M, Scholze M, Gramatica P, Vighi M, Grimme LH (2004) Joint algal toxicity of phenylurea herbicides is equally predictable by concentration addition and independent action. *Environ Toxic Chem An Inter J* 23:258–264
 55. Cedergreen N, Christensen AM, Kamper A, Kudsk P, Mathiassen SK, Streibig JC, Sørensen H (2008) A review of independent action compared to concentration addition as reference models for mixtures of compounds with different molecular target sites. *Environ Toxic Chem An Inter J* 27:1621–1632
 56. Belden JB, Gilliom RJ, Lydy MJ (2007) How well can we predict the toxicity of pesticide mixtures to aquatic life? *Integr Environ Assess Manag* 3:364–372
 57. Norwood W, Borgmann U, Dixon D, Wallace A (2003) Effects of metal mixtures on aquatic biota: a review of observations and methods. *Hum Ecol Risk Assess* 9:795–811
 58. Kümmerer K (2008) Pharmaceuticals in the environment: sources, fate, effects and risks. Springer Science & Business Media, Berlin
 59. Warne MSJ (2003) In A review of the ecotoxicity of mixtures, approaches to, and recommendations for, their management. Proceedings of the fifth national workshop on the assessment of site contamination, National Environmental Protection Council Service Corporation, Adelaide, pp 253–276
 60. Faust M, Altenburger R, Backhaus T, Blanck H, Boedeker W, Gramatica P, Hamer V, Scholze M, Vighi M, Grimme L (2001) Predicting the joint algal toxicity of multi-component s-triazine mixtures at low-effect concentrations of individual toxicants. *Aquat Toxicol* 56:13–32
 61. Parvez S, Venkataraman C, Mukherji S (2009) Nature and prevalence of non-additive toxic effects in industrially relevant mixtures of organic chemicals. *Chemosphere* 75:1429–1439
 62. Solomon KR, Brock TC, De Zwart D, Dyer SD, Posthuma L, Richards S, Sanderson H, Sibley P, van den Brink PJ (2008) Extrapolation practice for ecotoxicological effect characterization of chemicals. CRC Press, New York
 63. Posthuma L, Suter GW II, Traas TP (2001) Species sensitivity distributions in ecotoxicology. CRC Press, New York
 64. Baas J, Jager T, Kooijman B (2010) A review of DEB theory in assessing toxic effects of mixtures. *Sci Total Environ* 408:3740–3745
 65. Dyer S, Warne MSJ, Meyer JS, Leslie HA, Escher BI (2011) Tissue residue approach for chemical mixtures. *Integr Environ Assess Manag* 7:99–115
 66. Raies AB, Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 6:147–172
 67. Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Kuz'min VE (2012) Existing and developing approaches for QSAR analysis of mixtures. *Mol Inform* 31:202–221
 68. Roy K, Kar S, Das RN (2015) A primer on QSAR/QSPR modeling: fundamental concepts. Springer, Cham
 69. Khan PM, Rasulev B, Roy K (2017) Chemometric modeling of refractive index of

- polymers using 2D descriptors: a QSPR approach. *Comput Mater Sci* 137:215–224
70. Khan PM, Roy K, Benfenati E (2019) Chemometric modeling of *Daphnia magna* toxicity of agrochemicals. *Chemosphere* 224:470–479
71. Khan PM, Roy K (2018) QSPR modelling for prediction of glass transition temperature of diverse polymers. *SAR QSAR Environ Res* 29:935–956
72. Oecd (2007) Guidance document on the validation of (quantitative) structure activity relationship [(Q) SAR] models. Organisation for Economic Co-operation and Development Paris, France
73. Tropsha A (2012) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29:476–488
74. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
75. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press, London
76. Golbraikh A, Tropsha A (2002) Beware of q2! *J Mol Graph Model* 20:269–276
77. Yasri A, Hartsough D (2001) Toward an optimal procedure for variable selection and QSAR model building. *J Chem Inf Comput Sci* 41:1218–1227
78. Shahlaei M (2013) Descriptor selection methods in quantitative structure-activity relationship studies: a review study. *Chem Rev* 113:8093–8103
79. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010
80. Fourches D, Pu D, Tassa C, Weissleder R, Shaw SY, Mumper RJ, Tropsha A (2010) Quantitative nanostructure- activity relationship modeling. *ACS Nano* 4:5703–5712
81. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466–1474
82. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: an easy approach to molecular descriptor calculations. *Match* 56:237–248
83. Alvadesc (2019) <https://www.alvascience.com/alvadesc/>
84. Khan PM, Rasulev B, Roy K (2018) QSPR modeling of the refractive index for diverse polymers using 2D descriptors. *ACS Omega* 3:13374–13386
85. Rasulev B, Jabeen F, Stafslien S, Chisholm BJ, Bahr J, Ossowski M, Boudjouk P (2017) Polymer coating materials and their fouling release activity: a cheminformatics approach to predict properties. *ACS Appl Mater Interfaces* 9:1781–1792
86. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure activity relationships (QSAR). *Expert Opin Drug Discov* 13:1075–1089
87. Roy K, Kar S (2016) In silico models for ecotoxicity of pharmaceuticals. In: *In silico methods for predicting drug toxicity*. Springer, NY, pp 237–304
88. Halling-Sørensen B, Nielsen SN, Lanzky PF, Ingerslev F, Lützhøft HCH, Jørgensen SE (1998) Occurrence, fate and effects of pharmaceutical substances in the environment-a review. *Chemosphere* 36:357–393
89. Bialk-Bielinska A, Caban M, Pieczyńska A, Stepnowski P, Stolte S (2017) Mixture toxicity of six sulfonamides and their two transformation products to green algae *Scenedesmus vacuolatus* and duckweed *Lemna minor*. *Chemosphere* 173:542–550
90. Zou X, Zhou X, Lin Z, Deng Z, Yin D (2013) A docking-based receptor library of antibiotics and its novel application in predicting chronic mixture toxicity for environmental risk assessment. *Environ Monit Assess* 185:4513–4527
91. Zou X, Lin Z, Deng Z, Yin D, Zhang Y (2012) The joint effects of sulfonamides and their potentiator on *Photobacterium phosphoreum*: differences between the acute and chronic mixture toxicity mechanisms. *Chemosphere* 86:30–35
92. Long X, Wang D, Lin Z, Qin M, Song C, Liu Y (2016) The mixture toxicity of environmental contaminants containing sulfonamides and other antibiotics in *Escherichia coli*: differences in both the special target proteins of individual chemicals and their effective combined concentration. *Chemosphere* 158:193–203
93. Wang D, Shi J, Xiong Y, Hu J, Lin Z, Qiu Y, Cheng J (2018) A QSAR-based mechanistic study on the combined toxicity of antibiotics and quorum sensing inhibitors against *Escherichia coli*. *J Hazard Mater* 341:438–447

94. Wang D, Wu X, Lin Z, Ding Y (2018) A comparative study on the binary and ternary mixture toxicity of antibiotics towards three bacteria based on QSAR investigation. *Environ Res* 162:127–134
95. Clever M (2004) Mixture toxicity of the anti-inflammatory drugs diclofenac, ibuprofen, naproxen, and acetylsalicylic acid. *Eco-toxicol Environ Saf* 59:309–315
96. Verhaar HJM, Van Leeuwen CJ, Hermens JLM (1992) Classifying environmental pollutants. *Chemosphere* 25:471–491
97. Van Leeuwen CJ, Van Der Zandt PTJ, Aldenberg T, Verhaar HJM, Hermens JLM (1992) Application of QSARs, extrapolation and equilibrium partitioning in aquatic effects assessment. I. Narcotic industrial pollutants. *Environ Toxicol Chem* 11:267–282
98. Escher BI, Bramaz N, Richter M, Lienert J (2006) Comparative ecotoxicological hazard assessment of beta-blockers and their human metabolites using a mode-of-action-based test battery and a QSAR approach. *Environ Sci Technol* 40:7402–7408
99. Lienert J, Güdel K, Escher BI (2007) Screening method for ecotoxicological hazard assessment of 42 pharmaceuticals considering human metabolism and excretory routes. *Environ Sci Technol* 41:4471–4478
100. De García SAO, Pinto GP, García-Encina PA, Irusta-Mata R (2014) Ecotoxicity and environmental risk assessment of pharmaceuticals and personal care products in aquatic environments and wastewater treatment plants. *Ecotoxicology* 23:1517–1533
101. Mahmoud WMM, Toolaram AP, Menz J, Leder C, Schneider M, Kümmeler K (2014) Identification of phototransformation products of thalidomide and mixture toxicity assessment: an experimental and quantitative structural activity relationships (QSAR) approach. *Water Res* 49:11–22
102. Villa S, Vighi M, Finizio A (2014) Experimental and predicted acute toxicity of antibacterial compounds and their mixtures using the luminescent bacterium *Vibrio fischeri*. *Chemosphere* 108:239–244
103. Hernández AF, Parrón T, Tsatsakis AM, Requena M, Alarcón R, Olga López O (2013) Toxic effects of pesticide mixtures at a molecular level: their relevance to human health. *Toxicology* 307:136–145
104. Tsatsakis AM, Zafiroopoulos A, Tzatzarakis MN, Tzanakakis GN, Kafatos A (2009) Relation of PON1 and CYP1A1 genetic polymorphisms to clinical findings in a cross-sectional study of a Greek rural population professionally exposed to pesticides. *Toxicol Lett* 186:66–72
105. Zeliger H (2011) Human toxicology of chemical mixtures. William Andrew, NY
106. Arnold SF, Klotz DM, Collins BM, Vonier PM, Guillette LJ, McLachlan JA (1996) Synergistic activation of estrogen receptor with combinations of environmental chemicals. *Science* 272:1489–1492
107. Liu S-S, Song X-Q, Liu H-L, Zhang Y-H, Zhang J (2009) Combined photobacterium toxicity of herbicide mixtures containing one insecticide. *Chemosphere* 75:381–388
108. Gutowski L, Baginska E, Olsson O, Leder C, Kümmeler K (2015) Assessing the environmental fate of S-metolachlor, its commercial product Mercantor Gold® and their photo-products using a water-sediment test and in silico methods. *Chemosphere* 138:847–855
109. Jansen E, Michels M, Van Til M, Doelman P (1994) Effects of heavy metals in soil on microbial diversity and activity as shown by the sensitivity-resistance index, an ecologically relevant parameter. *Biol Fertil Soils* 17:177–184
110. Nweke CO, Umeh SI, Ohale VK (2018) Toxicity of four metals and their mixtures to *Pseudomonas fluorescens*: an assessment using fixed ratio ray design. *Ecotox Environ Contam Toxicol* 13:1–14
111. Gikas P (2008) Single and combined effects of nickel (Ni (II)) and cobalt (Co (II)) ions on activated sludge and on other aerobic microorganisms: a review. *J Hazard Mater* 159:187–203
112. Su L-m, Xing Y, Mu C-f, Yan J-c, Zhao Y-h (2008) Evaluation and QSAR study of joint toxicity of substituted phenols and cadmium to *Photobacterium phosphoreum*. *Chem Res Chinese U* 24:281–284
113. Su L, Zhang X, Yuan X, Zhao Y, Zhang D, Qin W (2012) Evaluation of joint toxicity of nitroaromatic compounds and copper to *Photobacterium phosphoreum* and QSAR analysis. *J Hazard Mater* 241:450–455
114. Aggerbeck M, Blanc EB (2018) Role of mixtures of organic pollutants in the development of metabolic disorders via the activation of xenosensors. *Curr Opin Toxicol* 8:57–65
115. Lu GH, Wang C, Wang PF, Chen ZY (2009) Joint toxicity evaluation and QSAR modeling of aromatic amines and phenols to bacteria. *Bull Environ Contam Toxicol* 83:8–14
116. Zeng M, Lin Z, Yin D, Yin K (2008) QSAR for predicting joint toxicity of halogenated

- benzenes to *Dicrateria zhanjiangensis*. *Bull Environ Contam Toxicol* 81:525–530
117. Wang B, Yu G, Zhang Z, Hu H, Wang L (2006) Quantitative structure-activity relationship and prediction of mixture toxicity of alkanols. *Chin Sci Bull* 51:2717–2723
118. Chen CY, Chen SL, Christensen ER (2005) Individual and combined toxicity of nitriles and aldehydes to *Raphidocelis subcapitata*. *Environ Toxicol Chem* 24:1067–1073
119. Lin Z, Niu X, Wu C, Yin K, Cai Z (2005) Prediction of the toxicological joint effects between cyanogenic toxicants and aldehydes to *Photobacterium phosphoreum*. *QSAR Comb Sci* 24:354–363
120. Huang H, Wang X, Shao Y, Chen D, Dai X, Wang L (2003) QSAR for prediction of joint toxicity of substituted phenols to tadpoles (*Rana japonica*). *Bull Environ Contam Toxicol* 71:1124–1130
121. Altenburger R, Schmitt H, Schüürmann G (2005) Algal toxicity of nitrobenzenes: combined effect analysis as a pharmacological probe for similar modes of interaction. *Environ Toxic Chem An Inter J* 24:324–333
122. Pohl HR, Ruiz P, Scinicariello F, Mumtaz MM (2012) Joint toxicity of alkoxyethanol mixtures: contribution of in silico applications. *Regul Toxicol Pharmacol* 64:134–142
123. Jin H, Wang C, Shi J, Chen L (2014) Evaluation on joint toxicity of chlorinated anilines and cadmium to *Photobacterium phosphoreum* and QSAR analysis. *J Hazard Mater* 279:156–162
124. Swartz RC, Schults DW, Ozretich RJ, Lamberson JO, Cole FA, Ferraro SP, Dewitt TH, Redmond MS (1995) ΣPAH: a model to predict the toxicity of polynuclear aromatic hydrocarbon mixtures in field-collected sediments. *Environ Toxicol Chem* 14:1977–1987
125. Choi K, Sweet LI, Meier PG, Kim P-G (2004) Aquatic toxicity of four alkylphenols (3-tert-butylphenol, 2-isopropylphenol, 3-isopropylphenol, and 4-isopropylphenol) and their binary mixtures to microbes, invertebrates, and fish. *Environ Toxicol* 19:45–50
126. Wang T, Lin Z, Yin D, Tian D, Zhang Y, Kong D (2011) Hydrophobicity-dependent QSARs to predict the toxicity of perfluorinated carboxylic acids and their mixtures. *Environ Toxicol Pharmacol* 32:259–265
127. Hoover G, Kar S, Guffey S, Leszczynski J, Sepúlveda MS (2019) In vitro and in silico modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line. *Chemosphere* 233:25–33
128. Kar S, Ghosh S, Leszczynski J (2018) Single or mixture halogenated chemicals? Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach. *Chemosphere* 210:588–596
129. Bucher JR, Lucier G (1998) Current approaches toward chemical mixture studies at the National Institute of Environmental Health Sciences and the US National Toxicology Program. *Environ Health Perspect* 106:1295–1298



Chapter 20

QSPR Modeling of Adsorption of Pollutants by Carbon Nanotubes (CNTs)

Probir Kumar Ojha, Dipika Mandal, and Kunal Roy

Abstract

Harmful effects produced by hazardous chemicals/pollutants toward the environment have been a serious issue of concern since the past. Therefore, a cherished goal of chemists lies in applying novel methods to control the harmful effects of hazardous chemicals/pollutants toward the environment. There are several traditional techniques which are widely used to make the environment free from all types of toxic/hazardous contaminants. Among these processes, adsorption is widely used as an efficient technique to remove various toxic contaminants from the environment due to its low-cost process and because it is easy to perform. Nanotechnology has introduced a new generation of adsorbents like carbon nanotubes (CNTs), which have drawn a widespread interest due to their outstanding ability for the removal of various inorganic and organic pollutants from the environment. CNTs have been widely investigated as alternative adsorbents for the pollution management due to their high surface area and high adsorption affinity toward the organic contaminants, and that they can be modified (functionalized) in different ways to enhance their selectivity toward specific target pollutants. Estimation of adsorption property of environmental pollutants like organic materials, heavy metal ions, radioactive elements, etc. is necessary for both single-walled carbon nanotubes (SWCNTs) and multi-walled carbon nanotubes (MWCNTs). However, considering a sufficient number of such chemicals synthesized in factories and industries, it will be very impractical to carry out an exhaustive testing of chemical hazard. To investigate the toxic property of hazardous chemicals using nanoparticles like CNTs is time-consuming, and it needs animal experimentation. According to Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), use of laboratory animals is causing ethical, scientific, and logistical problems that would be incompatible with the time-schedule envisaged for testing. In this perspective, the non-animal methods like quantitative structure-activity relationships (QSARs) could be used in a tiered approach to provide a rapid and scientifically justified basis to evaluate the adsorption property of different hazardous organic chemicals onto the CNTs. The QSAR modeling investigates the chemical features or structural properties of organic chemicals which are essential for adsorption of hazardous chemicals onto CNTs. The present chapter reviews the information regarding source of hazardous chemicals which are toxic to the environment, risk assessment and management of toxic chemicals, basic information of CNTs, and mechanism of adsorption of organic chemicals into the CNTs. Finally, an overview about the necessity of in silico methods like QSPR modeling for prediction of adsorption property of toxic chemicals as well as successfully reported QSPR models regarding adsorption of hazardous chemicals onto both SWCNTs and MWCNTs are discussed.

Key words Adsorption, CNTs, SWCNTs, MWCNTs, Hazardous chemicals, QSAR, REACH

1 Introduction

Pollutants are the substances, which when introduced into the environment, affect adversely to the human health and ecosystem. Pollutants are produced from various sources like burning of fossil fuels, wastes from incineration, exhausts from automobiles, agricultural processes, and industrial sectors. Some pollutants are biodegradable and cannot persist in the environment for long time. But, most of the pollutants are resistant to environmental degradation processes (biological, chemical, and photolytic) and bioaccumulate in the food chain and affect environment as well as human health. One of the major sources of pollutants are industrial effluents, and it is a challenging job for the environmentalists and industries for proper disposal of the by-products. Polycyclic aromatic hydrocarbons (PAHs) like naphthalene, phenanthrene, *p*-nitrophenol, etc. are very common pollutants, and they have carcinogenic, mutagenic, and toxic properties [1, 2]. These hydrophobic materials are produced from combustion of coal and oil, exhaust from motor vehicles, and effluents from petrochemical plants [3–5]. By normal physico-chemical methods such as coagulation, flocculation, sedimentation, filtration, and osmosis process, they cannot be easily removed from the environment. Chlorobenzenes (1,2,4,5-tetrachlorobenzene, 1,2,4-trichlorobenzene, 1,2-dichlorobenzene, chlorobenzene) are mainly used as solvents, degreasing agents, and chemical intermediates. They cause necrosis, restlessness, tremors, and muscle spasms on little exposure and cause numbness, cyanosis, and hyperesthesia in humans on long-term contact. Among the perfluorinated compounds, perfluorooctanesulfonates (PFOS) are the most common, and they have been utilized as surfactants, fire retardants, paints, adhesives, waxes, and polishes [6]. PFOS are also important contaminants due to their high concentration, global distribution, environmental persistence, and bioaccumulation. These are highly soluble in ground water and causing water pollution. By conventional water purification methods, these compounds cannot be removed easily because of their stability. Dialkyl phthalate esters (DPEs) are most commonly used as plasticizers in polyvinyl chloride, polyvinyl acetates, cellulosics, and polyurethanes and as nonplasticizers in products like lubricating oils, automobile parts, paints, glues, insect repellents, photographic films, perfumes, and food packaging materials (paperboard and cardboard). They have been observed worldwide in food, water and soil, marine ecosystems, affecting human, or other living organisms. They enter the environment during production processes or by leaching from plastic products after disposal and interfere with human hormone-regulated physiological processes [7]. Chlorophenols are generally used for the production of pesticides, dyes, and biocides. Chlorophenols are most

priority environmental pollutants proposed by the US Environmental Protection Agency (US EPA), [8] because they are highly carcinogenic and toxic in nature [9]. Chlorophenols are also produced during water disinfection with chlorine, which make unpleasant taste and odor of drinking water at very low concentration (less than 0.1 mg/L) [10, 11]. Heavy metals like arsenic, cadmium, chromium, mercury, zinc, copper, and lead are the common contaminants in waste water in the recent years. The major sources of these metals are modern chemical industries such as metal plating facilities, battery manufacturing, fertilizer, mining, paper and pesticides, metallurgical, fossil fuel, tannery, and production of different plastics such as polyvinyl chloride, etc. The exposure of these metals causes high blood pressure, speech disorders, fatigue, sleep disabilities, aggressive behavior, poor concentration, irritability, mood swings, depression, increased allergic reactions, autoimmune diseases, vascular occlusion, and memory loss in human [12]. Pharmaceuticals are one of the most important products with undeniable benefits to human health and lifestyle. Unfortunately, since 1970, due to overuse of these products together with their inappropriate disposal, surplus residues of active pharmaceutical ingredients (APIs) have been found in different compartments of environment [13, 14]. Among the pharmaceuticals classes, antibiotics are widely used in healthcare systems which are poorly metabolized after intake; about 25–75% may leave the bodies in an unaffected form after ingestion [15]. In 30 states in 139 rivers of the USA, under nationwide survey of “emerging pollutants,” biologically active compounds of diverse therapeutic classes were detected by the US Geological Survey (USGS) [16]. Hence, it is essential to remove pharmaceuticals like antibiotics, contrast medium, and other contaminants from the environment. Tetracycline antibiotics are mostly used as veterinary therapeutics and growth promoters for animals. After usage by farming industry, tetracyclines are excreted as unmodified parent compounds through feces and urine. Only a small portion of them is metabolized. Large amount of antibiotics like sulfamethoxazole and lincomycin and contrast medium (iopromide) come from effluents of hospitals or radiological clinics to the environment. These pollutants are detected in water and wastewaters. Tetracyclines are commonly noticed in surface water, groundwater, and water which have toxic effect [17]. Therefore, it is essential to remove antibiotics as well as pharmaceuticals and contrast medium to get purified water [18, 19]. Personal care products (PCPs) are generally used to improve the quality of daily life [20]. The use of these PCPs has been increasing from last few years, and these are present in high concentration in the aquatic environment (e.g., water, sediments, and biota) which causes harmful effect to the aquatic organisms. Among other classes of chemicals, dyes are one of the most severe environmental pollutants produced by the textile,

dyeing, printing, ink, and related industries. Dyes are the substances which provide color and mostly used in the textile, pharmaceutical, food, cosmetics, plastics, photographic, and paper industries. Around 800,000 tons of dyes are produced per year worldwide, and out of these, 10–15% of dyes are lost during the dyeing and finishing processes in textile industry. This industry utilizes above 800 chemicals for various processes. During the dyeing processes, the unfixed portion of dyes is washed out, and it remains present in the textile effluents. Thus, the textile and finishing industry presents a large amount of pollution to the environment. Most of the chemicals are harmful and affect adversely to the human health directly or indirectly [21–23]. Dyes not only affect aesthetic merit but also reduce light penetration and photosynthesis. These are also carcinogenic and toxic to the environment and human health [24]. Again, from last two decades, the use of various herbicides has been increased in plant agriculture in controlling bacterial diseases. Studies have suggested that the use of these pesticides is increasing day by day; 2.5 million ton pesticides are used in a year worldwide [25–27]. Generally, organic food involves no residue of synthetic fertilizers, chemical pesticides, genetically modified organisms (GMOs), hormones, and antibiotics [28]. In 2016, the European Food Safety Authority (EFSA) reported that the most detected residues like copper, spinosad (a natural toxin), and bromide ion in organic food are of low concern [29]. However, both conventional and organic foods may contain banned pesticides like hexachlorobenzene, dichlorodiphenyltrichloroethane (DDT), lindane, and dieldrin [30]. Hence, organic foods could contain the same amount or even more of various environmental pollutants, like polychlorinated dibenzo-p-dioxins, polychlorinated dibenzofurans (PCDD/Fs), polychlorinated biphenyls (PCBs), PAHs, and heavy metals, than conventional food [31].

The US EPA has set maximum contamination levels (MCLs) and maximum contamination level goals (MCLG) for each pollutant, with no ill effects on health. In the literature, there are large numbers of pollutants reported, but adsorption data of only around 70,000 pollutants are available, because determination of experimental data of huge number of pollutants is a lengthy, laborious, and expensive process [32]. In the recent years, nanomaterials have gained priority for pollution management, because they contain high surface area, high adsorption affinity toward organic and inorganic pollutants, and they can be modified in various ways to enhance their selectivity toward specific pollutants [33]. Many researchers have special attention to CNTs due to their large specific surface area, small size, inertness toward chemicals, hollow and layered structures, and strong interaction between CNTs and pollutant molecules [34]. CNTs were first discovered by Sumio Iijima in 1991 [35]. CNTs are formed by rolling up of graphene sheets into (concentric) cylinder with nanosize diameter. Generally, their

length is few micrometers. CNTs, as a new kind of adsorbents, have been proven to be of very potential for removal of many types of pollutants like small molecules [36–38], heavy metal ions [39–41], radionuclides [42, 43], and organic chemicals [44, 45]. Several functional groups such as hydroxyl, carboxyl, amine, and ligands can be introduced on the surface of the CNTs for functionalization to make them selective and specific for certain pollutants.

The risk assessment of hazardous organic chemicals requires a huge number of experimental data resulting in high costs, time consumption, and animal testing for *in vivo* testing. Unfortunately, the number of available experimental data is very few. In this regard, quantitative structure-activity/property/toxicity relationship (QSAR/QSPR/QSTR) approach may be a suitable alternative to predict the probable hazards from their chemical structure information [46]. Thus, to fill the data gaps, government and nongovernment regulatory authorities suggest the use of *in silico* methods for prediction of the physicochemical properties, toxicological activity, distribution, fate, etc. of organic chemicals along with their effects on environment and living systems much before they enter into the market for usage. Thus, usage of QSAR as one of the nonexperimental methods is noteworthy in order to minimize animal usage, time, and cost involvement in toxicity prediction of organic chemicals [47, 48]. In the recent years, QSAR/QSPR modeling has been observed to be useful for modeling response of novel chemicals like ionic liquids, nanoparticles, CNTs, etc., thus increasing the area of applications manifolds. QSPR modeling has also been found to be beneficial in agricultural sciences, in nanotoxicology, and in treating environmental pollution. The pollutants discharged into the water bodies from the industries could be modeled against the CNTs (carbon nanotubes) to determine the features which could be essential for uptake by the CNTs. QSPR modeling of organic chemicals/pollutants using adsorption properties by CNTs can be of great importance for environmental scientists to understand the structural or physicochemical properties which are the key features for adsorption of organic chemicals onto CNTs; thus this knowledge can be applied for pollution-free environment.

2 Sources of Pollutants and their Effects

- *Industrial effluents:* These are the major sources of generation of pollutants. Industrial pollutants are formed during production and synthesis of plastics, dyes, polymers, pharmaceuticals, detergents, pesticides, disinfectants, fertilizers, etc. Many literatures [7, 12, 29] suggested that these substances are highly toxic to the animals and plants. Even at very low concentrations, phenolic compounds cause genotoxicity and mutagenicity and reduce

photosynthesis, respiration, and various enzymatic reactions. According to European Commission and the US Environmental Protection Agency (US EPA), phenols and their derivatives are named as hazardous pollutants and are listed as hazardous materials [49].

- *Pharmaceuticals:* Pharmaceuticals and personal care products (PPCPs) including antibiotics, analgesics, steroids, antidepressants, antipyretics, stimulants, antimicrobials, disinfectants, fragrances, cosmetics, and many other chemicals are used to improve human health and lifestyle. After administration, some amount of drugs is metabolized. The unmetabolized active substances are excreted through urine (generally 55–80% of the total administered dose) and partially in feces, thus entering into the environment. Concentration levels of ng/L to mg/L of pharmaceutical wastes have been identified in surface water and groundwater in Asian countries [50]. These pharmaceutical wastes, with continuous accumulation, may cause serious adverse effects on human beings [51].
- *Hospital effluent:* Hospitals and their effluents are important sources of environmental pollution. Various micro-contaminants are obtained from diagnostic, laboratory, and research activities and also from medicine excretion by patients. These include active drugs and their metabolites, chemicals, heavy metals, disinfectants and sterilizants, specific detergents for endoscopes and other instruments, radioactive markers, and iodinated contrast media. Hospital effluents are resulting from their improper removal by conventional systems and detected in hospital waste water which causes toxic effects to the human health [52].
- *Agricultural waste:* Farmers are regularly utilizing fertilizers in their fields which provide necessary nutrients as a form of nitrogen and phosphorus for growing the crops and fruits. The excess nitrogen and phosphorus are drained out to the groundwater resulting in eutrophication of water bodies and degradation of ecosystem. Farmers also use different pesticides to protect the crops from pests. These pesticides are harmful for the environment and living organisms due to their hazardous nature. These pesticides, when released to the environment, may be degraded either by microorganisms or chemical processes. The transformed products may be more toxic than the parent compounds, and consequently these substances may cause a greater risk to the environment. So, their use does involve potential risks to the human health and the environment.
- *Household waste:* The amount of household wastes is increasing gradually due to increasing the world's population. Americans produce 71 million tons of paper garbage, 31 million tons of

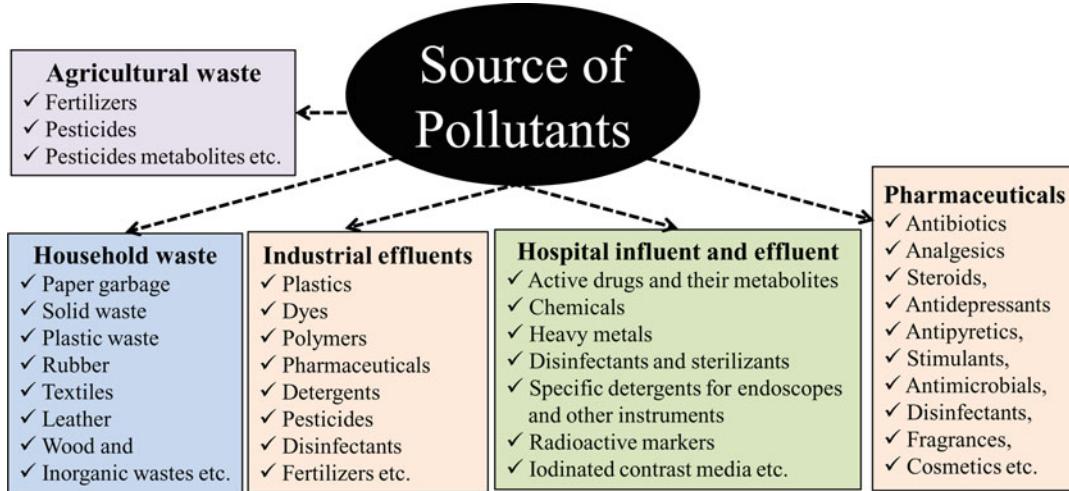


Fig. 1 Sources of pollutants with some examples

solid waste, 14 million tons of plastic waste, and 20 million tons of other materials like rubber, textiles, leather, wood, and inorganic wastes. Many of these are recycled and reused. Proper recycling is necessary to protect the environment and human health. But, due to incomplete combustion of household waste, various pollutants such as polychlorinated dibenzo-p-dioxins, dibenzofurans, biphenyls, chlorobenzene, chlorophenols, and polycyclic aromatic hydrocarbons are formed. These compounds are highly toxic and produce carcinogenic effects and are retained in the environment for longer time [53].

The pollutants obtained from different sources are illustrated in Fig. 1.

3 Risk Assessment of Environmental Pollutants

Risk assessment is a process to determine the concentration, occurrence, and level of exposure of the pollutants to the environment and human health [54]. The main objective of the environmental risk assessment (ERA) is risk mitigation and risk management. The stages of doing an environmental risk assessment are discussed below:

- *Hazard Identification:* It is the first step of risk assessment which identifies the sources and occurrence of environmental hazards and describes whether a compound is able to cause adverse health outcomes at any level of exposure of environmental pollutants. This comprises discussion of any toxicological and epidemiological information [55]. Though many of the research

scientists very much trust on in vivo data, but due to massive deficiency of appropriate data for majority of hazardous chemicals, greater effort should be offered on the proficient use of in vitro analysis, in silico analysis, and computational technique in system biology [56].

- *Dose-Response Assessment:* This can be defined as the “*estimation of the relationship between dose, or level of exposure to a substance, and the incidence and severity of an effect*”[57]. The dose-response relationship is determined from epidemiological and toxicological data.
- *Exposure Assessment:* Exposure assessment is defined as the measurement of the magnitude, frequency, and extent of exposure of the hazardous material to the specified target group in the environment. It detects the source of pollutants, their pathways, and their outdoor exposure to the environment and biomonitoring under numerous exposure scenarios [58]. Environmental exposure to a pollutant may be direct, where pollutants enter into the environment after emission from the industry, or indirect, in which pollutants are coming from drinking water or food chain.
- *Risk Characterization:* It is the last step of risk assessment which describes the nature, likelihood, and magnitude of adverse effects after gathering toxicological and exposure information. It is qualitative and sometime quantitative measurement of possibility of adverse effects of environmental pollutants at specified exposure conditions. This process depends on the results of the previous steps, i.e., environmental hazard and environmental exposure assessment [59].

4 Risk Management of Environmental Pollutants

Environmental risk management is defined as the method of identification, evaluation, selection, and implementation actions to decrease the risk to human health and to the ecosystems. The process of risk management occurred by environmental pollutants need to be well-adjusted and should be balanced with cost benefit and practicality to implement. The policies and environmental strategies that manage the endogenous and exogenous environmental risks are explained below:

- *Waste Prevention and Management:* Waste from various sources including water and energy must be reduced or removed at the source or by practices such as modifying production, maintenance and facility processes, or replacing, preserving, recycling, and reusing materials [60].

- *Hazardous Substance Management:* Risk management process can be controlled by identifying and managing the hazardous chemicals and other substances present into the environment. It is necessary to practice safe handling, movement, storage, use, recycling or reuse, and disposal of the hazardous substances [60].
- *Greenhouse Gas (GHG) Management:* It is essential to record and maintain the greenhouse gas emissions at the facility and corporate level. Carbon oxides production at greenhouse can be reduced by taking initiatives and practices such as use of renewable energy/alternative fuels, filters, freight consolidation, driver efficiency, reduce fuel consumption, etc.
- *Relationship with Suppliers and Customers:* It is one of the most important steps for risk management of pollutants. There is a need to audit and monitor suppliers, and there should be a good relationship between suppliers and customers in order to encourage and consciousness about environmental pollution, cooperation with suppliers to meet environmental objectives [61].
- *Compliance:* The environmental management system is planned according to environmental safety and health regulations.

5 Carbon Nanotubes (CNTs)

5.1 Types of Carbon Nanotubes

5.1.1 Single-Walled CNTs (SWCNTs)

Based on the number of concentrically rolled-up graphene layers, carbon nanotubes are classified into two main categories as follows:

SWCNTs are composed of cylindrical shape single sheet of graphene with a diameter from 0.4 to 2.5 nm [62]. Depending on the chiral indices (n_1, n_2), these nanotubes have two designs such as armchair and zigzag [63] (as shown in Fig. 2). This design depends on the method of wrapping of the graphene sheets into a cylinder. For example, rolling of a sheet of paper from its corner can form one design, and rolling of the sheet from its edge can form another design. The pair of indices (n_1, n_2) are called chiral vector. The chiral indices n_1 and n_2 are equal for armchair CNTs, whereas the chiral indices are zero for zigzag CNTs. The electrical properties of SWCNTs depend on their structural design. The armchair nanotubes are always metallic, while zigzag nanotubes are either metallic or semiconductor [64]. When $n_1 - n_2 = 3i$, where i is a nonzero integer, then the nanotube is called metallic (highly conducting); otherwise the nanotube is semiconducting in nature. SWCNTs have distinctive mechanical, electrical, and thermal properties but possess low solubility as well as poor dispersibility in aqueous and other common organic solvents [65]. They have hydrophobic surface and show high polarizability along with van der Waals

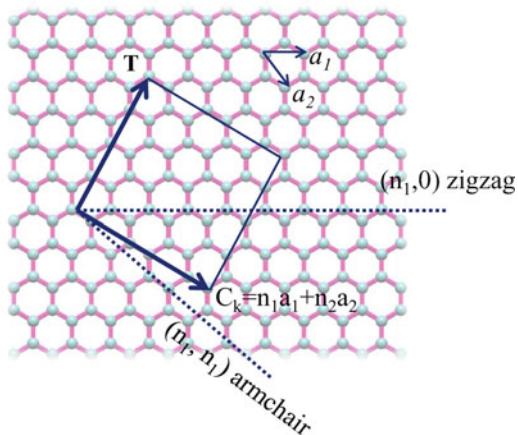


Fig. 2 Chiral structure of SWCNTs. Here, C_k and T are the two vectors, a_1 and a_2 are the two basis vectors, n_1 and n_2 are integers (also called indexes which determines the chiral angle)

interactions; thus they are forming aggregates with each other and with other biological and chemical substances in water [66, 67]. Chen et al. reported that SWCNTs have superior adsorption property than MWCNTs because of the molecular sieving effect. Due to this effect, larger molecules could not enter into the innermost layers of the MWNTs [33, 68]. Chen et al. [69] also observed that SWCNTs have higher adsorption capacity for perfluorooctanesulfonates than MWCNTs, which is due to larger specific surface area (SSA) and smaller diameter of SWCNTs [70, 71]. The molecular structures of SWCNTs are given in Fig. 3a.

5.1.2 Multi-walled CNTs (MWCNTs)

The first discovered CNTs were multi-walled carbon nanotubes (MWCNTs). MWCNTs (Fig. 3b) contain two or more numbers of concentric layers of graphene with different diameters of up to 100 nm. These nanotubes may be few nanometers to a few micrometers long [62]. MWCNTs have two structural models like Russian Doll model and Parchment model. In the Russian Doll model, the graphene sheets are rolled as concentric cylinders (the inner tube has small diameter as the outer nanotubes). On the other hand, in the Parchment model, a single graphene sheet is wrapped around itself several times, resembling a rolled paper. The Russian Doll structure is observed more commonly [72] than the Parchment model. The interlayer distance in MWCNTs is close to the distance between graphene layers in graphite, approximately 3.4 Å. In MWCNTs, the outer layers protect the inner layer from any chemical interactions with outside materials. These nanotubes have higher tensile strength and cheaper than single-walled nanotubes.

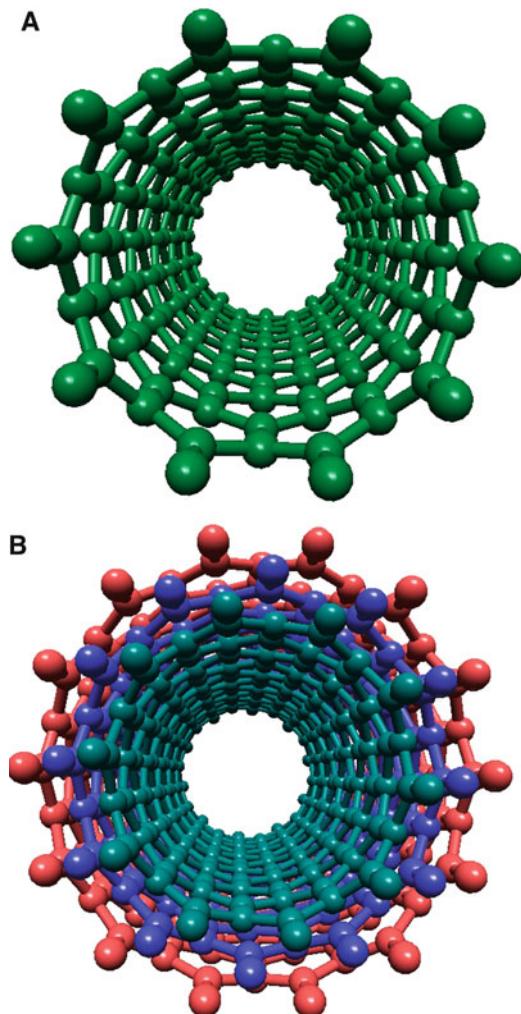


Fig. 3 Molecular structure of CNTs. (a) SWCNT; (b) MWCNT

5.2 Application of CNTs

Nowadays, nanotechnology is one of the most attractive and developing fields which offers many advantages. After discovery in 1991 by Iijima [35], CNTs are growing quickly, and many researchers have given much effort for elucidation of their novel properties and novel applications in different fields. Due to their unique properties and mechanical strength, CNTs are useful in various areas which are discussed below:

5.2.1 Structural

Due to amazing structural properties of CNTs, they can be applicable for [73]:

- *Textiles*: CNTs are useful for production of waterproof and tear-resistant fabrics.

- *Body armor*: CNTs are used to prepare combat jackets. These jackets are bullet proof and comfortable in any season.
- *Concrete*: CNTs can enhance the tensile strength of concrete and reduce the proliferation of halt crack.
- *Polyethylene*: CNT fibers can be used as polyethylene. This type of polyethylene can enhance 30% of elastic modulus of the polymers.
- *Sports equipment*: CNTs are used for production of golf balls, golf clubs, stronger and lighter tennis rackets, bicycle parts, and baseball bats.

5.2.2 Electromagnetic

CNTs can be used for manufacturing of electrical conductors, semiconductors, and insulators:

- *Buckypaper*: CNT sheets are 250 times stronger and 10 times lighter than steel. They can be used as heat sink for chipboards, backlight for LCD screens, or Faraday cage to protect electrical devices/airplanes [74].
- *Light bulb filament*: CNTs are used as substitute of tungsten filaments in incandescent lamps [75].
- *Magnets*: MWCNTs coated with magnetite can produce a strong magnetic field [75].
- *Solar cells*: CNTs can exchange indium tin oxide (ITO) in several solar cells, and light can pass through the graphene layers which generate photocurrent [76].
- *Electromagnetic antenna*: CNTs can also be used as an antenna for radio and other electromagnetic devices because of their durability, lightweight, and conductive properties [75].

5.2.3 Electroacoustic

CNTs are applicable in the field of electroacoustic as:

- *Loudspeaker*: CNTs also find their applications in loudspeakers manufacturing. Such a loudspeaker is able to produce sound having frequency similar to the sound of lightning producing thunder.

5.2.4 Chemical

CNTs have remarkable applications in the chemical field as follows:

- *Air pollution filter*: CNTs provide more adsorption capacity and large specific surface area; therefore, they are used for filtration of air. When polluted air moves toward the CNTs, the conductance will alter, and this is helpful for detecting and filtering the polluted air [77]. CNT membranes can successfully filter carbon dioxide from emissions of different factories and industries.

- *Water filter:* In the recent years, CNTs are used for purification of drinking water. Tangled CNT sheets serve mechanically and electrochemically in strong arrangement with controlled nano-scale porosity. The thin tubes resist the large particles and allow the smaller one to pass through CNTs. These nanotubes have been used for removal of electrochemically oxidized organic contaminants [78], bacteria, and viruses [79]. Portable filters containing CNT meshes are used for purification of contaminated drinking water. Membranes attached with the open ends of the CNTs improve flow properties for both gases and liquids [80].
- *Chemical nanowires:* CNTs are also used for the production of nanowires using gold, zinc oxide, gallium arsenide, etc. The gold-based CNT nanowires can specifically detect hydrogen sulfide (H_2S), and zinc oxide (ZnO)-based CNT nanowires can be used for light-emitting devices and harvesters of vibrational energy [77].
- *Sensors:* Sensors are recently used in different fields as detecting devices such as detection of temperature, air pressure, chemical gases (carbon monoxide, ammonia), molecular pressure, strain, etc. CNTs are attached to increase the efficiency of biosensors and molecular sensors. The working principle of these sensors is generally dependent on the generation of current/voltage. The electric current is generated by the flow of free charged carrier induced on any material. The major advantages of CNTs are the small size of the nanotubes sensing element as well as little amount of material required for a response. At first, Wong et al. [81] revealed that it is possible to sense functional chemical groups connected onto the ends of CNTs by using chemical force microscopy techniques, and Collins et al. [82] reported

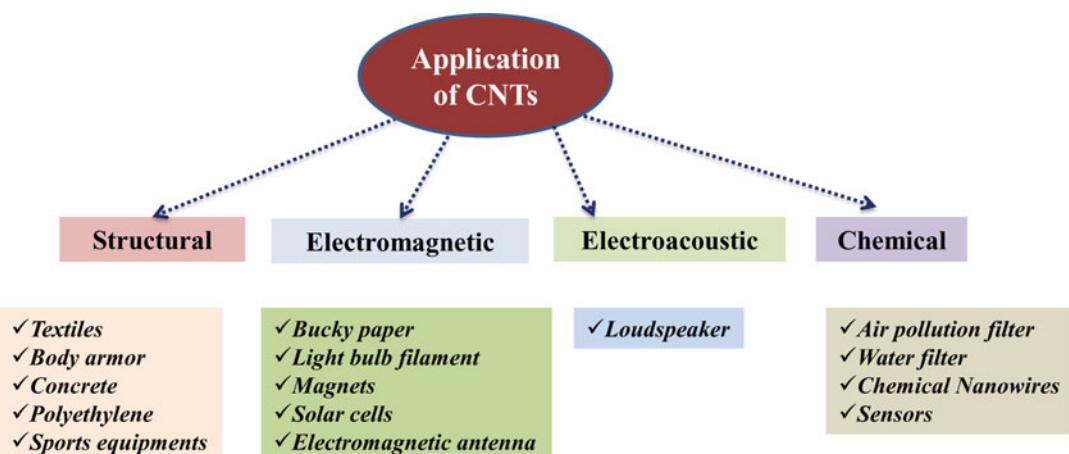


Fig. 4 Applications of CNTs on various fields

that SWCNTs are very sensitive to air and vacuum conditions, and they also suggested that MWNTs can be used as efficient sensors for NH₃, H₂O, CO₂, and CO.

Applications of CNTs on different fields are depicted in Fig. 4.

5.3 Role of CNTs as a Nanomaterial in Pollution Management

Development of agricultural production and fast industrialization make the water resources contaminated with heavy metals. The quality of water is a major concern worldwide because of wastewater discharge from various sources like domestic, industrial, or agricultural. The existence of heavy metals in water resources is a serious issue for both environment and ecosystem. There are several traditional techniques like reverse osmosis, chemical precipitation, filtration, ion exchange, coagulation, and adsorption which are widely used to make the water as well as environment free from all types of toxic contaminants [83]. Among all of these traditional processes, adsorption is widely used as an efficient technique to remove various toxic contaminants from both water and environment due to its low-cost process which is also easy to perform. In this perspective, CNTs are one of the most efficient and widely studied adsorbents due to their high surface area with light mass density, ease of synthesis, and interactions with toxic environmental contaminants [84, 85]. The contaminants mainly found in wastewater are heavy metal ions which are non-biodegradable, highly toxic, and carcinogenic causing accumulative poisoning, cancer, and damaging nervous system. CNTs show a wider adsorption affinity for heavy metals as well as various types of environmental pollutants including organic materials and radioactive elements [77].

5.4 Mechanism of Adsorption of CNTs

Adsorption is the process where atoms, molecules, or ions from a substance like gas, liquid, or dissolved solid adhere to a surface of the adsorbent. Pollutants get adhered to the surface of graphene layers of CNTs through adsorption. The adsorption of various materials by carbon nanotubes takes place due to van der Waals forces, electrostatic interactions, π - π electron donor-acceptor interaction, hydrogen bonding, ion exchange, electrophobic interaction, and mesopore filling. There are several properties which make the CNTs capable of adsorbing many pollutants; for example, (1) the total CNTs surface area is high (100–300 m²/g) which enhance the adsorption capacity, (2) the pore volume of fibrous material is high for easy accessibility, and (3) their surface charge offers a control to select a specific pollutant. Long and Yang [86] reported that adsorption of dioxin on carbon nanotubes could be due to π - π stacking between the two benzene rings of dioxin and the graphite sheets of carbon nanotubes. Ji et al. [17] proposed that a strong interaction between tetracycline and MWCNTs could be caused by van der Waals forces and π - π electron donor-acceptor

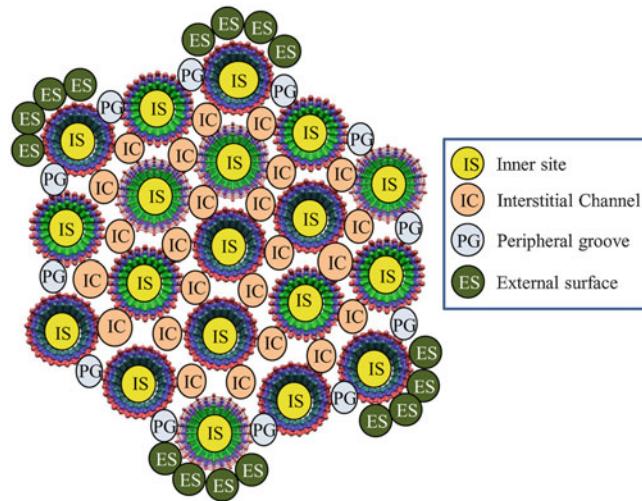


Fig. 5 Major adsorption sites of CNTs in bundle. IS inner site, IC interstitial channel, PG peripheral groove, ES external surface

interaction between MWCNTs and tetracycline. On the other hand, MWCNTs act as the electron donor and conjugated enone structures of tetracycline as the electron acceptor and cation- π bonding between the protonated amino group and the graphene π -electrons also occurred. Pan and Xing [87] suggested that adsorption of 17 α -ethinyl estradiol and bisphenol A on CNTs occur by π - π electron donor-acceptor and hydrogen bonding interactions.

CNTs have four possible sites for adsorption of different pollutants such as (1) open-ended hollow interiors of nanotubes, (2) interstitial pore spaces between the tube bundles, (3) groves present at the boundary of nanotube bundles, and (4) external surface of the outermost CNTs (Fig. 5) [88, 89]. In the interior space of CNTs, the adsorption of pollutants is difficult, because the caps of the individual tubes are generally closed, and if the tubes have open ends, the smaller diameter of the tubes cannot adsorb larger-sized pollutants. Interstitial pore spaces between the tube bundles of CNTs are the good adsorption sites for low-molecular-weight adsorbates (e.g., metal ions) [90]. The grove edges of nanotube bundles and the external surface are the superior adsorption sites for most of the pollutants. Various functional groups can be introduced on the surface of the CNTs for functionalization to increase their colloidal stability and chemical reactivity which make them selective and specific for certain pollutants. The most of the organic and inorganic pollutants are adsorbed at the external surface of the functionalized CNTs [91].

6 Role of Predictive QSPR Models on the Adsorption of CNTs

CNTs are associated with considerable adsorption affinity. Assessment of adsorption property of environmental pollutants (organic materials, heavy metal ions, and radioactive elements) is necessary for both SWCNTs and MWCNTs. However, considering a large number of such chemicals (pesticides, herbicides, fungicides, etc.) synthesized in factories and industries, it will be impracticable to perform an exhaustive testing. Thus, alternative strategies using limited experimental data can be of much use. In this regard, the non-testing methods, i.e., QSPRs could be used in a tiered approach to provide a rapid and scientifically justified basis to evaluate the adsorption property of different hazardous organic chemicals onto CNTs. The predictive QSPR modeling paradigm investigates the chemical features of the hazardous organic chemicals responsible for their high adsorption toward CNTs. Apart from that, QSPR modeling provides an understanding of the important structural requirements or essential molecular properties and the requisite features of molecules that are important to increase or decrease the adsorption of organic contaminants. QSPR models also provide an important guidance for the chemists to increase the efficient application of CNTs which may be useful for reducing the environmental pollution. QSPR models are also supported in the REACH legislation [92]. A few studies have reported predictive QSPR models on adsorptive property of organic chemicals toward CNTs. Modeling physicochemical properties enables design and development of purpose-specific efficient analogues and allows the user to capture specific information on the adsorption coefficient. However, considering the scope of this book chapter, we would like to present an account on some of the representative published QSPR models on adsorption of chemicals onto CNTs.

6.1 Successful QSPR Modeling of Adsorption of Pollutants by SWCNTs

Wang et al. [93] developed two predictive QSPR models with multiple linear regression (MLR) and support vector machine (SVM) algorithms using the adsorption data ($\log K$ values) of 61 organic pollutants onto SWCNTs employing theoretical molecular descriptors. They validated the models extensively using different validation parameters like determination coefficient (R^2), root mean square error for the training set and validation set (RMSE_t and RMSE_v), leave-one-out cross-validated correlation coefficient (Q^2_{LOO}), and external explained variance (Q^2_v) to evaluate the goodness of fit, robustness, and predictive ability. The nonlinear SVM model was developed using the same dataset and molecular structural descriptors as used in the MLR model. The authors also checked the applicability domain of the models using Williams plot (the plot of standardized residuals (d^*) versus leverage values(h)) and claimed that based on the molecular structures of the

compounds in the training set, the applicability domain for the developed models covers diverse compounds with functional groups including $>\text{C}=\text{C}<$, $-\text{C}\equiv\text{C}-$, $-\text{C}_6\text{H}_5$, $>\text{C}=\text{O}$, $-\text{COOH}$, $-\text{C}(\text{O})\text{O}-$, $-\text{OH}$, $-\text{O}-$, $-\text{F}$, $-\text{Cl}$, $-\text{Br}$, $-\text{NH}_2$, $-\text{NH}-$, $>\text{N}->\text{N-N}<$, $-\text{NO}_2$, $>\text{N-C(O)-NH}_2$, $>\text{N-C(O)-NH-}$, $-\text{S-}$, and $-\text{S(O)(O)-}$. Based on the model results, the authors suggested that (1) cyclic compounds having substituents could be well adsorbed by SWNTs, (2) hydrophobic compounds prefer to interact with SWNTs than with water, (3) compounds with higher α values (($100 \times$ molecular polarizability)/volume) tend to have stronger dispersion interactions and dipole-induced dipole forces with SWNTs, and (4) fluorine, oxygen, and chlorine atom were influential toward the adsorption of organic pollutants by SWCNTs. The authors also depicted that the electrostatic interactions between hydrogen atoms and π electrons in SWCNTs as well as the hydrogen bonding interactions between hydrogen atoms and oxygen atoms of water also affected the $\log K$ values. The authors also claimed that molecular dipole moment (μ) encoding the dipole-dipole interactions had the slightest effect on the $\log K$ values among all the theoretical molecular descriptors. Finally, they concluded that the adsorption for organic pollutants onto SWCNTs was influenced by the van der Waals, hydrophobic, electrostatic, and hydrogen bonding interactions. Among these interactions, the van der Waals and hydrophobic interactions contributed most to the adsorption of organic pollutants onto SWCNTs.

Recently, Ghosh et al. [94] reported partial least squares (PLS) regression-based QSPR modeling for adsorption of 40 hazardous synthetic organic chemicals (SOCs) by SWCNTs to identify the significant structural features essential for effective adsorption in SWCNTs, the adsorption behavior of diverse SOCs onto SWCNTs, and to give a deep insight to understand the mechanisms and factors behind the adsorption of hazardous SOCs onto SWCNTs/functionalized SWCNTs. Prior to development of the final models, the authors had applied a variable selection strategy to reduce the noise in the input. The authors validated the models extensively using different validation parameters. Based on the statistical quality, the authors claimed that the models were statistically significant. The authors also checked the consensus predictivity of the developed PLS models using “Intelligent consensus predictor” tool [95] to find out whether the quality of the test set prediction could be enhanced through “Intelligent” selection of models. They found that the consensus predictivity of the models were better than the individual models ($Q^2_{\text{F1}} = 0.938$, $Q^2_{\text{F2}} = 0.937$). From the insights obtained from the PLS regression-based QSPR models, the authors claimed that the hazardous SOCs might get adsorbed onto the SWCNTs through hydrophobic interaction as well as hydrogen bonding interactions and electrostatic interaction to the functionally modified SWCNTs.

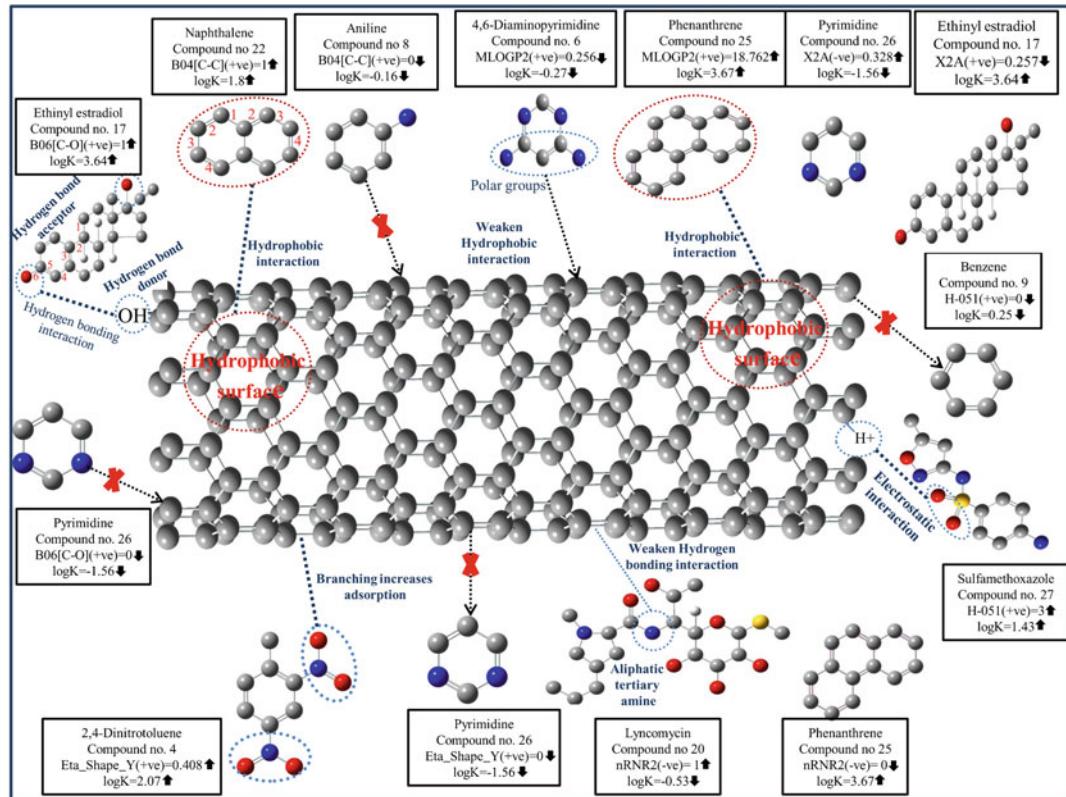


Fig. 6 Model descriptors with their probable interactions patterns onto the SWCNTs as proposed by Ghosh et al. [94]

According to the analysis, the authors interpreted that hydrophobic surface of the molecules, molecular shape and degree of branching, presence of two carbon atoms at topological distance 4, number of H atom attached with α -C atom, and presence of carbon and oxygen atom at the topological distance 6 could enhance the adsorption of hazardous SOCs to the SWCNTs, while number of tertiary aliphatic amine and presence of carbon and sulfur at topological distance 7 might be detrimental for the adsorption of hazardous SOCs to the SWCNTs. The adsorption mechanisms reported by Ghosh et al. of contributing descriptors are depicted in Fig. 6. The authors also suggested that among all the modeled descriptors, MLOGP2 had the strongest impact on the adsorption of hazardous SOCs onto SWCNTs. Finally, the authors claimed that the developed models might provide knowledge to scientists to boost the efficient application of SWCNTs as adsorbents, which might be useful for the management of pollution free environment.

Lata and Vikas [96] reported QSPR models for the adsorption coefficient of 40 aromatic organic compounds by SWCNTs using quantum-mechanical descriptors to identify the key structural

information at the electronic level which could be important for the adsorption of aromatic organic chemicals by SWCNTs. The authors examined the real predictivity of the existing linear solvation energy relationship (LSER) models for the adsorption prediction of OCs by SWCNTs and compared them with the developed quantum-mechanical models by using state-of-the-art statistical procedures. For this purpose, they employed an external set compound which was not used for the model development purposes. Based on the insights obtained from the models, they found that the mean polarizability was the key structural information among all, which affect most for the adsorption of OCs by SWCNTs. This contribution of polarizability was due to the interactions between electrons of parallel spin. The authors also proposed that results obtained from the models developed from the mixture of quantum-mechanical descriptors and solvatochromic descriptors were found to be better than the models developed from individual descriptors. Finally, the authors used the proposed models to predict the adsorption efficiency of nucleobases, steroid hormones, and selective agrochemicals like insecticides, herbicides, pesticides, and endocrine disrupting chemicals. They claimed that based on the model predictions, Guanine and Progesterone should be strongly adsorbed by the SWCNTs. The authors finally proposed that the models could be used to predict the nanotoxicity associated with the adsorption of biomolecules and other environmental pollutants by SWCNTs. These authors [96] developed the LSER models based on the algorithm of Abraham and coworkers [97, 98] as follows:

$$\log K_d = rR + pP + \alpha A + bB + \nu V + c \quad (1)$$

where K_d is the adsorbent-water distribution coefficient (in L/g), while A , B , V , P , and R are solvatochromic descriptors of adsorbate (solute) molecules representing their interaction with adsorbent and solvent [99, 100]. The parameter, R (in $\text{cm}^3 \text{ mol}^{-1}/10$), stands for the excess molar refractivity relative to a compound of the same molar volume. Another parameter, P , signifies dipolarity or polarizability; the parameter, A , depicts effective hydrogen bond donating ability (the acidity); and B represents the effective hydrogen bond accepting ability (the basicity) of the adsorbate. The parameter V in ($\text{cm}^3 \text{ mol}^{-1}$)/100 is the characteristic McGowan volume that is known to generalize the dispersion interactions [101]. In equation, r , p , α , b , and ν are the regression coefficients, and c is the regression constant.

Liu et al. [102] reported a QSAR model and DFT simulation of 25 benzene derivatives to explore the adsorption characteristics and to see the key interactions pattern to SWCNTs. To illustrate the preferential molecule-SWCNTs conformations, the authors built an armchair SWCNT (3, 3) with a diameter of 4.07 Å, a length of 12.30 Å along the tube axis, and added terminating hydrogen

atoms to both ends of the SWCNs using Materials Studio (<http://accelrys.com/products/materialsstudio/>). They designed 6 to 10 interaction modes to explore the potential interaction configuration for each molecule-SWCNT system in the gas phase. All geometries were completely optimized in all internal degrees of freedom using DFT at the M062X/6-31G(d) levels of theory. They performed DFT simulations to optimize the structural, dynamic, and energetic aspects of the molecule-SWCNT complexes and to determine the adsorption mechanisms. The authors developed a QSAR model containing 8 descriptors by stepwise regression method using 79 three-dimensional SurVolSha (surface area, volume and shape) descriptors. Based on the statistical results, they claimed that the model had favorable predictive capability and could be utilized to predict the molecule-SWCNT adsorption. From the insights obtained from the model, they reported that π - π stacking was dominating the adsorption of benzene derivatives onto the SWCNTs, whereas the substituents played a secondary effect on the adsorption process. They also investigated from the final optimized molecule-SWCNT configurations that the parallel modes have higher occurrence probabilities than the perpendicular modes. From this observation, they concluded that the parallel modes are more stable than the perpendicular modes for the same molecule-SWCNT system. Thus, face-to-face π - π packing dominated the adsorption interaction of the aromatic molecules-SWCNTs compared to face-to-side π - π packing. From the benzene-SWCNT simulation results, they suggested that benzene preferred to interact with the SWCNTs by a bridge configuration with a relatively low energy. Finally, they suggested that one could modify the functional group of the derivatives to acquire expected adsorption on the SWCNTs and could carry out non-covalent functionalization of CNTs which could be utilized for the organic compounds from the environment.

6.2 Successful QSPR Modeling of Adsorption of Pollutants by MWCNTs

Roy et al. [68] have recently reported predictive QSPR models for adsorption of diverse organic pollutants by MWCNTs using two different datasets containing 59 and 69 organic pollutants with multiple end points to explore the key structural features essential for adsorption to multi-walled CNTs employing only easily computable 2D descriptors. The first dataset contained defined adsorption affinity properties (k_∞) of 59 diverse organic pollutants by MWCNTs, and the second dataset contained adsorption affinity of 69 organic pollutants related to specific surface area (k_{SA}) of MWCNTs. They have converted all the end point values into logarithmic scale for the modeling purpose. The authors have employed a variable selection approach prior to development of the final models to reduce the noise in the input. The models were extensively validated (both internal and external) using different stringent statistical validation parameters like R^2 , adjusted

determination coefficient (R_a^2), variance ratio (F), standard error of estimate (s), and leave-one-out cross-validated correlation coefficient (Q^2_{LOO}); external predictivity parameters like R^2_{pred} , Q^2_{F2} , and concordance correlation coefficient (CCC); and r_m^2 parameters like $r_m^2(LOO)$ and Δr_m^2 for internal validation and $r_m^2(test)$ and $\Delta r_m^2(test)$ for external validation, mean absolute error (MAE) criteria for both external and internal validation tests. The authors have also tried to explore whether the quality of predictions of test set compounds can be enhanced through an “intelligent” selection of multiple MLR models using the “Intelligent consensus predictor” tool. Based on both internal and external validation parameters, the authors suggested that the statistical results of the reported models showed good predictivity (Dataset 1: $R^2 = 0.893\text{--}0.920$, $Q^2_{(LOO)} = 0.863\text{--}0.895$, $Q^2_{F1} = 0.887\text{--}0.919$; Dataset 2: $R^2 = 0.793\text{--}0.845$, $Q^2_{(LOO)} = 0.743\text{--}0.811$, $Q^2_{F1} = 0.783\text{--}0.890$). The authors claimed that the consensus predictivity for the test set compounds showed better results than those from the individual MLR models based on not only the MAE-based criteria but also the other external validation metrics they used (Dataset1, $Q^2_{F1} = 0.935$, $Q^2_{F2} = 0.935$, $MAE_{(95\%)} = \text{Good}$; Dataset2, $Q^2_{F1} = 0.887$, $Q^2_{F2} = 0.879$, $MAE_{(95\%)} = \text{Good}$.). From the information obtained from the developed models, the authors suggested that higher number of aromatic rings, high unsaturation or electron richness of molecules, polar groups substituted in aromatic ring, presence of oxygen and nitrogen atoms, size of the molecules, and hydrophobic surface of the molecules could enhance the adsorption of the organic pollutants to the CNTs, while presence of C-O group, aliphatic primary alcohol, and presence of chlorine atoms might retard the adsorption of organic pollutants. From these observations, the authors concluded that the organic pollutants might get adsorbed on to the CNTs through hydrogen bonding, $\pi\text{-}\pi$ stacking, hydrophobic, and electrostatic interactions. Finally, the authors claimed that the reported models might be helpful in the process of removal of the harmful and toxic contaminants/disposals of the by-products from the various industries by increasing the adsorption of pollutants, hence making a pollution free environment.

Lata and Vikas [103] reported the role of quantum-mechanical descriptors in the concentration-dependent adsorption of 64 diverse aromatic compounds by MWCNTs using QSPR modeling. They used a dataset containing aromatic organic compounds comprising drugs, herbicides, pesticides, cosmetic constituents, dyes, and so forth, at five different adsorbate equilibrium concentrations (C_e) such as 10^{-1} , 10^{-2} , 10^{-3} , 10^{-4} , and 10^{-5} of the adsorbate (in mg/L). The authors developed three types of models using quantum-mechanical descriptors, solvatochromic descriptors (poly-parameter linear-solvation-energy relationships (pp-LSERs)

models), and mixed types of descriptors (using both quantum-mechanical descriptors and solvatochromic descriptors) and compared the results among them. Different strategies were employed to develop the models like single parametric models, multiple parametric models, etc. They validated the models both internally and externally. Based on the results, the authors claimed that the models developed from the mixture of descriptors performed better than the individual type of descriptors. Based on the performance of the models, the authors claimed that the most influencing descriptor for the adsorption of aromatic organic compounds by MWCNTs were mean polarizability but that arising from the quantum-mechanical exchange interactions between electrons of the same spin. Finally, the models developed from the combination of the quantum-mechanical descriptors and pp-LSER's descriptors were used to predict the adsorption of nucleobases and steroid hormones.

Ahmadi and Akbari [104] explored predictive QSPR models of the surface area normalized adsorption coefficients of 69 aromatic organic compounds on MWCNTs using the Monte Carlo method. The descriptors were calculated with the simplified molecular-input line-entry system (SMILES) and hydrogen-suppressed molecular graphs (HSGs). The authors developed the models using the CORAL (CORrelation and Logic) software [105]. The authors divided the whole dataset randomly into three sets, namely, training, calibration, and validation sets (balance of correlations method). They also used an invisible training set which were used for the confirmation of good correlation coefficients for compounds that are not involved in the training set. The calibration set was used for the optimization of the Monte Carlo search. Finally, the validation set was used to predict the properties of the aromatic organic compounds which were not used for the development of the models. For the division of the dataset, the authors maintained the ratio of 35% training, 35% invisible training, 15% calibration, and 15% validation. The authors claimed that the results obtained from three random splits were robust, very simple, predictable and reliable for the training, invisible training, calibration, and validation sets. Finally, they proposed that the reported QSPR model could be used for the prediction of the adsorption coefficient of numerous aromatic compounds on MWCNTs.

Paszkiewicz et al. [106] reported principal component analysis (PCA) to explore the similarities between the studied polycyclic aromatic hydrocarbons (PAHs) based on quantum-mechanical (QM) descriptors calculated using the second-order Møller-Plesset (MP2) perturbational method and 6-311++G(d,p) basis set and constitutional descriptors. The authors claimed that the first two principal components explained 77% (PC1) and 11% (PC2), respectively, of the total variance in the data. Based on the PCA score plot, the authors found a high correlation between PC1 and ring count

defining descriptors like nCIC (number of rings) which represented the size of the PAHs. The authors observed that the key descriptors like Rbrid (ringbridge count), nCIR (number of circuits), RCI (ring fusion density), nAT (total number of atoms), and nC (total number of carbon atoms) are also closely related to the size of the particles. Thus, they claimed that the features related to the size of the PAHs were important to regulate the adsorptive property of PAHs to the MWCNTs. They also observed that the cumulative electronegativity, polarizability of the atoms, and the energy gap between HOMO and LUMO orbitals ($\text{GAP} = E_{\text{LUMO}} - E_{\text{HOMO}}$) were also important. The energy gap between HOMO and LUMO orbitals ($\text{GAP} = E_{\text{LUMO}} - E_{\text{HOMO}}$) was an important indicator of kinetic stability. On the other hand, PC2 was mostly related to the number of hydrogen atoms. Further, to understand the adsorbing characteristics of PAHs to the CNTs, they performed a theoretical investigation on the interaction mechanisms between PAH molecules and MWCNTs using the PM6 method.

Wang et al. [107] reported quantitative nanostructure-property relationship (QNPR) modeling of adsorption coefficients data (represented by $\log K_\infty$ and $\log K_{\text{SA}}$) of diverse organic compounds (two datasets: one containing 59 organic compounds with adsorption coefficient data ($\log K_\infty$) and another containing 69 organic compounds with adsorption coefficient data related to specific surface area ($\log K_{\text{SA}}$)) for multi-walled CNTs (MWCNTs). They developed QNPR models using norm index descriptors to predict the adsorption affinity of OCs to MWNTs. In their work, after energy minimization of all compounds at the STO-3G level, they calculated norm index descriptors based on a series of distance matrix including step matrix DM1, adjacent matrix DM2, and Euclidean distance matrix DM3. Acceptable statistical results corresponding to measures of fitness, robustness, and predictivity were reported for the developed models (squared correction coefficient for the training set and the test set of 0.9500 and 0.9792 for $\log K_\infty$ and 0.9258 and 0.9770 for $\log K_{\text{SA}}$, respectively). The authors also checked the domain of applicability of the models using the plot of standardized residuals versus leverage values and claimed that all the compounds were present within acceptable domain. Furthermore, the authors also performed Y-randomization analysis for seven times and reported that all the seven random models showed lower R^2 and $Q^2(\text{LOO}_{\text{CV}})$ values than the actual models. They also claimed that among different norm index metrics, property metrics along with some atomic properties played an important role for the adsorption of OCs by MWNTs. After mechanistic interpretation of the developed models, they concluded that the OCs get adsorbed to the MWCNTs through different physicochemical interactions such as hydrophobic, π - π interaction, hydrogen bonding, and electrostatic interactions. The authors claimed that norm index descriptors were

suitable for adsorption of OCs; thus, the interpretation of descriptors based on atomic contribution of the molecule was comparatively easier. The authors suggested that the widespread and prospective applications of norm index descriptors in future in the field of nanotechnology might be possible.

A quantitative structure-property relationship (QSPR) model was reported by Heidari and Fatemi [108] based on adsorption coefficient of 40 different small aromatic organic chemicals to MWCNTs using CORAL software. The authors claimed that CORAL software tool was used for the first time for the development of QSPR models of adsorption coefficients of chemicals on CNTs. The authors developed two models by using hydrogen-filled graph-based descriptors (for model 1) and hybrid descriptors (for model 2) which were the combination of SMILES and hydrogen-filled graph-based (HFG) descriptors. They used a newer technique which involves three-dimensional response surfaces of all subsets to optimize the Monte Carlo parameters of models. For model development, the whole dataset was divided into sub-training (developer of the model), validation (avoider of overtraining), and test (an estimator of predictability) sets. The results portrayed acceptability of both models. Based on the statistical results, the authors claimed that HFG descriptor-based model was better than the hybrid-based descriptor model. The authors also suggested that CORAL software tool could be very much useful in future for the modeling of adsorption coefficients on nanoparticles. The authors also reported that descriptors generated by CORAL software tool were important and easy to interpret.

Wang et al. [109] reported a computational study on the interaction of functionalized MWCNTs and bisphenol AF (BPAF) to find out different kind of non-covalent interactions. Fluorescence spectra technique was applied to evaluate binding mechanism between carboxylic MWCNTs and BPAF. The authors also performed an experimental process to evaluate adsorption of BPAF onto carboxylic MWCNTs. At first, they used theoretical data to recognize the interaction between MWCNTs-COOH with BPAF at the molecular level. The electronic transition of BPAF was calculated with the help of density functional theory. From the molecular modeling, the authors concluded that two types of binding modes can exist between MWCNTs-COOH and BPAF: one was insert-binding mode and another was surface-binding mode. The binding ability of the insert-binding mode was claimed to be stronger than the surface-binding mode. Secondly, from the fluorescence experimental data, the authors reported that the interaction between MWCNTs-COOH and BPAF happened because of several non-covalent binding forces like hydrophobic, π - π stacking, and hydrogen bonding. From the experimental data, they also confirmed that adsorption equilibrium followed pseudo-second-order model and could be obtained within 5 min; hence, MWCNTs-

COOH shows good adsorption toward BPAF. Finally, the authors concluded that all the experimental and theoretical results could be useful for designing a new process to remove endocrine-disrupting chemicals from water.

Toropova et al. [110] performed a QSPRs study on adsorption ($\log K_\infty$) of 59 organic contaminants by MWCNTs using CORAL software package to develop predictive QSPR models which was based on the Monte Carlo technique. In this work, they divided the dataset randomly into the sub-training, calibration, test, and validation sets. Using those distributions together with the distribution from the work of Apul et al. [111], five various QSPR models for the $\log K_\infty$ were built. The authors divided the dataset into training and test sets in the form of different splits, i.e., Split1, Split2, Split3, and Split4. Among them, they claimed that Split4 was interpreted as a “successful split,” in other words, a “successful random event,” from the point of view of the user of the CORAL software. The Monte Carlo optimization generated four molecular features (S_{Ak} , NOSP, BOND) such as (1) the features, which have only positive values of the correlation weights, could be classified as promoters of $\log K_\infty$ increase; (2) the features, which had only negative values of the correlation weights, could be classified as promoters of $\log K_\infty$ decrease; (3) the features, which had both positive and negative correlation weights for different runs of the Monte Carlo optimization, could be classified as features with unclear role; and, finally, (4) the features which were blocked according to the used threshold. After the mechanistic interpretation of the models, the authors claimed that the presence of aromaticity, absence of double and triple bonds, presence of nitrogen, and presence of the branching should be classified as stable promoters of the $\log K_\infty$ increase, whereas the presence of single cycle (“1”) and the presence of oxygen should be classified as stable promoters of $\log K_\infty$ decrease.

QSPR models were reported by Rahimi-Nasrabadi et al. [112] to predict the adsorption of 59 aromatic organic compounds by MWCNTs. They performed linear and nonlinear QSPR models by employing K_∞ as the dependent variable. The values of the dataset were transformed into logarithmic scale, and the relationship between $\log K_\infty$ and molecular descriptors was examined. The training set, test set, and validation set consist of 43, 5, and 11 compounds, respectively. The authors used HyperChem program (ver. 7) for the calculation of descriptors. Along with it, Austin Model1 (AM1) was used as the semiempirical method to optimize the molecular geometry. Dragon descriptors [113] and four Abraham descriptors were also added to the pool of descriptors in order to study the possible nonlinear relationship between $\log K_\infty$ and the mentioned descriptors. In order to choose the most significant descriptors among the molecular descriptors, they performed numerous stepwise MLR models and nonlinear multilayered

perceptron neural network (MLP-NN) models. They selected five descriptors by using self-organizing map for model development purposes. The authors employed both linear and nonlinear techniques to connect the structure of the studied chemicals with their adsorption descriptor (K_∞) using stepwise multiple linear regression (MLR) techniques. Both the models (linear and nonlinear models) showed statistically good results and were validated well internally and externally. Based on the results, they suggested that MLP-NN model was better than the MLR one. From this, the author proposed that the relationship between the structures of the organic compounds and their adsorption on MWCNTs was nonlinear. Based on the applicability domain (leverage approach), they found that there was no high leverage compound in the training, test, and validation sets. After mechanistic interpretation of the models, the authors suggested that molar volume and hydrogen bond accepting ability, molecular mass and size, amount of branching, and three-dimensional structure were the essential features to characterize and control the adsorption of the organic compounds by MWCNTs.

Hassanzadeh et al. [114] reported QSPR modeling of adsorption of diverse organic chemicals by MWCNTs for two datasets consisting of 40 and 69 compounds with multiple end points. The end points of both the dataset were taken in the logarithm scale. They reported QSPR models for dataset 1 containing 40 diverse organic pollutants with defined adsorption affinity properties (k_∞) for MWCNTs using solvatochromic descriptors as independent variables, whereas dataset 2 contained adsorption affinity of 69 organic pollutants related to specific surface area (K_{SA}) for MWCNTs, and here 3D molecular descriptors were used as independent variables. The authors found that 39 compounds from dataset 2 were common with dataset 1. The authors developed nonlinear models using a combination of radial basis function neural network (RBFN) and genetic algorithm (GA) called as whole space GA-RBFN (wsGA-RBFN) which was introduced for better description of QSPR models. The authors validated the models both internally and externally using various stringent statistical parameters. Based on the statistical results, the authors claimed that the approach called whole space GA-RBFN (wsGA-RBFN) was significant to predict adsorption coefficient ($\log K_\infty$) of dataset 1 and $\log K_{SA}$ for dataset 2 as compared to GA-RBFN and MLR models. The authors also claimed that it is not required to keep the row of independent variables as centers for RBFN as mentioned by other authors, any point in the whole space of independent matrix could be used as the center. After comparison of the results of the developed models using solvatochromic descriptors and physicochemical descriptors, the authors concluded that the results obtained from the solvatochromic descriptors were better than the physicochemical descriptors.

6.3 Successful QSPR Modeling of Adsorption of Pollutants by Both SWCNTs and MWCNTs

Ersan et al. [115] applied linear solvation energy relationship (LSER) technique to develop predictive models using adsorption isotherm data of both aromatic and aliphatic organic compounds (OCs) by graphene and graphene oxide (GO). They further compared the results with those of single-walled carbon nanotubes (SWCNTs) and multi-walled carbon nanotubes (MWCNTs). They used a database of 38 OCs (28 aromatic and 10 aliphatic) for graphene and 69 OCs (59 aromatic and 10 aliphatic) for GO for development of LSER models. The authors developed the model using a single-point adsorption descriptor (K_d) values (L/g) for each OC at specific levels of chemical saturation of 10^{-4} , 10^{-3} , and 10^{-2} which was denoted as $\log K_{d,0.0001}$, $\log K_{d,0.001}$, and $\log K_{d,0.01}$. The solvatochromic parameters used to develop the models were hydrogen bond donating capability (acidity) (denoted by A), hydrogen bond accepting capability (basicity) (denoted by B), McGowan's molecular volume (denoted by V), polarizability and dipolarity term (denoted by P), and excess molar refractivity (denoted by R). The authors found that the r^2 value of LSER models increased with decreasing equilibrium concentration ($\log K_d$) of aromatic OCs for GO, while the changes were minor for graphene. They investigated the impact of molecular weight for adsorption of aromatic OCs in graphene sheet and carbon nanotubes. The authors observed that the higher r^2 values of LSER models were obtained in case of graphene sheet with wider molecular weight range (up to 950 g/mol) of OCs, while in case of CNTs, higher r^2 values were observed with the models developed from lower range of molecular weight (<200 g/mol). They also found that the linearity of LSER models for GO gradually decreased with an increase in molecular weight above 400 g/mol, and they suggested that this was due to complex intermolecular interactions of OCs with polar GO surfaces. These complex intermolecular interactions of OCs with polar GO surfaces were due to the presence of oxygen containing functional groups. From the LSER models, they observed that McGowan's molecular volume (denoted by V) and hydrogen bond accepting capability (basicity) (denoted by B) terms were the most influential descriptors in the LSER equations for adsorption of aromatic and aliphatic OCs by graphene and GO. Based on the regression coefficient of the LSER model, they also claimed that hydrogen bond donating capability (acidity) (denoted by A) and hydrogen bond accepting capability (basicity) (denoted by B) properties contributed negatively, while McGowan's molecular volume (denoted by V) property contributed positively for the adsorption of all tested OCs and aliphatics by graphene, GO, and SWCNTs. On the other hand, in case of the GNS model, polarizability and dipolarity term (denoted by P) and excess molar refractivity (denoted by R) did not show a clear trend when compared to CNTs as reported by the authors.

Wang et al. [116] successfully developed 3D QSPR models for the adsorption coefficient of 39 aromatic organic chemicals onto MWCNTs using three learning approaches, namely, MLR, artificial neural network (ANN), and support vector machine (SVM) to investigate the essential physicochemical properties which were important for the adsorption of organic compounds to MWCNTs. The authors also compared the results obtained from the models constructed by using three different approaches. Based on the statistical results, the authors found that the nonlinear SVM and ANN models were far better than the MLR model. From the insights obtained from the models, the authors suggested that number of nitrogen and oxygen atoms (#NandO), octanol/water partition coefficient ($\log K_{ow}$), and number of atoms in rings of a ring molecule (#ringatoms) contributed positively toward the adsorption of OCs to MWCNTs, while dipole moment of the molecule and estimated number of hydrogen bonds that would be accepted by the solute from water molecule in an aqueous solution (accptHB) had the detrimental effects toward the adsorption of OCs to MWCNTs.

6.4 Successful QSPR Modeling of Adsorption of Heavy Metal Ions by MWCNTs

Salahinejad and Zolfonoun [117] reported Quantitative Ion Character-Activity Relationship (QICAR) model using the maximum adsorption capacity (q_{max}) of 25 heavy metal ions on MWCNTs to find out the important property/properties which could remove the heavy metals and to understand the probable adsorption mechanism of heavy metals on MWCNTs. Prior to the development of final models, the authors employed variable selection strategies to select the best subset of independent variables using three methods, namely, genetic algorithms (GA) (an optimization method and heuristic search technique based on natural evolution and selection), enhanced replacement method (ERM) (an optimal desired number of descriptors is selected based on searching the pool of descriptors which produce a linear model with minimize standard deviation(s)), and successive projection algorithm (SPA) (a forward selection method which uses a robust recursive algorithm to minimize variable collinearities problems). The final model was developed by using the PLS regression technique. They validated the models both internally and externally using different validation parameters. Based on the model results, the authors claimed that the models were robust. From the insights obtained from the QICAR models, the authors concluded that for the adsorption of heavy metal ions on MWCNTs, electronegativity, ionic radius, and atomic number of the heavy metal ions were the important parameters.

6.5 Molecular Docking of Organic Compound to CNTs

To investigate the mechanism behind the adsorption of fenuron to CNTs, Ali et al. [118] reported a molecular docking study using AutoDock 4.2 software. From the docking study, the authors suggested that fenuron might get adsorbed on to the CNTs through $\pi\text{-}\sigma$, $\pi\text{-}\pi$ stacked, $\pi\text{-}\pi$ T-shaped, and $\pi\text{-}$ alkyl types of hydrophobic interactions. They also reported that the binding energy and binding affinity between fenuron and CNTs were $6.5 \text{ kcal mol}^{-1}$ and $5.85 \times 10^4 \text{ M}^{-1}$, respectively.

6.6 Overview

In overview, as evident from the reported in silico models reviewed above (Subheading 6), the pollutants get adsorbed to the CNTs through different physicochemical forces like van der Waals forces, electrostatic interactions, $\pi\text{-}\pi$ stacking, electron donor-acceptor interaction, hydrogen bonding, ion exchange, and hydrophobic interactions.

7 Conclusion

In general, this chapter deals with an overview of hazardous chemicals and their effects on environment and the strategy of removal of those chemicals from the environment using the adsorption property of CNTs. Apart from these, we have also tried to highlight the necessity of in silico models for removal of pollutants from the environment as well as discussed the reported QSPR models for the adsorption data of hazardous chemicals onto different types of CNTs like SWCNTs and MWCNTs. A sufficient number of chemicals belonging to different categories like pesticides, herbicides, fungicides, organic materials, heavy metal ions, radioactive elements, etc. are synthesized routinely in factories and industries. The environmental concern starts when these contaminants enter into different compartments of environment and make them polluted. The laboratory testing of these wide varieties of pollutants is impracticable because it is time-consuming and due to cost, involvement of animals, and involvement of large number of labor. As discussed previously that adsorption is widely used as an efficient technique to remove various toxic contaminants from the environment due to its low-cost process and easy execution. Different types of CNTs have introduced a new generation of adsorbents which have drawn a widespread interest due to their outstanding ability for the removal of various inorganic and organic pollutants from the environment. In this regard, researchers have come up with an alternative method like QSPR for prediction of adsorption property of organic pollutants by CNTs. Using QSPR, one can utilize the available adsorption data of pollutants onto CNTs in order to predict the same for untested or not yet synthesized chemicals. Insufficient adsorption data related to a definite class of chemicals has slowed down the computational approaches to some

extent. A limited number of adsorption data on CNTs are available to make meaningful QSPR models. Therefore, it is the time to develop properly documented databases for environmentalists. The government and nongovernment authorities should carefully handle the risk assessment and management of hazardous chemicals with proper regulation to make the pollution free environment. It is obvious that *in silico* models cannot substitute the experimental approaches, but combination of both these approaches can give us the better understanding and quantification of adsorption property of pollutants or hazardous chemicals by CNTs.

Acknowledgments

P.K.O. acknowledges the financial support from UGC, New Delhi, India, in the form of a fellowship (Letter number and date: F./PDFSS-2015-17-WES-11996; dated: 06/04/2016). K.R. wishes to thank CSIR, New Delhi for financial assistance under a Major Research project (CSIR ProjectNo.01IJ2895)/17/EMR-II).

References

- May WE, Wasik SP, Freeman DH (1978) Determination of the solubility behavior of some polycyclic aromatic hydrocarbons in water. *Anal Chem* 50:997–1000
- Walters RW, Luthy RG (1984) Equilibrium adsorption of polycyclic aromatic hydrocarbons from water onto activated carbon. *Environ Sci Technol* 18:395–403
- Nielsen T (1996) Traffic contribution of polycyclic aromatic hydrocarbons in the center of a large city. *Atmos Environ* 30:3481–3490
- Harrison RM, Smith DJT, Luhana L (1996) Source apportionment of atmospheric polycyclic aromatic hydrocarbons collected from an urban location in Birmingham, UK. *Environ Sci Technol* 30:825–832
- Domeno C, Nerin C (2003) Fate of polycyclic aromatic hydrocarbons in the pyrolysis of industrial waste oils. *J Anal Appl Pyrolysis* 67:237–246
- Moody CA, Field JA (2000) Perfluorinated surfactants and the environmental implications of their use in fire-fighting foams. *Environ Sci Technol* 34:3864–3870
- Wang F, Yao J, Sun K, Xing B (2010) Adsorption of dialkyl phthalate esters on carbon nanotubes. *Environ Sci Technol* 44:6985–6991
- Ahmed Adam OEA, Al-Dujaili AH (2003) The removal of phenol and its derivatives from aqueous solutions by adsorption on petroleum asphaltene. *J Chem* 2013:694029
- Okolo B, Park C, Keane MA (2000) Interaction of phenol and chlorophenols with activated carbon and synthetic zeolites in aqueous media. *J Colloid Interface Sci* 226:308–317
- Suffet IHM, Khiari D, Bruchet A (1999) The drinking water taste and odor wheel for the millennium: beyond geosmin and 2-methylisoborneol. *Water Sci Technol* 40:1–13
- Ahmaruzzaman M (2008) Adsorption of phenolic compounds on low-cost adsorbents: a review. *Adv Colloid Interface Sci* 143:48–67
- Qu X, Alvarez PJJ, Li Q (2013) Applications of nanotechnology in water and wastewater treatment. *Water Res* 47(12):3931–3946
- Aherne G, English J, Marks V (1985) The role of immunoassay in the analysis of microcontaminants in water samples. *Ecotoxicol Environ Saf* 9:79–83
- Richardson M, Bowron J (1985) The fate of pharmaceutical chemicals in the aquatic environment. *J Pharm Pharmacol* 37:1–12
- Rivas J, Encinas A, Beltran F, Grahan N (2011) Application of advanced oxidation processes to doxycycline and Norfloxacin removal from water. *J Environ Sci Health A Tox Hazard Subst Environ Eng* A 46:944–951

16. Kolpin DW, Furlong ET, Meyer MT, Thurman EM, Zaugg SD, Barber LB, Buxton HT (2002) Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999–2000: a national reconnaissance. *Environ Sci Technol* 36:1202–1211
17. Ji L, Chen W, Duan L, Zhu D (2009) Mechanisms for strong adsorption of tetracycline to carbon nanotubes: a comparative study using activated carbon and graphite as adsorbents. *Environ Sci Technol* 43:2322–2327
18. Rand-Weaver M, Margiotta-Casaluci L, Patel A, Panter GH, Owen SF, Sumpster JP (2013) The read-across hypothesis and environmental risk assessment of pharmaceuticals. *Environ Sci Technol* 47:11384–11395
19. Michael I, Rizzo L, McArdell CS, Manaia CM, Merlin C, Schwartz T, Dagot C, Fatta-Kassinos D (2013) Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: a review. *Water Res* 47:957–995
20. Ebele AJ, Abdallah MAE, Harrad S (2017) Pharmaceuticals and personal care products (PPCPs) in the freshwater aquatic environment. *Emerg Contam* 3:1–16
21. Hassaan MA, El Nemr A (2017) Health and environmental impacts of dyes: mini review. *Am J Environ Eng* 1:64–67
22. Chequer FD, de Oliveira GAR, Ferraz EA, Cardoso JC, Zanoni MB, de Oliveira DP. Textile dyes: dyeing process and environmental impact. <https://doi.org/10.5772/53659>
23. Kant R (2012) Textile dyeing industry an environmental hazard. *Nat Sci* 4:22–26. <https://doi.org/10.4236/ns.2012.41004>
24. Wang S, Boyjoo Y, Choueib A, Zhu ZH (2005) Removal of dyes from aqueous solution using fly ash and red mud. *Water Res* 39:129–138
25. Pimentel D (1995) Amounts of pesticides reaching target pests: environmental impacts and ethics. *J Agric Environ Ethics* 8:17–29
26. Tariq MI, Afzal S, Hussain I, Sultana N (2007) Pesticides exposure in Pakistan: a review. *Environ Int* 33:1107–1122
27. Carter AD (2000) Herbicide movement in soils: principles, pathways and processes. *Weed Res* 40:113–122
28. González N, Marquès M, Nadal M, Domingo JL (2019) Occurrence of environmental pollutants in foodstuffs: a review of organic vs. conventional food. *Food Chem Toxicol* 125:370–375
29. European Food Safety Authority (2016) The 2014 European Union report on pesticide residues in food. *EFSA J* 14:4611. <https://doi.org/10.2903/j.efsa.2016.4611>
30. Gomiero T (2018) Food quality assessment in organic vs. conventional agricultural produce: findings and issues. *Appl Soil Ecol* 123:714–728
31. Domingo JL, Nadal M (2015) Human dietary exposure to polycyclic aromatic hydrocarbons: a review of the scientific literature. *Food Chem Toxicol* 86:144–153
32. Luehrs DC, Hickey JP, Nilsen PE, Godbole KA, Rogers TN (1996) Linear solvation energy relationship of the limiting partition coefficient of organic solutes between water and activated carbon. *Environ Sci Technol* 30:143–152
33. Chen W, Duan L, Zhu D (2007) Adsorption of polar and nonpolar organic chemicals to carbon nanotubes. *Environ Sci Technol* 41:8295–8300
34. Chen CL, Hu J, Shao DD, Li JX, Wang XK (2009) Adsorption behavior of multiwall carbon nanotube/iron oxide magnetic composites for Ni(II) and Sr(II). *J Hazard Mater* 164:923–928
35. Iijima S (1991) Helical microtubules of graphitic carbon. *Nature* 354:56–58
36. Chen CH, Huang CC (2009) Hydrogen adsorption in defective carbon nanotubes. *Sep Purif Technol* 65:305–310
37. Gaur A, Shim M (2008) Substrate-enhanced O₂ adsorption and complexity in the Raman G-band spectra of individual metallic carbon nanotubes. *Phys Rev B* 78:125422
38. Masenelli-Varlot K, McRae E, Dupont-Pavlovsky N (2002) Comparative adsorption of simple molecules on carbon nanotubes dependence of the adsorption properties on the nanotube morphology. *Appl Surf Sci* 196:209–215
39. Li YH, Wang SG, Wei JQ, Zhang XF, Xu CL, Luan ZK, Wu DH, Wei BQ (2002) Lead adsorption on carbon nanotubes. *Chem Phys Lett* 357:263–266
40. Li YH, Ding J, Luan ZK, Di ZC, Zhu YF, Xu CL, Wu DH, Wei BQ (2003) Competitive adsorption of Pb²⁺, Cu²⁺ and Cd²⁺ ions from aqueous solutions by multiwalled carbon nanotubes. *Carbon* 41:2787–2792
41. Chen CL, Wang XK (2006) Adsorption of Ni (II) from aqueous solution using oxidized multiwall carbon nanotubes. *Ind Eng Chem Res* 45:9144–9149
42. Chen CL, Wang XK, Nagatsu M (2009) Europium adsorption on multiwall carbon

- nanotube/iron oxide magnetic composite in the presence of polyacrylic acid. *Environ Sci Technol* 43:2362–2367
43. Chen CL, Hu J, Xu D, Tan XL, Meng YD, Wang XK (2008) Surface complexation modeling of Sr(II) and Eu(III) adsorption onto oxidized multiwall carbon nanotubes. *J Colloid Interface Sci* 323:33–41
 44. Goering J, Kadossow E, Burghaus U (2008) Adsorption kinetics of alcohols on singlewall carbon nanotubes: an ultrahigh vacuum surface chemistry study. *J Phys Chem C* 112:10114–10124
 45. Hyung H, Kim JH (2008) Natural organic matter (NOM) adsorption to multi-walled carbon nanotubes: effect of NOM characteristics and water quality parameters. *Environ Sci Technol* 42:4416–4421
 46. Cassani S, Gramatica P (2015) Identification of potential PBT behavior of personal care products by structural approaches. *Sustain Chem Pharm* 1:19–27
 47. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, San Diego
 48. Roy K, Kar S, Das RN (2015) A primer on QSAR/QSPR modeling: fundamental concepts (SpringerBriefs in molecular science). Springer, New York
 49. Mehndiratta P, Jain A, Srivastava S, Gupta N (2013) Environmental pollution and nanotechnology. *Environ Pollut* 2:49–58
 50. Li D, Yang M, Hu J, Ren L, Zhang Y, Li K (2008) Determination and fate of oxytetracycline and related compounds in oxytetracycline production wastewater and the receiving river. *Environ Toxicol Chem* 27:80–86
 51. Shore RF, Taggart MA, Smits J, Mateo R, Richards NL, Fryday S (2014) Detection and drivers of exposure and effects of pharmaceuticals in higher vertebrates. *Philos Trans R Soc Lond B Biol Sci* 369:20130570
 52. Kummerer K (2001) Drugs in the environment: emission of drugs, diagnostic aids and disinfectants into wastewater by hospital in relation to other sources- a review. *Chemosphere* 45:957–969
 53. Edo M, Ortuno N, Persson PE, Conesa JA, Jansson S (2018) Emission of toxic pollutants from co-combustion of demolition and construction wood and household waste fuel blends. *Chemosphere* 203:506–513
 54. Rappaport SM (2011) Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol* 21:5–9
 55. ADB (1990) Environmental risk assessment: dealing with uncertainty in environmental impact assessment. ADB environment paper no. 7. Asian Development Bank, Manila
 56. NRC (National Research Council) (2007) Toxicity testing in the 21st century: a vision and a strategy. National Academies Press, Washington, D.C. Available: https://www.nap.edu/login.php?record_id=11970&page=%2Fdownload.php%3Frecord_id%3D11970. Accessed 4 July 2015
 57. Calabrese EJ, Baldwin LA (1993) Performing ecological risk assessments. Lewis Publishers, Michigan
 58. ENHEALTH (2012) Environmental health risk assessment-guidelines for assessing human health risks from environmental hazards. Environmental Health Standing Committee, Canberra
 59. Environmental Risk Assessment. ISBN 978-0-12-811989-1. <https://doi.org/10.1016/B978-0-12-811989-1.00008-7>
 60. Chi KT, Hsu CW, Li JY (2015) Developing a green supplier selection model by using the DANP with VIKOR. *Sustainability* 7:1661–1689
 61. Zhu Q, Joseph S, Lai KH (2008) Confirmation of a measurement model for green supply chain management practices implementation. *Int J Prod Econ* 111:261–273
 62. Zhang Y, Bai Y, Yan B (2010) Functionalized carbon nanotubes for potential medicinal applications. *Drug Discov Today* 15:428–435
 63. Ray HB, Zakhidov AA, DeHeer AW (2002) Carbon nanotubes-the route toward applications. *Science* 297:787–792
 64. Ali J, Kong J (2009) Carbon nanotube electronics. Springer Science & Business Media, NY
 65. Nakashima N (2005) Soluble carbon nanotubes: fundamental and applications. *Int J Nanosci* 4:119–137
 66. Britz DA, Khlobystov AN (2006) Noncovalent interactions of molecules with single walled carbon nanotubes. *Chem Soc Rev* 35:637–659
 67. Girifalco LA, Hodak M, Lee RS (2000) Carbon nanotubes, buckyballs, ropes, and a universal graphitic potential. *Phys Rev B Condens Matter Mater Phys* 62:13104
 68. Roy J, Ghosh S, Ojha PK, Roy K (2019) Predictive quantitative structure-property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). *Environ Sci Nano* 6:224–247
 69. Chen X, Xia XH, Wang XL, Qiao JP, Chen HT (2011) A comparative study on sorption

- of perfluorooctanesulfonate (PFOS) by chars, ash and carbon nanotubes. *Chemosphere* 83:1313–1319
70. Apul OG, Karanfil T (2015) Adsorption of synthetic organic contaminants by carbon nanotubes: a critical review. *Water Res* 68:34–55
 71. Pan B, Lin DH, Mashayekhi H, Xing BS (2008) Adsorption and hysteresis of bisphenol A and 17 alpha-ethinyl estradiol on carbon nanomaterials. *Environ Sci Technol* 42:5480–5485
 72. Madani SY, Naderi N, Dissanayake O, Tan A, Seifalian AM (2011) A new era of cancer treatment: carbon nanotubes as drug delivery tools. *Int J Nanomedicine* 6:2963–2979
 73. Jorio A, Dresselhaus G, Dresselhaus MS (2008) Carbon nanotubes: advanced topics in the synthesis, structure, properties and applications. Springer, Berlin
 74. Ji Y, Lin YJ, Wong JSC (2006) Bucky paper's fabrication and application to passive vibration control. In: Proceedings of 1st IEEE international conference on nano/micro engineered and molecular systems (NEMS '06), Zhuhai, China, pp 725–729
 75. Jornet JM, Akyildiz IF (2010) Graphene-based nano-antennas for electromagnetic nano communications in the terahertz band. In: Proceedings of IEEE 4th European conference on antennas and propagations (EuCAP 2010), Barcelona, Spain, pp 1–5
 76. Laplaze D, Bernier P, Journet C, Vié V, Flamant G, Lebrun M (1997) Carbon sublimation using a solar furnace. *Synth Met* 86:2295–2296
 77. Ong YT, Ahmad AL, Zein SHS, Tan SH (2010) A review on carbon nanotubes in an environmental protection and green engineering perspective. *Braz J Chem Eng* 27:227–242
 78. Gao G, Vecitis CD (2011) Electrochemical carbon nanotube filter oxidative performance as a function of surface chemistry. *Environ Sci Technol* 45:9726–9734
 79. Rahaman MS, Vecitis CD, Elimelech M (2012) Electrochemical carbon-nanotube filter performance toward virus removal and inactivation in the presence of natural organic matter. *Environ Sci Technol* 46:1556–1564
 80. Holt JK, Park HG, Wang Y, Stadermann M, Artyukhin AB, Grigoropoulos CP, Noy A, Bakajin O (2006) Fast mass transport through sub-2-nanometer carbon nanotubes. *Science* 312:1034–1037
 81. Wong SS, Joselevich E, Woolley AT, Cheung CL, Lieber CM (1998) Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology. *Nature* 394:52
 82. Collins PG, Bradley K, Ishigami M, Zettl DA (2000) Extreme oxygen sensitivity of electronic properties of carbon nanotubes. *Science* 287:1801–1804
 83. Krishnan A, Dujardin E, Ebbesen TW, Yianilos PN, Treacy MMJ (1998) Young's modulus of single-walled nanotubes. *Phys Rev B* 58:14013
 84. Yu MF, Files BS, Arepalli S, Ruoff RS (2000) Tensile loading of ropes of single wall carbon nanotubes and their mechanical properties. *Phys Rev Lett* 84:5552
 85. Kang I, Heung YY, Kim JH, Lee JW, Gollapudi R, Subramanian S, Narasimhadvara S, Hurd D, Kirikera GR, Shanov V, Schulz MJ (2006) Introduction to carbon nanotube and nanofiber smart materials. *Compos Part B Eng* 37:382–394
 86. Long RQ, Yang RT (2001) Carbon nanotubes as superior sorbent for dioxin removal. *J Am Chem Soc* 123:2058–2059
 87. Pan B, Xing B (2008) Adsorption mechanisms of organic chemicals on carbon nanotubes. *Environ Sci Technol* 42:9005–9013
 88. Agnihotri S, Mota JP, Rostam-Abadi M, Rood MJ (2005) Structural characterization of single walled carbon nanotube bundles by experiment and molecular simulation. *Langmuir* 21:896–904
 89. Kang S, Herzberg M, Rodrigues DF, Elimelech M (2008) Antibacterial effects of carbon nanotubes: size does matter. *Langmuir* 24:6409–6413
 90. Yan XM, Shi BY, Lu JJ, Feng CH, Wang DS, Tang HX (2008) Adsorption and desorption of atrazine on carbon nanotubes. *J Colloid Interface Sci* 321:30–38
 91. Das R (2017) Nanohybrid catalyst based on carbon nanotube. In: Carbon nanostructures. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-58151-4_2
 92. European Commission, Directive 2006/121/EC of the European Parliament and of the Council of 18 December 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No. 1907/2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) and establishing a European Chemicals Agency.

- Off J Eur Union, L 396/850 of 30.12.2006, Office for Official Publications of the European Communities (OPOCE), Luxembourg
93. Wang Y, Chen J, Tang W, Xia D, Liang Y, Li X (2019) Modeling adsorption of organic pollutants onto single-walled carbon nanotubes with theoretical molecular descriptors using MLR and SVM algorithms. *Chemosphere* 214:79–84
 94. Ghosh S, Ojha PK, Roy K (2019) Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. *Chemosphere* 228:545–555
 95. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J Chemometr* 32:e2992
 96. Lata S (2018) Concentration dependent adsorption of aromatic organic compounds by SWCNTs: Quantum-mechanical descriptors for nano-toxicological studies of biomolecules and agrochemicals. *J Mol Graph Model* 85:232–241
 97. Fuguet E, Ràfols C, Bosch E, Abraham MH, Rosés M (2002) Solute solvent interactions in micellar electrokinetic chromatography. *J Chromatogr A* 942:237–248
 98. Platts JA, Abraham MH, Zhao YH, Hersey A, Ijaz L, Butina D (2001) Correlation and prediction of a large blood-brain distribution data set—an LFER study. *Eur J Med Chem* 36:719–730
 99. Ding H, Chen C, Zhang X (2016) Linear solvation energy relationship for the adsorption of synthetic organic compounds on single-walled carbon nanotubes in water. *SAR QSAR Environ Res* 27:31–45
 100. Ersan G, Apul OG, Karanfil T (2016) Linear solvation energy relationship (LSER) for adsorption of organic compounds by carbon nanotubes. *Water Res* 98:28–38
 101. Yu X, Sun W, Ni J (2015) LSER model for organic compounds adsorption by single walled carbon nanotubes: comparison with multi-walled carbon nanotubes and activated carbon. *Environ Pollut* 206:652e660
 102. Liu Y, Zhang J, Chen X, Zheng J, Wang G, Liang G (2014) Insights into the adsorption of simple benzene derivatives on carbon nanotubes. *RSC Adv* 4:58036–58046
 103. Lata S (2019) Exploring the role of quantum-mechanical descriptors in the concentration-dependent adsorption of aromatic organic compounds by multiwalled carbon nanotubes. *Int J Quantum Chem* 119:e25825
 104. Ahmadi S, Akbari A (2018) Prediction of the adsorption coefficients of some aromatic compounds on multi-wall carbon nanotubes by the Monte Carlo method. *SAR QSAR Environ Res* 29:895–909
 105. <http://www.insilico.eu/coral>
 106. Paszkiewicz M, Sikorska C, Leszczyńska D, Stepnowski P (2018) Helical multi-walled carbon nanotubes as an efficient material for the dispersive solid-phase extraction of low and high molecular weight polycyclic aromatic hydrocarbons from water samples: theoretical study. *Water Air Soil Pollut* 229:253
 107. Wang Y, Yan F, Jia Q, Wang Q (2017) Assessment for multi-endpoint values of carbon nanotubes: quantitative nanostructure-property relationship modeling with norm indexes. *J Mol Liq* 248:399–405
 108. Heidari A, Fatemi MH (2017) A theoretical approach to model and predict the adsorption coefficients of some small aromatic molecules on carbon nanotube. *JCCS* 64:289–295
 109. Wang Y, Xing F, Zhang H, Lou K (2016) Experimental and theoretical investigation on the interaction of carboxylic multi-walled carbon nanotubes with bisphenol AF. *Colloids Surf A Physicochem Eng Asp* 497:45–52
 110. Toropova AP, Toropov AA (2016) Assessment of nano-QSPR models of organic contaminant absorption by carbon nanotubes for ecological impact studies. *Mater Dis* 4:22–28
 111. Apul OG, Wang Q, Shao T, Rieck JR, Karanfil T (2013) Predictive model development for adsorption of aromatic contaminants by multi-walled carbon nanotubes. *Environ Sci Technol* 47:2295–2303
 112. Rahimi-Nasrabadi M, Akhoondi R, Pourmortazavi SM, Ahmadi F (2015) Predicting adsorption of aromatic compounds by carbon nanotubes based on quantitative structure property relationship principles. *J Mol Struct* 1099:510–515
 113. <http://www.talete.mi.it/products/dragondescription.htm>
 114. Hassanzadeh Z, Kompany-Zareh M, Ghavami R, Gholami S, Malek-Khatabi A (2015) Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. *J Mol Struct* 1098:191–198
 115. Ersan G, Apul OG, Karanfil T (2019) Predictive models for adsorption of organic compounds by Graphenenanosheets: comparison with carbon nanotubes. *Sci Total Environ* 654:28–34

116. Wang QL, Apul OG, Xuan P, Luo F, Karanfil T (2013) Development of a 3D QSPR model for adsorption of aromatic compounds by carbon nanotubes: comparison of multiple linear regression, artificial neural network and support vector machine. *RSC Adv* 3:23924–23934
117. Salahinejad M, Zolfonoun E (2018) An exploratory study using QICAR models for prediction of adsorption capacity of multi-walled carbon nanotubes for heavy metal ions. *SAR QSAR Environ Res* 29:997–1009
118. Ali I, Alharbi OM, ALOthman ZA, Al-Mohaimeed AM, Alwarthan A (2019) Modeling of fenuron pesticide adsorption on CNTs for mechanistic insight and removal in water. *Environ Res* 170:389–397



Chapter 21

Ecotoxicological QSAR Modeling of Organophosphorus and Neonicotinoid Pesticides

Alina Bora, Luminita Crisan, Ana Borota, Simona Funar-Timofei, and Gheorghe Ilia

Abstract

Organophosphorus and neonicotinoid pesticides are important agrochemicals used worldwide. The beginning of the quantitative structure-activity/toxicity relationship (QSAR/QSTR) field, after the 1960s, is related to the study of the organophosphorus pesticide activity. QSARs have been recognized as an important research direction in the field of medicinal, analytical chemistry, toxicology, pharmaceutical, and environmental chemistry. The main aim of QSAR/QSTR models is to find reliable relationships between the biological activity/toxicity and the experimental or theoretical compound molecular descriptors, to design new structures with improved target properties and safety profile. In this chapter, successful QSAR models are presented for the ecotoxicological data of organophosphorus and neonicotinoid pesticides. In particular, QSAR models for organophosphorus aquatic and terrestrial organism ecotoxicity; for the neonicotinoid toxicity against the honeybees, *Musca domestica L.*, *American cockroach*, and aphids (*Aphis craccivora* and *Myzus persicae*); and for the inhibition ability of acetylcholinesterase and other enzymes by organophosphorus pesticides are presented. The literature data indicate a large variety of QSAR approaches employed in these published studies. In case of organophosphorus pesticides, many available ecotoxicity data for human beings and animals were employed in the computational studies. For the neonicotinoid pesticides a limited number of QSAR models were reported, especially due to the lack of the degradability and aquatic organism toxicity data. The ligand-based combined with structure-based approaches remain a powerful tool in the design of new environment-friendly and less toxic organophosphorus and neonicotinoid pesticides.

Key words QSAR, QSTR, Toxicity, Organophosphorus, Neonicotinoid, Environment, Acetylcholinesterase, Agrochemical, Ecology, Bioconcentration factor

1 Introduction

Computational chemistry is a very rapid and inexpensive choice prior to experimental tests used in order to avoid synthesizing excessive chemical compounds. The research of computational

Authors “Alina Bora, Luminita Crisan and Ana Borota” are contributed equally to this work.

chemistry through QSAR (quantitative structure-activity relationship) can rationally predict the biological activity/toxicity of untested analogue compounds, reducing significantly the experimental laboratory costs and very important being friendly and safe for the environment. An in silico QSAR evaluation can allow understanding of a potential relationship between the chemical structure and its activity/toxicity. QSAR techniques have evolved since the 1960s, after the published paper of Corwin Hansch, the founder of modern QSAR, presenting the correlation between biological activity and chemical structure [1].

The general QSAR equation has the formula:

$$\begin{aligned} \text{Biological activity} = & \text{function (parameters/molecular descriptors)} \\ & + \text{error} \end{aligned} \quad (1)$$

The parameters used in the QSAR equation are lipophilic, electronic, steric, polarizability, and various other calculated descriptors derived from the chemical structure of compounds.

In agrochemistry, various QSAR techniques were successfully applied in the pesticide development, since the 1960s [2]. In different QSAR studies, organophosphorus (OP) pesticides have received special attention. In agricultural chemistry, the different QSAR techniques are generally based on toxicity prediction. This way, the risk of the OP compounds to human health and the environment may be assessed. OP pesticides can be soaked up by skin absorption, ingestion, and/or inhalation and can cause dysfunctions of the nervous, renal, immune, endocrine, reproductive, cardiovascular and respiratory systems. Most of the observed toxicological effects of the organophosphorus pesticides are related to the inhibition of acetylcholinesterase (AChE) [3].

In modern society, chemical pesticides are extensively used for controlling pests. Annually, approximately six million tons of pesticides are used in agriculture, but a very small percentage attain the target, whereas the large percentage expands to nontarget species, inducing toxicological effect concerning environmental and health [4–6]. As a result of using pesticides, the population is vulnerable to them; small quantities are detected in fruits, vegetables, fish, cereals, tea, honey, milk, etc. The evaluation of toxicological and ecotoxicological of pesticides risks was traditionally accomplished by laboratory experiments. These experiments on animals are expensive and raise a major ethical problem. Nowadays, the scientific community and legislation authorities propose alternative methods for experimental techniques. Computational chemistry tries to comprise all areas from *chemoinformatics* to *molecular modeling*, and it seems to be used on a larger scale as a predictive tool to guide experimentalists in the synthesis of new compounds and in the investigation of complex physicochemical processes.

2 Organophosphorus Pesticides: An Overview

Pesticides include large-scale organic and inorganic chemicals used against weeds, insects, fungi, rodents, etc. Organophosphorus (OP) compounds are a class of pesticides subjected to integrated risk assessment, which share exposure characteristics for different species [7], and are among the most commonly used chemicals in agriculture around the world. The OP pesticides have relatively low persistency and high efficiency, being broadly used in the world (around 140 OPs are or were used as practical pesticides). Several biological effects have been attributed to the organophosphorus compounds [8]. Most OP pesticides were used as insecticides, but they were employed as plant growth regulators, acaricides, anthelmintics, nematocides, chemosterilants, and rodenticides as well. Small changes in the chemical structures of OP agrochemicals modified considerably the toxicity from species to species. Therefore, compounds having similar chemical structures were frequently used for different tasks. An important development in the agricultural practice and scientific knowledge on the structure-activity relationship of organophosphorus insecticides were reached by the invention of parathion, by Schrader in 1944. Despite its highly toxic effects to mammals and insects, many less toxic insecticides have been developed by small structural modifications, like chlorthion, fenthion, and fenitrothion.

OPs inhibit progressively the AChE enzyme through phosphorylation of the active site serine, by covalently binding it to the hydroxyl group of serine, then the compound is split and the AChE is phosphorylated. The general chemical structure of the class of OP pesticides is shown in Fig. 1 [9–11], where the R1 and R2 substituents are the side groups of OPs, and usually they can be hydrogen atoms, alkyl, aryl, etc. The R3 substituent represents the leaving group (e.g., cyano, halogens, alkyl, alkylthio, or aryl groups). This leaving group is replaced through nucleophilic substitution, by the oxygen atom of serine in the AChE active site. The active form of OP compounds has the oxygen atom linked to the phosphorus atom (P=O). The OP compounds, which represents the majority of novel OP pesticides, have the sulfur atom linked to the phosphorus atom (P=S). In this case, the thiono group must be metabolized to an oxono group and then can be bound to the AChE active site [10, 12]. The process of dephosphorylation is very slow [13], such that the neurotransmission of the acetylcholine receptors is hindered resulting in the symptoms of poisoning. Sometimes, the resulted organophosphorus group could be involved in a dealkylation reaction, which causes the occurrence of the non-reactivable AChE form (so-called aged enzyme) [14].

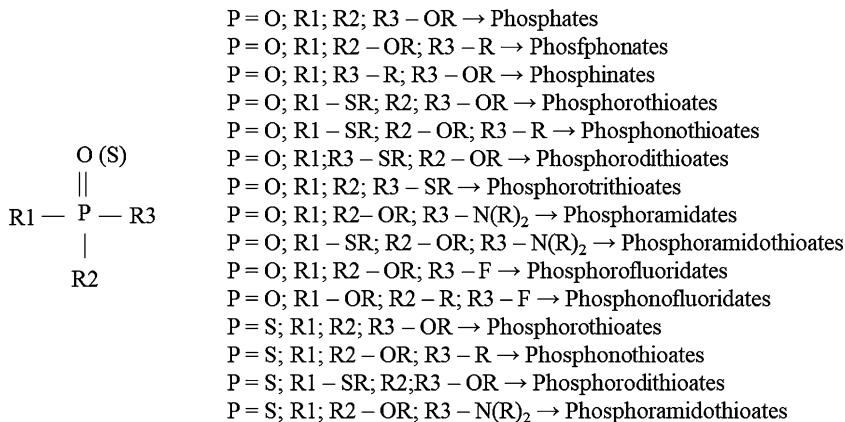


Fig. 1 General chemical structure of OP pesticides

The irreversible AChE inhibition with OP pesticides leads to acute toxicity [15]. The acute toxicity is, according to the IUPAC Gold Book [16], the “ability of a substance to cause adverse effects within a short time of dosing or exposure.” The lethal dose (LD₅₀) that causes the death of 50% of the test group is generally used to measure of the acute toxicity of chemicals.

3 The Transition from Organophosphates to Neonicotinoids

To sum up the pre-neonicotinoids era, the pesticides have been in a continuous evolution starting from the inorganic classes to the organochlorines, which eventually evolved into organophosphates, carbamates, and synthetic pyrethroids. Each class of pesticides has positively contributed to the increase of insecticidal properties and selectivity and reduced their persistence in ecosystems. Despite the positive contributions, some drawbacks have, also, been found regarding the nontarget species toxicity, the human health risks, and the evolved resistance of pests to the harmful substances. In order to overcome these disadvantages, a new class of pesticide-like compounds with improved insecticide profile was required. These new pesticides should have a different mode of action than the previous generations of pesticides, high insecticidal potential, low toxicity for nontarget species, and no harmful effect on the environment. The synthetic neonicotinoid class was designed to meet these requirements [17].

4 Neonicotinoid Pesticides: An Overview

Neonicotinoids or neonic are the most widely used insecticides introduced to the global market with a major impact in the

economy and the ecosystem of any country [18–20]. They are registered in more than 120 countries accounting over 26% of the global insecticide market [21, 22]. Since their first appearance in the mid-1990s, they have become the most successful class of chemical insecticides, having large-scale applications ranging from plant protection (e.g., crops, vegetables, fruits), consumer/professional/veterinary products, animal health, and biocides to invertebrate pests (e.g., insects such as aphids, whiteflies, nematodes, parasites) control [22, 23]. Due to their systemic nature, neonicotinoids are rapidly absorbed through the leaves or roots of the developing plant and subsequently transported through its tissues. This property confers them many advantages in pest control. Plant protection and controlling the action of pests are very important clues in order to improve the quality and quantity of the products [24].

Neonicotinoids, which are structurally similar to nicotine, act as nicotinic acetylcholine receptor (nAChR) agonists in the central brain of insects, causing excitation at low concentration, followed by paralysis and death at higher concentrations of compound [25, 26].

A number of research papers discussed the high preference of neonics for binding to insect nAChRs rather than those of vertebrates, so they are classified as being more toxic to insects [27, 28].

The neonicotinoid family contains three generations of neonicotinoids, such as (1) imidacloprid, acetamiprid, thiacloprid, and nitenpyram; (2) thiamethoxam and clothianidin; and (3) dinotefuran, which are marketed under different trade names (Fig. 2). Imidacloprid (IMI), the first commercial neonicotinoid, is the world's largest selling insecticide with certified uses for more than 140 crops. In the early twenty-first century, the sulfoxaflor, a fourth-generation neonicotinoid, has been licensed for use or under consideration for licensing in several worldwide countries including China, the USA, Canada, Mexico, Argentina, and Australia [29]. As one of the largest pesticide markets, China plays a significant role in the development and commercialization of the new neonicotinoids. In this light, throughout the years of 2005 to 2013, new and promising generation of neonicotinoid-like insecticides, namely, guadipyr, huanyanglyn, paichongding, cycloaprid, and imidaclothiz, were synthesized and tested by the researcher groups in China [30, 31]. The development of the second, third, and fourth generation of the neonicotinoids illustrates the extent to which structural modifications at the well-known IMI core impact the neonicotinoid compound activity, chemical properties, and its potential uses [32, 33]. All commercial neonicotinoid compounds possess in their chemical structures an aromatic heterocycle (e.g., pyridine), a flexible linkage, a hydroheterocycle or guanidine/amidine fragment, and an electron-

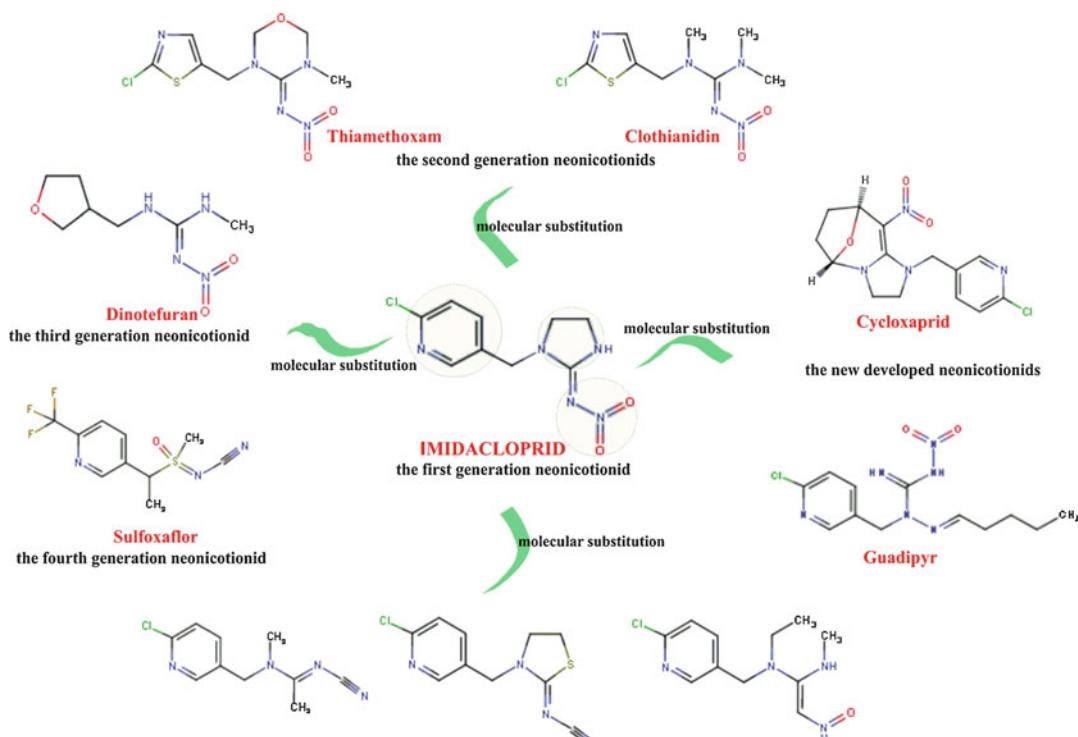


Fig. 2 Commercial and new designed neonicotinoid insecticides

withdrawing unit such as nitro ($-NO_2$) or cyano ($-CN$). These features contribute directly to their potency and selectivity.

5 QSAR Models for the Ecotoxicology of Organophosphorus and Neonicotinoid Pesticides

5.1 QSAR Models for Organophosphorus Aquatic and Terrestrial Organisms Ecotoxicity

In the modern society, chemical pesticides are extensively used for controlling pests. Annually, approximatively, six million tons of pesticides are used in agriculture, but a very small percentage attain the target, whereas the large percentage expands to nontarget species, inducing toxicological effect concerning environmental and health [4, 5]. As a result of using pesticides, the population is vulnerable to them, small quantities are detected in fruits, vegetables, fish, cereals, tea, honey, milk, etc. The evaluation of toxicological and ecotoxicological risks of pesticides was traditionally accomplished by laboratory experiments. These experiments on animals are expensive and raise a major ethical problem. Nowadays, the scientific community and legislation authorities propose alternative methods for experimental techniques. Computational chemistry tries to comprise all areas from *chemoinformatics* to *molecular modeling*, and it seems to be used on a larger scale as a predictive

tool to guide experimentalists in the synthesis of new compounds and in the investigation of complex physicochemical processes.

MLR and PLS methods were employed by Verhaar et al. [34], in order to model the bioconcentration factor ($\log BC_F$), the uptake rate constant from water ($\log k_1$), the elimination rate constant to water ($\log k_2$) measured in guppy (*Poecilia reticulata*), and the biotransformation, expressed by the rate constants in either NADPH- or GSH-enriched rainbow trout liver homogenates for the 12 organophosphorothionates. Additionally the dissociation constant for the reversible binding of substrate to AChE ($\log K_D$); the phosphorylation constant for the irreversible phosphorylation of AChE by the (reversibly) bound substrate ($\log k_2'$); the overall inhibition of AChE by substrate, expressed as a bimolecular reaction rate constant ($\log k_i$); and the acute aquatic median lethal concentration ($\log LC_{50}$) toward the guppy were modeled too. The quantum-chemical descriptors were used to designate the reactive components of the aquatic toxicity of organophosphorothionates and were calculated using the AM1 semiempirical Hamiltonian inside the MOPAC program. The results hint that those compounds with an intermediate hydrophobicity, which usually is quantified by the logarithm of the respective octanol/water partition coefficient ($\log K_{ow}$), the charge density difference between the central phosphorus atom, the oxygen leaving group, and a large absolute hardness or the nucleophilic delocalizability on the central phosphorus atom of the organophosphorothionates, should display enhanced toxicity. By investigating the scores and loading plots of the latent variables, significant information can be extracted from the PLS model, which is a better way to choose a set of relevant descriptors than traditional MLR.

11 O,O-Dimethyl O-phenyl phosphorothionates with substituents at the 2-, 4-, and 5-positions of the phenyl ring, having fish toxicity, expressed as the 14-d acute fish toxicity to the guppy ($\log LC_{50}$) and the alkylation rate constant toward 4-nitrobenzylpyridine ($\log k_{NBP}$), were involved into linear two-parameter quantitative structure-activity relationship (QSAR) models [35]. Quantum-chemical descriptors such as the electronegativity (EN), the energies of the highest occupied molecular orbital (E_{HOMO}), the energies of the lowest unoccupied molecular orbital (E_{LUMO}), the absolute hardness (η), the nucleophilic delocalizability ($D^N(r)$), and the electrophilic delocalizability ($D^E(r)$) were calculated. Additionally, the calculated pK_a values [36] of the protonated leaving groups, the hydrophobicity, which was calculated as $\log K_{ow}$, according to Leo's scheme [37] and the approximate effective diameters (D_{eff}) [38] were used to develop a QSAR model. The $\log K_{ow}$, D_{eff} , and pK_a parameters were selected in order to model the partitioning and reactivity, respectively, of the selected phosphorothionates.

The genetic partial least squares (G/PLS) technique was applied by Drew et al. [39] in order to predict ecotoxicity data of 20 organophosphorus insecticides. The activity data for these OP compounds were determined in the form of 14-day LC₅₀ toxicity values against Guppy (*Poecilia reticulata*). The molecular descriptors were calculated using the CERIUS2 [40] and TSAR [41] software tools, and the electronic parameters were calculated using the Gaussian 94 [42] program, which resulted from ab initio quantum mechanics calculations using the 6-311G* basis set. Drew et al. [39] highlighted that their resulted equation, which was derived from G/PLS analysis, is trustful for predicting the activity of organophosphates having similar structure and unknown activity.

Yan et al. [43] used the multivariate linear regression analysis to predict the lethal toxicity (LC₅₀ values) to fish (*Cyprinus carpio*), for 43 OP compounds. These OP compounds are divided into six subclasses: phosphate, phenylphosphonothioate, phosphorodithioate, phosphorothioate, phosphorothiolothionate, and phosphordiamide. The experimental LC₅₀ (48 h) toxicity data of the 43 OP compounds, taken from the paper of Li [44], were utilized as the dependent variables, and 1381 molecular descriptors were used as Dragon independent variable in a stepwise variable selection technique. The statistical parameters for the prediction of lethal toxicity values were (1) for OP compounds having low toxicity, $R = 0.942$, $R^2_{adj} = 0.862$, and SEE = 2.899, and (2) for the OP compounds having high toxicity, $R = 0.977$, $R^2_{adj} = 0.937$, and SEE = 0.143. The results suggested that hydrophobicity, steric, and electronic features play an important role in predicting the toxicity of OP compounds to fish.

Classification techniques (linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), k -nearest neighbors classification (KNN), nearest mean classifier (NMC), soft independent modeling of class analogy (SIMCA), and classification and regression tree (CART)) were performed by Mazzatorta P. et al. [45] for toxicity prediction. Each model employed individual datasets, different descriptors and algorithms, and specific toxicological endpoints. For a set of 235 agrochemical compounds, the toxicity was expressed as LC₅₀ values, which corresponds to the water concentration which kills 50% of the aquatic animals (trout and daphnia) and the LD₅₀ values, which represent the lethal dose for 50% of the test animals, for rat and birds. Fifty-seven OP compounds having toxicity endpoints for trout, 59 OP compounds having toxicological endpoints for rat, 49 OP compounds having toxicological endpoints for daphnia, 37 OP compounds having toxicological endpoints for quail, and 28 OP compounds having toxicological endpoints for duck were studied. It was noticed that among the most widely used descriptors for the toxicity prediction, the hydrophobicity and the

HOMO and LUMO energy descriptors were repeatedly selected. Mainly, the descriptors that were related to the hydrophobicity are important, because they indicate how the molecule can pass through the cell membrane.

A set of 38 OP compounds was used by Guo et al. to develop a QSAR model for the prediction of the LD₅₀ values, on male rats, exposed orally [46]. The authors took into account the precursor metabolic effects and the primary AChE inhibition, suggesting the ADME effects, ligand binding, and the phosphorylation process. To correctly predict the acute toxicity of the OP compounds, the ADME and thermochemical analysis have been combined with the CoMFA approach. The statistical results have considerably increased from 0.49 to 0.9 (R^2 values), when the CoMFA term (representing the calculated affinity score from the molecular interaction field) was included, besides the energy of the parent OP compound, the energy of the metabolite compound, the negative accessible surface area of the metabolite compound, the energy of the Ser203-OP compound, the energy of the aged Ser203-OP, the energy of the aging leaving group, and the energy of the phosphorylation leaving group. Using only the CoMFA terms, the correlation between the predicted AChE-binding affinity and the experimental LD₅₀ gave the 0.892 value for R^2 and only 0.571 value for Q^2 . The slightly predictivity of the COMFA model has been improved to the 0.82 value for Q^2 , when the ADME effects and subsequent covalent binding processes have been taken into account.

Successful regression equations were obtained by García-Domenech et al. [47], between topological descriptors and the acute intraperitoneal (LD₅₀-i.p.) and oral (LD₅₀-oral) toxicities to rat of a group of 39 organophosphorus pesticides. The importance of Subgraph Randic'-Kier-Hall-like indices and topological charge indices, for the toxicity prediction, was highlighted.

Can [48] proposed a QSTR model using the multiple linear regression method, which allows simulating the acute oral toxicity to rats of 27 organophosphate insecticides (20 OP compounds in the training set and 7 OP compounds in the test set). Electronic, hydrophobic, steric, polar, and geometric descriptors were used as independent variables in order to predict the LD₅₀ data to rat. The final QSTR equation, having $R^2 = 0.901$, indicates a strong correlation between the toxicity of the organophosphate insecticides and their descriptors, e.g., molecular mass, polarizability, logP, and molar refractivity. The positive contributions to the toxicity of the organophosphate insecticides are related to the molar refractivity and logP, while the molecular mass and polarizability have a negative influence. According to the Can's model, the toxicity of insecticides is getting higher if the logP values, molar refractivity are increasing, and, the toxicity of insecticides is getting lower if compounds have higher polarizability and molecular mass.

35 OP analogues, having acute toxicity data ($\log LD_{50}$ values) for the housefly (*Musca nebulo L.*), were selected by Zhao and Yu [49]. The multi-block partial least squares (MBPLS) method were applied to correlate the 24 h acute toxicities, assessed for the housefly with the molecular interaction field (MIF) descriptors. Very good statistical results: $R^2 = 0.995$ and $Q^2 = 0.865$ were obtained. The hydrogen bond and hydrophobic effects of OP compounds were selected to have a significant influence on the insecticidal toxicity against the housefly. The QSAR model presented in this paper, based on the MIF descriptors may be useful to interpret the mechanisms of ligand-receptor interactions.

2D-QSAR and 3D-QSAR calculations using the CoMFA methodology were carried out by Niraj et al. [50] for the prediction of the insecticidal activity against *Musca domestica*, for the same series of 35 OP pesticides [49]. A 2D-QSAR model was built, based on the stepwise method for variable selection, combined with the multiple regression approach. The statistical results of the 2D-QSAR equation have been exceeded by the statistical results of the 3D-QSAR equation obtained using the CoMFA approach, in terms of the correlation coefficient value. The squared correlation coefficient for the 2D-QSAR equation was of 0.7797, while for CoMFA, the squared correlation coefficient value was of 0.958. The 2D-QSAR model shows that the insecticidal activity (pLD_{50} values) of the 35 OP compounds is related to the physicochemical-type descriptor and atom-type count descriptors. The total polar surface area including phosphorus and sulfur atoms (polar surface area including P and S), the most hydrophilic value on van der Waals surface (SA average hydrophilicity), and delta alpha type A (delta alpha A) descriptors, which are included in the Vlife MD Suite, were observed to contribute positively to the activity, while the H-acceptor count descriptor was revealed to contribute negatively. The resulted CoMFA model reflects the major contribution to the insecticidal activity of the electrostatic field of about 81.77%, while the steric contribution was of 18.23%.

22 OP pesticides having toxicity data for *Daphnia magna* and 15 OP pesticides having toxicity data for honeybees (*Apis mellifera*) were the subjects of a QSTR study [51]. The statistical results obtained from a stepwise multiple regression analysis were satisfactory ($R^2 = 0.895$ and $R^2 = 0.920$, respectively), underlining the role of lipophilicity for describing the toxicity to *Daphnia*, while for honeybees the lipophilicity descriptor has been selected after volumetric and electronic characteristics.

Additionally, the QSAR modeling approach is a reliable method to estimate the degradation processes of compounds, under different environmental conditions. QSAR models help us to determine the typical behavior of pesticides in the environment, to replace the costly and time-consuming analytical procedures. The degradation of OP compounds in water can arise by a variety

of pathways. Tanji and Sullivan [52] obtained a linear model equation for estimating the rates of chemical hydrolysis of several OP pesticides in natural river waters. The rates of hydrolysis of 11 OP pesticides (phosphates, phosphorothioates, and phosphorothiolates) were evaluated with respect to first-order connectivity indices, derived from the graph theory. These descriptors are some of the most successful topological indices accessible in QSAR methodologies.

Taking into account the environmental risk of the OP pesticides, the study of their degradability is a major requirement. To this end, the estimation of the metabolism of OP agrochemicals by a chloroperoxidase enzymatic process was achieved by Lu et al. [53] with the aid of QSAR models, by using the PLS analysis. The enzymatic activity of chloroperoxidase on metabolizing the selected OPs, expressed as moles of OPs oxidized per second per mole of chloroperoxidase, was used as a dependent variable. Two QSAR models have been obtained based on chemical descriptors computed on a small number of 9 OP pesticides, one with 18 and another with 12 independent variables (including 10 quantum-chemical descriptors). The last one is the optimal model, having a correlation coefficient R of 0.918. The study concludes that chloroperoxidase may metabolize faster the OPs having high absolute values of atomic charges on the sulfur and phosphorus atoms, while OPs having higher polarity will be slowly metabolized.

5.2 QSAR Models of Neonicotinoid Terrestrial Organism Ecotoxicity

The increased success of neonics is due to their (1) physicochemical properties which offer many advantages over the traditional pesticides such as pyrethroids, chlorinated hydrocarbons, organophosphates, and carbamates, (2) high target specificity and efficiency, (3) relatively low toxicity for nontarget organism and environment, and (4) high versatility in application methods [18, 22, 29, 54, 55]. They have a favorable toxicological profile for birds, fish, and most environmental organism. Despite their indisputable advantages, the neonicotinoids along with many other factors were connected to potential adverse ecological effects on bees and other beneficial insects even with low levels of contact [19, 56–58]. These findings have led to a critical need for the development of new control strategies and new insecticide candidates with an improved toxicity profile. A successful alternative solution to overcome the adverse effects arising from exposure to hazardous agents, to carry out the chemical risk assessment steps, and to minimize the demand of animal tests is the use of computational in silico techniques such as QSAR, pharmacophore modeling, molecular docking, etc. [59–61]. To date, a large number of QSARs models which manage the acute toxicity of pesticides have been developed [59–63], but a limited number are devoted to neonicotinoids. Currently, a wide variety of neonicotinoids are continuously synthesized and tested against a broad array of insect species [63–67]. In this context, the

toxicity prediction of neonicotinoids remains a source of interest in QSAR modeling. Computational toxicity models have earned broad acceptance for assessing the chemical toxicity of pesticide or environmental safety as well as for designing new powerful and greener candidates by minimizing the time and cost of the research.

5.2.1 QSAR Models of Neonicotinoids with Insecticidal Activity Against *Apis mellifera* L.

Despite their widespread use around the world, the pesticides can affect a range of species of ecological importance, such as honeybees (*Apis mellifera* L.). In this context, particular attention should be paid to their protection, not only for their ecological importance but also for their economic value, as honey producers and pollinators. Moreover, the neonicotinoids were directly linked as potentially harmful elements for both managed and wild bees [68, 69]. Despite the importance of bees and the possible toxic effect of neonicotinoids against them, only a few studies have reported QSARs regarding this subject, most of them using a various class of pesticides and organophosphorus derivatives [51, 70–72] and only one for neonicotinoids [73].

Vighi et al. [51] developed a QSTR model for estimating the acute toxicity of organophosphorus pesticides to *Apis mellifera* (discussed in Subheading 5.1). Devillers et al. [70] proposed a neural network-based QSAR model to estimate the toxicity of pesticides to *Apis mellifera*. Toropov and Benfenati [71] predicted the toxicity of pesticides in bees by a QSAR model using the SMILES (simplified molecular-input line-entry system) descriptors. Singh et al. [72] developed global QSTR (quantitative structure-toxicity relationship) models to predict both qualitative and quantitative the toxicity of 237 structurally diverse pesticides in honeybees.

Zhao and Li [73] proposed a three-dimensional quantitative structure-activity of neonicotinoid insecticides. First, a PLS analysis was used to establish the relationship between the structures of 30 diverse neonicotinoids and their pLC₅₀ biological activities employing the leave-one-out method for cross-validation. Also, CoMFA and CoMSIA methods were used to explain this relationship. The electrostatic and steric fields were computed by the CoMFA method, while the hydrophobic, the hydrogen bond donor, and acceptor fields by CoMSIA method. The results indicated that the CoMSIA analysis explained more intuitively the structure-activity relationship of compounds than the CoMFA analysis. The CoMFA and CoMSIA contour maps highlighted that the introduction of bulky or electropositive groups at the 2-position of the chosen template (having the highest pLC₅₀ value) could enhance the pLC₅₀ values. In this context, the template structure substituted at 1-, 2-, 4-, and 12-positions with eight units (-OH, -COOH, -OCH₃, -C₂H₂, -C₂H₄, NH₂, -NO, and -Br) was used to derive new 37 neonicotinoid analogues. The pLC₅₀ and the logarithm of the bioconcentration factor (the ratio

of the concentration of a particular chemical in a living organism to the chemical's concentration in the surrounding water, expressed as logBCF) values were predicted for 37 designed analogues in order to evaluate the toxicity effects and their accumulative capacities in the environment. The outcomes suggested that all new 37 derivatives had higher toxicity than the experimental value of the template compound (increased by 0.04–9.75% and 0.23–11.45% by the CoMFA and CoMSIA models, respectively). The bioconcentration factor (expressed as logBCF values) was modeled for the 37 new compounds using the CoMSIA approach and the same template compound. These results indicated that the logBCF values of 17 new analogues were higher than that of template compound (increased by 0.87–85.29%). These enhanced bioconcentration factor values indicate an easier accumulation of these 17 compounds in various environments. By contrast, the other 20 analogues had lower bioconcentration factor values than that of the template compound (decreased by 0.38–147.88%), and therefore they would not readily accumulate in the environment. Further, these 20 analogues and the template compound were subjected to molecular docking analysis to study the bi-directional selective toxicity on pests and bees resistance (potential chronic sublethal effects on pollinating insects such as bees). As a result, 10 out of these 20 neonicotinoids showed bi-directional selective toxicity effects, and 7 showed bi-directional selective resistance-inducing effects on both bees and pests. From these ten derivatives, one compound was designated as a new insecticide with potentially toxic effects on pests and resistance-inducing effects on pests and bees. These features will be useful for the design of new environmental friendly neonicotinoid pesticides.

*5.2.2 QSAR Models of Neonicotinoids with Insecticidal Activity Against *Musca domestica* L. and American Cockroach*

A comprehensive literature survey showed that new classes of neonicotinoids were developed by structural modification of the lead compounds, imidacloprid. By chain-opening ring or by introducing or replacing different parts of the IMI scaffold with fragments which resemble the natural neurotransmitter acetylcholine, more closely than nicotine is expected to deliver highly effective neonicotinoids. In order to evaluate the fragments replacement effects, quantitative analyses of the structure-activity relationship of neonicotinoids and analogous have been performed. The outcomes of this quantitative analysis have provided useful information for further structural modification and new insecticide development. Furthermore, the structural changes of the marketed neonicotinoids can be an effective strategy to combat resistance.

In this light, Nishimura et al. [74] analyzed quantitatively the neuroblocking activity (expressed as the concentrations of each compound required for the blocking, in terms of $\log(1/BC)$) of imidacloprid analogues with various substituents at the 5-position

of the pyridine ring by employing physicochemical substituent parameters. The multiple linear regression approach related the neuroblocking activity to the calculated pesticide descriptors. The QSAR outcomes indicate that the introduction of various substituents such as halogens, alkoxy groups, and alkyls at the 5-position of the pyridine ring of imidacloprid diminished the neuroblocking activity. The same effect was explained by the use of steric and electronic parameters. Also, two nerve-binding activities of the imidacloprid analogues were measured: the conduction blockage in the excised central nerve cord of the American cockroach and the radioligand [³H]IMI binding to the housefly head membrane. The results showed that the higher the binding activity to the housefly head membrane, the higher the blocking activity in the American cockroach. It was assumed that the tested neonicotinoids initially bind to the nAChR leading to blockage of the nervous system, followed by the insect death.

To find the substantial information for the design of better neuroblocking insecticides, 3D-QSAR approaches were performed to evaluate the neuroblocking activity ($\log(1/BC)$) of 16 imidacloprid analogues substituted at the 5-position, using CoMFA and CoMSIA methodologies [74]. The statistical results of CoMFA (A5: $Q^2 = 0.707$, $R^2 = 0.986$) and CoMSIA (A3: $Q^2 = 0.715$, $R^2 = 0.961$) models for neuroblocking activity exhibited good prediction abilities. The contribution of steric, H-bond donor and H-bond acceptor fields was 68.2%, 0.8%, and 31.0%, respectively. This suggests that the steric and H-bond acceptor nature of a compound is essential for high neuroblocking activity. The CoMFA and CoMSIA contour plots analysis showed that the introduction of sizable and alkoxy substituents was unfavorable for activity. This observation is in accordance with the results of Nishimura et al. [75]. The same contour plots indicated that H-bond acceptor region located at nitrogen atom on the pyridine ring and nitro group on the imidazolidine ring contribute positively to activity.

A quantitative relationship between the neuroblocking activity using the cockroach ganglion ($\log(1/BC)$ values) and the insecticidal activity against American cockroaches ($\log(1/MLD)$ values; MLD—the minimum lethal dose) of 23 neonicotinoid variants of the key pharmacophore, constructed with the central ring conjugated to an NCN, CHNO₂, or NNO, was examined by Kagabu et al. [76]. Analyzing the outcomes of their in silico studies, the authors suggested that the neuroblocking potency is proportional to the Mulliken charge on the nitro oxygen atom or cyano nitrogen atom. Also, the variation of fragments at the pharmacophore structure of neonicotinoids allowed insecticidal activity against American cockroaches at the nanomolar level in the presence of synergists. For cyanoimino variants, the neuroblocking effect was observed at the

micromolar level. The equation for neuroblocking activity and the insecticidal activity showed that both potencies are proportionally related when other factors are the same. In a previous paper [77], the group of Kagabu analyzed the relationship of the neuroblocking potencies of variants of the central ring imidazolidine of imidacloprid-related nitroimine and nitromethylene compounds to the physicochemical parameters such as the Mulliken charge on the nitro oxygen atom and the octanol-water partition coefficient, logP. The quantitative equation indicated that the neuroblocking activity was proportional with both physicochemical factors.

Okazawa and co-workers [78] predicted the binding mode ($\log(1/K_i)$) of imidacloprid and related compounds to housefly head (*Musca domestica* L.) acetylcholine receptors using 3D-QSAR approaches. The CoMFA maps facilitate insight into the binding mode from the ligand side. These contour maps showed that the area around the fifth and sixth positions on the pyridine ring should be sterically and electrostatically permissible. The authors highlight that for successful interaction with the nAChR receptor, a specific chemical structure of the compounds and a specific conformation of the nitroimino unit is required.

Suzuki et al. [79] modeled the relationship between the insecticidal activities (expressed as pKi values) against the housefly, *Musca domestica*, of 26 N3-substituted imidacloprid analogues [80] and their binding activity toward the nAChR receptor using the multiple linear regression (MLR) approach. In this regard, 2D-structural descriptors of the IMI analogues were correlated with the pKi values to find out the key structural features that influence the binding activity. The researchers found two significant predictors in the best MLR models including the squared Ghose-Crippen octanol-water partition coefficient and the number of tautomers, which increased the binding activity. A detrimental effect on activity was induced by higher values of the lopping centric index descriptor. It can be concluded that the insecticidal activity was a function of the lipophilic character of the compounds.

The Kiriyama group [81] examined quantitatively the relationship between insecticidal ($\log(1/IC_{50})$) and the binding activities ($\log(1/EC_{50})$) of dinotefuran and 23 related analogues tested against housefly, *Musca domestica* L. The binding affinities were measured using housefly head membrane preparation and two radioligands [^3H]imidacloprid and [^{125}I] α -bungarotoxin. The binding activity measured with [^3H]imidacloprid shows a better correlation with the insecticidal activity. Multiple linear regressions between the binding activity (used as a dependent variable) and the insecticidal activity (as the independent variable), together with other parameters (e.g., hydrophobicity descriptor), indicate that the higher the binding activity, the higher is the insecticidal activity. These QSAR approaches support the key neonicotinoid

pharmacophore and also clarify the overwhelming role of pharmacokinetic factors in the activity of neonicotinoids.

Li and co-workers [82] constructed a pharmacophore model based on the 34 neonicotinoids having neuroblocking activity. The best pharmacophore model, obtained by 3D-QSAR, consists of a hydrogen bonding acceptor, a hydrogen bond donor, a hydrophobic aliphatic, and a hydrophobic aromatic center. One out of the 34 neonicotinoids showed the highest neuroblocking activity ($1 \mu\text{mol/L}$) against the American cockroach species. Based on this pharmacophore, a series of heterocyclic compounds was designed and synthesized. It can be concluded that this pharmacophore is a useful tool for the development of novel neuroblocking insecticides targeting the nAChR receptor.

Analyzing the presented QSAR studies [74–82], it could be pointed out that most of them have applied 3D-QSAR methods and CoMFA analysis on a relatively small number (16 to 34) of compounds to indicate the electrostatic potential, steric potential, and permeability coefficient as significant parameters for the design of new pesticides.

In this regard, a total of 78 imidacloprid-based derivatives tested against *Drosophila melanogaster* nAChR (Dm-nAChR) and *Musca domestica* nAChR (Md-nAChR) were analyzed using 3D-QSAR (CoMFA and CoMSIA) methods, in conjunction with the homology modeling, molecular dynamic (MD) simulation, and molecular docking [83]. In this study two optimal 3D-QSAR models with reliable predictive abilities were developed, having $Q^2 = 0.64$, $R^2_{\text{pred}} = 0.72$ for Dm-nAChR, and $Q^2 = 0.63$, $R^2_{\text{pred}} = 0.672$, for Md-nAChR. The graphical analysis of the 3D-contour maps is highly consistent with the docking results. The additional three methods were performed to better understand the ligands—Dm-/Md-nAChR receptor interactions and to provide some key structural features (e.g., small, electropositive, and hydrophobic substituents at the tetrahydroimidazole nitrogen vs larger, electronegative, and polar groups at nitro region) for further design of new potent inhibitors against the Dm-/Md-nAChR.

Nagaoka and co-workers [84] synthesized a number of 18 nitromethylene neonicotinoid derivatives possessing substituents that contain a sulfur atom, oxygen atom, or aromatic ring at 5-position on the imidazolidine ring. The ethylene moiety of the imidazolidine ring is considered to be an important metabolic position in the housefly *Musca domestica* [85]. For these 18 derivatives, the insecticidal activities against adult female houseflies and the affinity for nAChR were evaluated by means of the CoMFA and the Hansch-Fujita QSAR methods and the docking approach. The insecticidal activity, expressed as ED₅₀, was tested only for the 18 synthesized compounds. The statistical analysis of the CoMFA model for the receptor affinities of 50 compounds (18 novel

neonicotinoids and 32 compounds) reported by Nishiwaki et al. [86] indicated a relative contribution of the steric and electrostatic effects of 64% and 36%, respectively. The CoMFA contour maps highlighted that the more positive electrostatic features of compounds increase the activity. The relationship between the insecticidal activity and the receptor affinity was quantitatively analyzed, considering the number of sulfur and oxygen atoms, the hydrophobicity, and logP parameters, using the conventional Hansch-Fujita method. The resulted equation indicated that higher receptor affinity contributed positively to insecticidal activity, while higher hydrophobicity and the introduction of heteroatoms (S atom for this case) influenced negatively the insecticidal activity. The receptor affinity of the alkylated derivative compared with the receptor affinity of compounds possessing ether or thioether groups suggested that changing the carbon atom to a sulfur atom has no influence on the receptor affinity, while conversion to an oxygen atom was unfavorable for the receptor affinity. Furthermore, a docking model of the housefly nAChR bound to nitromethylene analogues recommended that the ligand-binding region becomes larger as the size of the substituent increases. This study completes the list of factors which influence insecticidal activity, besides the receptor affinity.

5.2.3 QSAR Models of Neonicotinoids with Insecticidal Activity Against Aphids

Aphids are among the most destructive pests causing significant economic damages and lower agriculture yields, either directly by sucking saps from various aerial tissues causing withering and death or indirectly or by transmitting several plant viruses. Blackman and Eastop [87] estimated that only 100 out of 450 aphids species show significant economic problems. From these 100 species, the authors discussed in detail 14 aphids species (13 on Aphidinae subfamily and 1 on Myzocallidinae subfamily), as being the most serious agricultural pests. The control of aphid's negative effects is achieved almost exclusively by employing insecticides. Two out of 13 species, namely, *Aphis craccivora* (pea aphids) and *Myzus persicae*, were intensively used to measure the insecticidal activity of various neonicotinoids, and QSAR models were developed based on this data. Additional QSARs developed on neonicotinoids tested against the cabbage aphid (*Brevicoryne brassicae*), armyworm (*Pseudaletia separata* Walker), and brown planthopper (*Nilaparvata lugens*) insect species and *Tetranychus cinnabarinus* (carmine spider mite of Acari: Tetranychidae) were found in the literature.

In the light of the published theoretical studies for aphids, Tian et al. [88] developed a QSAR model to predict the insecticidal activity of ten novel nitromethylene neonicotinoids containing a tetrahydropyridine ring with exo-ring ether modifications. The 22 new synthesized compounds were confirmed and analyzed by means of ¹H-NMR, high-resolution mass spectroscopy, elemental

analysis, and IR methods. From the 22 nitromethylene derivatives, only 10 exhibited good insecticidal activity (expressed as LC₅₀) against *Aphis craccivora*. It has been observed that the insecticidal activity of the compounds was correlated with the nature of the substituents introduced at the 5- and 7-positions of the tetrahydropyridine ring as follows: the activity increases when the substituents are of short alkyl types such as H, methyl, ethyl, or propyl and decreases as the groups expand. The bioactivities were quantitatively analyzed using physicochemical parameters and mono- and bi-parameter regression analysis. The QSAR results suggested that the volumes of the substituents together with the hydrophobic and electrostatic properties are essential requirements for the insecticidal activity.

The Wang group [89] designed three novel series of N3-substituted (with sulfonylamidino or sulfonyltriazolo fragments) imidacloprid analogues, which were structurally characterized by NMR spectroscopy, mass spectrometry, elemental analysis, and single-crystal X-ray diffraction analysis. Their insecticidal activities tested against *Aphis craccivora* were used to develop a 2D-QSAR model with six selected descriptors by employing a genetic algorithm-multiple linear regression (GA-MLR) method. The N3-substituted derivatives exhibited moderate to significant insecticidal activities, with LC₅₀ values ranging from 0.00895 to 0.49947 mmol/L, being comparable to that of the control, imidacloprid. Moreover, 1 out of 64 derivatives showed an approximately fourfold higher activity than that of IMI, based on the LC₅₀ value of 0.00895 mmol/L. The QSAR outcomes showed that the size, shape, and distribution of the substituents at the N3-position of IMI were significant for activity. The docking study of titled ligands into the active site of the acetylcholine-binding protein receptor indicated that all compounds realized similar hydrophobic and van der Waals interactions with Trp53, Met114, Trp143, Tyr185, and Tyr192 residues. Moreover, the presence of stronger hydrogen bond interactions between the nitro and sulfonyl group with the residues of the binding site increases the insecticidal activity. The analyzed results are useful to understand the ligand-receptor interaction mechanism of these analogues and to further optimize new neonicotinoid scaffolds. The observation regarding the key role of the size and nature of the substituents for the insecticidal activity of the N3-substituted IMI derivatives was also supported by the computational study of Bora et al. [90].

The influence of the ring size and the conversion of the bridge from O to N atom of 37 novel seven-membered azabridged neonicotinoids on the insecticidal activities, tested against *Aphis craccivora*, *Pseudaletia separata* Walker, and *Nilaparvata lugens* species, were evaluated by Xu et al. [91]. The synthesized new neonicotinoids were subjected to crystal structure development,

insecticidal assay, molecular docking, and SAR analysis. A pH value of 2–3 of the hydrolyzed succinaldehyde solution was a key condition to synthesize compounds with excellent yields. The insecticidal assay measurements against all three pest species suggested that with few exceptions, the title compounds exhibited high insecticidal activities compared to that of imidacloprid and cycloxyprid, used as controls, and of the eight-membered compounds. The bioassay showed that introducing a seven-membered azabridge unit contributes to a great improvement of the neonicotinoid analogues activities and can be considered an excellent lead structure to develop new potential insecticides. The docking study and the binding mode evaluations highlighted that the introduction of a methyl unit at position 2 of the phenyl ring is also a key feature to get high neonicotinoid insecticidal activity.

In the literature, it is mentioned that the problem of resistance and cross-resistance for various species could be overcome by modifying the structure of the existing neonicotinoids. Having this purpose in mind, the group of Yang [92] has modified the hydro-heterocycle or guanidine/amidine, and electron-withdrawing segments of the existing neonicotinoid pharmacophore with sulfonylamidine moiety to develop two new series of sulfonylamidine analogues. Their structures were subjected to chemical characterization by $^1\text{H-NMR}$, $^{13}\text{C-RMN}$, and HR-RMS, crystal structure development, insecticidal/acaricidal activities, and molecular docking study. The insecticidal and acaricidal activities (expressed as mortality (%)) were measured against the *Tetranychus cinnabarinus* and *Brevicoryne brassicae* species. The bioassay results indicated excellent acaricidal and moderate insecticidal activity for the titled compounds. The effect of different substituent group at the sulfonylamidine level was investigated by structure-activity relationships (SARs). The SAR results underlay that insecticidal/acaricidal activity difference could be ascribed to the combination of the substituent length, flexibility, and electronic characteristics. In addition, the docking simulation of the representative compound, which exhibited the highest acaricidal activity against *Tetranychus cinnabarinus* (66.7% and 83.3% of mortality in vivo, at concentration of 0.5 g/L and 1 g/L, respectively), has demonstrated a good placement into the nAChR binding site which is consistent with its high activity. The docking simulation provided key features for the structure-based design of new sulfonylamidine neonicotinoids.

Two papers of Funar and Bora [93, 94] have modeled the insecticidal activity of two different neonicotinoid scaffolds, tested against *Aphis Craccivora*, using the MLR approach. In the first paper [93], the structures of 30 neonicotinoid insecticides, bearing nitroconjugated double bond and five-membered heterocycles and, also, nitromethylene compounds containing a tetrahydropyridine ring with exo-ring ether modification were optimized at the PM7

semiempirical level, and several descriptors were generated from the minimum energy conformers. The correlation of the structural descriptors with the insecticidal activities (pLC_{50} values) against *Aphis Craccivora* leads to an MLR model with good statistical results and predictive power. Structural features, such as the number of six-membered rings, the basic pK_a capacity, and the number of ring secondary C(sp³), are beneficial for insecticidal activity of these neonicotinoids. In a second paper [94], the structures of 24 dihydropyrrole-fused and phenylazo neonicotinoid derivatives were investigated using molecular mechanics calculations based on the 94s variant of the Merck Molecular force field (MMFF94s) and the conformational search abilities of the OMEGA software. As for the previous study, the minimum energy conformers were used to derive descriptors, which were further related to the insecticidal activity (pLC_{50} values) against *Aphis craccivora*. The best derived MLR model has presented robustness ($R^2 = 0.880$, $Q^2 = 0.827$) and predictive power abilities. The MLR equation indicated three descriptors, namely, Galvez topological charge index of order 2, the leverage-weighted total index/weighted by atomic van der Waals volumes, and *R*-autocorrelation of lag 3/weighted by atomic masses, as favorable for high insecticidal activity. The developed MLR model can be confidently used to design new neonicotinoids, saving time and resources.

In a recently published paper of Bora and co-workers [95], molecular docking in conjunction with QSAR approaches was combined together in order to explore the common binding mode of 42 neonicotinoid insecticides into the nAChR active site, tested against *Aphis craccivora* (cowpea aphids), and to predict the toxicity of untested chemicals. Based on the best conformation selected by molecular docking, a high number of molecular descriptors have been computed and further employed to derive QSAR models by linear (MLR and PLS (partial least squares)) and non-linear (artificial neural networks (ANN)) and support vector machine (SVM) methods. Robust models with predictive power were generated for the titled neonicotinoids. The MLR/ANN/SVM/PLS models indicated that the presence in the neonicotinoid structure of more than five-membered rings, of the = CHR and/or the R≡N, and of R=N- fragments are favorable for the insecticidal activity. The analysis of QSAR and docking outcomes allowed the prediction of five novel insecticide compounds, which fulfill the requirements of the model applicability domain, ligand efficiencies, and binding orientations. In conclusion, these predictive models can be applied to other similar untested chemicals, active against *Aphis craccivora*, saving time, resources, and money as well as for the chemical risk assessment.

A crucial role in the binding of neonicotinoids to nAChR receptors was demonstrated to play the water-bridged hydrogen

bonds. The water-bridged importance was, also, observed in many other crystals structure of proteins such as HIV-1 protease, EGFR, etc. [96, 97]. To better understand the influence of water bridges on the insecticidal activity, Xia et al. [98] proposed an approach of heterodimeric aggregation with water. To accomplish this goal, QSAR and pharmacophore models were applied to a series of 19 neonicotinoid derivatives and their aggregates containing water bridge. For comparison, the models were, also, realized for compound monomers. The CoMSIA, pharmacophore, and linear QSAR models clarified the significant role of water molecule of the active site, while the CoMFA analysis was not considered as a good choice to elucidate the contribution of water bridges to activity. CoMSIA analysis, also, illustrated that increasing the hydrogen bond donor ability of water bridge could be essential for activity. All three aggregate-based models presented good statistical and predictive abilities than the monomer models ones.

The same essential role of water-bridged hydrogen bonds in the neonicotinoid-nAChR receptor recognition was emphasized in the paper of Zhu and co-workers [99]. For this purpose, 24 neonicotinoid compounds, having 9 fragments (1H-1,2,3-triazole, CN, COOMe, CONHNH₂, CONHMe, NO₂, NH₂, NHCOMe, and NHCSNH), which mimic the water bridges, were designed, synthesized, bioassayed against *Aphis craccivora*, and modeled by molecular docking. The compounds substituted with the cyano fragment displayed good insecticidal activity, compared with other fragment-substituted compounds, suggesting the cyano group as being optimal for mimicking the water bridges. The docking outcomes indicated that cyano fragments act only as H-bond donor, while the water bridges operate as both donor and acceptor. The other eighth fragments could operate as both donor and acceptor. The reduced insecticidal activity of these fragments could be attributed to their length, compared with that of the cyano group. These facts revealed that the water site could not be occupied by those fragments, even by the cyano group. So, this affirmation illustrates again the significant role of water-bridged hydrogen bonds in the neonicotinoid—nAChR recognition.

The two new developed neonicotinoids, sulfoxaflor and guadipyr, have proven to be very effective for the control of sap-feeding pest insects including those resistant to other insecticides. In this regard, the study of Loso et al. [100] presents a detailed structure-activity relationship of the 3-pyridyl ring of sulfoxaflor. The SAR study revealed the key role of the 3-pyridyl ring and methyl substituent on the methylene bridge connecting the pyridine and sulfoximine unit to obtain enhanced *Myzus persicae* insecticidal activity. A QSAR model, using a genetic algorithm-multiple linear regression approach, was developed to evaluate the effect of pyridine ring substituents, of 18 sulfoximine derivatives

including sulfoxaflor, on the *Myzus persicae* insecticidal activity. The model equation indicates a strong correlation between SlogP (the calculated log octanol/water partition coefficient) descriptor and the insecticidal activity, expressed as $\log(\text{LC}_{50})$. The QSAR model was highly predictive and explains the optimized pyridine substitution arrangement for sulfoxaflor.

Crisan et al. [101] proposed an approach to identify new insecticides against *Myzus persicae*, starting from the newly launched neonicotinoid, guadipyr. Thus, a series of 31 guadipyr analogues, active against *Myzus persicae*, was investigated using linear (MLR and PLS) and nonlinear (ANN and SVM) methods, together with pharmacophore modeling. Robust MLR/PLS/ANN/SVM models with predictive power were found for guadipyr analogues by correlating the insecticidal activity (pLC_{50}) with molecular descriptors generated from the energy optimized structure. For all four model types, three descriptors such as the number pyridine rings (including acceptor nitrogen atom), JGI7 (7-ordered mean topological charge), and R6p (R autocorrelation of lag 6/weighted by atomic polarizabilities) are considered to be significant to explain the insecticidal activity against *Myzus persicae*. These three descriptors confirm the positive effect of molecule geometry, size, and shape on the insecticidal activity. Based on the QSARs and pharmacophore outcomes, four new insecticide inhibitors were predicted, according to the model applicability domain and the binding mode.

5.3 QSAR Models for the Inhibition Ability of Acetylcholinesterase and Other Enzymes by Organophosphorus (OP) Pesticides

OPs induce the poisoning of a vast array of species, including humans, especially through the action on the AChE enzyme, by which a phosphorylation process undergoes [102, 103]. However, this site of action of OPs has not been identified in plants and microorganisms [7]. AChE inhibitors are classified into two categories: reversible and irreversible. While reversible inhibitors generally have therapeutic applications, the irreversible AChE modulators (like the OP compounds) produce toxic effects [104]. Thus, the toxicodynamics of OPs are seemingly irreversible through the accumulation of acetylcholine and desensitization of the cholinergic receptor, by overstimulating it [11], resulting in damage to the peripheral and central nervous system [105].

There is a multitude of toxicological effects resulting from the interaction of OP pesticides with human targets. It was observed that OP compounds can produce the neurodegenerative disorder, named organophosphate-induced delayed neuropathy (OPIDN), by triggering the neuropathy target esterase (NTE) enzyme [106, 107] when a suprathreshold dose exposure of the subject occurs, with severe symptoms which may include weakness, sensory loss, paralysis, and coma [108]. However, unlike AChE and cholinergic toxicity, NTE inhibition produces neuropathy only if an aging inhibitor is used [109]. Stallone and co-workers stressed in a study

among farm residents that high concentration of OP pesticides can produce severe depressive symptoms, which were correlated with poisoning symptoms [110].

Not only high OP concentrations affect human health but also chronic low-dose exposure leads to a number of health problems, such as neurological disorders: Parkinson's [111] and Alzheimer's diseases [112], cancers [113, 114], endocrine disruptors [115], genotoxicity [116], respiratory complications [117], etc. Another interesting study made by Bouchard et al. [118] shows the correlation between the organophosphate exposure of a representative sample of 1139 children in the USA exposed at common levels of these pesticides and the attention-deficit/hyperactivity disorder prevalence.

OP compounds inhibit AChE through a process which comprises two steps: the first one, represents the binding step, which is described by the binding constant (K_a), and the second one, the phosphorylation step, which is described by the phosphorylation rate constant (K_p) [119]. The K_p/K_a ratio, named bimolecular rate constant or inhibition constant (K_i), is one of the most significant determinants of the toxicity of OP pesticides and therefore is frequently used as dependent variable in QSAR models. Mastrantonio et al. [119] developed QSARs using electronic, steric, hydrophobic, and conformational descriptors derived from 10 OP compounds for predicting the inhibition kinetic values for K_a , K_p , and K_i determined for the AChE activity of Wistar rat brain extracts. The possible conclusions of this work are the following: the hydrophobic interactions play an interesting role, leading to the increase of the K_p values, which are associated with a good capacity of interaction between the ligand and the target molecules. The same characteristic increases the K_a values, meaning the difficulty to situate the ligand in the active site properly. It is important to note that the binding affinity of the inhibitors was observed to be determined by the conformational freedom of OPs, which for this research appears to be sufficient for the quantification of kinetic magnitudes.

The prediction of OP AChE inhibition for a dataset of eight compounds, using in silico methods, consisting in pharmacophore and 3D-QSAR modeling, using the Catalyst Program, was realized by Yazal et al. [120]. This group found a very good correlation coefficient R^2 of 0.994 between the experimental and predicted values of activity (IC_{50}), for the best pharmacophore hypothesis, consisting in one hydrogen bond acceptor site, two hydrophobic sites, and one aromatic ring. The model shows information about the structural and steric requirements of the compounds, in order to inhibit the AChE enzyme, a feature which is correlated with their neurotoxicity.

A much bigger dataset of 278 OP AChE inhibitors and their pentavalent organophosphate oxon human acetylcholinesterase bimolecular rates has been involved in the generation of a consensus QSAR model [121]. The consensus QSAR model was obtained based on averages of predictions of individual models, using molecular descriptors computed with CODESSA (topological, topographical, and geometric parameters) and AMPAC (electrostatic, quanto-chemical, and thermodynamic descriptors) software. The correlation between the experimental and predicted values for the inhibitory bimolecular rates of the human AChE was good, and also R^2 correlations values for the training, internal, external, and y-randomized tests were found to have reasonable values.

The same aforementioned series of OPs was then used by Veselinović et al. [122] in the generation of QSAR models, using the Monte Carlo method, with the bimolecular rate constants, as activity, and 2D descriptors derived from SMILES codes, as independent variables. The QSARs were obtained using two different approaches: (1) the classic scheme of training, test, and validation set and (2) the balance of correlation (sub-training, calibration, test, and validation system). The statistical parameters resulted for all the models were very good. The average values obtained for three Monte Carlo runs in the classical approaches were between 0.871 and 0.926 for the coefficient of determination R^2 and between 0.863 and 0.921, for the cross-validated correlation coefficient Q^2 . In the case of the correlation balance based QSARs, the smallest R^2 value was of 0.8471 for the validation set and 0.919 the biggest value for the test set. The most significant Q^2 value, of 0.914, was obtained for the test set.

Given that OPs are not selective pesticides, there are studies that have highlighted, also, non-cholinergic pathways, like the digestive enzymes, trypsin and alpha-chymotrypsin, which belong to the serine protease family, being targeted by OP compounds [123]. These receptors may not play a role in acute toxicity but are significant in chronic low-level toxicity. However, trypsin was found in a much higher concentration in the pancreas, compared with AChE in blood, which by inhibition is likely to lead to acute pancreatitis [124]. Ruark [125] in a master thesis study developed integrated QSAR biologically based dose-response (BBDR) models, using the heuristic regression procedure, with the aim to predict bimolecular rate constants of OP binding to trypsin and alpha-chymotrypsin, respectively. The descriptors used for the models were those which describe electrophilicity, lipophilicity, hydrogen bonds, steric hindrance, connectivity, van der Waals interactions, London dispersion, and electrostatic forces. The R^2 value obtained for global trypsin was of 0.94 and for global α -chymotrypsin of 0.92. These QSAR-BBDR models can be used in predicting the toxicodynamics of OP pesticides with their aforementioned targets

from different species, as well as for prediction of OP bimolecular rate constants for different proteins of the serine proteases family.

6 Conclusions

In this work, an analysis of publications dedicated to the ecotoxicological QSAR modeling of organophosphorus and neonicotinoid pesticides is portrayed. To date, various QSAR techniques published in the literature have generally as goal the prediction of improved target properties and knowledge on the compound mechanisms of action. By using computational chemistry facilities time, human and financial resources can be saved. Organophosphorus and neonicotinoid agrochemicals have received special attention not only because of their pesticidal activity but also due to their ecotoxicological properties. Classical Hansch and 3D QSAR (e.g., CoMFA, CoMSIA) methods, as well as several statistical approaches, like multiple linear regression, partial least squares, artificial neural networks, support vector machines, and classification techniques were employed in the QSAR modeling studies of organophosphorus and neonicotinoid pesticides. The QSAR models developed for these compounds are very promising theoretical alternative (nonanimal) approaches in that they revealed high significant predictive capabilities of less harmful and eco-friendly agrochemicals, which are less expensive. However, the applicability of QSAR models and their implementation in practice remains a very interesting subject and requires continuous improvement. Few QSAR studies were reported for some neonicotinoid ecotoxicity properties. Future research on the neonicotinoid degradability and other animal and plant toxicities will ameliorate their environment-friendly features. The combination of QSARs with virtual screening approaches will bring useful information in the design of new pesticides with improved mechanism of action and less ecotoxic properties.

Acknowledgments

This work was financially supported by the Project No. 1.1/2018 of the “Coriolan Dragulescu” Institute of Chemistry of the Romanian Academy. Alina Bora, Luminita Crisan, and Ana Borota contributed equally to this work.

Glossary

AChE	Acetylcholinesterase
ADME	Absorption, distribution, metabolism, excretion

Analogue	A chemical compound that differs from another compound by one or more atoms. These compounds form a congeneric series of the compound, which have similar structures and similar physicochemical properties.
ANN	Artificial neural networks
CART	Classification and regression tree
CoMFA	Comparative molecular field analysis
COMSIA	Comparative molecular similarity index analysis
Descriptor	A numeric representation of molecules based on their chemical structures, e.g., steric descriptors, which are related to shape or molecular size, hydrophobic descriptors (usually quantified by $\log P$ —the partition coefficient between hydrophilic and hydrophobic phases), and electronic descriptors such as atomic charge, etc.
EC ₅₀	Half maximal effective concentration
ED ₅₀	Effective dose (for inducing paralysis or death in 50% of the tested population)
F	Fischer test
IMI	Imidacloprid
In silico	An expression, which denotes, “performed on the computer or via computer simulation.”
K_a	The binding constant
K_i	The inhibition constant or bimolecular rate constant
K_p	The phosphorylation rate constant
KNN	K-nearest neighbors classification
LC ₅₀	The median lethal concentration (the concentration of a substance expected to induce death of 50% of the members of a tested population)
LD ₅₀	The median lethal dose (the single dose necessary to kill 50% of the members of a tested population)
LDA	Linear discriminant analysis
logBCF	Logarithm of the bioconcentration factor
MLR	Multiple linear regression
NMC	Nearest mean classifier
nAChR	Nicotinic acetylcholine receptor
OP	Organophosphorus
p	Significance level of regression
PLS	Partial least squares
Q^2	Cross-validation correlation coefficient
QDA	Quadratic discriminant analysis
QSAR	Quantitative structure-activity relationship
QSTR	Quantitative structure-toxicity relationship
R^2	The coefficient of correlation/determination
R^2_{adj}	The adjusted R^2
RDA	Regularized discriminant analysis
RMSE	Root-mean-square error
SEE	Standard error of the estimate
SIMCA	Soft independent modeling of class analogy
SVM	Support vector machine

References

1. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180
2. Hansch C, Fujita T (1964) *p*-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 86:1616–1626
3. Pope CN (1999) Organophosphorus pesticides: do they all have the same mechanism of toxicity? *J Toxicol Environ Health B Crit Rev* 2:161–181
4. Gavrilescu M (2005) Fate of pesticides in the environment and its bioremediation. *Eng Life Sci* 5:497–526
5. Coppage DL, Bradeich E (1976) River pollution by anti-cholinesterase agent. *Water Res* 10:19–24
6. Kumar SK (2000) Biological basis of assessment of ecotoxicology of pesticides on soil organisms. Dissertation master's thesis, Centre for Environment Jawaharlal Nehru Technological University, Hyderabad
7. Vermeire T, McPhail R, Waters M (2003) Integrated human and ecological risk assessment: a case study of organophosphorous pesticides in the environment. *Hum Ecol Risk Assess* 9:343–357
8. Morifusa E (1979) Organophosphorus pesticides: organic and biological chemistry. CRC Press, Boca Raton
9. Marrs TC (1993) Organophosphate poisoning. *Pharmacol Ther* 58:51–66
10. Gupta RC (2006) Classification and uses of organophosphates and carbamates. In: Gupta RC (ed) Toxicology of organophosphate & carbamate compounds. Elsevier, Amsterdam
11. Bajgar J (2004) Organophosphates/nerve agent poisoning: mechanism of action, diagnosis, prophylaxis, and treatment. *Adv Clin Chem* 38:151–216
12. Elersek T, Filipic M (2011) Organophosphorous pesticides - mechanisms of their toxicity. In: Stoytcheva M (ed) Pesticides – the impacts of pesticides exposure. IntechOpen, London
13. Boublík Y, Saint-Aguet P, Lougarre A et al (2002) Acetylcholinesterase engineering for detection of insecticide residues. *Protein Eng Des Sel* 15:43–50
14. Curti C, Masson P (1993) Aging of cholinesterase after inhibition by organophosphates. *Ann Pharm Fr* 51:63–77
15. Aldridge WN, Reiner E (1972) Acylated amino acids in inhibited B-esterases. In: Neuberger A, Tatum EL (eds) Enzyme inhibitors as substrates. North-Holland Publishing Company, Amsterdam
16. IUPAC. Compendium of Chemical Terminology, 2nd ed. (the “Gold Book”). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: <http://goldbook.iupac.org> (2006) created by Nic M, Jirat J, Kosata B; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. <https://doi.org/10.1351/goldbook>. Last update: 2014-02-24; version: 2.3.3. <https://doi.org/10.1351/goldbook.AT06800>
17. Nauen R, Denholm I (2005) Resistance of insect pests to neonicotinoid insecticides: current status and future prospects. *Arch Insect Biochem Physiol* 58:200–215
18. Jeschke P, Nauen R, Schindler M, Elbert A (2011) Overview of the status and global strategy for neonicotinoids. *J Agric Food Chem* 59:2897–2908
19. Casida JE, Durkin KA (2013) Neuroactive insecticides: targets, selectivity, resistance, and secondary effects. *Annu Rev Entomol* 58:99–117
20. Bonmatin JM, Giorio C, Girolami V et al (2015) Environmental fate and exposure; neonicotinoids and fipronil. *Environ Sci Pollut Res* 22:35–67
21. Shao X, Liu Z, Xu X, Li Z, Qian X (2013) Overall status of neonicotinoid insecticides in China: production, application and innovation. *J Pestic Sci* 38:1–9
22. Jeschke P, Nauen R (2008) Neonicotinoids – from zero to hero insecticide chemistry. *Pest Manag Sci* 64:1084–1098
23. Casida JE (2018) Neonicotinoids and other insect nicotinic receptor competitive modulators: progress and prospects. *Annu Rev Entomol* 63:125–144
24. Elbert A, Haas M, Springer B et al (2008) Applied aspects of neonicotinoid uses in crop protection. *Pest Manag Sci* 64:1099–1105
25. Matsuda K, Kanaoka S, Akamatsu M, Sattelle DB (2009) Diverse actions and target-site selectivity of neonicotinoids: structural insights. *Mol Pharmacol* 76:1–10
26. Nauen R, Bretschneider T (2002) New modes of action of insecticides. *Pest Outlook* 13:241–245

27. Casida JE, Quistad GB (2004) Why insecticides are more toxic to insects than people: the unique toxicology of insects. *J Pestic Sci* 29:81–86
28. Liu GY, Ju XL, Cheng J (2010) Selectivity of imidacloprid for fruit fly versus rat nicotinic acetylcholine receptors by molecular modeling. *J Mol Model* 16:993–1002
29. Simon-Delso N, Amaral-Rogers V, Belzunces LP et al (2015) Systemic insecticides (neonicotinoids and fipronil): trends, uses, mode of action and metabolites. *Environ Sci Pollut Res Int* 22:5–34
30. Li C, Xu X-Y, Li J-Y et al (2011) Synthesis and chiral purification of ¹⁴C-labeled novel neonicotinoids, paichongding. *J Label Comp Radiopharm* 54:775–779
31. Shao X, Swenson TL, Casida JE (2013) Cycloexaprid insecticide: nicotinic acetylcholine receptor binding site and metabolism. *Agric Food Chem* 61:7883–7888
32. Tomizawa M, Durkin KA, Ohno I et al (2011) N-haloacetylimino neonicotinoids: potency and molecular recognition at the insect nicotinic receptor. *Bioorg Med Chem Lett* 21:3583–3586
33. Zhang WW, Yang XB, Chen WD et al (2010) Design, multicomponent synthesis, and bioactivities of novel neonicotinoid analogues with 1,4-dihydropyridine scaffold. *J Agric Food Chem* 58:2741–2745
34. Verhaar HJM, Eriksson L, Sjostrom M et al (1994) Modelling the toxicity of organophosphates: a comparison of the multiple linear regression and PLS regression methods. *Quant Struct-Act Relat* 13:133–143
35. Schuurmann G (1990) QSAR analysis of the acute fish toxicity of organic phosphorothionates using theoretically derived molecular descriptors. *Environ Toxicol Chem* 9:417–428
36. Perrin DD, Dempsey B, Serjeant EP (1981) pK_a prediction for organic acids and bases. Chapman and Hall, New York
37. Leo A (1986) CLOGP-3.42 MedChem Software, Medicinal Chemistry Project, Pomona College, Claremont
38. Opperhuizen A, Volde EW, Gobas FAPC et al (1985) Relationship between bioconcentration in fish and steric factors of hydrophobic chemicals. *Chemosphere* 14:1871–1896
39. Drew MGB, Lumley JA, Price NR (1999) Predicting ecotoxicology of organophosphorous insecticides: successful parameter selection with the genetic function algorithm. *Quant Struct-Act Relat* 18:573–583
40. Cerius2 (1997) Molecular Simulations Inc., San Diego. <http://www-jmg.ch.cam.ac.uk/cil/SGTL/cerius2.html>. Accesed 12 Mar 2019
41. Tsar V2.4, Oxford Molecular Ltd., Magdalen Centre, Oxford Science Park, Standford-on-Thames, Oxford
42. Frisch MJ, Trucks GW, Schlegel HB et al (1995) Gaussian 94, revision E.1, Gaussian, Inc., Pittsburgh
43. Yan D, Jiang X, Yu G et al (2006) Quantitative structure-toxicity relationships of organophosphorous pesticides to fish (*Cyprinus carpio*). *Chemosphere* 63:744–750
44. Li F (1991) The metabolism and toxicity of the agrochemical. Chemical Industry Press, Beijing
45. Mazzatorta P, Benfenati E, Lorenzini P, Vighi M (2004) QSAR in ecotoxicity: an overview of modern classification Techniques. *J Chem Inf Comput Sci* 44:105–112
46. Guo JX, Wu JJ, Wright JB, Lushington GH (2006) Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: a molecular modeling study. *Chem Res Toxicol* 19:209–216
47. Garcia-Domenech R, Alarcon-Elbal P, Bolas G et al (2007) Prediction of acute toxicity of organophosphorus pesticides using topological indices. *SAR QSAR Environ Res* 18:745–755
48. Can A (2014) Quantitative structure-toxicity relationship (QSTR) studies on the organophosphate insecticides. *Toxicol Lett* 230:434–443
49. Zhao J, Yu S (2013) Quantitative structure-activity relationship of organophosphate compounds based on molecular interaction fields descriptors. *Environ Toxicol Pharmacol* 35:228–234
50. Niraj RR, Saini V, Kumar A (2015) QSAR analyses of organophosphates for insecticidal activity and its in-silico validation using molecular docking study. *Environ Toxicol Pharmacol* 40:886–894
51. Vighi M, Garlanda MM, Calamari D (1991) QSARs for toxicity of organophosphorous pesticides to *Daphnia* and honeybees. *Sci Total Environ* 109–110:605–622
52. Tanji K, Sullivan J (1995) QSAR analysis of the chemical hydrolysis of organophosphorus pesticides in natural waters. Technical completion report, project number W-843. <https://escholarship.org/content/qt44p7338k/qt44p7338k.pdf>. Accesed 23 Jan 2019

53. Lu GN, Dang Z, Tao XQ et al (2007) Quantitative Structure-Activity Relationships for enzymatic activity of chloroperoxidase on metabolizing organophosphorus pesticides. *QSAR Comb Sci* 26:182–188
54. Bass C, Denholm I, Williamson MS, Nauen R (2015) The global status of insect resistance to neonicotinoid insecticides. *Pestic Biochem Physiol* 121:78–87
55. Tomizawa M, Casida JE (2003) Selective toxicity of neonicotinoids attributable to specificity of insect and mammalian nicotinic receptors. *Annu Rev Entomol* 48:339–364
56. Neonicotinoids: risks to bees confirmed (2018) European Food Safety Authority. <https://www.efsa.europa.eu/en/press/news/180228>, <https://www.efsa.europa.eu/sites/default/files/news/180228-QA-Neonics.pdf>. Accesed 15 Feb 2019
57. Goulson D, Nicholls E, Botías C, Rotheray EL (2015) Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science* 347(6229):1255957
58. Thompson H, Maus C (2007) The relevance of sub-lethal effects in honey bee testing for pesticide risk assessment. *Pest Manag Sci* 63:1058–1061
59. Basant N, Gupta S (2017) QSAR modeling for predicting mutagenic toxicity of diverse chemicals for regulatory purposes. *Environ Sci Pollut Res* 24:14430–14444
60. Hamadache M, Benkortbi O, Hanini S, Amrane A (2018) QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (*Apis mellifera* L.). *Environ Sci Pollut Res* 25:896–907
61. Wang X, Chu Z, Yang J, Li Y (2017) Pentachlorophenol molecule design with lower bioconcentration through 3D-QSAR associated with molecule docking. *Environ Sci Pollut Res* 24:25114–25125
62. Hamadache M, Benkortbi O, Hanini S et al (2016) A quantitative structure-activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. *J Hazard Mater* 303:28–40
63. Cronin MTD, Walker JD, Jaworska JS et al (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ Health Perspect* 111:1376–1390
64. Shao X, Li Z, Qian X, Xu X (2009) Design, synthesis, and insecticidal activities of novel analogues of neonicotinoids: replacement of nitromethylene with nitroconjugated system. *J Agric Food Chem* 57:951–957
65. Shao X, Fu H, Xu X et al (2010) Divalent and oxabridged neonicotinoids constructed by dialdehydes and nitromethylene analogues of imidacloprid: design, synthesis, crystal structure, and insecticidal activities. *J Agric Food Chem* 58:2696–2702
66. Xu R, Xia R, Luo M et al (2014) Design, synthesis, crystal structures, and insecticidal activities of eight-membered azabridge neonicotinoid analogues. *J Agric Food Chem* 62:381–390
67. Lei C, Geng L, Xu X, Shao X, Li Z (2018) Isoxazole-containing neonicotinoids: design, synthesis, and insecticidal evaluation. *Bioorg Med Chem Lett* 28:831–833
68. van der Sluijs JP, Simon-Delso N, Goulson D et al (2013) Neonicotinoids, bee disorders and the sustainability of pollinator services. *Curr Opin Environ Sustain* 5:293–305
69. Goulson D (2013) An overview of the environmental risks posed by neonicotinoid insecticides. *J Appl Ecol* 50:977–987
70. Devillers J, Pham-Delegue MH, Decourtye A et al (2003) Modeling the acute toxicity of pesticides to *Apis mellifera*. *Bull Insectol* 56:103–109
71. Toropov A, Benfenati E (2007) SMILES as an alternative to the graph in QSAR modeling of bee toxicity. *Comput Biol Chem* 31:57–60
72. Singh KP, Gupta S, Basant N, Mohan D (2014) QSTR modeling for qualitative and quantitative toxicity predictions of diverse chemical pesticides in honey bee for regulatory purposes. *Chem Res Toxicol* 27:1504–1515
73. Zhao Y, Li Y (2018) Modified neonicotinoid insecticide with bi-directional selective toxicity and drug resistance. *Ecotoxicol Environ Saf* 164:467–473
74. Nishimura K, Kiriyama K, Kagabu S (2006) Quantitative structure-activity relationships of imidacloprid and its analogs with substituents at the C5 position on the pyridine ring in the neuroblocking activity. *J Pestic Sci* 31:110–115
75. Sung N-D, Jang S-C, Choi K-S (2006) CoMFA and CoMSIA on the neuroblocking activity of 1-(6-chloro-3-pyridylmethyl)-2-nitroiminoimidazolidine analogues. *Bull Kor Chem Soc* 27:1741–1746
76. Kagabu S, Nishimura K, Naruse Y, Ohno I (2008) Insecticidal and neuroblocking potencies of variants of the thiazolidine moiety of thiacloprid and quantitative relationship study

- for the key neonicotinoid pharmacophore. *J Pestic Sci* 33:58–66
77. Kagabu S, Ishihara R, Hieda Y, Nishimura K, Naruse Y (2007) Insecticidal and neuroblocking potencies of variants of the imidazolidine moiety of imidacloprid-related neonicotinoids and the relationship to partition coefficient and charge density on the pharmacophore. *J Agric Food Chem* 55:812–818
78. Okazawa A, Akamatsu M, Ohoka A et al (1998) Prediction of the binding mode of imidacloprid and related compounds to house fly head acetylcholine receptors using three-dimensional QSAR analysis. *Pestic Sci* 54:134–144
79. Suzuki T, Avram S, Borota A, Funar-Timofei S (2014) QSAR modeling of N3-substituted imidacloprid insecticides used against the housefly *musca domestica*. *J Toyo Univ Natural Sci* 8:83–95. <http://jairo.nii.ac.jp/0236/00004962/en>
80. Nishiwaki H, Nagaoka H, Kuriyama M et al (2011) Affinity to the nicotinic acetylcholine receptor and insecticidal activity of chiral imidacloprid derivatives with a methylated imidazolidine ring. *Biosci Biotechnol Biochem* 75:780–782
81. Kuriyama K, Nishiwaki H, Nakagawa Y, Nishimura K (2003) Insecticidal activity and nicotinic acetylcholine receptor binding of dinotefuran and its analogues in the housefly, *Musca domestica*. *Pest Manag Sci* 59:1093–1100
82. Li J, Ju XL, Jiang FC (2008) Pharmacophore model for neonicotinoid insecticides. *Chin Chem Lett* 19:619–622
83. Li Q, Kong X, Xiao Z et al (2012) Structural determinants of imidacloprid-based nicotinic acetylcholine receptor inhibitors identified using 3D-QSAR, docking, and molecular dynamics. *J Mol Model* 18:2279–2289
84. Nagaoka H, Nishiwaki H, Kubo T et al (2015) Docking model of the nicotinic acetylcholine receptor and nitromethylene neonicotinoid derivatives with a longer chiral substituent and their biological activities. *Bioorg Med Chem* 23:759–769
85. Nishiwaki H, Sato K, Nakagawa Y et al (2004) Metabolism of imidacloprid in houseflies. *J Pestic Sci* 29:110–116
86. Nishiwaki H, Kuriyama M, Nagaoka H et al (2012) Synthesis of imidacloprid derivatives with a chiral alkylated imidazolidine ring and evaluation of their insecticidal activity and affinity to the nicotinic acetylcholine receptor. *Bioorg Med Chem* 20:6305–6312
87. Blackman RL, Eastop VF (2007) Taxonomic issues. In: van Emden HF, Harrington R (eds) *Aphids as crop pests*. CABI, Wallingford
88. Tian Z, Shao X, Li Z et al (2007) Synthesis, insecticidal activity, and QSAR of novel nitromethylene neonicotinoids with tetrahydro-pyridine fixed cis configuration and exo-ring ether modification. *J Agric Food Chem* 55:2288–2292
89. Wang MJ, Zhao XB, Wu D et al (2014) Design, synthesis, crystal structure, insecticidal activity, molecular docking, and QSAR studies of novel N3substituted imidacloprid derivatives. *J Agric Food Chem* 62:5429–5442
90. Bora A, Avram S, Funar-Timofei S, Halip L (2018) Computational electronic profile of the insecticide imidacloprid and analogues. *Rev Roum Chim* 63:861–867
91. Xu R, Luo M, Xia R et al (2014) Seven-membered azabridged neonicotinoids: synthesis, crystal structure, insecticidal assay, and molecular docking studies. *J Agric Food Chem* 62:11070–11079
92. Yang L, Zhao Y-L, Li H-H et al (2014) Design, synthesis, crystal structure, bioactivity, and molecular docking studies of novel sulfonylamidine-derived neonicotinoid analogues. *Med Chem Res* 23:5043–5057
93. Funar-Timofei S, Bora A (2017) QSAR study of neonicotinoid insecticidal activity against cowpea aphids by the MLR approach. In: Proceedings of the 21st international electronic conference on synthetic organic chemistry, 4727. <https://doi.org/10.3390/ecsoc-21-04727>
94. Funar-Timofei S, Bora A (2019) Insecticidal activity evaluation of phenylazo and dihydro-pyrrole-fused neonicotinoids against cowpea aphids using the MLR approach. *Proceedings* 9:18. <https://doi.org/10.3390/ecsoc-22-05664>
95. Bora A, Suzuki T, Funar-Timofei S (2019) Neonicotinoid insecticide design: molecular docking, multiple chemometric approaches, and toxicity relationship with Cowpea aphids. *Environ Sci Pollut Res Int* 26:14547–14561. <https://doi.org/10.1007/s11356-019-04662-9>
96. Hidaka K, Kimura T, Abdel-Rahman HM et al (2000) Small-sized human immunodeficiency virus type-1 protease inhibitors containing allophenylnorstatine to explore the S2' pocket. *Biochemistry* 39:12534–12542

97. Wissner A, Berger DM, Boschelli DH et al (2000) 4-Anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors. *J Med Chem* 43:3244–3256
98. Xia S, Cheng J, Feng Y et al (2014) Computational investigations about the effects of hetero-molecular aggregation on bioactivities: a case of neonicotinoids and water. *Chin J Chem* 32:324–334
99. Zhu C, Li G, Xiao K et al (2019) Water bridges are essential to neonicotinoids: insights from synthesis, bioassay, and molecular modeling studies. *Chin J Chem* 30:255–258
100. Loso MR, Benko Z, Buysse A et al (2016) SAR studies directed toward the pyridine moiety of the sap-feeding insecticide sulfoxaflof (Isoclast™ active). *Bioorg Med Chem* 24:378–382
101. Crisan L, Borota A, Suzuki T, Funar-Timofei S (2019) An approach to identify new insecticides against *Myzus Persicae*. In silico study based on linear and non-linear regression techniques. *Mol Inform.* <https://doi.org/10.1002/minf.201800119>
102. Namba T (1971) Cholinesterase inhibition by organophosphorus compounds and its clinical effects. *Bull World Health Organ* 44:289–307
103. Koureas M, Tsakalof A, Tsatsakis A, Hadjichristodoulou C (2012) Systematic review of biomonitoring studies to determine the association between exposure to organophosphorus and pyrethroid insecticides and human health outcomes. *Toxicol Lett* 210:155–168
104. Colovic MB, Krstic DZ, Lazarevic-Pasti TD et al (2013) Acetylcholinesterase inhibitors: pharmacology and toxicology. *Curr Neuropharmacol* 11:315–335
105. Mearns J, Dunn J, Lees-Haley PR (1994) Psychological effects of organophosphate pesticides: a review and call for research by psychologists. *J Clin Psychol* 50:286–294
106. Johnson MK (1969) A phosphorylation site in brain and the delayed neurotoxic effect of some organophosphorus compounds. *Biochem J* 111:487–495
107. Johnson MK (1969) The delayed neurotoxic effect of some organophosphorus compounds. Identification of the phosphorylation site as an esterase. *Biochem J* 114:711–717
108. Kobayashi S, Okubo R, Ugawa Y (2017) Delayed polyneuropathy induced by organophosphate poisoning. *Intern Med* 56:1903–1905
109. Richardson RJ, Hein ND, Wijeyesakere SJ et al (2013) Neuropathy target esterase (NTE): overview and future. *Chem Biol Interact* 203:238–244
110. Stallone L, Beseler C (2002) Pesticide poisoning and depressive symptoms among farm residents. *Ann Epidemiol* 12:389–394
111. Wang A, Cockburn M, Ly TT et al (2014) The association between ambient exposure to organophosphates and Parkinson's disease risk. *Occup Environ Med* 71:275–281
112. Yan D, Zhang Y, Liu L, Yan H (2016) Pesticide exposure and risk of Alzheimer's disease: a systematic review and meta-analysis. *Sci Rep* 6:32222
113. Guyton KZ, Loomis D, Grosse Y et al (2015) Carcinogenicity of tetrachlorvinphos, para-thion, malathion, diazinon, and glyphosate. *Lancet Oncol* 16:490–491
114. Lerro CC, Koutros S, Andreotti G et al (2015) Organophosphate insecticide use and cancer incidence among spouses of pesticide applicators in the agricultural health study. *Occup Environ Med* 72:736–744
115. Kitamura S, Sugihara K, Fujimoto N, Yamazaki T (2011) Organophosphates as endocrine disruptors. In: Satoh T, Gupta RC (eds) *Anticholinesterase pesticides: metabolism, neurotoxicity, and epidemiology*. Wiley, New York
116. Sutris JM, How V, Sumeri SA, Muhammad M et al (2016) Genotoxicity following organophosphate pesticides exposure among Orang Asli children living in an agricultural island in Kuala Langat, Selangor, Malaysia. *Int J Occup Environ Med* 7:42–51
117. Hulse EJ, Davies JO, Simpson AJ et al (2014) Respiratory complications of organophosphorus nerve agent and insecticide poisoning. Implications for respiratory and critical care. *Am J Respir Crit Care Med* 190:1342–1354
118. Bouchard MF, Bellinger DC, Wright RO, Weisskopf MG (2010) Attention-deficit/hyperactivity disorder and urinary metabolites of organophosphate pesticides. *Pediatrics* 125:e1270–e1277
119. Mastrantonio G, Mack HG, Della Védova CO (2008) Interpretation of the mechanism of acetylcholinesterase inhibition ability by organophosphorus compounds through a new conformational descriptor. An experimental and theoretical study. *J Mol Model* 14:813–821
120. Yazal JE, Rao SN, Mehl A, Slikker W Jr (2001) Prediction of organophosphorus

- acetylcholinesterase inhibition using three-dimensional quantitative structure-activity relationship (3D-QSAR) methods. *Toxicol Sci* 63:223–232
121. Ruark CD, Hack CE, Robinson PJ et al (2013) Quantitative structure-activity relationships for organophosphates binding to acetylcholinesterase. *Arch Toxicol* 87:281–289
122. Veselinovic JB, Nikolic GM, Trutic NV et al (2015) Monte Carlo QSAR models for predicting organophosphate inhibition of acetylcholinesterase. *SAR QSAR Environ Res* 26:449–460
123. Schaffer NK, Lang RP, Simet L, Drisko RW (1958) Phosphopeptides from acid-hydrolyzed P32-labeled isopropyl methylphosphonofluoridate-inactivated trypsin. *J Biol Chem* 1:185–192
124. Somogyi L, Martin SP, Venkatesan T, Ulrich CD (2001) Recurrent acute pancreatitis: an algorithmic approach to identification and elimination of inciting factors. *Gastroenterology* 120:708–717
125. Ruark CD (2010) Quantitative structure-activity relationships for organophosphates binding to trypsin and chymotrypsin. Dissertation, Miami University



Chapter 22

QSARs and Read-Across for Thiochemicals: A Case Study of Using Alternative Information for REACH Registrations

Monika Nendza, Jan Ahlers, and Dirk Schwartz

Abstract

A case study on acute aquatic toxicity of thiochemicals shows the possibilities and limitations of filling data gaps with alternative information in accordance with the requirements of REACH. It is the objective of this study to extract as much information as possible from available experimental studies with fish, daphnia, and algae to estimate required data by QSARs and read-across.

Thiochemicals are considered to be toxic with an unspecific reactive mode of action (MoA) causing so-called excess toxicity, i.e., the effects are much higher than estimated from $\log K_{\text{OW}}$ -dependent baseline QSARs. Differences in toxicity between groups of thiochemicals, for example, thioglycolates or mercapto-propionates, are thought to be due to differences in reactivity of the respective sulfur moiety, i.e., toxicodynamic differences. Thiochemicals within each group are different with regard to partitioning between biophases related to, e.g., increasing aliphatic chain length, i.e., toxicokinetic differences.

Due to the toxicodynamic and toxicokinetic differences, QSARs and read-across are limited to thiochemicals within the same group. Since the database per group of thiochemicals is too small to derive scientifically valid QSARs, most of the 36 data gaps for 16 thiochemicals to be registered by 2018 were closed by read-across. Testing strategies to fill remaining data gaps include tests with algae (six substances) and daphnia (six substances). Only for two substances, experimental (limit) fish studies are recommended. Overall, a substantial (>60%) reduction of tests by predictive in silico methods is possible.

Key words REACH, QSARs, Read-across, Category approaches, Acute aquatic toxicity, Unspecific reactive mode of action (MoA), Excess toxicity, Integrated testing strategies (ITS), 3Rs

1 Acute Aquatic Toxicity of Thiochemicals for REACH Registrations

Substances produced in or imported into the EU at more than 1 t/y have to be registered, and physicochemical, (eco)toxicological, as well as exposure-relevant information have to be supplied [1–3]. The minimal ecotoxicological information requirements for chemicals to be registered by 2018 (1–100 t/y) are data on acute toxicity to algae, daphnia, and fish (the latter only for substances ≥ 10 t/y). Any available relevant additional information has to be presented in order to achieve an optimal assessment of possible hazards to men and the environment. The information is used for

classification and labelling according to the CLP regulation [4], chemical safety assessment (CSA) including derivation of predicted environmental concentration (PEC) and predicted no-effect concentration (PNEC), and evaluation of persistent, bioaccumulative, and toxic/very persistent and very bioaccumulative (PBT/vPvB) properties.

Among the main aims of REACH¹ is the promotion of alternative methods for the assessment of hazards of substances avoiding animal testing where possible. Alternative information may be deviations from the standard test guidelines (e.g., limit tests), test results obtained with nonstandard organisms, in vitro test data, intra- or extrapolation from analogues (read-across), predictions from (quantitative) structure-activity relationships (QSARs), and extrapolations from acute to chronic data and vice versa [2, 5]. Alternative information is acceptable for REACH registrations, if it is equivalent to the results that would be obtained by standard testing and adequate to draw conclusions for classification and labelling, PNEC derivation, and PBT/vPvB assessment. The equivalence and adequacy have to be substantiated by a weight-of-evidence (WoE) approach, making best use of all available data [5–7]. Integrated testing strategies (ITS) can increase the efficiency of hazard and risk assessment and at the same time reduce the use of animals by targeted testing of chemicals [8, 9]. Lombardo et al. [10] presented a comprehensive ITS approach for organizing and using existing aquatic toxicity data to fulfill the requirements of REACH.

The aim of this case study on acute aquatic toxicity of thiochemicals is to close data gaps observed during the process of REACH registrations. First, we want to use alternative information and extract as much information as possible from available experimental studies with fish, daphnia, and algae to estimate required data by QSARs and read-across. The case study thiochemicals, a group of relatively homogeneous substances with a limited number of functional groups, are produced in amounts between 1 t/y and more than 1000 t/y and are used as reducing agents (antioxidants) in cosmetics, cleaners, and polymers; as (co)binders or hardeners in coatings, adhesives, and sealants; and in optical applications (films, lenses). Thiochemicals with production rates above 100 t/y are already registered, and therefore a number of data on aquatic toxicity are available for QSARs and read-across to fill data gaps for the lower-tonnage thiochemicals.

Established ITS and WoE [5, 11] support a stepwise procedure for obtaining as much information as possible from available (experimental) data and to fill data gaps for analogues with alternative information in accordance with the requirements of REACH. The main components of the approach are collection and

¹ REACH: EU regulation on registration, evaluation, authorization, and restriction of chemicals [1]

evaluation of available information, identification of suitable in silico methods, calculation of multiple predictions, and, finally, overall assessment of the available information to conclude about the (eco)toxicity of a substance. In the following sections, we will outline the principal procedures and illustrate them with case study examples, directed by relevant guidance documents on alternatives to animal testing [2] and the Read-Across Assessment Framework (RAAF) [3]. In those cases, where neither appropriate standard test results nor equivalent and adequate alternative information could be obtained, additional testing may become necessary.

2 Exploratory Data Analysis (EDA): Existing Information and Data Gaps

Predictive (eco)toxicology and ITS rely on careful collection of all available information both in terms of the actual endpoint, for example, acute fish toxicity, and any properties that may affect it, for example, physicochemical properties, such as water solubility, reactivity, and degradation as well as relevant metabolites [5]. Very important is the confirmation of the same chemical structures including type and amount of impurities [12]. Furthermore, factors such as the quality of the data, consistency of results, nature and severity of effects, and relevance of the information will have an influence on the weight given to the available evidence.

2.1 Test Substances and Available Data on Acute Aquatic Toxicity

The case study is based on 36 thiochemicals representing 6 chemical classes (Table 1) with quality-controlled information on octanol/water partition coefficient ($\log K_{ow}$) ($n = 36$), water solubility (S_w) ($n = 35$), biodegradability (OECD 301) ($n = 36$), acute algae toxicity (OECD 201) ($n = 17$), acute daphnia toxicity (OECD 202) ($n = 19$), and acute fish toxicity (OECD 203) ($n = 22$). The available data have been published recently [13, 14].

The test data on aquatic toxicity have been evaluated according to Klimisch et al. [15] based on study reports with plausibility checks regarding the stability of the substances over the duration of the experiments and comparing the toxic concentrations with the water solubility of the substances. Only experimental data with Klimisch code 1 (reliable without restrictions) or 2 (reliable with restrictions) have been selected to provide a sufficiently valid basis for the derivation of PNECs and for input (source) data for QSARs and read-across. Experimental data with Klimisch code 3 (not reliable) have not been used. Some problems arise from the fact that a number of thiochemicals are not sufficiently stable during the course of the tests. In these cases the decrease in concentrations had to be followed analytically, and the aquatic toxicity is estimated based on geometric mean test concentrations.

Table 1
Chemical categories of the thiochemicals

Chemical category	Characteristic fragments of thiochemical categories
Thioglycolates ($n = 10$)	
Mercaptopropionates ($n = 11$)	
Thiolactates ($n = 2$)	
Thiodiglycolates ($n = 2$)	
Thiodipropionates ($n = 5$)	
Mercaptanes ($n = 6$)	

2.2 Toxicological and Chemical Grouping of Thiochemicals and Mode of Action (MoA)

The toxicological grouping of the thiochemicals follows the chemical grouping (Table 1) in terms of functional similarity [16–19] leading to either common or different MoA.² All the thiochemicals are considered to be toxic due to the reactivity of the sulfur groups. Interactions with biogenic structures result in reactive toxicity that can significantly exceed so-called narcotic effects, i.e., the effects are much higher than estimated from log K_{ow} -dependent baseline QSARs [19]. Differences in toxicity between the groups of thiochemicals are thought to be due to differences in reactivity of the respective sulfur moiety, i.e., toxicodynamic differences. Within each group, the thiochemicals are different with regard to partitioning between biophases related to hydrophobicity, i.e., toxicokinetic differences. Comparisons of the experimental acute toxicities of thiochemicals with log K_{ow} -based baseline QSARs for algae [20], daphnia [21, 22], and fish [23, 24] reveal excess toxicities of more than one order of magnitude with distinct pattern for the different

²The concept of functional similarity can support the MoA classification of chemicals by combining toxicological knowledge (which toxicity pathways can happen in which species under which exposure conditions) with chemical expertise (which parts of the chemical structures and physicochemical properties are involved in which interactions) [16–19].

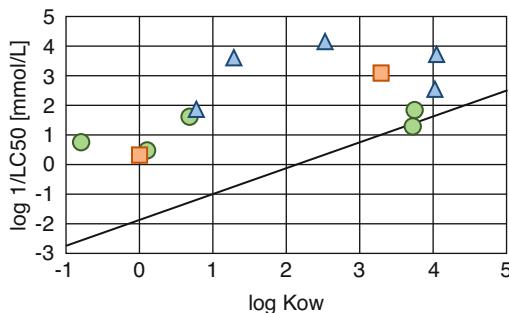


Fig. 1 MoA-related excess toxicity to fish of three categories of thiochemicals: thioglycolates (circles), mercaptopropionates (triangles), and mercaptans (squares) relative to the baseline QSAR by Könemann [23]

groups of thiochemicals [13]. Figure 1 illustrates the reactive excess toxicities of three groups of thiochemicals toward fish. Similar results are obtained with algae and daphnia.

2.3 Physicochemical Descriptors of Thiochemicals

Parameterization of the thiochemicals is related to (1) the reactivity of the substances due to the respective sulfur moiety (Table 1) and (2) the hydrophobicity and size of the molecules expressed, for example, in terms of molecular weight (MW), chain length (#C), or $\log K_{\text{OW}}$. MW values were collected from ChemSpider [25]. #C values were counted from SMILES. Multiple $\log K_{\text{OW}}$ values were calculated for the undissociated thiochemicals with EPISuite [26], ACD/Labs and ChemAxon from ChemSpider [25], XlogP and AlogP from TEST [27], and consensus, read-across and LSER from ChemProp [28]. The mean of the results from the different independent algorithms (consolidated $\log K_{\text{OW}}$) was calculated.

2.4 Data Gaps

Among the substances to be registered under REACH in 2018 were seven thiochemicals with production rates between 1 and 10 t/y, i.e., information on acute toxicity to algae and daphnia are required, and nine thiochemicals with production rates between 10 and 100 t/y, i.e., information on acute toxicity to algae, daphnia, and fish are necessary. Only for two substances, sufficient experimental data were already available. Information on acute fish toxicity existed for four substances, which had to be completed with estimates for daphnia and algae. No experimental data were available for ten substances, for four of them data gaps for acute toxicity to algae and daphnia and for six of them additionally to fish needed to be filled.

3 QSARs, Read-Across, and Testing Strategies for Acute Aquatic Toxicity of Thiochemicals

QSARs and read-across can be used for predictions depending on the amount of available data. Substantial numbers of similar substances with the same MoA are required to derive and validate QSAR models for predictive purposes. Read-across is feasible with at least one similar substance with the same MoA to extrapolate toxic concentrations between compounds (analogue approach); several similar substances allow for category approaches [3]. Since none of the methods perform in a superior manner throughout, we strongly recommend to obtain multiple independent predictions with different methods and to deal with the variability and uncertainty of estimated data by using consensus toxicity estimates [29].

3.1 QSARs

QSAR models that fulfill the OECD criteria³ for scientific validity of QSARs [30] are not available for thiochemicals from suitable inventories (e.g., JRC QSAR Model Database [31], QSAR Data-Bank repository [32]), software platforms (e.g., VEGAHub [33], ChemProp [28], Chemistry Dashboard [34]), and the literature (e.g., [29]). Due to MoA considerations (see Subheading 2.2, Fig. 1), joint QSAR modelling of all thiochemicals is not appropriate, and the available database is too small to derive new statistically valid QSAR models for different groups of thiochemicals. Therefore, it was not possible to fill data gaps regarding the aquatic toxicity of thiochemicals with QSAR predictions. Instead, read-across had been used.

3.2 Read-Across

Read-across predicts endpoint information for one substance (target substance) by using data from the same endpoint from (an) other substance(s) (source substance(s)). Depending on the number of source substances, category and analogue approaches are feasible. Category approaches use several substances that are likely to be similar or follow a regular trend as a result of structural similarity, i.e., a group (category) of substances [3]. If several similar substances with the same MoA reveal a regular pattern related to structural and physicochemical properties, trend analyses offer differentiated predictions. If no robust and reliable trends are evident, a worst-case prediction is possible among similar substances with alike (eco)toxicities. Analogue approaches use at least one similar substance with the same MoA to extrapolate toxic concentrations between compounds. Prerequisite is sufficient similarity of the target substance and the source substance(s) with

³ OECD criteria for the scientific validity of QSAR models [30]: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness of fit, robustness, and predictivity; and (5) a mechanistic interpretation, if possible

regard to chemical structure, physicochemical properties, reactivity, (eco)toxicological MoA, toxicokinetics, metabolic fate, and degradation pattern. The following stepwise procedures allow for category and analogue approaches with some differentiation of specific elements:

1. *Exploratory data analysis:*

- (a) Confirm identity and chemical structure of the target chemical.
- (b) Collect suitable analogues (category members) with sound experimental data for the respective endpoint, for example, 96-h LC₅₀ fish, from suitable databases, for example, ECHA's registered substances [35], OECD Toolbox [36], ECOTOX Knowledgebase [37], eChem-Portal [38], and the literature.
- (c) Confirm similarity of source and target chemicals in terms of chemical structures [12], physicochemical properties, MoA, stability of active ingredients, (common) degradation products, etc. [5, 36, 39–42].

2. *Read-across:*

- (a) If several analogues are available, look for trends based on structural and physicochemical descriptors, for example, log K_{OW}, MW, and chain length. Often a graphical inspection of the relationships is very helpful.
- (b) Perform predictions, for category approaches, based on robust and reliable trends (is there a possible mechanistic interpretation of the trend?). If no resilient trends are evident, make a worst-case prediction. Predictions by analogue approaches extrapolate the effect data from the source to the target compound, preferably on a molar basis (correction for different molecular weights).⁴ The rationale is that there is mostly one predominant toxic principle per molecule. The same numbers of sufficiently similar molecules are likely to cause the same effects. Extrapolations on a weight basis may include inactive parts of the molecules to variant extents and, hence, introduce unnecessary uncertainty into read-across.
- (c) Describe the uncertainty of estimates by read-across depending on the uncertainty of the experimental source values and the chemical and toxicological similarity between the source and target compounds. Determine propagated error of estimates by category approaches considering variability of experimental input data and descriptors of source compounds. Uncertainty and

⁴ Note that QSARs are always performed on a molar basis.

variability of estimates may be visualized with a plot of the trend including the target chemical together with the all other category members.

(d) Document steps 1 and 2.

3. *Consensus toxicity estimates*:

- (a) Determine consensus toxicity estimates based on the information of all available (experimental and calculated) sound and valid data for WoE, for example, using (geometric) mean, median, or Bayesian networks [9, 29].
- (b) Describe supporting/confounding evidence in terms of structural (dis)similarity, applicability domain (AD) considerations, common or different MoA, (same) degradation products, metabolites, other effect data (non-guideline studies, in vitro studies, studies with non-standard organisms), so-called adverse outcome pathways (AOP),⁵ etc.

The *exploratory data analysis* of the case study thiochemicals used the available experimental data for several chemicals within the same category, for example, thioglycolates or mercaptopropionates (Table 1), to fill data gaps based on a trend related to MW, #C, or $\log K_{\text{OW}}$. Exploratory data analyses indicated #C to be the best descriptor within the categories, outperforming $\log K_{\text{OW}}$ and MW (Fig. 2). An explanation could be that $\log K_{\text{OW}}$ and MW are obscured by the different thiol functions, while the chain length better captures the mechanistic differences within the categories.

Read-across using category and analogue approaches is performed individually for each example, and uncertainties are described accordingly. The following examples of read-across illustrate different degrees of success: *Consensus toxicity estimates* range from robust and reliable predictions that are equivalent to guideline

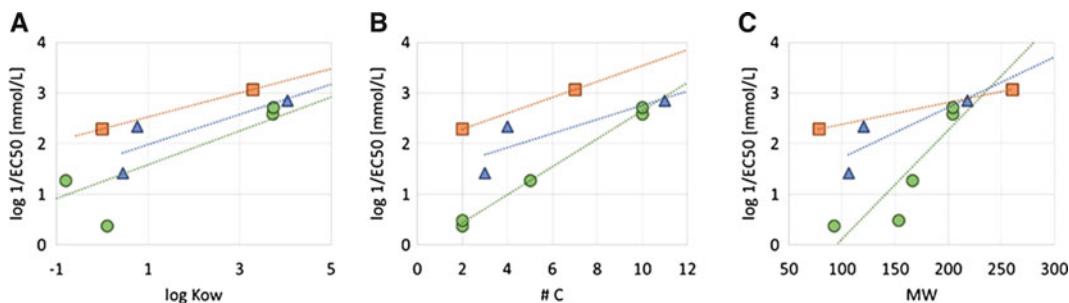


Fig. 2 Trends in daphnia toxicity related to $\log K_{\text{OW}}$ (a), #C (b), and MW (c) for three categories of thiochemicals: thioglycolates (circles), mercaptopropionates (triangles), and mercaptans (squares)

⁵ <https://aopwiki.org/>

study results and adequate for hazard assessment to highly uncertain estimates that prompt ITS considerations about the best next test with maximum information gain [43].

3.2.1 Thiolactic Acid (TLA)

Exploratory data analysis reveals no experimental data for TLA (Fig. 3) and information requirements for registration; thus, data gaps exist regarding acute toxicities on algae, daphnia, and fish. TLA belongs to the group of thiolactates. For one of the TLA salts, ammonium thiolactate, experimental data for fish, daphnia, and algae are available. Up to a concentration of 100 mg/L test substance (70 mg/L active ingredient), no effects could be observed.

Read-across from ammonium thiolactate to TLA is possible since in both cases, the thiolactate ion is responsible for the toxic effect. Correction for different molecular weights results in EC₅₀ for algae >120 mg/L and EC₅₀ for daphnia and fish >60 mg/L, respectively.

Consensus toxicity estimates rely on calculated values for daphnia and fish as basis for PNEC derivation as well as C&L. The WoE of the read-across results for TLA is sufficient to refrain from animal testing.

3.2.2 Glycol Dimercaptoacetate (GDMA)

Exploratory data analysis reveals an experimental EC₅₀ for fish (4.8 mg/L) for GDMA (Fig. 4) and information requirements for registration regarding acute toxicities on algae, daphnia, and fish; thus, data gaps exist for algae and daphnia. GDMA belongs to the group of thioglycolates, and #C = 6 allows extrapolating trends within this group.

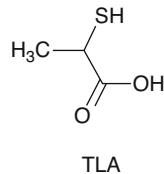


Fig. 3 Chemical structure of thiolactic acid (TLA), CAS 79-42-5

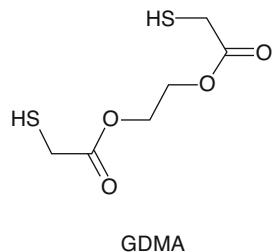


Fig. 4 Chemical structure of glycol dimercaptoacetate (GDMA), CAS 123-81-9

Trend analyses on the basis of experimental data for analogues in the range of 5–10 C atoms lead to the following estimates: EC₅₀ (algae), 6.7 mg/L; EC₅₀ (daphnia), 5.9 mg/L; and LC₅₀ (fish), 13 mg/L. The estimated fish toxicity is in reasonable agreement with the measured one. The estimations for algae and daphnia toxicity are obtained by extrapolation and considered reliable. Therefore, it was concluded that these data are sufficient for the requirements of REACH.

Consensus toxicity estimates result in PNEC derivation as well as classification and labelling (C&L) performed on the basis of the experimental fish toxicity.

3.2.3 2-Ethylhexyl 3-Mercaptopropionate (EHMP)

Exploratory data analysis reveals an experimental 48-h LC₅₀ for fish (0.63 mg/L) for EHMP (Fig. 5), but the value was not considered sufficiently reliable as the test result is based on nominal concentrations and the substance is relatively unstable. According to the information requirements for registration, data on acute toxicity with algae and daphnia have to be presented. EHMP belongs to the group of mercaptopropionates, and #C = 11 allows extrapolating trends within this group.

Trend analyses are possible on the basis of experimental data for analogues in the range of 4–11 #C. The estimates are obtained by extrapolation for algae (EC₅₀, 0.05 mg/L), daphnia (EC₅₀, 0.38 mg/L), and fish (LC₅₀, 0.11 mg/L) and, therefore, considered reliable.

Read-across of the fish acute toxicity from isoctyl 3-mercaptopropionate (iOMP), having different branching (Fig. 5), results in LC₅₀ = 0.04 mg/L. The two estimates for fish are in sufficient agreement, and the geometric mean (0.07 mg/L) is applied.

Consensus toxicity estimates rely on calculated values. Algae and fish are most sensitive. The lowest EC₅₀ (for algae) serves as the basis for PNEC derivation as well as C&L. Due to remaining uncertainties of the predictions, an algae test was proposed to verify the results.

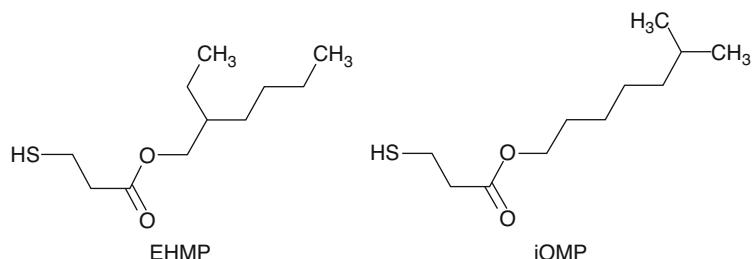


Fig. 5 Chemical structures of 2-ethylhexyl 3-mercaptopropionate (EHMP), CAS 50448-95-8, and isoctyl 3-mercaptopropionate (iOMP), CAS 30374-01-7

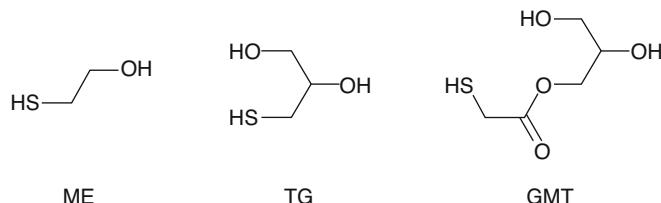


Fig. 6 Chemical structures of 2-mercaptopropan-1-ol (ME), CAS 60-24-2; thioglycerol (TG), CAS 96-27-5; and glyceryl monothioglycolate (GMT), CAS 30618-84-9

3.2.4 Thioglycerol (TG)

Exploratory data analysis reveals no experimental data for TG (Fig. 6) and information requirements for registration; thus, data gaps exist regarding acute toxicities on algae, daphnia, and fish. TG belongs to the group of mercaptans.

Read-across of EC₅₀/LC₅₀ from 2-mercaptopropan-1-ol (ME), being a substructure of TG (Fig. 6), results in 26 mg/L for algae, 0.55 mg/L for daphnia, and 51 mg/L for fish. Extrapolations from glyceryl monothioglycolate (GMT), sharing 2-hydroxy- and 1-thiol group (Fig. 6), give EC₅₀/LC₅₀ of 5.7 mg/L for algae and daphnia and 19 mg/L for fish. Since the very limited database and variable level of impurities do not allow to establish equivalent reactivity of the thiol functions of the source and target substances, the results of the read-across are not reliable.

Consensus toxicity estimates are not yet feasible. It is recommended to obtain experimental data for algae and daphnia toxicity. A comparison of the test results with the estimates will show whether it is also necessary to conduct a fish test.

3.2.5 Lauryl/Stearyl Thiodipropionate (E1218)

Exploratory data analysis reveals information requirements for registration regarding acute toxicities on algae, daphnia, and fish, but no valid data are available. E1218 belongs to the group of thiodipropionates (Fig. 7). Given the low water solubility (<1 mg/L), long-term toxicity data should be used for E1218.

Read-across from dilauryl thiodipropionate (E12) and distearyl thiodipropionate (E18) is justified by the composition of E1218 (E12, 21–31%; E18, 18–28%; E1218, 35–57%). E18 long-term studies with algae, daphnia, and fish conclude that no effects are observed up to the limit of S_w. In acute tests with E12, there were also no effects in the range of S_w, and long-term testing can be waived.

Consensus toxicity estimates assume that E1218 behaves comparably due to similar structure and similar physicochemical properties, i.e., no effects in the range of S_w. Thus, tests for aquatic toxicity required for registration can be covered by existing test results from read-across source substances.

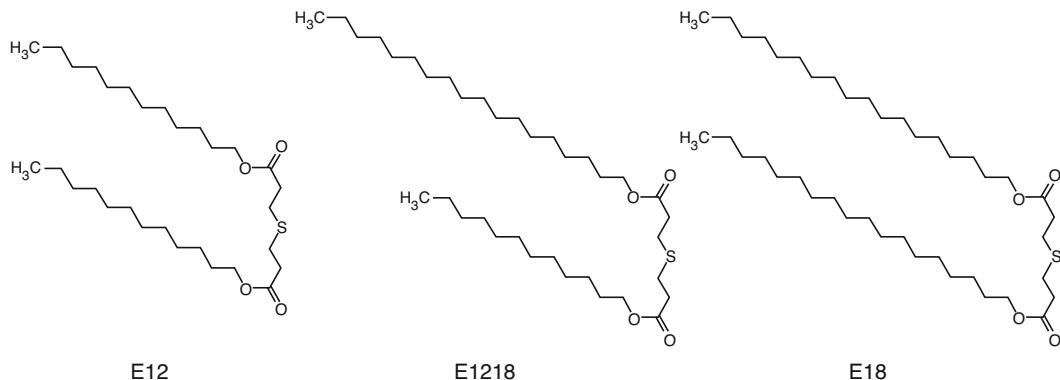


Fig. 7 Chemical structures of dilauryl thiodipropionate (E12), CAS 123-28-4; lauryl/stearyl thiodipropionate (E1218), CAS 13103-52-1; and distearyl thiodipropionate (E18), CAS 693-36-7

3.3 Overall Assessment of the Available Information

For the 16 thiochemicals of this case study, similar to the examples presented in Subheading 3.2, all data gaps were considered, and it was discussed whether the required information could be obtained by read-across or if tests have to be performed. In each case it was examined whether the number and quality of the information are equivalent to the standard information required by REACH.

Among the 16 thiochemicals to be registered in 2018 were 14 substances with data gaps. For five substances the data gaps on aquatic toxicity were closed by read-across. For the remaining nine thiochemicals, testing strategies were developed in order to obtain information that is sufficient to achieve a sound and reliable assessment. Starting with 36 data gaps, only 14 tests (6 algae, 6 daphnia, 1 limit fish test, and 1 acute fish test) have been proposed. Thus, a substantial (>60%) reduction of tests by predictive *in silico* methods is possible.

With new experimental data becoming available, iterative improvements of the above-described extrapolations can be achieved.

4 Implications of Data Quality

ITS use all available information for hazard assessment. In a WoE approach, it has to be decided whether this information is equivalent to the standard information or which additional tests are required [5, 11]. This decision is not an easy task, as a number of uncertainties have to be taken into account. It starts with the question how well defined are the chemical structures including type and amount of impurities [12] and how certain are the physicochemical properties (e.g., S_w , $\log K_{ow}$). Problematic are aquatic toxicity tests at concentrations above water solubility. Especially for thiochemicals, tests without analytical control have to be regarded

critically. Another issue is the stability of thiochemicals, which differs considerably depending on duration and conditions of the tests. Therefore all thiochemicals of our case study are more or less “difficult substances” and had to be regarded very carefully before a final conclusion could be drawn [44]. Some aspects of the wide range of quality of information are represented by our examples.

4.1 Structural and Functional Similarity

In Subheadings 2.2 and 3, we discussed why the same MoA between source and target compounds [16–19] is a prerequisite for sound read-across. This requirement is fulfilled in case of our examples TLA (identical active substructures of the molecule) and GDMA (interpolation of trends in the same group), and no further tests are required. For TG, read-across was performed with two source substances, which appear to be similar. However, GMT is a thioglycolate with a different MoA and therefore unsuitable for read-across.

Another problem arises if source and target compound have the same MoA, but the target substance is outside the applicability domain. In case of TMPMP⁶ (16 C atoms) and TEMPIC⁷ (18 C atoms), extrapolations of a trend observed for mercaptopropionates with #C range 4–11 are too uncertain. In this case tests are necessary.

Moreover, extrapolations from acute to chronic data have to be applied with care as quite often the MoAs leading to acute and chronic toxicities are different [45]. An example are endocrine disruptors like nonylphenol that are often unspecific toxicants at the acute level but act very specifically on the long term [46]. This issue, however, seems to be not relevant in case of thiochemicals.

4.2 Experimental Difficulties and Variability of Source Data

A number of thiochemicals of our case study have a very low S_W , and thus toxicities obtained from nominal concentrations (often above S_W) cannot be used (e.g., EHMP). Moreover, many of these compounds are rather unstable (e.g., EHMP, PETMA,⁸ TMPMP,⁶ TEMPIC,⁷ TG). In these cases further studies are required to clarify whether the estimated toxicities arose from the compound itself or from its decomposition products. As (probably) the source substances are also unstable and in that case the toxicities have—according to the guidelines—been estimated from the geometric mean between the concentrations at the beginning and at the end of the assay, an inherent error is propagated to the target compound.

⁶ TMPMP: Trimethylolpropane trimercaptopropionate, CAS 33007-83-9

⁷ TEMPIC: Tris[2-(3-mercaptopropionyloxy)ethyl]isocyanurate, CAS 36196-44-8

⁸ PETMA: Pentaerythritol tetrakis(mercptoacetate), CAS 10193-99-4

4.3 ITS and WoE

ITS are based on a large number of information of quite different value, which have to be weighted appropriately. The more information is available, the better are the estimates. For example, the valid estimated toxicities for GDMA (and also GDMP⁹) are supported by valid experimental ones, whereas for EHMP the non-valid experimental value could not support the estimated ones, and an algae test was recommended. However, not only validity of tests and estimates influences the quality of the results but also the variations of the individual data for the same regulatory endpoint.

4.4 Further Tests

One important target of REACH is a reduction of vertebrate testing. Therefore, in the consensus toxicity estimates of Subheading 3.2, fish tests were only suggested, if it was quite clear that this information is essential (e.g., TMPMP). In most cases, it was proposed to first perform tests with algae and/or daphnia (e.g., TG). Based on the outcome of these tests and in accordance with all other data, it has to be discussed if fish tests are still necessary or can be waived or if limit tests (step-down approach) are sufficient (e.g., TEMPIC).

References

- European Commission (2006) REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. European Commission, Brussels
- ECHA (2016) Practical guide How to use alternatives to animal testing to fulfil your information requirements for REACH registration. ECHA, Helsinki
- ECHA (2017) Read-across assessment framework (RAAF). ECHA, Helsinki
- European Commission (2009) Regulation (EC) 1272/2008 on classification, labelling and packaging of substances and mixtures (CLP) from January 20, 2009. European Commission, Brussels
- Ahlers J, Stock F, Werschkun B (2008) Integrated testing and intelligent assessment – new challenges under REACH. Environ Sci Pollut Res Int 15:565–572
- EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttki R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne JL, Dumont AF, Hempen M, Martínez SV, Martino L, Smeraldi C, Terron A, Georgiadis N, Younes M (2017) Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J 15(8):e04971. <https://doi.org/10.2903/j.efsa.2017.4971>
- Rovida C, Alépée N, Api AM, Baskett DA, Bois FY, Caloni F, Corsini E, Daneshian M, Eskes C, Ezendam J, Fuchs H, Hayden P, Hegele-Hartung C, Hoffmann S, Hubesch B, Jacobs MN, Jaworska J, Kleensang A, Kleinstreuer N, Lalko J, Landsiedel R, Lebreux F, Luechtefeld T, Locatelli M, Mehling A, Natsch A, Pitchford JW, Prater D, Prieto P, Schepky A, Schüürmann G, Smirnova L, Toole C, van Vliet E, Weisensee D, Hartung T (2015) Integrated Testing Strategies (ITS) for safety assessment. ALTEX 32(1):25–40

⁹GDMP: Glycol di(3-mercaptopropionate), CAS 22504-50-3

8. Gabbert S, Benighaus C (2012) Quo vadis integrated testing strategies? Experiences and observations from the work floor. *J Risk Res* 15 (6):583–599. <https://doi.org/10.1080/13669877.2011.646291>
9. Jaworska JS, Gabbert S, Aldenberg T (2010) Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. *Regul Toxicol Pharmacol* 57:157–167
10. Lombardo A, Roncaglioni A, Benfenati E, Nendza M, Segner H, Jeram S, Paunée E, Schüürmann G (2014) Optimizing the aquatic toxicity assessment under REACH through an integrated testing strategy (ITS). *Environ Res* 135:156–164
11. ECHA (2017) Guidance on information requirements and chemical safety assessment Chapter R.7b: Endpoint specific guidance. ECHA, Helsinki
12. Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? *QSAR Combi Sci* 27:1337–1345
13. Ahlers J, Nendza M, Schwartz D (2019) Environmental hazard and risk assessment of thiochemicals. Application of integrated testing and intelligent assessment strategies (ITS) to fulfil the REACH requirements for aquatic toxicity. *Chemosphere* 214:480–490. <https://doi.org/10.1016/j.chemosphere.2018.09.082>
14. Rücker C, Mahmoud WMM, Schwartz D, Kümmeler K (2018) Biodegradation tests of mercaptocarboxylic acids, their esters, related divalent sulfur compounds and mercaptans. *Environ Sci Pollut Res* 25:18393–18411. <https://doi.org/10.1007/s11356-018-1812-x>
15. Klimisch HJ, Andreae E, Tillmann U (1997) A systematic approach for evaluating the quality of experimental and ecotoxicological data. *Regul Toxicol Pharmacol* 25:1–5
16. Nendza M, Müller M, Wenzel A (2014) Discriminating toxicant classes by mode of action: 4. Baseline and excess toxicity. *SAR QSAR Environ Res* 25(5):393–405. <https://doi.org/10.1080/1062936X.2014.907205>
17. Nendza M, Wenzel A (2006) Discriminating toxicant classes by mode of action: 1. (Eco) toxicity profiles. *Environ Sci Pollut Res Int* 13:192–203
18. Nendza M, Müller M (2000) Discriminating toxicant classes by mode of action: 2. Physico-chemical descriptors. *Quant Struct-Act Relat* 19:581–598
19. Nendza M, Müller M, Wenzel A (2017) Classification of baseline toxicants for QSAR predictions to replace fish acute toxicity studies. *Environ Sci: Processes Impacts* 19 (3):429–437. <https://doi.org/10.1039/C6EM00600K>
20. Shigeoka T, Sato Y, Takeda Y, Yoshida K, Yamauchi F (1988) Acute toxicity of chlorophenols to green algae, *Selenastrum capricornutum* and *Chlorella vulgaris*, and quantitative structure-activity relationships. *Environ Toxicol Chem* 7:847–854
21. Hermens JLM, Canton H, Janssen P, de Jong R (1984) Quantitative structure-activity relationships and toxicity studies of mixtures of chemicals with anaesthetic potency: acute lethal and sublethal toxicity to *Daphnia magna*. *Aquat Toxicol* 5:143–154
22. Deneer JW, van Leeuwen CJ, Maas-Diepeveen JL, Hermes JLM (1989) QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. *Aquat Toxicol* 15:83–98
23. Könemann H (1981) Quantitative structure-activity relationships in fish toxicity studies. Part I: relationship for 50 industrial pollutants. *Toxicology* 19:209–221
24. Nendza M, Russom CL (1991) QSAR modeling of the ERL-D fathead minnow acute toxicity database. *Xenobiotica* 21:147–170
25. Royal Society of Chemistry (2018) ChemSpider. <http://www.chemspider.com>
26. US EPA (2012) KOWWIN v1.68 from EPI-Suite, Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, D.C. <https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>
27. US EPA (2012) T.E.S.T. v4.1 <http://www.epa.gov/nrmrl/std/qsar/qsar.html>
28. ChemProp (2019) <http://www.ufz.de/ecochem/chemprop>
29. Nendza M, Gabbert S, Kühne R, Lombardo A, Roncaglioni A, Benfenati E, Benigni R, Bossa C, Stremmel S, Scheringer M, Fernández A, Rallo R, Giralt F, Dimitrov S, Mekyan O, Bringezu F, Schüürmann G (2013) A comparative survey of chemistry-driven in silico methods to identify hazardous substances under REACH. *Regul Toxicol Pharmacol* 66(3):301–314
30. OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models, vol OECD Environment Health and Safety Publications, Series on testing and assessment, vol 69. OECD, Paris
31. JRC QSAR Model Database. <https://qsardb.jrc.ec.europa.eu/qmrf/>

32. Ruusmann V, Sild S, Maran U (2015) QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models. *J Cheminform* 7(1):32. <https://doi.org/10.1186/s13321-015-0082-6>
33. VEGAHub. <https://www.vegahub.eu/>
34. Chemistry Dashboard. <https://comptox.epa.gov/dashboard/about>
35. Registered substances (2019) <https://echa.europa.eu/information-on-chemicals/registered-substances>
36. QSAR Toolbox version 4.3 (2019) <http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm>
37. ECOTOX Knowledgebase (2019) <https://cfpub.epa.gov/ecotox/>
38. eChemPortal (2019) <https://www.echemportal.org/echemportal/index.action>
39. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz GY, Niemela J, Mekyan OG (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. *J Chem Inf Model* 45:839–849
40. Kühne R, Ebert RU, Schüürmann G (2009) Chemical domain of QSAR models from atom-centered fragments. *J Chem Inf Model* 49:2660–2669
41. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith GD, Yang C (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA* 33:155–173
42. Jaworska JS, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *ATLA* 33:445–459
43. Gabbert S, Weikard HP (2013) Sequential testing of chemicals when costs matter: a value-of-information approach. *Human Ecol Risk Assess* 19(4):1067–1088
44. OECD (2019) Guidance document on aqueous-phase aquatic toxicity testing of difficult test substances, Series on testing and assessment. OECD Publishing, Paris
45. Ahlers J, Riedhammer C, Vogliano M, Ebert R-U, Kühne R, Schüürmann G (2006) Acute to chronic ratios in aquatic toxicity—variation across trophic levels and relationship with chemical structure. *Environ Toxicol Chem* 25 (11):2937–2945. <https://doi.org/10.1897/05-701R.1>
46. Nendza M, Wenzel A, Müller M, Lewin G, Simetska N, Stock F, Arning J (2016) Screening for potential endocrine disruptors in fish: evidence from structural alerts and in vitro and in vivo toxicological assays. *Environ Sci Eur* 28 (1):26. <https://doi.org/10.1186/s12302-016-0094-5>



Chapter 23

In Silico Ecotoxicological Modeling of Pesticide Metabolites and Mixtures

Chia Ming Chang, Chiung-Wen Chang, Fang-Wei Wu, Len Chang, and Tien-Cheng Liu

Abstract

Prior to registration, careful assessment of transformation products (TPs) that are more toxic than their parent compounds is required, and EU regulations require greater use of non-animal test methods and risk assessment strategies. Predicting the toxicity of transformation products and chemical mixtures is a major challenge for modern toxicology. Since the metabolic processes of transformation products and toxic effects of chemical mixtures involve complex mechanisms, it is essential to use in silico modeling methods to consider different chemico-biological interactions of metabolic transformation and mixture toxicity. This chapter reviews previous modeling methods used to study pesticide metabolites and mixtures.

Although various metabolites are emitted into the environment, there are few ways to interpret metabolites by predicting their ecotoxicological potential, so their formation and environmental fate are largely unknown. In vitro testing has limited coverage of metabolic processes present throughout the organism and may not always predict in vivo results. For systematically assessing the metabolic activation of persistent organic pollutants, researchers designed a comprehensive metabolic simulator to generate the metabolic profile of the POPs. In order to analyze and evaluate parent compounds and transformation products in the environment, data generation based on quantitative structure-activity relationship (QSAR) is becoming more and more important. Besides these, a process-based multimedia multi-species model allows us to quantitatively estimate the environmental exposure and fate of parent compounds and transformation products.

Pollutants in the environment usually appear in a joint form, and the biological effects of the mixture are different from the single separated components, so the risk assessment criteria for a single compound cannot accurately infer the actual complex environmental assessment. The interaction between the components of the mixture promotes significant changes in compositional characteristics and complications leading to synergistic or antagonistic effects. The covalent bonding, ionic bonding, van der Waals force, and hydrophilicity are important intermolecular forces that affect the interaction of chemical mixtures and are associated with four types of descriptors. This relationship has been able to study the reaction mechanisms of various environmental characteristics of organic pollutants.

Key words Pesticide, Transformation product, Chemical mixture, Ecotoxicity, Environmental fate, In silico modeling

1 Introduction: The Ecotoxicity of Transformation Products

Assessment of ecotoxicity data for parent compounds and transformation products (TP) based on chemical structure can be performed at an early stage of the risk assessment process to identify those chemical substances that require further testing [1, 2]. According to the European Plant Protection Products Regulation 1107/2009, the risk of pesticide metabolites to animals and the environment needs to be assessed prior to registration [3]. This regulation requires more use of non-animal testing methods and risk assessment strategies to minimize vertebrate testing. Since it is impossible to evaluate the ecotoxicity of each transformation product by experiment, Sinclair and Boxall used quantitative structure-activity relationships (QSARs), read-across methods, and expert systems to estimate the ecotoxicity of the transformation products based on chemical structure [4]. In addition, previous literature integrated prediction methods for biodegradation products, estimation of physicochemical properties and degradation half-life, persistence metrics, and joint persistence calculations to identify transformation products that makes a significant contribution to the joint persistence of the parent compounds [5].

Galassi et al. investigated the risks in detail based on the occurrence of metabolites of priority pesticides in surface water and groundwater in Italy and estimated their persistence based on field and ecotoxicity data [6]. An ecotoxicological endpoint, the 96-hour acute LC₅₀ for rainbow trout, was used as the appropriate database when developing QSAR [7]. Potentially persistent transformation products were known in the freshwater ecotoxicity studies of 15 pesticides and perchloroethylene. It is important to incorporate the potential effects of the transformation products into the characteristic factor (CF) calculations [8].

Escher and co-workers assumed two scenarios for the phytotoxicity endpoints of β -blocker mixtures and their associated human metabolites: metabolites lose their specific activity and act as baseline toxicants, and metabolites exhibit the same identity as the specific mode of action of their parent drug and used QSAR to simulate its total toxicity potential [9]. They employed toxic ratio (TR) to indicate whether a compound acts according to baseline toxicity or a specific toxic mode of action [10]. The predicted baseline effect concentration EC₅₀_{baseline,i} for a given compound i, the ratio of i to the experimentally determined EC₅₀_{experimental,i}, is the toxicity ratio TR. EC₅₀_{baseline,i} can be derived from QSAR for baseline toxicity in the corresponding test system. Baseline QSAR can be determined by 24-hour chlorophyll fluorescence. TR_i < 10 corresponds to baseline toxicity, and TR_i ≥ 10 indicates a specific mode of toxic effect [10]. Moreover, Escher and Fenner integrated the study into a framework that emphasized the data from the

parent compound to the read-across of the transformation product to determine the priority of the contribution of transformation products to overall environmental risk [11].

There are many possible explanations for transformation products that are more toxic than their parent compounds: The transformation product has the same toxicity mechanism as the parent; the transformation product is the active ingredient of the insecticide; the bioconcentration factor of the transformation product is greater than the parent; the product produced by the transformation pathway has a different and more effective mode of action than the parent compound [1].

Characterization factors (CFs) are used in product life cycle impact assessment (LCIA) to determine the impact of stressors on humans and ecosystems. When degradation products are more toxic, more durable, more mobile, or more bioaccumulative than their parent compounds, it is important to include the effect of these TPs on chemical characterization factors in LCIA. Through this work, the durability, mobility, and toxicity of the transformation product are solved by its parent compound. Uncertainty analysis can be used to quantify the uncertainty of the characterization factor, with and without transformation products [8, 12].

The environmental risk assessment of most human drugs is based on parent drugs. However, most drugs are widely metabolized by the body, and only a small fraction is released into the wastewater stream through non-metabolic forms. Although various metabolites are emitted into the environment, there are few ways to interpret metabolites by predicting their ecotoxic potential, and little is known about the ecotoxic potential of metabolite mixtures. Metabolism is generally considered to make the parent compound more hydrophilic and therefore less toxic [9, 13–15].

2 Integrated Software for Modeling Metabolites

The widely used ECOSAR software was developed by the US Environmental Protection Agency (EPA), which was listed as a useful non-test method by EFSA [16] and incorporated into the OECD QSAR Toolbox. The method has been used in regulatory authorities. Reuschenbach et al. evaluated the ECOSAR software for QSAR prediction of chemical toxicity of aquatic organisms. The ECOSAR predictions and experimentally derived toxicity data cover the acute effects of growth inhibition on fish, *Daphnia*, and algae [17]. To predict acute fish toxicity of pesticide metabolites, Burden et al. used ECOSAR software for prediction of 150 metabolites. The experimental fish LC₅₀ values were obtained from the Pesticide Properties Database (<http://sitem.herts.ac.uk/aeru/ppdb/en/atot.htm>) [2].

UM-PPS has been used for computer-aided prediction of microbial metabolites to detect multiple compounds in complicated environmental samples. This is an effective procedure for comprehensive screening of a large number of potential transformation products (TP) in environmental samples [18]. The advantage of using UM-PPS [19] is that its transformation rules are obtained from a collection of known microbial degradation pathways of approximately 1100 chemicals. (UM-BBD) [20]. A computational prediction of possible microbial TP was performed, predicting two generations of TP, allowing for aerobic and anaerobic transformation, demonstrating that UM-PPS is superior to other similar tools, such as META [21] or CATABOL [22]. Gutowski et al. used a different set of QSAR software to predict the physicochemical properties and toxicity of S-metolachlor (SM) and stable transformation products (TP). The software used includes CASE Ultra V.1.5.0.1 (MultiCASE Inc.) [23] and LeadsScope software V.3.2.3-1, as well as a training set for the 2012 SAR Genetox database provided by LeadsScope [24]. The SMILES code was used to input the TP structure of the molecule [25]. Juan José Villaverde et al. recently discovered that the clethodim photodegradation solution is more toxic to the *Vibrio fischeri* than the parent compound, and QSAR analysis can provide physicochemical properties, fate, and ecotoxicological endpoints of degradation products [26]. They used six QSAR modeling methods with T.E.S.T., namely, grading, FDA, single model, group contribution, nearest neighbor, and consensus methods, in order to have greater confidence in the predictions performed. Kern et al. developed a systematic and efficient method for screening a large number of potential TPs in environmental water samples to more fully understand the presence of TP in the environment. The study used a fairly new high-resolution mass spectrometry (HR-MS) analysis technique that overcomes the list of targets that lack analytical reference standards, including the most comprehensive potential TP possible. They use the University of Minnesota pathway prediction system (UM-PPS), a rule-based system for predicting microbial metabolites [18, 19].

Pesticides entering the water environment undergo different hydrolysis, oxidation, biodegradation, or photolysis pathways, which results in a higher pesticide TP concentration than the parent compound. In the water treatment of ozone, the presence of pesticide TP in drinking water may cause new problems [27–30]. TP is usually formed in complex matrices. Because of its low concentration, separation and purification are very difficult, so its formation and environmental fate are largely unknown. In order to analyze and evaluate TP in the environment, data generation based on quantitative structure-activity relationships (QSAR) is becoming more and more important [25, 31–33].

3 Multimedia Multi-species Models

The time range of environmental exposure and the relative concentration of the parent compound and its transformation products in surface waters can be obtained from a process-based multimedia multi-species model, enabling us to quantitatively estimate the environmental fate of the transformation product [34]. Kern et al. measured pesticides in a small river that discharged from the Switzerland agricultural watershed, using dynamic multimedia multi-species models for TP prioritization, and comparing predicted relative surface water exposure potential with experimental data [35]. Since the transformation products have different types of environmental fate models, Fenner et al. have introduced multimedia multi-species models that are generally applicable to chemical risk assessment and environmental resource quality assessment [36].

Because the half-life of different amide pesticides depends on the chemical class and experimental parameters, Latino et al. simultaneously encode the reaction and the corresponding half-life in Eawag-Soil. Consider the initial transformation reaction to establish a meaningful quantitative-structural biotransformation relationship (QSBR) [37]. In the Eawag-Soil package, metadata under experimental conditions (e.g., soil texture, soil moisture, pH, etc.) are stored in the physical scene. Pathway information is stored in a biotransformation reaction scheme in a physical pathway. Compounds and reactions involved in a given pathway are stored separately in the physical compound and reaction. The high proportion of toxic metabolites of biocide and the scarcity of data on these compounds suggest that further research into their effects in the aquatic compartment is needed. Europe plans LIFE-COMBASE to build computational tools to predict the acute toxicity of biocide actives and their environmental degradation products to aquatic organisms, including fish, invertebrates, algae, and sewage treatment plant (STP) microorganisms [38]. Lienert et al. evaluated each of the parent drugs and their metabolites as a mixture of similar compounds to assess the potential hazards of ecotoxicology, including metabolites formed in humans. In the absence of literature data for physicochemical properties or effects, baseline toxicity was estimated using a quantitative structure-activity relationship (QSAR), and each parent drug and its metabolites were treated as a mixture of similar compounds. Input data from the literature (lipophilic, baseline effect concentration (EC_{50}), estimates of excretion scores) were used to mimic the toxic potential of the parent drug and metabolite (TP mixture). In addition, the use of simple drug concentration predictions in Swiss wastewater produced a risk quotient (RQ mixture) [39].

Biotransformation of microbial communities in environmental systems is a very effective mechanism to reduce the persistence of their environment, but it can also lead to the formation of potentially dangerous transformation products [11, 40, 41]. Pathway prediction systems typically rely on a dictionary of biotransformation rules that recognize complex functional groups and convert them into product substructures [42–45]. These biotransformation rules are intended to reflect known microbial transformation pathways of chemical contaminants. They are based primarily on data collected from the Eawag biodegradation/biocatalytic database (Eawag-BBD), formerly known as the University of Minnesota biodegradation/biocatalysis database (UM-BBD) [46]. However, most of the data for Eawag-BBD comes from studies of pure microbial cultures or laboratory cultures with extended adaptation periods.

For microbial communities in different environments, different enzyme-catalyzed reactions may occur at very different rates. This suggests that not only chemical structures but also specific environmental conditions can be considered, which can greatly improve biotransformation prediction. Latino et al. introduced enviPath to the new database and path prediction system. enviPath provides a database environment to facilitate annotation of biotransformation half-life and pathway information and allows half-life and pathway information to be supplemented by metadata for environmental and/or experimental conditions of different agricultural soils, aquatic sediments, and activated sludge, simulated transformation pathway [37, 47, 48].

Lienert et al. proposed a screening tool for identifying drugs with high environmental risks, including their human metabolites. The tool uses drug data on human metabolism and excretion, drug sales data, and physicochemical properties of parent drugs and metabolites. For many cases where ecotoxicological data are lacking, QSAR and lipophilicity can be used to estimate baseline toxicity [39].

4 The Metabolic Pathway Software System MetaPath

To systematically assess the metabolic activation of parent POP chemicals and metabolites for hazard identification, Mekenyan et al. also designed an integrated metabolic simulator to generate a metabolic profile for parent POP chemicals [49]. The toxic mode of action (MOA) of the transformed product does not necessarily exhibit the same MOA as the parent compound, and the correct assignment of MOA is a weakness of any application of the QSAR method [50]. MetaTox software can be used to predict metabolites, which are formed by nine types of reactions (aliphatic and aromatic hydroxylation, N- and O-glucuronidation, N-, S- and C-oxidation,

and N- and O-de Alkylation). The probability calculation for generating metabolites is based on the analysis of the “structure-bio-transformation reaction” and “structural modified atom” relationships using the Bayesian method. The parent compound and each of the produced metabolites can then be used to predict the value of acute rat toxicity (LD_{50}) for intravenous administration using the QSAR model [51]. In the study, GUSAR uses self-consistent regression based on regularized least-squares method. Quantitative neighborhoods of atoms and multilevel neighborhoods of atoms descriptors are used to create the QSAR models. GUSAR uses three methods (similarity, leverage, and accuracy assessment) to estimate the applicability domain (AD) of the QSAR model during the prediction of acute toxicity [52–55]. Using the metabolic pathway software system MetaPath, the differences in analytical methods for metabolites in each study and the relative amounts of quantified metabolites can be identified to compare metabolic maps between rat, goat, and fish (bluebird or rainbow trout) species [56].

The assessment of metabolite structure in the early stages of drug development can be performed in two different ways, predicting metabolic sites and without preliminary prediction of metabolic sites. The first approach focuses on different enzymes, and an understanding of the mechanism of action of the enzyme as a result of SOM prediction provides an assumption about the possible structure of the metabolite. The second approach is primarily to implement predictive metabolite structures in expert systems that use biotransformation dictionaries to predict metabolites. Such a system contains rules for converting a parent compound to its metabolite [51, 57].

5 Prediction Models for the Human Toxicity of Transformation Products

Madden et al. used a series of computational tools to address the major challenge of predicting toxicity after skin exposure, notifying the prediction of skin metabolism by understanding the differences in the enzymatic landscapes between skin and liver [58]. In describing the role of pesticides in breast milk, Agatonovic-Kustrin et al. used sensitivity analysis to select descriptors and applied artificial neural network modeling to correlate selected descriptors (inputs) with M/P ratios (outputs) to develop predictive QSARs [59]. Xiao et al. used QSAR to mimic the binding affinity of the metabolite 3,4,3',4'-tetrachloroazobenzene (TCAB) to certain human receptors. TCAB was found to have strong binding affinity for AhR in EROD and micro EROD induction assays [60]. VirtualToxLab predicts the toxic potential of chemicals by simulating and quantifying their interaction with a range of proteins, using automated multidimensional QSAR analysis to know that these proteins can

cause adverse effects [61–63]. The model of VirtualToxLab comes from the free open source www.Biograf.ch. Pinto et al. trained 18 high-throughput ER assays on approximately 1600 ToxCast chemicals, covering different chemical structure categories, including known reference ER ligands and various chemicals with known estrogen-like activities. They used three QSAR models to predict the ER agonist bioactivity of parent compounds and their metabolites: the ER agonist model is available through the Online Chemical Database with Modeling Environment (OCHEM) Web platform at (<https://ochem.eu/home/show.do>), an ER agonist model developed by Lockheed Martin (LM), and the developed ER model is available at (<http://infochim.u-strasbg.fr/webserv/VSEngine.html>) [64, 65]. These ER QSARs estimate that most known estrogen metabolites have stronger estrogenic activity than their parent compounds [66]. Dekant et al. proposed a stratified test strategy that does not or only partially retain the targeted toxicity of active ingredients (AI) when the drinking water concentration is >3.0 µg/L for the degradation product, the “non-relevant metabolites” of AI. A detailed toxicity database for parent AI and conclusions based on structure-activity relationships should be included [67]. Further assessment of the relevance of metabolites should include the collection of all available information about “non-relevant metabolites,” the nature of the parent AI, and the toxicological information of the structure-related compounds. This information is integrated into hazard assessment methods by QSAR analysis to predict biotransformation into potentially toxic metabolites and as a “smart” and the foundation of targeted approach to design toxicity testing [67, 68]. However, for reproductive toxicity, only limited structural alert information (with respect to certain receptors and/or enzyme binding/inhibition properties) makes most QSAR models unreliable. Clark reviewed how computer models predict bacterial mutagenicity in humans and rats, human cytochrome P450 (CYP) metabolism, and bioavailability [69].

Environmental chemicals induce adverse reactions by binding to receptors and are thought to demonstrate valid QSARs for receptors such as AhR, ER, and AR [60, 70]. Using QSAR, VirtualToxLab was applied with a series of 16 proteins to mimic the binding affinity of 3,4,3',4'-tetrachloroazobenzene (TCAB) to human receptors (www.biograf.ch) [60, 61].

In vitro testing has a limited coverage of metabolic processes present throughout the organism and may not always predict in vivo results, as these chemicals may be false negatives when tested in assays without metabolic activity. Previous studies have developed QSAR modeling methods for predicting chemical metabolites and estimating chemical interactions with ER. This computer simulation analysis has proven to be a fast and inexpensive method for detecting environmental chemicals with estrogen metabolites,

thereby reducing the potential for false negative results in HTS analysis [66, 71–74].

Hydrolytic degradation products from the parent active ingredient are not toxic or highly toxic (T , T^+), carcinogenic, mutagenic or reproductive toxicity, significantly reduced or inactive against pest activity, known as “non-relevant metabolites.” A toxicological-based risk assessment of the presence of “non-relevant metabolites” is required without the use of large numbers of animals. Toxicity testing is only required if the animal exceeds the threshold caused by thresholds of toxicological concern TTC and cannot assess the hazard based on other information [67].

6 Quantitative Structure-Activity Relationship Models for Transformation Products

The uncertain results of the training set indicate that there is an inherent weakness in the molecular connectivity theory in the complex reaction of OP insecticides [75]. Ortiz-Hernández et al. proposed a mechanism for the hydrolysis of pesticides by *Flavobacterium sp.* at the bond between the phosphorus and the heteroatom to produce phosphoric acid and three metabolites [76]. The second-order rate constant k value for oxidative transformation of various emerging organic micro-pollutants can be predicted using the QSAR and group contribution methods developed by Lee and von Gunten [77]. In this QSAR analysis, the descriptors Hammett σ (σ , σ^+ and σ^-) and Taft σ^* constants of the most common substituents in physical organic chemistry are utilized, with a view to the relative convenience and application [77, 78].

Due to widespread use, it is not uncommon for humans to be poisoned by pesticides. Once organophosphorus pesticides enter the body, they are metabolized by cytochrome P-450, producing toxic metabolites that react with acetylcholinesterase. Previous literature proposed a new biooxidation mechanism of organophosphorus pesticides. Under this mechanism, any drug or procedure that reacts with the phosphorus atom of the pesticide may help prevent the gradual progression of pesticide metabolism and toxicity [76, 79].

Oxidation processes are widely used in water treatment for disinfection and oxidation purposes. QSAR (quantitative structure-activity relationship) can be used to predict the reaction k value of various oxidants and organic compounds to predict the conversion efficiency of micro-contaminants during oxidized water treatment. Especially when considering the huge number and large structural diversity of synthetic organic compounds commonly found in various water resources, QSAR-based prediction methods are very useful as screening tools [77].

Lo Piparo et al. used rule-based toxicity predictions to compare the empirical and theoretical results of three allelochemicals

(DIMBOA, BOA, and MBOA) with their metabolites and found that only degraded metabolites showed significant ecotoxic effects. The generated QSAR model showed good internal prediction ability ($R_{cv}^2 > 0.6$) [80]. They describe the microenvironment around the molecule based on the comparative molecular field analysis (CoMFA) of Chem-X software. This technique measures the spatial and electrostatic interaction energy between small probes at a series of regular grid locations around a molecule, studying the magnitude and direction of interactions between electrons and three-dimensional space [80, 81]. QSAR modeling of basic properties supports the hypothesis that halogenated substituents (meta-Br; meta-I; ortho-Cl) may hinder the degradation of three amides by *Variovorax* sp. [82] However, it is not possible to infer from the simple chemical reaction in solution that the cause of degradation may be insufficient. To elaborate on this aspect, the electrostatic potential model of the molecule was prepared in previous studies. It has been found that the hindrance of enzymatic degradation may be related to properties such as relative polarity and spatial properties in the molecular region away from the location where the actual degradation occurs. When experimental toxicity data is difficult to determine or not available at all, computer simulation methods are based on theoretical knowledge gained in different scientific fields, supplemented by the powerful computing power of modern computers to derive models for predicting chemical properties [80].

In order to fully characterize the fate of bromoxynil and iodobenzonitrile in soil and groundwater environments, it is necessary to study the mobility and persistence of degradation products. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) has become a widely used method for the analysis of polar organic compounds in a variety of matrices, which is ideal for the analysis of iodobenzonitrile, bromoxynil, and possible transformation products [82–84].

7 Quantum-Chemical Modeling Methods for Transformation Products

Lewis proposed a broader concept of acid and base. An acid is defined as a molecule, ion, or group of atoms that can accept foreign electrons, also known as an electron acceptor. Typically, Lewis acids are systems with unoccupied molecular orbitals. A base is defined as a molecule, ion, or group of atoms that can provide electrons, also known as an electron donor. The nature of the acid-base reaction is that the electrons of the base enter the unoccupied molecular orbitals of the acid. It can also be the molecular orbital of an unoccupied acid that accepts the electrons of the base to form a coordination bond.

It can be expressed as a general formula: $A +: B \rightarrow A: B$ wherein A represents an acid and B represents a base. An important feature

of the Lewis acid-base theory is that many aprotic compounds are included in the acid range. The electrophile in the organic reaction can be considered as a Lewis acid which is easy to add a nucleophile. The nucleophile is a Lewis base.

According to the frontier molecular orbital (FMO) theory, the formation of chemical bonds is mainly determined by the interaction of the frontier molecular orbitals. The frontier molecular orbital of the nucleophile is the highest occupied molecular orbital (HOMO), the electrophilic frontier molecular orbital is the lowest unoccupied molecular orbital (LUMO), and the acid-base reaction is the interaction between HOMO and LUMO.

The use of standard quantum chemical methods allows for a more detailed study of pesticides and their metabolites. Villaverde et al. explored the potential of quantum chemistry in the toxicity and environmental behavioral simulation of pesticides and their by-products, including certain chemical reaction mechanisms and their degradation pathways [85]. Duirk et al. used the QSAR model to predict the rate of hydrolysis rate under drinking water treatment conditions (in the presence of chlorine-containing water) to determine the exposure risk of conversion products of OP pesticides in drinking water [86]. Frontier molecular orbital theory has been used to correlate oxidation rate coefficients with the highest occupied molecular orbital energy (E_{HOMO}) [87]. E_{HOMO} is a good molecular descriptor describing the oxidation of OP pesticides in each subgroup, so the subgroup differences are quickly understood. The phosphorothioate subgroup mainly contains ethyl and phenyl esters. Due to the sulfur linkage and the methyl ester on the tetrahedral phosphorus atom, the detected phosphorodithioate is more susceptible to chlorine oxidation than the thiosulfate subgroup, and the phosphorodithioate subgroup has a specificity than the phosphorothioate subgroup. TP is more difficult to remove by PAC adsorption and ozonation than its parent insecticide. Compounds with relatively high energy level E_{HOMO} can more easily transfer electrons to the lowest unoccupied molecular orbital of ozone, and thus the energy level of E_{HOMO} is positively correlated with removal by ozonation [88]. We can use density functional theory (DFT) calculations to investigate the most stable conformer of aloxydim herbicides, the factors controlling its stability, and the mechanism of mutual transformation between the most relevant conformers [89]. Density functional theory calculations can be used as a preliminary strategy for estimating pesticide degradation pathways and by-product formation. There are four structural features with strong intramolecular hydrogen bonds. The stability of the resulting fragment is controlled by the presence of intramolecular hydrogen bonds. The high stability is compared to other possible forms in the gas phase. The most stable species identified may play an important role in the environment as a by-product of

long-term existence throughout the degradation mechanism of the herbicide.

Degradation products (DP) can exhibit very different physicochemical and toxicological properties from its parent compound, and their behavior in the environment can vary widely. Sinclair and Boxall used different computational methods to predict the aquatic acute ecotoxicity of fish, daphnids, and algae from 485 DPs of 60 pesticides. This study shows that 30% DP is more toxic than its parent compound. The transformation reaction usually results in a smaller, more polar, and thus less hydrophobic molecule. Due to the increase in water solubility, DP can be easily transported to environmental water, and the presence of pesticide DP in the aqueous medium causes deterioration of water quality [1, 26, 90, 91].

DP from aloxydim is more toxic than the parent active compound. However, current experimental results usually provide only a partial and very limited overview of the problem, so intermediates are difficult to characterize, and the mechanism of degradation has not been clearly defined. A proper understanding of aloxydim DP is essential to prevent adverse effects from improper use of pesticides. In this regard, computational studies of pesticides offer great potential for identifying the most relevant stable DP and its physicochemical properties. Villaverde et al. explored the degradation process of aloxydim by DFT calculations. The main purpose is not only to identify the structure and properties of the parent compound but also to identify those DPs formed after the most unstable N-O bond cleavage and loss of oxime ether groups. These computational simulations have minimized the animal testing performed on pesticide toxicology risk assessments, overcoming the challenges of modern legislation [89, 92].

All potential transformation pathways need to be addressed when assessing potential pesticide exposures caused by drinking water. Hydrolysis and chemical oxidation are the most relevant pathways for organophosphorus (OP) pesticides under drinking water treatment conditions. Some studies have shown that the chlorine reactivity of different types of pesticides may vary greatly due to changes in chemical structure [86, 93].

Neuwoehner et al. propose a method to promote and systematically assess the ecotoxicological risk of transformation products. To gain a complete picture of how the transformation products work, they used a QSAR and toxicity ratio (TR) analysis to perform an action pattern analysis of the experimental data. They evaluated the toxicity and mode of action relative to the parent compound and used the mixture toxicity test as a diagnostic tool to support the pattern analysis. The ultimate goal is to clarify whether the transformation product has a similar potential risk to the parent compound [94].

8 Mixture Toxicity of Pesticides

With the advancement of science and technology, human beings produce more and more kinds of chemical substances, so the environment is filled with many different pollutants. When organisms are exposed to a mixture of chemicals for a long time, their health is bound to be seriously threatened. In general, the biological effects and toxicity of a mixture differ from a single isolated component, so the risk assessment criteria for a single compound cannot accurately infer actual complex environmental assessments. Therefore, in order to evaluate the toxic effects of mixed forms of compounds on organisms, it is necessary to further study and establish a method for effectively evaluating and predicting the toxicity of mixtures to promote rapid health judgment and implementation of laws and regulations for disaster prevention.

In agriculture, due to the increase in the world's population, in order to solve the problem of insufficient food, a large number of pesticides and fertilizers are used to rapidly increase production. In addition, farmers often mix a variety of pesticides to save time and achieve better results, health hazards remain unclear. Therefore, the joint toxicity study of pesticide mixtures is very necessary.

The octanol-water partition coefficient (K_{ow}) is an effective parameter and is commonly used to assess the toxicity of a single organic chemical without observational data. The partition coefficient of the mixture can be used to develop a QSAR model to predict mixture toxicity. However, only K_{ow} of a single type of chemical can be measured by UV spectrophotometer or HPLC, and it is difficult to obtain a mixed-type K_{ow} . Until 1995, Verhaar et al. used the C₁₈-containing Empore™ disk/water to investigate the bioconcentration factor (BCF) of single and mixed-type chemicals, achieving very high correlation between BCF and C₁₈-containing Empore™ disk/water partition coefficients [95].

The acute toxicity (EC₅₀) of 36 substituted aromatic compounds to *Vibrio fischeri* was predicted using a QSAR model constructed from an octanol/water partition coefficient. The model used in the report of Verhaar et al. [95] was extended and used to calculate the octanol/water partition coefficient of the chemical mixture. The QSAR model is verified to be robust enough by the leave-one-out method. Furthermore, by classifying these chemicals as polar and nonpolar, the toxicity of the chemical mixture can be more accurately predicted from the partition coefficient.

9 Concentration Addition (CA) and Independent Action (IA) Modeling

A review of QSAR studies for the toxicity of mixtures in the QSAR method was published by Altenburger et al. [96]. Pollutants in the

environment usually appear in mixed form, and a single compound cannot predict the chemical action of the mixture. The interaction between the components of the mixture promotes significant changes in compositional characteristics and complications leading to synergistic or antagonistic effects, as opposed to the ideal reference hypothesis, where an additive addition refers to concentration addition (concentration addition, CA) and independent action (IA). It is a well-known reference model for assessing joint activity, and its mechanism of action can be confirmed by pharmacology. Most studies have shown that the mixing of the compounds uses only one anesthetic or a specific mode of action, and it is assumed that CA can satisfactorily simulate this mixing. However, the interaction of different reactive compounds tends to produce a combined effect that is less than CA. The molecular description parameters calculated from the composition of the mixture can be used as the characteristics of the mixture, and the toxicity of the anesthetic mixture can be predicted from the molecular description parameters.

A mixed toxicity prediction test was carried out on *Q67* luminescent bacteria against six organophosphorus insecticides. Organophosphorus pesticides are present as a mixture of surface waters. To determine the toxicity of the mixture in a multicomponent space, a uniform design (UD) was used to design the mixture, as changes in concentration can be studied from a small number of experimental results. The two mixed toxicity prediction models used in the study were concentration addition (CA) and independent action (IA). The results showed no specific differences observed between all CA-predicted and mixture toxicity. However, the toxicity of the IA-predicted mixture is also very good, especially in the low concentration fraction [97].

The basic idea of the concept of concentration addition (CA) is that if chemicals with the same toxicity mechanism are mixed at the same ratio, they can be considered to be the same chemical substance of the same biological target. The independent action (IA) concept is based on the idea that the components in the mixture assume their behavior different and the toxicity of each component is not affected by the toxicity of other compounds [98]. Neale et al. used experimental EC₅₀ values for various chemicals, only nonantibiotics, only antibiotics, and all chemicals to prepare equivalent mixtures for 0.5 and 16 h. All ingredients contribute the same to the effect of the mixture in the equivalent mixture, listing the proportion of each chemical contained in the mixture. The experimental results were compared with CA, IA, and TSP mixture toxicity predictions [99]. Since it is not clear whether chloroacetanilide has the same mode of action, Junghans et al. elucidated the combined effects of various chloroacetanilide herbicides with CA and IA, trying to understand whether it can be

predicted by understanding the concentration-response relationship of a single substance [100].

Deneer evaluated the usefulness of the concept of CA in terms of the combined action of pesticides on aquatic organisms [101]. When the concentration of the components of the mixture is lower than their respective NOEC values, the concept of CA provides a highly accurate prediction of the toxicity of the s-triazine mixture and is uncorrelated to the level of effect considered and the concentration of the components of the mixture. IA-based predictions tend to underestimate the overall toxicity of the s-triazine mixture [102].

In order to predict the toxicity of multicomponent mixtures with the highest possible accuracy and to give reliable statistical estimates of the low toxic effects of the individual mixture components, CA is clearly not a universal solution. In the case where the components are known to specifically interact with different molecular target sites, IA has proven to be superior [103]. The algal joint toxicity of the phenylurea mixture can be predicted by CA. However, the concept of IA has proven to be equally effective, both of which predict almost the same mixture toxicity [104].

The actual exposure scenarios in the field runoff water were studied. The 25 pesticide mixtures showed good CA predictability for the reproduction of freshwater algae *Scenedesmus vacuolatus*. However, IA slightly underestimated the toxicity of the actual mixture. The EC₅₀ values for each prediction are only 1.3 times different. In the so-called toxic units (TU), only a few components dominate the mixed scenario [105]. The toxicity unit (TU) method was used to test the synergistic relationship between atrazine and various organophosphorus pesticides. The response model was not always able to accurately predict the mixed toxicity of pesticides with different modes of action [106].

10 Modeling Deviation

If the joint effects of chemicals in simple or complex mixtures are inferred to deviate from CA and IA (or both), the presence of chemical A alters the toxicity of compound B in the mixture, which is the conceptual framework of chemical interactions. In order to provide a basis to support and explain the underlying mechanisms that lead to chemical interactions that affect the toxicity of mixtures, Spurgeon et al. proposed a biology-based framework. The framework combines (1) external exposure, including speciation, binding, and transport; (2) toxicokinetics, including absorption, distribution, metabolism, and excretion; and (3) toxicity kinetics, including mutual interaction with receptor site effect. In the case of mixtures, interactions can be classified as related to processes caused by the above reasons. Once the nature and type of

potential interactions that may occur are determined, it is easier to design the experimental method for studying the mechanism that can lead to interactions [107]. To assess the deviation potential of experimental toxicity predictions, the accuracy of the prediction method was quantified by applying model deviation ratio (MDR) to the CA and IA models. MDR is the ratio between observed and predicted mixture toxicity. If the experimental value falls within half or twice the predicted value ($0.5 \leq \text{MDR} \leq 2$), it is assumed that the prediction method is met [98]. If the observed toxicity value of the mixture falls within the 95% confidence interval for the expected value of the CA or IA model, then the mixture is considered to be in accordance with the model. If the observed toxicity value of the mixture exceeds the 95% confidence interval for the expected value, the mixture may not conform to the additive model (CA or IA model). However, models that are classified as antagonistic or synergistic must be avoided due to very small biologically insignificant biases, and therefore the expected and observed toxicity values are required to differ by at least 30%. The model deviation ratio (MDR) method was used to quantitatively estimate the difference between predicted toxicity and measured toxicity. For the CA model, MDR was derived by dividing the predicted toxicity value (IC_{25}) by the observed toxicity value. For the IA model, MDR is obtained by dividing the observed effect by the predicted effect, since concentration and toxicity are inversely proportional to the response. For both models, an MDR value greater than 1.3 means that the toxicity of the mixture is synergistic, while a value less than 0.7 is consistent with antagonism [108]. Di Nica et al. performed a concentration assessment (CA) and a predictive assessment of the independent effect (IA) model for the concentration-response curves of different binary and multicomponent mixtures of QAC. The consistency between the experimental and predicted IC_x was observed and confirmed by applying the model deviation ratio (MDR) [98].

The measure of deviation from CA is the corrected toxicity enhancement index cTEI, also known as the predictive quality index or the relative model deviation ratio or effect residual rate [109–111]. The ratio between CA prediction ($\text{EC}_{50,\text{CA}}$) and experimental EC_{50} is 2; the mixture produces a cTEI of -1 (if CA is more effective than the experiment) and +1 (if the effectiveness of CA is lower than the experiment) [112].

11 Computational Approach to the Toxicity Assessment

The ecological effects caused by pesticide mixtures are rarely considered in the regulatory process. However, precedents for mixtures related to human health during the pesticide registration process can be followed [113]. Measurement of the effects of exposure to

sublethal concentrations of organophosphate diazido, malathion, chlorpyrifos, and brain acetylcholinesterase inhibitors such as carbaryl and carbofuran in juvenile coho salmon (*Oncorhynchus kisutch*) were reported. It was assessed whether the chemicals in the mixture act alone (resulting in additive AChE inhibition) or whether the components interact to produce antagonistic or synergistic toxicity [114]. LeBlanc and Wang used the data for the website [Computational Approach to the Toxicity Assessment of Mixtures (CATAM)] to analyze response additivity. It is assumed that the observed effect is real, but it is not statistically significant due to the low power of the experimental design [115]. Feron et al. developed a hazard identification and risk assessment scheme for complex mixtures and a consistent method for generating total volatile organic compound values for indoor air. They used toxic equivalent factors or alternative methods, as well as quantitative structure-activity relationship analysis combined with lumping analysis and physiological-based pharmacokinetic/pharmacodynamic models to study complex mixtures [116]. Chèvre et al. proposed a method for defining the risk quotient of a herbicide mixture having a similar mode of action (RQ_m). This method has the advantage of being easy to calculate and communicate and is proposed as a substitute for the current limit of Switzerland herbicides of 0.1 µg/L. From the concentration addition model, RQ_m can be expressed as the sum of the measured environmental concentration and the WQC ratio of each herbicide. RQ_m should be less than 1 to ensure acceptable risks to aquatic organisms [117]. Competitive inhibition is often a concern at concentrations above ambient exposure levels and is the most common type of interaction in various types of mixtures. PBK modeling can play a central role in predicting interactions in chemical mixture risk assessment [118]. Boberg et al. suggest that chemicals for mixed risk assessment should be grouped before we can better understand the path of adverse outcomes. Grouping methods can be based on integrated in vivo and in vitro data, read-across, and computational methods such as QSAR models or integrated systems biology [119].

12 Quantitative Structure-Activity Relationship (QSAR) Modeling

Baseline toxicity is the minimum toxicity caused by each compound, that is, chemicals are inserted into the biological membrane, disrupting structure and function [120]. The hydrophobic descriptor commonly selected for baseline toxicity QSAR is the octanol-water partition coefficient K_{ow} , but the liposome-water partition coefficient K_{lipw} has been shown to be a better descriptor because it allows the development of common QSAR for polar and nonpolar baseline toxicants [121]. Since some of the chemicals

studied were acids or bases added at pH 7, K_{lipw} was replaced with a liposome-water partitioning ratio of pH 7, D_{lipw} (pH 7) when QSAR was applied [94]. Escher et al. developed a new QSAR covering a large chemical space in which the iono-corrected liposome-water partition ratio was used as the only descriptor for the chemical-independent quantitative structure-activity model of the Microtox assay. This baseline toxicity QSAR clearly includes major water pollutants and ionizable chemicals. This baseline toxicity QSAR can be used as a diagnostic tool to identify specific active chemicals to obtain baseline toxicity equivalents for environmental samples with unknown mixture compositions and to predict mixture effects for mixtures of known composition [120]. Ghafourian et al. compared the chemical space of their dataset with the skin permeability datasets in previous literatures. Stepwise regression analysis was used to develop the model. The predictability of the model has been tested by a leave-many-out procedure. The predicted mean absolute error (MAE) was calculated as a measure of model accuracy [122]. Toropova et al. used Monte Carlo techniques to calculate the best descriptor based on SMILES. The univariate correlation between the optimal descriptor and toxicity of the binary mixture was analyzed to develop a predictive model with satisfactory statistical quality [123].

Based on the comprehensive toxicity test results of benzene and its derivatives to *Vibrio fischeri* by Lin and co-workers [124–126], a QSAR model consisting of quantum-chemical parameters was established to predict mixture toxicity [127]. The logarithm of nuclear repulsion energy (log Enr) and HOMO-LUMO energy difference (GAP_{h-l}) were the significant descriptors. Furthermore, the molar volume difference parameter ($GAPV_m$) of the mixture can increase the correlation between the structure and mixture toxicity [127]. Chen et al. used molecular simulation techniques to identify mode of inhibition. The pesticides have the same binding sites at the bottom of the luciferin pocket, and the combined toxicity can be predicted by the concentration increase model. In addition, there is a linear relationship between the binding free energy of the mixture (ΔG_{mix}) and the median effective concentration of the mixture (EC_{50}) [128]. According to the information from protein-chemical and protein-protein interaction networks, Kim and co-workers trained machine learning models to classify chemical mixtures and proposed a new method to predict synergistic toxicity of binary mixtures against *Vibrio fischeri* [129].

In the study of Qin et al., a predictive QSAR model was developed to predict the additive and nonadditive toxicity of binary and multicomponent mixtures. The simple and accurate GA-MLR model was effectively presented. Internal and external validation was used to assess the predictive power of the QSAR model, which has predictive power for high additive and nonadditive effects on mixture toxicity. Furthermore, the proposed model provided a

more accurate prediction of the antagonistic and synergistic toxicity of the mixture compared to the CA and IA models. Therefore, the QSAR model can be used to predict additive and nonadditive toxicity of binary and multicomponent mixtures [130]. The toxicity interaction of the mixture can be assessed by predicting the results of CA ($pEC_{50,CA}$) and IA ($pEC_{50,IA}$) and the 95% CI of $EC_{50,Obs}$. the range of $pEC_{50,CA}$ or $pEC_{50,IA}$ values between the upper and lower limits of 95% CI is a mixture exhibiting an additive effect. Synergy and antagonism are considered to be nonadditive effects. $pEC_{50,CA}$ or $pEC_{50,IA}$ less than the lower limit of 95% CI of $EC_{50,Obs}$ indicates synergy, while $pEC_{50,CA}$ or $pEC_{50,IA}$ greater than the upper limit of 95% CI of $EC_{50,Obs}$ indicates antagonism [130].

The molecular descriptors (calculated using the Dragon 7.0 software) include constituent descriptors, topological descriptors, connectivity indices, information indices, 2D autocorrelations, and atom-centered fragments. The original molecular descriptors of chemicals can be refined according to some principles [131, 132]. For the feature selection, selecting the best variable from the remaining descriptors can be obtained using the genetic algorithm (GA) and combining a multiple linear regression (MLR) model with several largest variables [133, 134]. The applicability domain of QSAR model is defined by the leverage method of the hat matrix. The method is based on the molecular descriptors of the mixture [131, 135] and the identification of a mixture of standard deviation residuals greater than 2.5 standard deviation units by LOO cross-validation. Outliers in the QSAR model are defined as h value greater than the warning level (h^*) and the LOO normalized residual greater than 2.5, which are graphically depicted in the Williams plot [130].

In the work of Sobati et al., the mixture descriptor was calculated using the molecular descriptors of the pure compounds constituting the mixture and their mole fractions according to several mixing rules. The authors used the enhanced replacement method (ERM) as an effective tool for subset variable selection [136]. The main statistical criteria in this study are the determination coefficient (R^2), the mean absolute relative deviation (AARD), and the root mean square deviation (RMSD). The main external statistical verification methods are leave-one-out (LOO) and leave-n-out (LNO) cross-validation technique, bootstrap technique, y-randomization technique, and external validation. These techniques have been briefly introduced in his article [136–139].

Gaudin et al. used a series of formulas to derive the mixture descriptors used to develop the QSPR model of the mixture. They considered the linear or nonlinear dependence of the flash point on the concentration of each compound. The proposed best model was a four-parameter model with a predicted average absolute error of 10.3 °C [140].

13 QSAR Modeling Based on Chemical Reactivity Theory

Here, we introduce the use of QSAR models based on chemical reactivity theory (quantum four-element method), which can be used to successfully predict the physicochemical properties and toxicological endpoints of organic and inorganic compounds. In practical applications, it provides a quick screening method for safe chemical mixtures and a mechanistic interpretation of the toxicity of chemical mixtures.

Chemical reactions are caused by potential molecular properties; therefore, some molecular parameters can be used as indicators of chemical reactions. Important intermolecular forces include: covalent bonding, ionic bonding, van der Waals force, and hydrophilicity. These four forces are important variables that affect the interaction of chemical mixtures and are associated with four types of descriptors. This relationship has been able to study the reaction mechanisms of various environmental characteristics of organic pollutants. The quantum four-element model classifies electronic attributes into four distinct properties (contact, non-contact, deformable, and non-deformable) (Fig. 1) [141–146]. By combining two adjacent electronic properties, four major chemical bonds or forces will be generated. The electrostatic interaction is a combination of non-contact and non-deformable properties; a combination of deformable and contact properties form electron flow; polarization is a combination of non-contact and deformable properties; non-deformable and contact electronic attributes form a hydrophilic interaction. Therefore, according to the model parameters of the quantum four-element model, it can be known which mechanism of action (electrostatic, electron flow, polarization, hydrophilic interaction) causes a chemical reaction. In the quantum four-element QSAR model, all descriptors are based on chemical reactivity theory (Fig. 2), having specific physicochemical significance and being independent of each other [147–149].

In a previous QSAR study of the present authors on mixture toxicity of organic pollutants [143], four types of quantum four-element mixture descriptors were used as initial parameter sets to determine the appropriate QSAR model to evaluate the organic compound mixture against *Vibrio fischeri* 15 min [124, 125, 127] and the toxicity of *Scenedesmus obliquus* after 48 h of exposure [150]. The 1/EC_{50M} defining biological activity was the dependent variable of the QSAR model. The combined toxic effects of the mixture of organic compounds can be expressed by the mixture descriptor, calculated as follows: $D = \sum x_i D_i$, where D_i is the value of the selected descriptor and x_i is the fractional concentration of the mixture components. The descriptor value (D) is a measure of the contribution of each component of the mixture to the overall activity. The four types of mixture descriptors D were the

HSAB	Quantum four Element		Descriptor		Interaction
			Electrophile	Nucleophile	
Hard	Non-contact	Non-deformable	ρ_{\max}^+ (local)	$-\rho_{\max}^-$ (local)	electrostatic
		Deformable	s_{\max}^+ (local)	s_{\max}^- (local)	Polarization
Soft	Contact	Deformable	$-\mu^+$ (global)	μ^- (global)	Electron flow
		Non-deformable	1/APSA (global)	1/APSA (global)	Hydrophilic

Fig. 1 The correspondence between four types of quantum four-element reactivity indices and the Pearson's hard/soft definition [141–146]

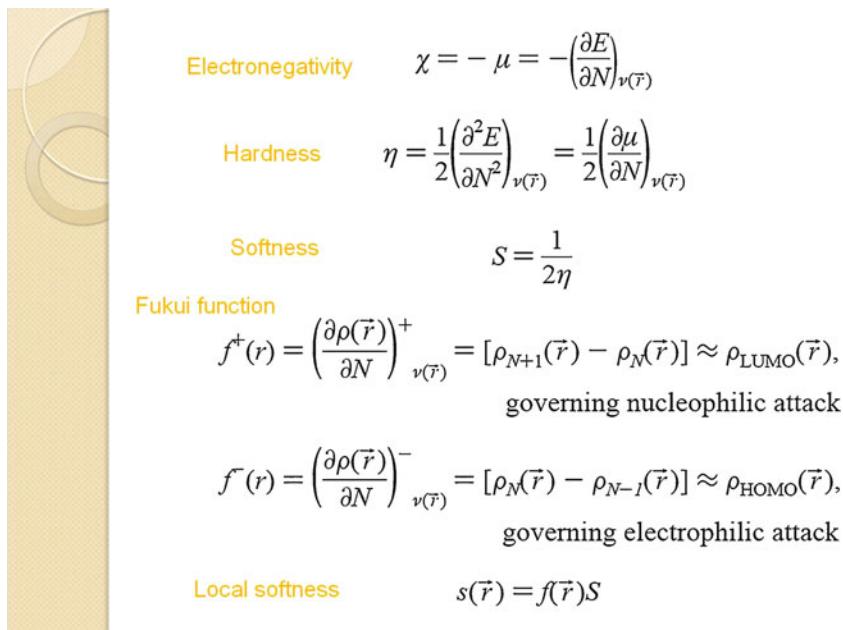


Fig. 2 Conceptual density functional theory [147–149]

independent variables: (1) charge acceptance and charge donation chemical potential, (2) the maximum positive charge of hydrogen atom and the maximum negative charge, (3) the maximum

nucleophilic and electrophilic condensed local softness, and (4) the inverse of the apolar surface area. Multiple regression analysis was performed using experimental measurements ($1/\text{EC}_{50\text{M}}$) and mixture descriptor D . The entire data set was divided into training sets and test sets. In order to select the interpretation parameters, an inverse elimination procedure is used to determine the basic parameters retained in the model.

The determination coefficient (R^2) and the adjusted determination coefficient (R^2_{adj}), leave-one-out cross-validated R^2 (R^2_{CV}), standard deviation (SD), and ANOVA F-statistic (F) were used to find the quality of the QSAR model. The results of MLR analysis indicate that the maximum positive charge of hydrogen atom and the inverse of the apolar surface area, representing electrostatic and hydrophilic interactions, are the most important descriptors of the mixture toxicity of benzene and its derivatives to *Vibrio fischeri*. It is found that the electron acceptance chemical potential and the maximum positive charge of hydrogen atom, representing electron flow and electrostatic interaction, are the most important descriptors of the joint toxicity of aromatic compounds to *Scenedesmus obliquus*. This method provides a basis for explaining the interactions that affect the toxicity of organic compound mixtures.

14 Conclusions

Transformation products are usually formed in complex matrices. Because of their low concentration, separation and purification are very difficult, and little is known about the ecotoxicological potential of parent compound and metabolite mixtures. In order to minimize the test of vertebrates, in silico modeling has proven to be a fast and inexpensive method. Data from the parent compounds and the transformation products have been integrated to assess the risk of the pesticide metabolite to the animal and the environment, thereby prioritizing the contribution of the transformation product to the overall environmental risk. The University of Minnesota pathway prediction system (UM-PPS) is a rule-based system for predicting microbial metabolites. For microbial communities in different environments, different enzyme-catalyzed reactions can occur at very different rates. This suggests that not only chemical structures but also specific environmental conditions should be considered, which can greatly improve biotransformation prediction. To provide support and explain the underlying mechanisms that lead to chemical interactions that then affect the toxicity of mixtures, previous researchers have combined external exposure, toxicokinetics, and toxicodynamics to propose a biology-based framework. The use of quantum chemistry allows for a more detailed study of pesticides and their metabolites, and it is possible

to mimic the chemical reaction mechanisms of pesticides and their transformation products and their degradation pathways.

References

1. Sinclair CJ, Boxall ABA (2003) Assessing the ecotoxicity of pesticide transformation products. *Environ Sci Technol* 37(20):4617–4625
2. Burden N, Maynard SK, Weltje L, Wheeler JR (2016) The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: a retrospective validation approach. *Regul Toxicol Pharmacol* 80:241–246
3. EC (2009) Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. *Off J Eur Union* 50:1–50
4. Sinclair CJ, Boxall ABA (2009) Ecotoxicity of transformation products. In: Boxall ABA (ed) Transformation products of synthetic chemicals in the environment. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 177–204
5. Ng CA, Scheringer M, Fenner K, Hungerbühler K (2011) A framework for evaluating the contribution of transformation products to chemical persistence in the environment. *Environ Sci Technol* 45(1):111–117
6. Galassi S, Provenzi A, Halfon E (1996) Risk assessment for pesticides and their metabolites in water. *Int J Environ Anal Chem* 65(1–4):331–344
7. Roncaglioni A, Benfenati E, Boriani E, Clook M (2004) A protocol to select high quality datasets of ecotoxicity values for pesticides. *J Environ Sci Health B* 39(4):641–652
8. van Zelm R, Huijbregts MAJ, van de Meent D (2010) Transformation products in the life cycle impact assessment of chemicals. *Environ Sci Technol* 44(3):1004–1009
9. Escher BI, Bramaz N, Richter M, Lienert J (2006) Comparative ecotoxicological hazard assessment of beta-blockers and their human metabolites using a mode-of-action-based test battery and a QSAR approach. *Environ Sci Technol* 40(23):7402–7408
10. Verhaar HJM, van Leeuwen CJ, Hermens JLM (1992) Classifying environmental pollutants. *Chemosphere* 25(4):471–491
11. Escher BI, Fenner K (2011) Recent advances in environmental risk assessment of transformation products. *Environ Sci Technol* 45(9):3835–3847
12. Fenner K, Scheringer M, Hungerbühler K (2000) Persistence of parent compounds and transformation products in a level IV multi-media model. *Environ Sci Technol* 34(17):3809–3817
13. Ferrari B, Mons R, Vollat B, Fraysse B, Paxēaus N, Giudice RL, Pollio A, Garric J (2004) Environmental risk assessment of six human pharmaceuticals: are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment? *Environ Toxicol Chem* 23(5):1344–1354
14. Huschek G, Hansen PD, Maurer HH, Krengel D, Kayser A (2004) Environmental risk assessment of medicinal products for human use according to European Commission recommendations. *Environ Toxicol* 19(3):226–240
15. Baselt RC (2000) Disposition of toxic drugs and chemicals in man: Chemical Toxicology Institute.
16. EFSA P. Panel (European Food Safety Authority Panel on Plant Protection Products and their Residues) (2013) Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA J* 11(7):3290
17. Reuschenbach P, Silvani M, Dammann M, Warnecke D, Knacker T (2008) ECOSAR model performance with a large test set of industrial chemicals. *Chemosphere* 71(10):1986–1995
18. Kern S, Fenner K, Singer HP, Schwarzenbach RP, Hollender J (2009) Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environ Sci Technol* 43(18):7039–7046
19. Ellis LBM, Gao J, Fenner K, Wackett LP (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res* 36(Suppl 2):W427–W432
20. Ellis LBM, Roe D, Wackett LP (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res* 34(Suppl 1):D517–D521
21. Klopman G, Tu M (1997) Structure–biodegradability study and computer-automated prediction of aerobic biodegradation of

- chemicals. *Environ Toxicol Chem* 16(9):1829–1835
22. Mekenyan O, Dimitrov S, Dimitrova N, Dimitrova G, Pavlov T, Chankov G, Kотов S, Vasilev K, Vasilev R (2006) Metabolic activation of chemicals: in-silico simulation. *SAR QSAR Environ Res* 17(1):107–120
 23. Saiakhov R, Chakravarti S, Klopman G (2013) Effectiveness of CASE ultra expert system in evaluating adverse effects of drugs. *Mol Inf* 32(1):87–97
 24. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE (2000) LeadScope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 40(6):1302–1314
 25. Gutowski L, Olsson O, Leder C, Kümmeler K (2015) A comparative assessment of the transformation products of S-metolachlor and its commercial product Mercantor Gold® and their fate in the aquatic environment by employing a combination of experimental and in silico methods. *Sci Total Environ* 506–507:369–379
 26. Villaverde JJ, Sevilla-Morán B, López-Goti C, Calvo L, Alonso-Prados JL, Sandín-España P (2018) Photolysis of clethodim herbicide and a formulation in aquatic environments: fate and ecotoxicity assessment of photoproducts by QSAR models. *Sci Total Environ* 615:643–651
 27. Transformation products of pesticides in the environment: analysis and occurrence. In: Transformation products of emerging contaminants in the environment.
 28. Martins PF, Martinez CO, Gd C, Carneiro PIB, Azevedo RA, Pileggi SAV, Melo IS, Pileggi M (2007) Selection of microorganisms degrading S-Metolachlor herbicide. *Braz Arch Biol Technol* 50:153–159
 29. Olsson O, Khodorkovsky M, Gassmann M, Friedler E, Schneider M, Dubowski Y (2013) Fate of pesticides and their transformation products: first flush effects in a semi-arid catchment. *Clean (Weinh)* 41(2):134–142
 30. Schmidt CK, Brauch H-J (2008) N, N-dimethylsulfamide as precursor for N-nitrosodimethylamine (NDMA) formation upon ozonation and its fate during drinking water treatment. *Environ Sci Technol* 42(17):6340–6346
 31. Mahmoud WMM, Toolaram AP, Menz J, Leder C, Schneider M, Kümmeler K (2014) Identification of phototransformation products of thalidomide and mixture toxicity assessment: an experimental and quantitative structural activity relationships (QSAR) approach. *Water Res* 49:11–22
 32. Rastogi T, Leder C, Kümmeler K (2014) Qualitative environmental risk assessment of photolytic transformation products of iodinated X-ray contrast agent diatrizoic acid. *Sci Total Environ* 482–483:378–388
 33. Rastogi T, Leder C, Kümmeler K (2014) Designing green derivatives of β-blocker Metoprolol: a tiered approach for green and sustainable pharmacy and chemistry. *Chemosphere* 111:493–499
 34. Gasser L, Fenner K, Scheringer M (2007) Indicators for the exposure assessment of transformation products of organic micropollutants. *Environ Sci Technol* 41(7):2445–2451
 35. Kern S, Singer H, Hollender J, Schwarzenbach RP, Fenner K (2011) Assessing exposure to transformation products of soil-applied organic contaminants in surface water: comparison of model predictions and field data. *Environ Sci Technol* 45(7):2833–2841
 36. Fenner K, Schenker U, Scheringer M (2008) Modelling environmental exposure to transformation products of organic chemicals. In: Boxall ABA (eds) Transformation products of synthetic chemicals in the environment. The handbook of environmental chemistry, vol 2P. Springer, Berlin, Heidelberg
 37. Latino DARS, Wicker J, Gütlein M, Schmid E, Kramer S, Fenner K (2017) Eawag-Soil in enviPath: a new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data. *Environ Sci: Processes Impacts* 19(3):449–464
 38. Hernández-Moreno D, Blázquez M, Andreu-Sánchez O, Bermejo-Nogales A, Fernández-Cruz ML (2019) Acute hazard of biocides for the aquatic environmental compartment from a life-cycle perspective. *Sci Total Environ* 658:416–423
 39. Lienert J, Güdel K, Escher BI (2007) Screening method for ecotoxicological hazard assessment of 42 pharmaceuticals considering human metabolism and excretory routes. *Environ Sci Technol* 41(12):4471–4478
 40. Cwiertny DM, Snyder SA, Schlenk D, Kolodziej EP (2014) Environmental designer drugs: when transformation may not eliminate risk. *Environ Sci Technol* 48(20):11737–11745
 41. Boxall ABA, Sinclair CJ, Fenner K, Kolpin D, Maund SJ (2004) Peer reviewed: when synthetic chemicals degrade in the environment. *Environ Sci Technol* 38(19):368A–375A
 42. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M (2010) PathPred: an enzyme-catalyzed

- metabolic pathway prediction server. *Nucleic Acids Res* 38(suppl_2):W138–W143
43. Dimitrov S, Pavlov T, Dimitrova N, Georgieva D, Nedelcheva D, Kesova A, Vasilev R, Mekenyany O (2011) Simulation of chemical metabolism for fate and hazard assessment. II CATALOGIC simulation of abiotic and microbial degradation. *SAR QSAR Environ Res* 22(7–8):719–755
 44. Finley SD, Broadbelt LJ, Hatzimanikatis V (2009) Computational framework for predictive biodegradation. *Biotechnol Bioeng* 104 (6):1086–1097
 45. Ellis LBM, Gao J, Fenner K, Wackett LP (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res* 36(Web Server issue): W427–W432
 46. Gao J, Ellis LBM, Wackett LP (2010) The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Res* 38(Database issue): D488–D491
 47. Bending GD, Lincoln SD, Edmondson RN (2006) Spatial variation in the degradation rate of the pesticides isoproturon, azoxystrobin and diflufenican in soil and its relationship with chemical and microbial properties. *Environ Pollut* 139(2):279–287
 48. Helbling DE, Johnson DR, Honti M, Fenner K (2012) Micropollutant biotransformation kinetics associate with WWTP process parameters and microbial community characteristics. *Environ Sci Technol* 46 (19):10579–10588
 49. Mekenyany OG, Dimitrov SD, Pavlov TS, Veith GD (2005) POPs: a QSAR system for developing categories for persistent, bioaccumulative and toxic chemicals and their metabolites. *SAR QSAR Environ Res* 16 (1–2):103–133
 50. Escher BI, Baumgartner R, Lienert J, Fenner K (2009) Predicting the ecotoxicological effects of transformation products. In: Boxall ABA (ed) Transformation products of synthetic chemicals in the environment. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 205–244
 51. Rudik AV, Bezhentsev VM, Dmitriev AV, Druzhilovskiy DS, Lagunin AA, Filimonov DA, Poroikov VV (2017) MetaTox: web application for predicting structure and toxicity of xenobiotics' metabolites. *J Chem Inf Model* 57(4):638–642
 52. Filimonov D, Poroikov V, Borodina Y, Glorizova T (1999) Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *J Chem Inf Comput Sci* 39(4):666–670
 53. Filimonov DA, Zakharov AV, Lagunin AA, Poroikov VV (2009) QNA-based 'star track' QSAR approach. *SAR QSAR Environ Res* 20 (7–8):679–709
 54. Zakharov AV, Varlamova EV, Lagunin AA, Dmitriev AV, Muratov EN, Fourches D, Kuz'min VE, Poroikov VV, Tropsha A, Nicklaus MC (2016) QSAR modeling and prediction of drug–drug interactions. *Mol Pharm* 13(2):545–556
 55. Kolanczyk RC, Schmieder P, Jones WJ, Mekenyany OG, Chapkanov A, Temelkov S, Kotov S, Velikova M, Kamenska V, Vasilev K, Veith GD (2012) MetaPath: an electronic knowledge base for collating, exchanging and analyzing case studies of xenobiotic metabolism. *Regul Toxicol Pharmacol* 63 (1):84–96
 56. Kolanczyk RC, Serrano JA, Tapper MA, Schmieder PK (2018) A comparison of fish pesticide metabolic pathways with those of the rat and goat. *Regul Toxicol Pharmacol* 94:124–143
 57. Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A, Glen RC (2012) Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J Chem Inf Model* 52 (3):617–648
 58. Madden JC, Webb S, Enoch SJ, Colley HE, Murdoch C, Shipley R, Sharma P, Yang C, Cronin MTD (2017) In silico prediction of skin metabolism and its implication in toxicity assessment. *Computat Toxicol* 3:44–57
 59. Agatonovic-Kustrin S, Morton DW, Celebic D (2013) QSAR: an in silico approach for predicting the partitioning of pesticides into breast milk. *Comb Chem High Throughput Screen* 16(3):223–232
 60. Xiao H, Kuckelkorn J, Nüßer LK, Floehr T, Hennig MP, Roß-Nickoll M, Schäffer A, Hollert H (2016) The metabolite 3,4,3',4'-tetrachloroazobenzene (TCAB) exerts a higher ecotoxicity than the parent compounds 3,4-dichloroaniline (3,4-DCA) and propanil. *Sci Total Environ* 551–552:304–316
 61. Vedani A, Smiesko M, Spreafico M, Peristera O, Dobler M (2009) VirtualToxLab - in silico prediction of the toxic (endocrine-disrupting) potential of drugs, chemicals and natural products. Two years and 2,000 compounds of experience: a progress report. *ALTEX* 26(3):167–176

62. Vedani A, Dobler M, Smieško M (2012) VirtualToxLab — a platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol Appl Pharmacol* 261(2):142–153
63. Vedani A, Dobler M, Hu Z, Smieško M (2015) OpenVirtualToxLab—A platform for generating and exchanging in silico toxicity data. *Toxicol Lett* 232(2):519–532
64. Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, Xia M, Huang R, Rotroff DM, Filer DL, Houck KA, Martin MT, Sipes N, Richard AM, Mansouri K, Setzer RW, Knudsen TB, Crofton KM, Thomas RS (2015) Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol Sci* 148(1):137–154
65. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS (2016) CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 124(7):1023–1033
66. Pinto CL, Mansouri K, Judson R, Browne P (2016) Prediction of estrogenic bioactivity of environmental chemical metabolites. *Chem Res Toxicol* 29(9):1410–1427
67. Dekant W, Melching-Kollmuß S, Kalberlah F (2010) Toxicity assessment strategies, data requirements, and risk assessment approaches to derive health based guidance values for non-relevant metabolites of plant protection products. *Regul Toxicol Pharmacol* 56(2):135–142
68. Pavan M, Worth AP (2008) Publicly-accessible QSAR software tools developed by the Joint Research Centre. *SAR QSAR Environ Res* 19(7–8):785–799
69. Clark RD (2018) Predicting mammalian metabolism and toxicity of pesticides in silico. *Pest Manag Sci* 74(9):1992–2003
70. Safe SH (1998) Hazard and risk assessment of chemical mixtures using the toxic equivalency factor approach. *Environ Health Perspect* 106(Suppl 4):1051–1058
71. Mekenyen OG, Kamenska V, Schmieder PK, Ankley GT, Bradbury SP (2000) A computationally based identification algorithm for estrogen receptor ligands: part 2. Evaluation of a hER α binding affinity model. *Toxicol Sci* 58(2):270–281
72. Chen Y, Cheng F, Sun L, Li W, Liu G, Tang Y (2014) Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. *Ecotoxicol Environ Saf* 110:280–287
73. Kirchmair J, Goller AH, Lang D, Kunze J, Testa B, Wilson ID, Glen RC, Schneider G (2015) Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov* 14(6):387–404
74. Wang P, Dang L, Zhu B-T (2016) Use of computational modeling approaches in studying the binding interactions of compounds with human estrogen receptors. *Steroids* 105:26–41
75. Tanji KK, Sullivan JJ (1995) Qsar analysis of the chemical hydrolysis of organophosphorus pesticides in natural waters. Technical Completion Report Project Number W-843, University of California Water Resource Center
76. Ortiz-Hernández ML, Quintero-Ramírez R, Nava-Ocampo AA, Bello-Ramírez AM (2003) Study of the mechanism of Flavobacterium sp. for hydrolyzing organophosphate pesticides. *Fundam Clin Pharmacol* 17(6):717–723
77. Lee Y, von Gunten U (2012) Quantitative structure–activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. *Water Res* 46(19):6177–6195
78. Hansch C, Leo A, Taft RW (1991) A survey of Hammett substituent constants and resonance and field parameters. *Chem Rev* 91(2):165–195
79. Bello-Ramírez AM, Carreón-Garabito BY, Nava-Ocampo AA (2000) A theoretical approach to the mechanism of biological oxidation of organophosphorus pesticides. *Toxicology* 149(2):63–68
80. Lo Piparo E, Fratev F, Lemke F, Mazzatorta P, Smiesko M, Fritz JI, Benfenati E (2006) QSAR models for *Daphnia magna* toxicity prediction of benzoxazinone allelochemicals and their transformation products. *J Agric Food Chem* 54(4):1111–1115
81. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110(18):5959–5967
82. Nielsen MK, Holtze MS, Svensmark B, Juhler RK (2007) Demonstrating formation of

- potentially persistent transformation products from the herbicides bromoxynil and ioxynil using liquid chromatography-tandem mass spectrometry (LC-MS/MS). Pest Manag Sci 63(2):141–149
83. Jeannot R, Sabik H, Sauvard E, Genin E (2000) Application of liquid chromatography with mass spectrometry combined with photodiode array detection and tandem mass spectrometry for monitoring pesticides in surface waters. J Chromatogr A 879(1):51–71
 84. Bossi R, Vejrup KV, Mogensen BB, Asman WAH (2002) Analysis of polar pesticides in rainwater in Denmark by liquid chromatography-tandem mass spectrometry. J Chromatogr A 957(1):27–36
 85. Villaverde JJ, López-Goti C, Alcamí M, Lamsabhi AM, Alonso-Prados JL, Sandín-España P (2017) Quantum chemistry in environmental pesticide risk assessment. Pest Manag Sci 73(11):2199–2202
 86. Duirk SE, Desetto LM, Davis GM (2009) Transformation of organophosphorus pesticides in the presence of aqueous chlorine: kinetics, pathways, and structure–activity relationships. Environ Sci Technol 43(7):2335–2340
 87. Hu J-y, Morita T, Magara Y, Aizawa T (2000) Evaluation of reactivity of pesticides with ozone in water using the energies of frontier molecular orbitals. Water Res 34(8):2215–2222
 88. Matsushita T, Morimoto A, Kuriyama T, Matsumoto E, Matsui Y, Shirasaki N, Kondo T, Takanashi H, Kameya T (2018) Removals of pesticides and pesticide transformation products during drinking water treatment processes and their impact on mutagen formation potential after chlorination. Water Res 138:67–76
 89. Villaverde JJ, Sandín-España P, Alonso-Prados JL, Lamsabhi AM, Alcamí M (2018) Computational study of the structure and degradation products of aloxydim herbicide. J Phys Chem A 122(15):3909–3918
 90. Sinclair CJ, Boxall ABA, Parsons SA, Thomas MR (2006) Prioritization of pesticide environmental transformation products in drinking water supplies. Environ Sci Technol 40(23):7283–7289
 91. Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P (2016) Trends in analysis of pesticide residues to fulfil the European Regulation (EC) No. 1107/2009. TrAC Trends Anal Chem 80:568–580
 92. Sandín-España P, Sevilla-Morán B, Calvo L, Mateo-Miranda M, Alonso-Prados JL (2013) Photochemical behavior of aloxydim herbicide in environmental waters. Structural elucidation and toxicity of degradation products. Microchem J 106:212–219
 93. Duirk SE, Collette TW (2006) Degradation of chlorpyrifos in aqueous chlorine solutions: pathways, kinetics, and modeling. Environ Sci Technol 40(2):546–551
 94. Neuwoehner J, Zilberman T, Fenner K, Escher BI (2010) QSAR-analysis and mixture toxicity as diagnostic tools: influence of degradation on the toxicity and mode of action of diuron in algae and daphnids. Aquat Toxicol 97(1):58–67
 95. Verhaar HJM, Busser FJM, Hermens JLM (1995) Surrogate parameter for the baseline toxicity content of contaminated water: simulating the bioconcentration of mixtures of pollutants and counting molecules. Environ Sci Technol 29(3):726–734
 96. Altenburger R, Nendza M, Schüürmann G (2003) Mixture toxicity and its modeling by quantitative structure–activity relationships. Environ Toxicol Chem 22(8):1900–1915
 97. Zhang Y-H, Liu S-S, Song X-Q, Ge H-L (2008) Prediction for the mixture toxicity of six organophosphorus pesticides to the luminescent bacterium Q67. Ecotoxicol Environ Saf 71(3):880–888
 98. Di Nica V, Gallet J, Villa S, Mezzanotte V (2017) Toxicity of Quaternary Ammonium Compounds (QACs) as single compounds and mixtures to aquatic non-target microorganisms: experimental data and predictive models. Ecotoxicol Environ Saf 142:567–577
 99. Neale PA, Leusch FDL, Escher BI (2017) Applying mixture toxicity modelling to predict bacterial bioluminescence inhibition by non-specifically acting pharmaceuticals and specifically acting antibiotics. Chemosphere 173:387–394
 100. Junghans M, Backhaus T, Faust M, Scholze M, Grimme LH (2003) Predictability of combined effects of eight chloroacetanilide herbicides on algal reproduction. Pest Manag Sci 59(10):1101–1110
 101. Deneer JW (2000) Toxicity of mixtures of pesticides in aquatic systems. Pest Manag Sci 56(6):516–520
 102. Faust M, Altenburger R, Backhaus T, Blanck H, Boedeker W, Gramatica P, Hamer V, Scholze M, Vighi M, Grimme LH (2001) Predicting the joint algal toxicity of multi-component s-triazine mixtures at low-effect concentrations of individual toxicants. Aquat Toxicol 56(1):13–32

103. Faust M, Altenburger R, Backhaus T, Blanck H, Boedeker W, Gramatica P, Hamer V, Scholze M, Vighi M, Grimme LH (2003) Joint algal toxicity of 16 dissimilarly acting chemicals is predictable by the concept of independent action. *Aquat Toxicol* 63 (1):43–63
104. Backhaus T, Faust M, Scholze M, Gramatica P, Vighi M, Grimme LH (2004) Joint algal toxicity of phenylurea herbicides is equally predictable by concentration addition and independent action. *Environ Toxicol Chem* 23(2):258–264
105. Junghans M, Backhaus T, Faust M, Scholze M, Grimme LH (2006) Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. *Aquat Toxicol* 76(2):93–110
106. Pape-Lindstrom PA, Lydy MJ (1997) Synergistic toxicity of atrazine and organophosphate insecticides contravenes the response addition mixture model. *Environ Toxicol Chem* 16(11):2415–2420
107. Spurgeon DJ, Jones OAH, Dorne J-LCM, Svendsen C, Swain S, Stürzenbaum SR (2010) Systems toxicology approaches for understanding the joint effects of environmental chemical mixtures. *Sci Total Environ* 408(18):3725–3734
108. Phyu YL, Palmer CG, Warne MSJ, Hose GC, Chapman JC, Lim RP (2011) A comparison of mixture toxicity assessment: examining the chronic toxicity of atrazine, permethrin and chlorothalonil in mixtures to Ceriodaphnia cf. dubia. *Chemosphere* 85(10):1568–1573
109. Altenburger R, Boedeker W, Faust M, Grimme LH (1996) Regulations for combined effects of pollutants: consequences from risk assessment in aquatic toxicology. *Food Chem Toxicol* 34(11):1155–1157
110. Warne MSJ, Hawker DW (1995) The number of components in a mixture determines whether synergistic and antagonistic or additive toxicity predominate: the funnel hypothesis. *Ecotoxicol Environ Saf* 31(1):23–28
111. Wang L-J, Liu S-S, Zhang J, Li W-Y (2010) A new effect residual ratio (ERR) method for the validation of the concentration addition and independent action models. *Environ Sci Pollut Res* 17(5):1080–1089
112. Tang JYM, McCarty S, Glenn E, Neale PA, Warne MSJ, Escher BI (2013) Mixture effects of organic micropollutants present in water: towards the development of effect-based water quality trigger values for baseline toxicity. *Water Res* 47(10):3300–3314
113. Lydy M, Belden J, Wheelock C, Hammock B, Denton D (2004) Challenges in regulating pesticide mixtures. *Ecol Soc* 9(6):1
114. Laetz CA, Baldwin DH, Collier TK, Hebert V, Stark JD, Scholz NL (2009) The synergistic toxicity of pesticide mixtures: implications for risk assessment and the conservation of endangered Pacific salmon. *Environ Health Perspect* 117(3):348–353
115. LeBlanc GA, Wang G (2006) Chemical mixtures: greater-than-additive effects? *Environ Health Perspect* 114(9):A517–A519
116. Feron VJ, Cassee FR, Groten JP (1998) Toxicology of chemical mixtures: international perspective. *Environ Health Perspect* 106 (suppl 6):1281–1289
117. Chèvre N, Loepfe C, Singer H, Stamm C, Fenner K, Escher BI (2006) Including mixtures in the determination of water quality criteria for herbicides in surface water. *Environ Sci Technol* 40(2):426–435
118. Desalegn A, Bopp S, Asturiol D, Lamon L, Worth A, Paini A (2019) Role of Physiologically Based Kinetic modelling in addressing environmental chemical mixtures – a review. *Comput Toxicol* 10:158–168
119. Boberg J, Dybdahl M, Petersen A, Hass U, Svingen T, Vinggaard AM (2019) A pragmatic approach for human risk assessment of chemical mixtures. *Curr Opin Toxicol* 15:1–7
120. Escher BI, Baumer A, Bittermann K, Henneberger L, König M, Kühnert C, Klüver N (2017) General baseline toxicity QSAR for nonpolar, polar and ionisable chemicals and their mixtures in the bioluminescence inhibition assay with Aliivibrio fischeri. *Environ Sci: Processes Impacts* 19(3):414–428
121. Vaes WHJ, Ramos EU, Verhaar HJM, Hermens JLM (1998) Acute toxicity of nonpolar versus polar narcosis: is there a difference? *Environ Toxicol Chem* 17(7):1380–1384
122. Ghafourian T, Samaras EG, Brooks JD, Riviere JE (2010) Modelling the effect of mixture components on permeation through skin. *Int J Pharm* 398(1):28–32
123. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) CORAL: models of toxicity of binary mixtures. *Chemom Intell Lab Syst* 119:39–43
124. Lin Z, Yu H, Wei D, Wang G, Feng J, Wang L (2002) Prediction of mixture toxicity with its total hydrophobicity. *Chemosphere* 46 (2):305–310
125. Lin Z, Zhong P, Yin K, Wang L, Yu H (2003) Quantification of joint effect for hydrogen bond and development of QSARs for

- predicting mixture toxicity. *Chemosphere* 52(7):1199–1208
126. Lin Z, Shi P, Gao S, Wang L, Yu H (2003) Use of partition coefficients to predict mixture toxicity. *Water Res* 37(9):2223–2227
127. Zhang L, Zhou P-j, Yang F, Wang Z-d (2007) Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives. *Chemosphere* 67(2):396–401
128. Chen F, Liu S-S, Duan X-T, Xiao Q-F (2014) Predicting the mixture effects of three pesticides by integrating molecular simulation with concentration addition modeling. *RSC Adv* 4(61):32256–32262
129. Kim J, Fischer M, Helms V (2018) Prediction of synergistic toxicity of binary mixtures to *vibrio fischeri* based on biomolecular interaction networks. *Chem Res Toxicol* 31(11):1138–1150
130. Qin L-T, Chen Y-H, Zhang X, Mo L-Y, Zeng H-H, Liang Y-P (2018) QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. *Chemosphere* 198:122–129
131. Qin L-T, Liu S-S, Chen F, Wu Q-S (2013) Development of validated quantitative structure–retention relationship models for retention indices of plant essential oils. *J Sep Sci* 36(9–10):1553–1560
132. Qin L-T, Liu S-S, Chen F, Xiao Q-F, Wu Q-S (2013) Chemometric model for predicting retention indices of constituents of essential oils. *Chemosphere* 90(2):300–305
133. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34(24):2121–2132
134. Gramatica P, Cassani S, Chirico N (2014) QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *J Comput Chem* 35(13):1036–1044
135. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* 22(1):69–77
136. Sobati MA, Aboozari D, Maghbooli B, Najafi H (2016) A new structure-based model for estimation of true critical volume of multi-component mixtures. *Chemom Intell Lab Syst* 155:109–119
137. Kiralj R, Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J Braz Chem Soc* 20:770–787
138. Wehrens R, van der Linden WE (1997) Bootstrapping principal component regression models. *J Chemom* 11(2):157–171
139. Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47(6):2345–2357
140. Gaudin T, Rotureau P, Fayet G (2015) Mixture descriptors toward the development of quantitative structure–property relationship models for the flash points of organic mixtures. *Ind Eng Chem Res* 54(25):6596–6604
141. Chang CM (2008) DFT-based linear solvation energy relationships for the infrared spectral shifts of acetone in polar and nonpolar organic solvents. *J Phys Chem A* 112(11):2482–2488
142. Chang CM, Lin TH, Chen YS, Chang CW, Huang KL, Wu FW, Hsu WJ, Yu MP, Lin C, Wang MK (2014) A quantum chemical approach using classical concepts to characterization and descriptive analysis of various reactions of metal ions and organic compounds. *Chemom Intell Lab Syst* 136:155–163
143. Chang CM, Ou YH, Liu TC, Lu SY, Wang MK (2016) A quantitative structure–activity relationship approach for assessing toxicity of mixture of organic compounds. *SAR QSAR Environ Res* 27(6):441–453
144. Ou YH, Chang CM, Chen YS (2016) A QSPR study on the solvent-induced frequency shifts of acetone and dimethyl sulfoxide in organic solvents. *Spectrochim Acta A Mol Biomol Spectrosc* 162:109–114
145. Yu Heng O, Len C, Chia MC (2018) A quantitative structure–property relationship study of the adsorption of amino acids on kaolinite surfaces. *Int J IJQSPR* 3(2):21–35
146. Len C, Chia MC (2019) A QSAR study on the persistence of fungicides in the environment. *Int J IJQSPR* 4(2):100–116
147. Pearson RG, Songstad J (1967) Application of the principle of hard and soft acids and bases to organic chemistry. *J Am Chem Soc* 89(8):1827–1836
148. Parr RG, Pearson RG (1983) Absolute hardness: companion parameter to absolute electronegativity. *J Am Chem Soc* 105(26):7512–7516
149. Torrent-Sucarrat M, De Proft F, Ayers PW, Geerlings P (2010) On the applicability of local softness and hardness. *Phys Chem Chem Phys* 12(5):1072–1080
150. Lu G, Wang C, Tang Z, Guo X (2007) Joint toxicity of aromatic compounds to algae and QSAR study. *Ecotoxicology* 16(7):485–490



Chapter 24

Combination of Read-Across and QSAR for Ecotoxicity Prediction: A Case Study of Green Algae Growth Inhibition Toxicity Data

Ayako Furuhama

Abstract

Effective prediction of the ecotoxicity of chemicals is important for environmental hazard and risk assessment. A previously reported three-step strategy for predicting 72-h growth inhibition toxicity against the green alga *Pseudokirchneriella subcapitata* has potential utility as a general framework for algal toxicity prediction. This strategy, which combines read-across and quantitative structure–activity relationship (QSAR), consists of a pre-screening process followed by three steps. At **Step 1**, an interspecies QSAR is used to predict the toxicities of chemicals that satisfy a log *D*-based criterion. At **Step 2**, the toxicities of nonpolar and polar narcotic chemicals (Class 1 and Class 2, respectively) are predicted with QSARs. At **Step 3**, read-across based on defined categories of chemicals is used for any remaining compounds. In this case study, the generalizability of the three-step strategy was evaluated by applying it to a recently published data set of 48-h growth inhibition toxicities against *Pseudokirchneriella subcapitata*. At the pre-screening stage, new category definitions were required for each endpoint having different test conditions used to obtain the data that were used to develop the strategy. Because the interspecies QSAR used at **Step 1** requires 48-h acute *Daphnia magna* toxicity (immobilization or mortality) as a descriptor, the fact that *Daphnia magna* data were lacking or unreliable for some of the compounds in the data set limited the utility of the three-step strategy. To circumvent this problem, read-across or local QSAR could be used instead of the interspecies QSAR at **Step 1**. At **Step 2**, the QSAR for nonpolar narcotic chemicals developed for the three-step strategy was applicable to the 48-h toxicity data set used in this case study; in contrast, the QSAR for polar narcotics showed unreliable predictivity when tested on the 48-h toxicity data set. Therefore, the polar narcotic QSAR was reconstructed so that it was applicable to the 48-h toxicity data. At **Step 3**, new categories for read-across were introduced to deal with the 48-h toxicity data; specifically, the chemical categories were classified into three types: Type A for toxic categories, Type B for categories applicable for read-across, and Type C for categories that were difficult to classify for read-across.

Key words Three-step strategy for algal toxicity prediction, Pre-screening, Interspecies QSAR, Non-polar/polar narcotic QSAR, Read-across, Log *D*-based criterion, Algae growth inhibition, *Pseudokirchneriella subcapitata*

1 Introduction

The use of fish, daphnia, and algal toxicity data obtained by means of standard ecotoxicity tests conducted according to test guidelines (TGs) such as those developed by the Organisation for Economic Co-operation and Development (OECD) [1–5], the International Organization for Standardization (ISO), or the US Environmental Protection Agency, combined with a scoring system to assess data reliability [6], is a straightforward method for assessing the environmental risks posed by chemicals. However, because adequate ecotoxicity data are not available for all chemicals, prediction of ecotoxicity is also important for risk assessment. Methods for rapid ecotoxicity prediction are particularly desirable for the prioritization and screening of chemicals for regulation. One practical strategy for risk assessment is to develop and apply integrated approaches to testing and assessment (IATA) [7]. As a part of IATA, chemicals can be assessed on the basis of knowledge about their modes or mechanisms of toxic action [8–12]. One of the simplest effective ways to rapidly predict ecotoxicity is to use quantitative structure–activity relationship (QSAR) models when a chemical falls into a category with a well-known mode of action. Examples include the QSAR models in widely used predictive systems, such as ECOSAR [13], which incorporate expert judgments and/or follow known principles [14]. Another method is the category approach, in which toxicity is predicted by means of read-across on a case-by-case basis. In addition, combining read-across and QSAR may be effective when implemented as a part of IATA that use existing (eco)toxicity data.

In the case study described herein, a previously reported three-step toxicity prediction strategy [15] that is based on algae growth inhibition toxicities determined in accordance with OECD TG 201 [3] against *Pseudokirchneriella subcapitata* (also known as *Selenastrum capricornutum* and *Raphidocelis subcapitata*) and that combines read-across and QSAR is evaluated to assess its applicability to a data set consisting of the 48-h *Pseudokirchneriella subcapitata* toxicities of 309 chemicals [16].

The three-step strategy involves the following steps [15]:

- Pre-screening (referred to as preparation in ref. [15]): Does the chemical contain certain specific structures? On the basis of previous analyses [15, 17, 18], structural alerts are used to identify outliers, such as pesticides, and chemicals with specific (or unknown) modes or mechanisms of action, which are omitted from the analyses used at **Steps 1** and **2**.
- **Step 1:** Does the chemical meet the criterion $\log P - \log D_{\text{pH}10} > 0$? If the difference between the log of the octanol–water partition coefficient ($\log P$) and the log of the octanol–

water distribution coefficient ($\log D$) at pH 10 ($\log D_{\text{pH}10}$) for a chemical is greater than zero, the chemical has the potential to exist in ionized form and therefore meets the $\log D$ -based criterion. Such chemicals can be expected to have similar toxicity mechanisms (e.g., mechanisms related to membrane penetration) [17, 18] for both algae and daphnia, and therefore an algae–daphnia interspecies QSAR can be used for algal toxicity prediction. That is, for chemicals that meet this criterion, acute *Daphnia magna* toxicity is likely to be a good descriptor for predicting algal toxicity.

- **Step 2:** Does the chemical fall into Class 1 or Class 2? Chemicals remaining after **Step 1** are categorized as Class 1 (nonpolar narcotic) or Class 2 (polar narcotic) according to the modified Verhaar scheme [19–21], and their algal toxicities are estimated by means of QSAR models.
- **Step 3:** Does the chemical fall into one of several defined categories? The chemicals omitted at the pre-screening stage and chemicals remaining after **Step 2** are subjected to a category approach on the basis of expert judgments or empirical rules. At this stage, if a chemical meets one of the defined categories in existing algal toxicity data sets, its algal toxicity is estimated quantitatively (by read-across) or qualitatively.

The main purpose of the three-step strategy is to predict algal toxicities that can be used for rapid screening and prioritizing of chemicals for further assessment. The QSAR models used in this strategy express relationships between activity (ecotoxicity) and physicochemical properties (either measured experimentally or calculated from structural information); that is, the QSARs do not involve read-across, even though (Q)SAR methods that do involve both QSAR models and category approaches have been used in regulatory contexts [14]. Interspecies correlations (i.e., interspecies QSAR or quantitative activity–activity relationship) models [22, 23] have also been used to predict the algal toxicities of ionized chemicals under certain conditions [17, 18].

One purpose of the case study described herein was to determine how the test conditions used to measure algae growth inhibition toxicity influence the efficacy of the three-step strategy, the predictivities of the models (e.g., interspecies QSARs) within it, and read-across with the categories used in the strategy.

Note that the three-step strategy combines read-across and QSAR, with categorization based on knowledges (Fig. 1). However, read-across and QSAR are not used simultaneously to predict the algal toxicity. That is, the three-step strategy does not integrate QSAR and read-across in the manner described by Benfenati et al., who assessed the bioconcentration factors of chemicals within a weight-of-evidence framework [24].

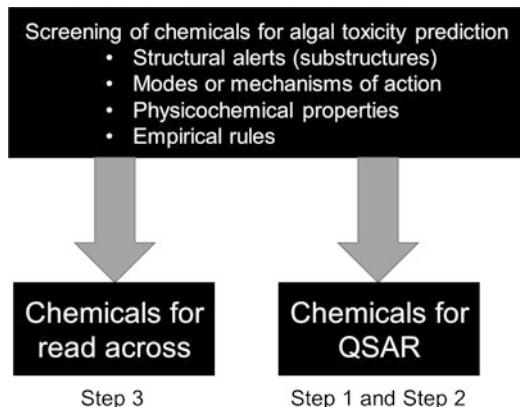


Fig. 1 Schematic of the algal toxicity prediction strategy using read-across and QSAR. On the basis of substructures, modes or mechanisms of action, physicochemical properties, and empirical rules, chemicals were categorized as suitable for algal toxicity prediction by read-across or QSAR. This strategy did not involve the integrational use of read-across and QSAR

2 Data Preparation

In this case study, the three-step strategy was evaluated with data reported by Kusk et al. [16], who recently determined 48-h 50% effect concentrations (EC_{50} [mg/L]) for 425 organic chemicals under identical conditions in growth inhibition tests against *Pseudokirchneriella subcapitata*. Data for 309 of the 425 chemicals were used in this study.

Most of the algal toxicity data used in the original work on the three-step strategy were obtained by means of 72-h exposure tests (except for some shorter-duration data obtained by means of OECD TG 201) and are expressed as growth rates. Kusk et al. carried out alga growth inhibition tests based on the procedures described in the standard TGs, OECD TG 201 [3], and ISO 8692 [25] and used average growth rate as the endpoint. However, these investigators modified the test conditions described in ISO 8692 as follows: the buffer capacity was higher, the initial biomass was low, and the test duration was shorter (48 h rather than 72 h). Specifically, mini-scale algae growth inhibition tests were conducted in closed glass vials containing 4 ml of test medium and 17 ml of CO_2 -enriched headspace. Instead of measuring the chemical concentration in the test medium, Kusk et al. corrected the nominal EC_{50} value by using the estimated phase distribution of the chemical: when the proportion of a chemical in the water phase was estimated to be <90%, the EC_{50} was corrected by the proportion in the water phase. For example, when 80% of a chemical was estimated to be in the water phase, the corrected value was calculated as $0.8 \times EC_{50}$ mg/L.

In this case study, the corrected EC₅₀ values of Kusk et al. were converted from units of milligrams per liter to units of millimoles per liter (using the molecular weights of the chemicals available in the supplementary data of ref. [16]) and then to the corresponding common logarithmic values, log (1/EC₅₀ [mM]). The modifications made by Kusk et al. to the standard test conditions—small, closed test system, shorter test duration, and so on [16]—may have influenced the measured algal toxicities. Even if the influence of the measured algal toxicities of chemicals was minor compared with the influence of differences between the tested algal species [26, 27], the former might affect the efficacy of the three-step strategy.

The 309 chemicals used for this case study, along with their algal log (1/48 h-EC₅₀ [mM]) values, are listed in the online resources (see Subheading 6 of this document). The data for 116 chemicals (425 minus 309) were excluded. Specifically, 36 chemicals with EC₅₀ values of >1000 mg/L were excluded. In addition, 80 chemicals for which an EC₅₀ could not be established were excluded; these were polyfluorinated compounds, compounds with low water solubilities, and compounds for which the percentage of the uncharged fraction in water was low. Note that although a corrected EC₅₀ for *N*-methyl-*N,N*-dioctyl-1-octanaminium chloride is listed in the supplemental data for ref. [16], this chemical was also excluded from this case study because the percentage of the uncharged fraction of this chemical in water was low.

3 Methods

The three-step strategy for predicting algal toxicities was developed with algae growth inhibition toxicities determined by the Japanese Ministry of the Environment in 2015 (<http://www.env.go.jp/chemi/sesaku/02e.pdf>) in accordance with OECD TG 201 [3]. In this case study, the three-step strategy was applied to the algal toxicities determined by means of the modified protocol reported by Kusk et al. [16]. The three-step strategy is a hierarchical one that comprises an interspecies QSAR model, two QSAR models, and read-across. Specifically, the strategy considers 48-h acute *Daphnia magna* EC₅₀ values (**Step 1**), mechanisms and modes of action (**Steps 2 and 3**), and structural profiles (**Step 3**), as well as physicochemical properties for read-across or local QSAR.

For calculation of the physicochemical properties of the 309 selected chemicals, the structure of each chemical was expressed as a SMILES string [28]. Minor components in the SMILES string, such as the hydrogen chloride in amycin hydrochloride (also referred to as tetracycline hydrochloride), were removed; the minor component means [H]Cl in O=C(N)C=1C(=O)[C@]2(O)C(O)=C3C(=O)c4c(O)cccc4[C@](O)(C)[C@H]3C[C@H]2[C@H](N(C)C)C=1O.[H]Cl. Salts were converted into their neutral forms, except in the case of ammonium and pyridinium ions, which were defined

by means of SMARTS notation [29] expressed as [#7v4+]. The modified SMILES strings used for the physicochemical property calculations are listed in Table I, which is available at <https://doi.org/10.6084/m9.figshare.8107646.v1>.

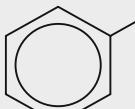
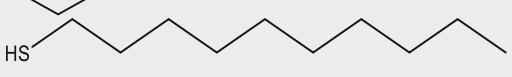
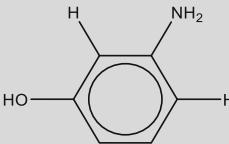
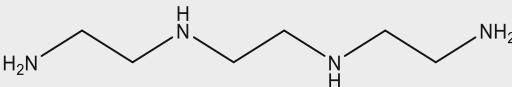
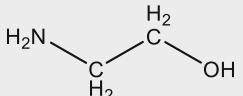
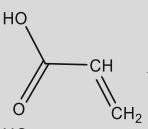
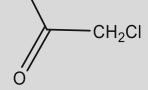
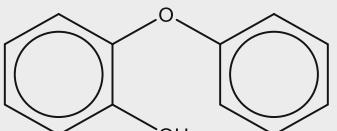
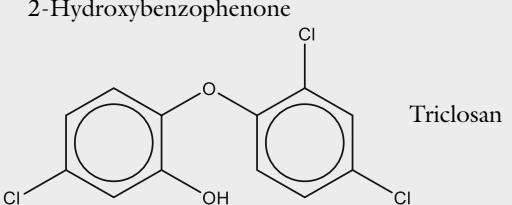
For selection of chemicals at **Step 1** and for quantitative analysis at **Steps 2** and **3**, $\log P$ values were calculated by means of the ACD/Labs software [30] using the Consensus LogP module and are referred to hereafter as $\log P(1)$ values. The values of $\log D_{\text{pH}10}$ were calculated by means of the ACD/LogD software [31] using the Consensus LogP and GALAS pKa modules. Additionally, the following physicochemical properties were estimated by means of the GALAS algorithm in the ACD/Labs software [30]: intrinsic solubility (i.e., solubility of the neutral form, $\log S_0$ [mol/L]) in water at 25 °C; and solubility in pure water along with the pH of the resulting solution ($\log S_w$ [mol/L] with a defined pH).

For evaluation of the interspecies QSAR model (**Step 1**) defined in the original work on the three-step strategy, 48-h immobilization toxicities (EC_{50} [mg/L]) or 50% lethal concentrations (LC_{50} [mg/L]) against *Daphnia magna* for the chemicals categorized at **Step 1** in this case study were collected from the QSAR Toolbox (ver. 4.3) [32]. EC_{50} or LC_{50} values that exceeded the solubility limit of the chemical in water and/or that were >100 mg/L were excluded from the data set. If, for a given chemical, acute *Daphnia magna* toxicity values determined by means of two or more than two tests were available, the geometric mean of all the available values was used. The test conditions under which the acute *Daphnia magna* toxicity data used here were obtained were inconsistent; however, it would be ideal to compare the results using data obtained by means of a standard TG such as OECD TG 202 [4], which was used for almost all the 48-h acute *Daphnia magna* EC_{50} values used in the development of the three-step strategy [15].

In accordance with the three-step strategy, at **Step 2** of this case study, chemicals were categorized as Class 1 (nonpolar narcotic) or Class 2 (polar narcotic) by means of the modified Verhaar scheme [19–21] as implemented in the Toxtree software [33, 34]. Additionally, in accordance with the original Verhaar scheme, chemicals for which $0 < \log P(1) < 6$ and chemicals with ionic groups were excluded from both Class 1 and Class 2. Values of the heat of formation (HF), a quantum chemical descriptor previously used for linear regression analyses [15], were calculated using the AM1 Hamiltonian [35] by means of the MOPAC 7 program [36]. For these HF calculations, the three-dimensional structure of each chemical was defined according to the method used in ref. [15]. Another descriptor used for the linear regression analyses at **Step 2** was the hydrophobicity parameter $\log P(1)$, as described in ref. [15].

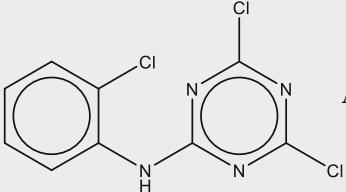
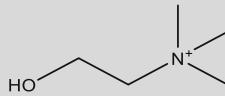
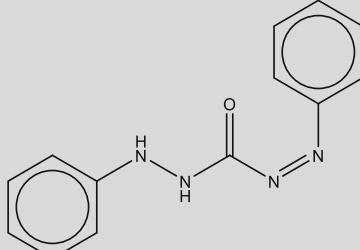
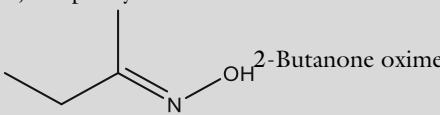
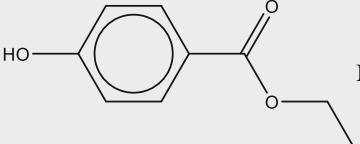
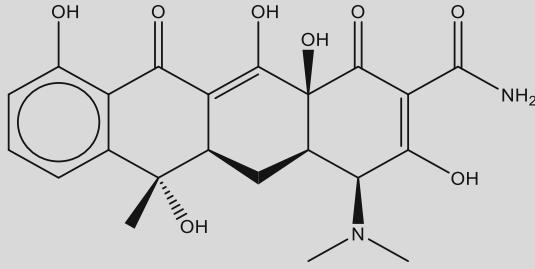
In the original work on the three-step strategy, six structural alerts (SA 1–SA 6) were applied during pre-screening: the SMARTS notations for these alerts are [SH][#6][\$([#6H3,#6H1]),\$([#6]([#6])([#6])),c2c([OH1])[cH1]c([Nv3H2])[cH1]c2,[Nv3H2][CH2][CH2][Nv3H1,OH1],[OH1]C(=O)[\$([CH1]=[CH2]),

Table 1
Structural alerts and categories used at the pre-screening stage

Alert or category	Example compounds
SA 1	 Thiophenol  1-Decanethiol
SA 2	 3-Aminophenol
SA 3	 Triethylenetetramine  Ethanolamine
SA 4	 Acrylic acid  Chloroacetic acid
SA 5 (includes hydroxybenzophenones and antibiotics related to triclosan)	 2-Hydroxybenzophenone  Triclosan

(continued)

Table 1
(continued)

Alert or category	Example compounds
SA 6	 <p>Anilazine</p>
Chemicals in the <i>Toxic</i> category in ref. [15] (includes ammonium, hydrazine, and oxime compounds; see Table 5 of ref. 15)	 <p>Ethanaminium, 2-hydroxy- N,N,N-trimethyl-</p>  <p>1,5-Diphenylcarbazone</p>  <p>2-Butanone oxime</p>
Parabens	 <p>Ethylparaben</p>
Tetracycline antibiotics	 <p>Tetracycline</p>
Pesticides	Pesticides listed in <i>The Pesticide Manual</i> [37, 38]

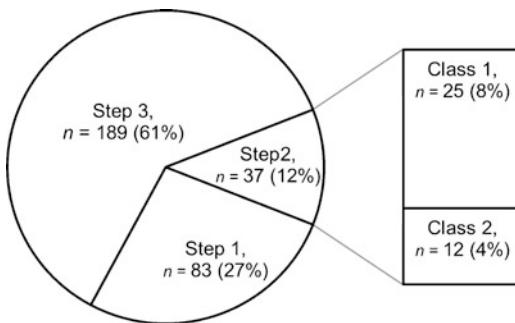


Fig. 2 Schematic showing the steps at which the 309 chemicals selected from the algal toxicity data set of Kusk et al. [16] were dealt with

$\$([CH_2]Cl)]$, $c[O,C]cc[OH]$, and n , respectively. Typical compounds containing these alerts are shown in Table 1. Chemicals listed in *The Pesticide Manual* [37, 38] were placed in the *Pesticides* category because such chemicals show specific modes or mechanisms of action. Chemicals with specific modes or mechanisms of action must be considered as possible outliers. Additionally, because the set of chemicals and the test conditions used by Kusk et al. differed from those used to develop the three-step strategy, additional structural alerts were designated by means of outlier analysis during pre-screening, and different categories were used at **Step 3**.

4 Results and Discussion

The proportions of chemicals categorized at **Steps 1–3** are depicted in Fig. 2. Most of the chemicals were categorized at **Step 3**, owing to the additional structural alerts required for the algal toxicity data set of Kusk et al. For example, the parabens (alkyl esters of *p*-hydroxybenzoic acid), which are widely used as preservatives [39], were categorized at **Step 1** in the original work on the three-step strategy, but in this case study, they were categorized at **Step 3** for the following reason. In accordance with OECD TG 201 [3], the experimentally determined 72-h algal EC₅₀ values for ethylparaben, propylparaben, and butylparaben reported by Yamamoto et al. are 52, 36, and 9.5 mg/L, respectively [40, 41]. In contrast, the 48-h algal EC₅₀ values for these chemicals measured by Kusk et al. were 0.4, 0.17, and 0.099 mg/L, respectively. The two orders of magnitude difference between the two sets of values is large enough to suggest that it may be due to something other than, or in addition to, the difference in test duration (such as the data analysis method or some other specific test conditions). Therefore, in this case study, these three chemicals were categorized at **Step 3**, having been identified as outliers by means of the additional structural alerts used at the pre-screening stage.

4.1 Pre-screening

The three-step strategy begins with a pre-screening process. When the strategy was applied to the algal toxicity data set of Kusk et al., structural alerts SA 1-SA 6, the category *Pesticides*, and, as mentioned in Subheading 4, additional structural alerts were applied to select outlier chemicals. The structural alerts and categories used for pre-screening of the chemicals in the data set of Kusk et al. are listed in Table 1, along with example compounds. The chemicals categorized as outliers in the pre-screening process are listed in Table VI, which is available at <https://doi.org/10.6084/m9.figshare.8107646.v1>.

In the original work on the three-step strategy, chemicals with ammonium, aromatic or aliphatic hydrazine, or oxime substructures were categorized as *Toxic* at **Step 3** (see Table 5 of ref. 15), with a note indicating that some of these chemicals would be ionized at pH 10. Chemicals in this category should be eliminated at the pre-screening stage so that they are not included among the **Step 1** chemicals.

Chlortetracycline hydrochloride and tetracycline hydrochloride (referred to as amycin hydrochloride) were categorized as tetracycline antibiotics with specific modes of action. The 48-h algal EC₅₀ values determined by Kusk et al. (0.11 and 0.35 mg/L, respectively) were lower (indicating higher toxicity) than the 72-h algal EC₅₀ values (3.1 and 2.2 mg/L) determined in accordance with ISO 8692 [25] by Halling-Sørensen [42]. In addition, Yang et al. also measured lower toxicities for these two compounds: the concentrations that caused 50% growth inhibition (72 h) were 1.8 and 1.0 mg/L for chlortetracycline hydrochloride and tetracycline, respectively [43]. Both of these antibiotics would be ionic at pH 10, as indicated by their pK_a values estimated by ACD/Labs [30], and could be categorized at **Step 1** initially. However, the 48-h *Daphnia magna* EC₅₀ of chlortetracycline hydrochloride [44], the 48-h *Daphnia magna* EC₅₀ of chlortetracycline [45], and the 48-h *Daphnia magna* no-observed-effect concentration (an EC₅₀ could not be determined for this compound) of tetracycline [46] are reported to be 127.4, 225, and 340 mg/L, respectively. Even though the tested compounds differed with regard to the presence or absence of hydrochloride, the difference between the 48-h algal EC₅₀ and acute *Daphnia magna* EC₅₀ values is approximately 1000-fold, and therefore, the **Step 1** interspecies QSAR (discussed in the Sect. 4.2) could not be applied to these compounds. Instead, structural information about the compounds (i.e., their categorization as tetracycline antibiotics) was used to identify them as outliers in the data set of Kusk et al.

In addition to chemicals in the *Pesticides* category, UV absorbers and antibiotics or biocides (see <http://www.oecd.org/chemicalsafety/pesticides-biocides/biocides.htm>) may be candidates for pre-screening owing to their specific modes or mechanisms of actions.

4.2 Step 1: Interspecies QSAR

Eighty-three of the chemicals remaining after pre-screening satisfied the log D -based criterion and were carried to **Step 1**. However, measured 48-h acute *Daphnia magna* toxicities were available for only 32 of the 83 chemicals (39%). The correlation between 48-h algal toxicity and acute *Daphnia magna* toxicity for this test set of 32 chemicals took the form

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.86da + 0.25, \\ -0.43 \leq da &\leq 4.94, \\ n = 32, r^2 &= 0.76, r^2_{\text{adj}} = 0.75, q^2_{\text{LOO}} = 0.73, s = 0.69, \text{RMSE} = 0.67, \end{aligned} \quad (1)$$

where da , n , r^2 , r^2_{adj} , q^2_{LOO} , s , and RMSE indicate the 48-h measured acute *Daphnia magna* $\log(1/E(L)C_{50} [\text{mM}])$, the number of chemicals, the coefficient of determination, the coefficient of determination adjusted for the number of degrees of freedom, the leave-one-out cross-validated coefficient of determination, the standard error, and the root-mean-square error, respectively.

The interspecies QSAR in the original work (Eq. 2 in Table 3 of ref. 15) and the statistical values for the prediction, which were estimated with the test set (i.e., the data used to construct Eq. 1), were as follows:

$$\begin{aligned} \log(1/72\text{h-algal EC}_{50} [\text{mM}]) &= 0.98da - 0.03, \\ -0.89 \leq da &\leq 4.27, \\ n = 103, r^2 &= 0.81, r^2_{\text{adj}} = 0.81, q^2_{\text{LOO}} = 0.80, s = 0.47, \\ n_{\text{test}} = 32, Q^2_{(\text{F1})} &= 0.77, Q^2_{(\text{F2})} = 0.74, \text{RMSEP} = 0.70, \end{aligned} \quad (2)$$

where n_{test} is the number of chemicals used for calculating the statistical values for the prediction (i.e., for external validation); $Q^2_{(\text{F1})}$ and $Q^2_{(\text{F2})}$ correspond to the correlation between the measured and predicted toxicities in the test set, as described below; and RMSEP is the RMSE of prediction for the test set. Algal toxicity data for two linear primary amines, hexadecylamine and pentadecylamine, were used to derive Eq. 1 and to calculate the statistical values for Eq. 2. However, the 48-h acute *Daphnia magna* toxicities (da) for these two amines were 4.8 and 4.9, respectively, which are outside the da range specified for Eq. 2. When these amines were omitted from the calculation of the statistical values for external validation ($n_{\text{test}} = 30$), $Q^2_{(\text{F1})}$, $Q^2_{(\text{F2})}$, and RMSEP were 0.72, 0.69, and 0.70, respectively. These values indicate that even though the da values for these two amines were outside the descriptor domain for Eq. 2, inclusion of these compounds seemed not to decrease the predictivity of the QSAR, as indicated by comparison of the $Q^2_{(\text{F1})}$ and $Q^2_{(\text{F2})}$ values.

According to Roy et al. [47], models with $Q^2_{(\text{F1})}$ or $Q^2_{(\text{F2})}$ values of >0.5 are regarded to have good predictivity, whereas Golbraikh and Tropsha [48] specified the threshold of predictive

models which is agreed to be 0.7 in the case of $Q^2_{(F1)}$ and $Q^2_{(F2)}$. The calculation of $Q^2_{(F1)}$ involves the average of the measured toxicities in the training data set (A), whereas the calculation of $Q^2_{(F2)}$ involves the average of the measured toxicities in the test data set (B): $Q^2_{(F1)} = 1 - \Sigma(Xm_i - Xp_i)^2 / \Sigma(Xm_i - A)^2$ and $Q^2_{(F2)} = 1 - \Sigma(Xm_i - Xp_i)^2 / \Sigma(Xm_i - B)^2$, where Xm_i and Xp_i indicate the measured and predicted toxicities, respectively, in the test data set. These definitions and criteria (thresholds) for other statistical values for external validation are available in the literature [47, 48]. In this case study, $Q^2_{(F1)}$ and $Q^2_{(F2)}$ met the >0.7 criterion (except that $Q^2_{(F2)}$ for $n_{\text{test}} = 30$ was 0.69), confirming that the interspecies QSAR model was predictive. It might be possible to improve the statistical values if uniform 48-h acute *Daphnia magna* toxicity values (*da*) were used—that is, if the data were generated at one laboratory by using a single test protocol or by following a harmonized protocol for *Daphnia* sp. acute immobilization tests, such as OECD TG 202 [4].

Upon implementation of the interspecies QSAR at **Step 1**, two problems became immediately apparent. First was the lack of measured acute *Daphnia magna* toxicity data; for 61% of the **Step 1** chemicals in the data set of Kusk et al., no data were available. If measured acute *Daphnia magna* toxicity data had been available for all 83 chemicals at **Step 1**, additional outlier chemicals that were unsuitable for the interspecies QSAR model might have been proposed, as discussed in the section on pre-screening (e.g., for the tetracycline antibiotics). Second, the uncertainty of the measured acute *Daphnia magna* toxicity data that were used for external validation in Eq. 2 remained. The model predictivity evaluated by means of the criterion for $Q^2_{(F1)}$ or $Q^2_{(F2)}$ could not guarantee the data quality. For example, some of the toxicity data were nominal concentrations; and others were measured concentrations or were corrected by means of a defined rule (e.g., Kusk et al. corrected the EC₅₀ values for some of the chemicals in their data set on the basis of the estimated concentration of the chemical in the water phase [16]; see Subheading 2). These problems could be solved by using acute *Daphnia magna* toxicity values estimated by QSAR instead of measured toxicity values. However, because most of the QSAR models in ecotoxicity prediction systems (e.g., ECOSAR [13] and KATE [49]) use log P as a descriptor, the predicted values for the **Step 1** chemicals that satisfy the $\log P(1) - \log D_{\text{pH}10} > 0$ criterion might be less stable than the predicted values for unionized chemicals. In addition, a study comparing eight software packages for modeling acute *Daphnia magna* toxicities suggested that the problems with the utility of the models for toxicity prediction were due to the unreliability of the toxicity data [50].

At **Step 1**, to overcome the lack of measured acute *Daphnia magna* data and the unreliability of some of the available data, subcategories were introduced so that read-across or local QSAR could be performed by using the distribution coefficients or the solubilities of the chemicals. For example, for the 31 **Step 1** chemicals categorized as primary aliphatic amines but not amides (as indicated by the SMARTS notation [NX3H2;!\$(NC=O)]; see Table II, which is available at <https://doi.org/10.6084/m9.figshare.8107646.v1>, for a list of these amines), the correlation between algal toxicity and $\log S_0$ took the form

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= -0.62 \log S_0 - 0.06, \\ -7.46 \leq \log S_0 &\leq 1.17, \\ n = 31, r^2 &= 0.80, r_{\text{adj}}^2 = 0.79, q^2_{\text{LOO}} = 0.77, s = 0.77, \text{RMSE} = 0.75. \end{aligned} \quad (3)$$

The statistical parameters indicate that the category designated “primary aliphatic amines ionized at pH 10” can be used for read-across prediction of algal toxicity. Additionally, Eq. 3 was constructed with more than 30 data points, and it can be reliably used as a local QSAR for a subcategory of chemicals in the data set of Kusk et al. In another example, for **Step 1** chemicals that were categorized as anilines with carboxylic acid moieties (but were not polycyclic aromatic compounds such as naphthalene and anthracene and did not have a methyl group attached to an aromatic ring; SMARTS notation c[NX3H2] combined with [CX3](=O)[OX2H1] and c [C!H3])—that is, the chemicals with IDs 94, 95, and 213 in Table II (available at <https://doi.org/10.6084/m9.figshare.8107646.v1>)—the correlation between algal toxicity and $\log S_0$ was given by

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= -1.14 \log S_0 - 1.95, \\ -2.19 \leq \log S_0 &\leq -0.96, \\ n = 3, r^2 &= 0.97, s = 0.17. \end{aligned} \quad (4)$$

Equation 4 represents read-across constructed under very specific conditions; if anilines containing either a polycyclic aromatic moiety or a moiety with SMARTS notation c[CH3] were included, the fit was poor (high r^2 and low s); and if $\log S_w$ was used as a descriptor instead of $\log S_0$, r^2 decreased to 0.36. These results are difficult to explain but may be due to the unreliability or ambiguity of results predicted by read-across, as discussed by Benfenati et al. [24].

4.3 Step 2: QSARs for Nonpolar and Polar Narcotic Chemicals

Thirty-seven (12%) of the 309 chemicals analyzed in this case study were categorized as either nonpolar narcotic (Class 1, $n = 25$) or polar narcotic (Class 2, $n = 12$) and were dealt with a **Step 2**.

For the Class 1 chemicals, the correlation between 48-h algal toxicity and $\log P(1)$ and the multiple linear regression model involving HF and $\log P(1)$ took the following forms:

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.71 \log P(1) - 1.16, \\ 0.72 \leq \log P(1) &\leq 5.85, \\ n = 25, r^2 &= 0.83, r^2_{\text{adj}} = 0.83, q^2_{\text{LOO}} = 0.80, s = 0.51, \text{RMSE} = 0.49; \end{aligned} \quad (5)$$

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.68 \log P(1) + 0.002 \text{HF} - 1.02, \\ 0.72 \leq \log P(1) &\leq 5.85, -112.7 \leq \text{HF} \leq 78.3, \\ n = 25, r^2 &= 0.84, r^2_{\text{adj}} = 0.82, q^2_{\text{LOO}} = 0.80, s = 0.51, \text{RMSE} = 0.48. \end{aligned} \quad (6)$$

For the Class 2 chemicals, the correlation between 48-h algal toxicity and $\log P(1)$ and the multiple linear regression model involving HF and $\log P(1)$ were given by Eqs. 7 and 8, respectively:

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.64 \log P(1) - 0.38, \\ 1.17 \leq \log P(1) &\leq 4.77, \\ n = 12, r^2 &= 0.78, r^2_{\text{adj}} = 0.76, q^2_{\text{LOO}} = 0.68, s = 0.39, \text{RMSE} = 0.36; \end{aligned} \quad (7)$$

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.53 \log P(1) - 0.009 \text{HF} - 0.04, \\ 1.17 \leq \log P(1) &\leq 4.77, -22.1 \leq \text{HF} \leq 37.8, \\ n = 12, r^2 &= 0.80, r^2_{\text{adj}} = 0.76, q^2_{\text{LOO}} = 0.68, s = 0.40, \text{RMSE} = 0.34. \end{aligned} \quad (8)$$

Notably, the introduction of the quantum mechanical descriptor HF slightly decreased the correlations (r^2_{adj}) and increased the standard errors (s) (compare Eq. 6 with Eq. 5 and Eq. 8 with Eq. 7). For both Class 1 and Class 2 chemicals, $\log P(1)$ correlated well with algal toxicity, and HF could be ignored or over-fitted.

The QSAR for the Class 1 chemicals in the original work on the three-step strategy (Eq. 5 of Table 3 in ref. 15) and the statistical values for the prediction, which were estimated by using the test set (i.e., the Class 1 chemicals in the data set of Kusk et al.), were as follows:

$$\begin{aligned} \log(1/72\text{h-algal EC}_{50} [\text{mM}]) &= 0.95 \log P(1) - 1.86, \\ 1.72 \leq \log P(1) &\leq 5.85, \\ n = 59, r^2 &= 0.82, r^2_{\text{adj}} = 0.82, q^2_{\text{LOO}} = 0.81, s = 0.35, \\ n_{\text{test}} &= 25, Q^2_{(\text{F1})} = 0.79, Q^2_{(\text{F2})} = 0.73, \text{RMSEP} = 0.62. \end{aligned} \quad (9)$$

Note that $Q^2_{(\text{F1})}$ and $Q^2_{(\text{F2})}$ for Eq. 9 were >0.7 ; that is, they satisfied the predictivity criterion specified above. The $\log P(1)$ values for 6 of the 25 chemicals in the test set were outside the descriptor range specified for Eq. 9 ($1.72 \leq \log P(1) \leq 5.85$): dichloromethane; 2-methyl-2-propanol; methanone, dicyclopropyl-; trichloroethanol; 1-propene,3-ethoxy-; and 2-hexanone. When these six chemicals were omitted ($n_{\text{test}} = 19$), $Q^2_{(\text{F1})}$, $Q^2_{(\text{F2})}$, and RMSEP were 0.75, 0.74, and 0.54, respectively. These $Q^2_{(\text{F1})}$ and $Q^2_{(\text{F2})}$ values also satisfied the >0.7 predictivity criterion. However, when only 1-propene,3-ethoxy- ($\log P(1) = 1.3$) was omitted, the RMSEP decreased significantly; the measured 48-h algal EC₅₀ of this compound is reported to be

6.7 mg/L, whereas the EC₅₀ predicted by Eq. 9 was 363 mg/L. The predictivity for Class 1 chemicals other than 1-propene,3-ethoxy- was not reduced if the hydrophobicity range for Class 1 compounds ($0 < \log P(1) < 6$) was used. Additionally, when 1-propene,3-ethoxy- was omitted from the test set, the resulting QSAR took the form

$$\begin{aligned} \log (1/48\text{-h-algal EC}_{50} [\text{mM}]) &= 0.75 \log P(1) - 1.35, \\ 0.72 \leq \log P(1) &\leq 5.85, \\ n = 24, r^2 &= 0.89, r^2_{\text{adj}} = 0.88, q^2_{\text{LOO}} = 0.87, s = 0.43, \text{RMSE} = 0.41. \end{aligned} \quad (10)$$

The statistical values related to goodness of fit and robustness for Eq. 10 were better than those for Eq. 5. Moreover, the coefficient for log $P(1)$ for Eq. 10 was higher than that for Eq. 5, and the intercept was lower; that is, the shape of Eq. 10 became more similar to that of typical nonpolar narcotic QSAR models with a log P descriptor: the hydrophobicity coefficients and the intercepts of simple linear regression models generated with 48-h and 96-h *Pseudokirchneriella subcapitata* toxicity data for nonpolar narcotic chemicals are usually >0.9 and <-1.5 , respectively [51–53]. A typical example is the model developed by Fu et al. [54] for *Pseudokirchneriella subcapitata* data, which uses a hydrophobicity parameter (calculated log P from KOWWIN, log K_{ow} , in the EPI Suite [55] rather than Consensus LogP) and data for a variety of endpoints, such as 48-h growth rate and yield and 72-h and 96-h growth rates:

$$\begin{aligned} \log(1/\text{algal EC}_{50} [\text{mM}]) &= 0.940 \log P - 1.85, \\ n = 76, r^2 &= 0.93, s = 0.34, F = 942, \end{aligned} \quad (11)$$

where F is Fisher's criterion. The units of toxicity— $\log (1/\text{algal EC}_{50} [\text{M}])$, not $\log (1/\text{EC}_{50} [\text{mM}])$ —and the notation for hydrophobicity, $\log K_{\text{ow}}$, not $\log P$, and other statistical values in the original paper [54] were changed for comparison of Eqs. 9 and 11. In Eq. 9, the hydrophobicity coefficient and the intercept are almost identical to those in Eq. 9 (the equation for the Class 1 chemicals in the original work on the three-step strategy). Because the reported nonpolar QSAR models (e.g., Eq. 11) were stable, QSARs for predicting the *Pseudokirchneriella subcapitata* toxicity of the Class 1 (nonpolar narcotic) chemicals could be used to screen untested chemicals.

Similarly, the QSAR models for Class 2 chemicals in the original work on the three-step strategy (Eqs. 7 and 8 in Table 3 of ref. 15), and the statistical values for the prediction, were as follows:

$$\begin{aligned} \log(1/72\text{h-algal EC}_{50} [\text{mM}]) &= 0.95 \log P(1) - 0.96, \\ 1.17 \leq \log P(1) &\leq 4.31, \\ n = 29, r^2 &= 0.71, r^2_{\text{adj}} = 0.70, q^2_{\text{LOO}} = 0.65, s = 0.49, \\ n_{\text{test}} &= 12, Q^2_{(\text{F1})} = 0.53, Q^2_{(\text{F2})} = 0.47, \text{RMSEP} = 0.56, \end{aligned} \quad (12)$$

$$\begin{aligned} \log(1/72\text{h-algal EC}_{50} [\text{mM}]) &= 0.86 \log P(1) + 0.016 \text{HF} - 1.03, \\ 1.17 \leq \log P(1) &\leq 4.31, -52.58 < \text{HF} < 76.73, \\ n = 29, r^2 &= 0.86, r^2_{\text{adj}} = 0.85, q^2_{\text{LOO}} = 0.84, s = 0.35, \\ n_{\text{test}} &= 12, Q^2_{(\text{F1})} = 0.67, Q^2_{(\text{F2})} = 0.63, \text{RMSEP} = 0.47. \end{aligned} \quad (13)$$

The $Q^2_{(\text{F1})}$ for Eq. 12 and the $Q^2_{(\text{F1})}$ and $Q^2_{(\text{F2})}$ values for Eq. 13 satisfied the predictivity criterion specified by Roy et al. [47] (>0.5) but not the criterion specified by Golbraikh and Tropsha (>0.7) [48]. The $Q^2_{(\text{F2})}$ for Eq. 12 was <0.5 , and the $Q^2_{(\text{F2})}$ and RMSEP values for Eq. 12 were smaller and larger than those for Eq. 13, which indicates that for the Class 2 chemicals, the predictivity of the simple linear regression model was poorer than that of the multiple regression model. When a test set chemical with a $\log P(1)$ value outside the range for Eqs. 12 and 13 (*n*-heptylaniline, $\log P(1) = 4.56$) was omitted ($n_{\text{test}} = 11$), $Q^2_{(\text{F1})}$, $Q^2_{(\text{F2})}$, and RMSEP for Eq. 12 were 0.50, 0.34, and 0.55, respectively; and the corresponding values for Eq. 13 were 0.61, 0.47, and 0.49. These values were worse than those for $n_{\text{test}} = 12$. That is, the predictivity of the model expressed by Eq. 12 with $n_{\text{test}} = 11$ was poorer than that with $n_{\text{test}} = 12$. Additionally, the use of the quantum chemical descriptor HF improved the goodness of fit and robustness of the QSAR in the original work on the three-step strategy (Eq. 13 herein). In contrast, comparison of Eqs. 7 and 8 herein suggests that HF can be ignored for the Class 2 QSAR model constructed with the data set of Kusk et al. Fu et al. [54] also evaluated correlations between algal toxicity and hydrophobicity using *Pseudokirchneriella subcapitata* toxicity data obtained at test durations of 48 and 72 h; differences in the test duration were found to affect the correlation for polar narcotic chemicals but not for nonpolar narcotic chemicals. Overall, the results described herein indicate that the Class 2 QSAR models in the original work on the three-step strategy (QSAR models based on 72-h algal toxicity tests) were not applicable to the 48-h algal toxicity data set of Kusk et al.

4.4 Step 3: Categorizations for Read-Across

Discrepancies between the 72-h algal toxicity data used in the original work on the three-step strategy and the 48-h algal toxicity data of Kusk et al. were revealed at the pre-screening stage and in the QSAR models for the Class 2 chemicals at **Step 2**. Therefore, for the 189 chemicals remaining at **Step 3**, new categories were proposed for the data set of Kusk et al.; these new categories were based in part on knowledge about the categories in the original work and in part on correlations between toxicity and $\log P$

(1) (hydrophobicity) or water solubility, which are applicable for read-across (or even for local QSARs).

Three types of new categories were introduced: Type A, Type B, and Type C (Table 2). Chemicals in the Type A categories have potentially high algal toxicity, and those in the Type B categories are chemicals for which a category approach with read-across is applicable. Chemicals in the Type C categories are somewhat toxic, but their toxicities could not be predicted by means of a simple descriptor. If data for some similar chemicals making a subcategory were available, read-across for chemicals in the Type C categories was possible in some cases. However, for some of the 48-h algal toxicity data of Kusk et al., the use of read-across with a simple descriptor or categories (without considering mechanisms and modes of action) was questionable. Notably, the chemical categories defined by the structural alerts in the section on pre-screening might be useful for read-across. For example, in the original work on the three-step strategy, chemicals having a thiol group (SA 1) were assigned to a read-across category, and a strong correlation was observed between their log (1/72-h algal EC₅₀ [mM]) and log P(1) values ($R^2 = 0.92$, $n = 5$; Eq. 10 in Table 3 of ref. 15). Additionally, Yamamoto et al. showed that for parabens, there is a clear correlation between log (1/72-h algal EC₅₀ [mM]) and log D at pH 7 ($R^2 = 0.75$, $n = 7$; Fig. 2 in ref. [40]).

Among the outliers discussed in the section on pre-screening, the parabens and the tetracycline antibiotics (Table 1) were potentially highly toxic. Some of the chemicals categorized as *Pesticides* showed low toxicity, but 60% of the pesticides had EC₅₀ values of <1 mg/L (“very toxic”), as indicated in the abstract of ref. [16]. In this case study, parabens, tetracycline antibiotics, and pesticides were designated as Type A.

The 48-h algal toxicities of the Type B chemicals can be used to read-across under the structure-based categories described in Table 2. For the Type B chemicals that would not be ionic at pH 10, log P(1) was used as a descriptor. In contrast, because some of the chemicals in the *Combined chemicals* Type B category (e.g., ammonium compounds) would be partially or fully ionic at pH 10, water solubility was introduced as a descriptor for those compounds.

The simple linear regression model for compounds in the *Aliphatic chemicals* Type B category took the form

$$\log(1/48\text{h-algal EC}_{50} \text{ [mM]}) = 0.56 \log P(1) - 0.82, \\ 1.17 \leq \log P(1) \leq 5.59, \\ n = 14, r^2 = 0.72, r^2_{\text{adj}} = 0.70, q^2_{\text{LOO}} = 0.44, s = 0.53, \text{RMSE} = 0.49. \quad (14)$$

The low robustness ($q^2_{\text{LOO}} < 0.5$) arose from the high residual between the predicted and measured values for *N,N*-diethyldodecanamide. However, when this chemical was omitted, the statistical

Table 2

List of Types A, B, and C categories for the 189 Step 3 chemicals from the 48-h algal toxicity data set of Kusk et al.^a

Type A (<i>n</i> = 55): Identified as potentially highly toxic at the pre-screening stage
Parabens (<i>n</i> = 3)
Tetracycline antibiotics (<i>n</i> = 2)
Pesticides (<i>n</i> = 50)
Notes: All the parabens and tetracycline antibiotics had EC ₅₀ values of <1.0 mg/L. Among the 50 pesticides, 30 had EC ₅₀ values of <1.0 mg/L. Four of the pesticides showed low toxicity (EC ₅₀ >50 mg/L and log (1/EC ₅₀ [mM]) <1): 2,2'-oxybis[1-chloropropane] (a nematicide, EC ₅₀ = 221 mg/L); 1-propanone,1,1,1,3,3,3-hexachloro- (a herbicide, 99 mg/L); triclopyr (a herbicide, 201 mg/L); and cyromazine (an insecticide, 47 mg/L)
Type B (<i>n</i> = 81): Showed good correlation between toxicity and hydrophobicity (or solubility), suitable for read-across
Aliphatic chemicals (<i>n</i> = 14)
Anilines (<i>n</i> = 14)
Aromatic chemicals (<i>n</i> = 32)
Combined chemicals (<i>n</i> = 21)
Notes: Aliphatic chemicals included simple aliphatic aldehydes, aliphatic amides, aliphatic ketones, aliphatic aldehydes, aliphatic sulfides, aliphatic esters, aliphatic vinyl ester, and aliphatic oxiranes. None of these compounds had aromatic substructures or substructures that were included in other categories. Nine chemicals had EC ₅₀ values of >50 mg/L, indicating that they were not highly toxic. Only dodecanamide, <i>N,N</i> -diethyl- (log <i>P</i> (1) = 5.59) had an EC ₅₀ of <1.0 mg/L Anilines included compounds with aromatic primary amine substructures (indicated by SMARTS notation c[Nv3Hv2]) but excluded anthraquinones. Some of these chemicals were very toxic (four chemicals had EC ₅₀ values of <1.0 mg/L) Aromatic chemicals included all aromatic compounds except those defined as Anilines or Combined chemicals. Five of these chemicals had EC ₅₀ values of <1.0 mg/L Combined chemicals included chemicals with SA 1, aliphatic and aromatic hydrazines, ammonium compounds, compounds with bromine atoms directly attached to an aromatic ring (aromatic bromine, as indicated by the SMARTS notation [cBr]), aromatic acetamides with aliphatic chloride, anthraquinones, and aliphatic disulfides Eight of these chemicals had EC ₅₀ values of <1.0 mg/L (indicating high toxicity), but six of the chemicals had EC ₅₀ values of >50 mg/L. There were three chemicals with SA 1, and read-across might be possible for these chemicals
Type C (<i>n</i> = 53): Categories that were difficult to classify for read-across
SA 6 (excluding pesticides) (<i>n</i> = 23)
Other aliphatic chemicals (<i>n</i> = 13)
Nitrogen-containing chemicals (<i>n</i> = 17)
Notes: Twenty-three chemicals had SA 6 but were not pesticides; only 3 of these 23 chemicals had EC ₅₀ values of <1.0 mg/L Other aliphatic chemicals consisted of chemicals with SA 3, aliphatic chlorides, aliphatic iodides, and aliphatic vinyl ketones. Six of these chemicals were very toxic (EC ₅₀ <1.0 mg/L), but they did not fall into well-defined categories Nitrogen-containing chemicals consisted of aliphatic and aromatic isothiocyanates, aromatic nitriles (with and without aromatic chloride), aliphatic and aromatic nitroso compounds, and nitrobenzenes. Seven of these chemicals had EC ₅₀ values of <1.0 mg/L, and five had EC ₅₀ values of >50 mg/L

^aEC₅₀ indicates 50% effective concentration for 48-h algae growth inhibition toxicity against *Pseudokirchneriella subcapitata*

values (including the correlation coefficient) deteriorated, and the upper limit of the $\log P(1)$ range decreased from 5.59 to 4.07:

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.34 \log P(1) - 0.45, \\ 1.17 \leq \log P(1) &\leq 4.07, \\ n = 13, r^2 &= 0.56, r^2_{\text{adj}} = 0.52, q^2_{\text{LOO}} = 0.53, s = 0.37, \text{RMSE} = 0.34. \end{aligned} \quad (15)$$

Equation 15 was not suitable as a model for quantitative prediction. However, the *Aliphatic chemicals* category could be redefined as a category consisting of chemicals with low toxicity ($\text{EC}_{50} \geq 7.0 \text{ mg/L}$) when the category was restricted to chemicals for which $\log P(1)$ was <5 .

The simple linear regression model for the *Anilines* Type B category of chemicals was as follows:

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.55 \log P(1) + 0.05, \\ 0.13 \leq \log P(1) &\leq 7.29, \\ n = 14, r^2 &= 0.85, r^2_{\text{adj}} = 0.83, q^2_{\text{LOO}} = 0.77, s = 0.47, \text{RMSE} = 0.44. \end{aligned} \quad (16)$$

Both the robustness (q^2_{LOO}) and the goodness of fit (r^2 and s) for Eq. 16 were better than those for Eq. 15. This result indicates that *Anilines* is a reliable category if applicability domains other than $\log P(1)$ range are considered. The *Anilines* category is defined simply as including anilines (aromatic primary amines) other than anthraquinones and excluding chemicals that are ionic at pH 10. Read-across (or even local QSAR) using the *Anilines* category and Eq. 16 should be applied to a chemical if its substructures do not overlap those that define the *Combined chemicals* category and if the chemical does not fall into one of the Type A or Type C categories.

The simple linear regression model for compounds in the *Aromatic chemicals* Type B category was given by

$$\begin{aligned} \log(1/48\text{h-algal EC}_{50} [\text{mM}]) &= 0.92 \log P(1) - 1.30, \\ 0.90 \leq \log P(1) &\leq 6.32, \\ n = 32, r^2 &= 0.78, r^2_{\text{adj}} = 0.77, q^2_{\text{LOO}} = 0.74, s = 0.59, \text{RMSE} = 0.57. \end{aligned} \quad (17)$$

Compared with the nonpolar narcotic QSAR model (discussed for the Class 1 chemicals at **Step 2**), Eq. 17 had a similar slope (~0.9) and an intercept that was ~0.5 log unit higher. This comparison indicates that aromatic compounds—other than those that have certain specific structures (such as structures indicated by the SMARTS notation c[Nv3H2], which fall into the *Anilines* category), or that have specific modes or mechanisms of action (such as pesticides)—would show algal toxicity ~0.5 log unit higher than the trend for nonpolar narcotic chemicals in the data set of Kusk et al.

Finally, the simple linear regression for chemicals in the *Combined chemicals* Type B category was given by

$$\log(1/48\text{h-algal EC}_{50} \text{ [mM]}) = -0.70 \log S_0 - 0.28, \\ -6.32 \leq \log S_0 \leq 0.40, \\ n = 21, r^2 = 0.75, r^2_{\text{adj}} = 0.74, q^2_{\text{LOO}} = 0.71, s = 0.79, \text{RMSE} = 0.75. \quad (18)$$

The goodness of fit and robustness of Eq. 18 were not as high as those for the Class 1 chemicals at **Step 2** or for the chemicals in the *Anilines* category. Nevertheless, for situations in which chemicals have various functional groups and similar chemicals are not easily defined, the *Combined chemicals* category can be used for including and excluding chemicals when applying read-across to a target chemical.

The general features of *SA 6 (excluding pesticides)*, *Other aliphatic chemicals*, or *Nitrogen-containing chemicals* (Type C; Table 2) categories were difficult to describe on the basis of a simple descriptor such as hydrophobicity. In addition to compounds with *SA 6 (excluding pesticides)*, compounds categorized as *Other aliphatic chemicals* and *Nitrogen-containing chemicals* could be described in terms of specific properties, e.g., protein binding, and read-across could be used to predict their algal toxicity. Chemicals in the Type C categories in this case study would have been designated as *Uncategorized* in the original work on the three-step strategy [15]. Generally, reactivity such as protein binding [56], the presence of toxic substructures described in the literature [57], or modes or mechanisms of toxic action [8–12] may be useful for generating categories applicable for read-across.

5 Conclusions

This case study of an algal toxicity data set obtained by Kusk et al. [16] revealed problems with generalizing a previously reported three-step strategy [15] for predicting algae growth inhibition toxicity. The lack of the measured acute *Daphnia magna* toxicity data and the unreliability of some of the available data hindered toxicity prediction by means of the interspecies QSAR used at **Step 1** of the three-step strategy; alternative strategies such as read-across could be used to circumvent this problem. At **Step 2** in this case study, new QSAR models were introduced for the data set of Kusk et al. Unlike the case for Class 1 (nonpolar narcotic) chemicals, in the case of the Class 2 (polar narcotic) chemicals, model reevaluation was necessary if test conditions used to obtain data for algal toxicity prediction differed from the conditions used in standard TGs. At **Step 3** of the case study, the categories used for read-across were grouped into three types for application to the data set of Kusk et al.

Although the combination of QSAR and read-across used in the three-step strategy was effective for predicting algal toxicity, whether or not previously developed QSARs (except that for

nonpolar narcotic chemicals) and read-across categories depended on whether the test conditions, such as test duration, deviated from those in standard TGs.

6 Note: Information

Data used in this chapter are available at <https://doi.org/10.6084/m9.figshare.8107646.v1> in the Figshare database. Table I lists the 309 chemicals (as named in ref. [16]) and the modified SMILES strings used to calculate their physicochemical properties ($\log P(1)$, $\log D_{\text{pH}10}$, $\log S_0$, and $\log S_w$) with the ACD/Labs [30] and ACD/LogD software [31]. Tables II, III, and IV list the $\log (1/48\text{-h algal EC}_{50} [\text{mM}])$ values, the physicochemical properties, and other data for the chemicals dealt with at **Steps 1, 2, and 3**, respectively.

The SMARTS notation examples used here are available at http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html.

Acknowledgments

The authors thank Drs. Y. Aoki, T.I. Hayashi, and H. Yamamoto, Professor N. Tatarazako, and Mr. K. Hasunuma for their helpful discussions about the interspecies QSAR and about the three-step strategy.

References

- OECD (2013) OECD guidelines for testing of chemicals. Test no. 210: fish, early-life stage toxicity test. OECD, Paris
- OECD (2012) OECD guidelines for testing of chemicals. Test no. 211: *Daphnia magna* reproduction test. OECD, Paris
- OECD (2011) OECD guidelines for testing of chemicals. Test no. 201: Freshwater alga and cyanobacteria, growth inhibition test. OECD, Paris
- OECD (2004) OECD guidelines for testing of chemicals. Test no. 202: *Daphnia sp.* acute immobilization test. OECD, Paris
- OECD (1992) OECD guidelines for testing of chemicals. Test no. 203: fish acute toxicity test. OECD, Paris
- Klimisch HJ, Andrae M, Tillmann U (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 25:1–5
- OECD (2016) OECD series on testing and assessment No. 260, Guidance document for the use of adverse outcome pathways in developing integrated Approaches to Testing and Assessment (IATA). OECD, Paris
- Kienzler A, Barron MG, Belanger SE, Beasley A, Embry MR (2017) Mode of action (MOA) assignment classifications for ecotoxicology: an evaluation of approaches. *Environ Sci Technol* 51:10203–10211
- Kienzler A, Barron MG, Belanger SE, Beasley A, Embry MR (2017) Response to “Comment on ‘Mode of action (MOA) assignment classifications for ecotoxicology: an evaluation of approaches’”. *Environ Sci Technol* 51:13511–13512
- McCarty LS, Borgert CJ (2017) Comment on “Mode of action (MOA) assignment classifications for ecotoxicology: an evaluation of approaches”. *Environ Sci Technol* 51:13509–13510

11. Bauer FJ, Thomas PC, Fouchard SY, Neunlist SJM (2018) A new classification algorithm based on mechanisms of action. *Computat Toxicol* 5:8–15
12. Scholz S, Schreiber R, Armitage J, Mayer P, Escher BI, Lidzba A, Leonard M, Altenburger R (2018) Meta-analysis of fish early life stage tests-association of toxic ratios and acute-to-chronic ratios with modes of action. *Environ Toxicol Chem* 37:955–969
13. U.S. Environmental Protection Agency ECO-SAR. <http://www.epa.gov/tsc-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model>. Accessed 25 Jan 2016
14. OECD (2015) OECD series on testing and assessment, No. 229, Fundamental and guiding principles for (Q)SAR analysis of chemical carcinogens with mechanistic considerations. OECD, Paris
15. Furuhama A, Hasunuma K, Hayashi TI, Tatarazako N (2016) Predicting algal growth inhibition toxicity: three-step strategy using structural and physicochemical properties. *SAR QSAR Environ Res* 27:343–362
16. Kusk KO, Christensen AM, Nyholm N (2018) Algal growth inhibition test results of 425 organic chemical substances. *Chemosphere* 204:405–412
17. Furuhama A, Hasunuma K, Aoki Y (2015) Interspecies quantitative structure–activity relationships (QSARs) for eco-toxicity screening of chemicals: the role of physicochemical properties. *SAR QSAR Environ Res* 26:809–830
18. Furuhama A (2016) Corrigendum. *SAR QSAR Environ Res* 27:245–247
19. Verhaar HJM, van Leeuwen CJ, Hermens JLM (1992) Classifying environmental pollutants. 1: structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* 25:471–491
20. Verhaar HJM, Solbe J, Speksnijder J, van Leeuwen CJ, Hermens JLM (2000) Classifying environmental pollutants: part 3. External validation of the classification system. *Chemosphere* 40:875–883
21. Enoch SJ, Hewitt M, Cronin MTD, Azam S, Madden JC (2008) Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree. *Chemosphere* 73:243–248
22. Cronin MTD (2010) Chapter 18 Biological read-across: mechanically-based species-species and endpoint-endpoint extrapolations. In: Cronin MTD, Madden JC (eds) *In silico toxicology: principles and applications*. The Royal Society of Chemistry, Cambridge, pp 446–477
23. Cronin MTD, Netzeva TI, Dearden JC, Edwards R, Worgan ADP (2004) Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database. *Chem Res Toxicol* 17:545–554
24. Benfenati E, Roncaglioni A, Petoumenou MI, Cappelli CI, Gini G (2015) Integrating QSAR and read-across for environmental assessment. *SAR QSAR Environ Res* 26:605–618
25. International Organization for Standardization (1997) Water quality – fresh water algal growth test with *Scenedesmus subspicatus* and *Raphidocelis subcapitata*. ISO Standard 8692. Geneva
26. Fu L, Li JJ, Wang Y, Wang XH, Wen Y, Qin WC, Su LM, Zhao YH (2015) Evaluation of toxicity data to green algae and relationship with hydrophobicity. *Chemosphere* 120:16–22
27. Wang XH, Yu Y, Fu L, Tai HW, Qin WC, Su LM, Zhao YH (2016) Comparison of chemical toxicity to different algal species based on inter-species correlation, species sensitivity, and excess toxicity. *Clean (Weinh)* 44:803–808
28. Weininger D (1988) SMILES, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36
29. Daylight Chemical Information Systems Inc. Daylight theory manual, 4. SMARTS^R – A language for describing molecular patterns. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed 13 Jun 2019
30. ACD/Labs, version 2018. Advanced chemistry development, Inc., Toronto, ON, Canada
31. ACD/LogD, version 2018. Advanced chemistry development, Inc., Toronto, ON, Canada
32. The QSAR toolbox version 4.3. <https://qsartoolbox.org/>. Accessed 10 Apr 2019
33. Patlewicz G, Jeliazkova N, Safford RJ, Worth AP, Aleksiev B (2008) An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* 19:495–524
34. Toxtree. <http://toxtree.sourceforge.net>. Accessed 10 May 2019
35. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc* 107:3902–3909
36. Stewart JJP (1993) MOPAC 7. <http://openmopac.net/Downloads/Downloads.html>. Accessed 13 Jun 2019

37. Council BCP (2015) The pesticide manual 17th edition: a world compendium. British Crop Protection Council, Alton, Hampshire, UK
38. Council BCP (2015) More on the pesticide manual: view supplementary entries. British Crop Protection Council. http://www.bcpc.org/page_Supplementary-Entries_102.html. Accessed 12 Nov 2015
39. Błędzka D, Gromadzińska J, Wąsowicz W (2014) Parabens. From environmental studies to human health. Environ Int 67:27–42
40. Yamamoto H, Tamura I, Hirata Y, Kato J, Kagota K, Katsuki S, Yamamoto A, Kagami Y, Tatarazako N (2011) Aquatic toxicity and ecological risk assessment of seven parabens: individual and additive approach. Sci Total Environ 410–411:102–111
41. Yamamoto H, Nakamura Y, Nakamura Y, Kitani C, Imari T, Sekizawa J, Takao Y, Yamashita N, Hirai N, Oda S, Tatarazako N (2007) Initial ecological risk assessment of eight selected human pharmaceuticals in Japan. Environ Sci 14(4):177–193
42. Halling-Sørensen B (2000) Algal toxicity of antibacterial agents used in intensive farming. Chemosphere 40:731–739
43. Yang LH, Ying GG, Su HC, Stauber JL, Adams MS, Binet MT (2008) Growth-inhibiting effects of 12 antibacterial agents and their mixtures on the freshwater microalga *Pseudokirchneriella subcapitata*. Environ Toxicol Chem 27:1201–1208
44. Ji K, Kim S, Han S, Seo J, Lee S, Park Y, Choi K, Kho Y-L, Kim P-G, Park J, Choi K (2012) Risk assessment of chlortetracycline, oxytetracycline, sulfamethazine, sulfathiazole, and erythromycin in aquatic environment: are the current environmental concentrations safe? Ecotoxicology 21:2031–2050
45. Park S, Choi K (2008) Hazard assessment of commonly used agricultural antibiotics on aquatic ecosystems. Ecotoxicology 17:526–538
46. Wollenberger L, Halling-Sørensen B, Kusk KO (2000) Acute and chronic toxicity of veterinary antibiotics to *Daphnia magna*. Chemosphere 40:723–730
47. Roy K, Kar S, Das R (2015) Statistical methods in QSAR/QSPR. In: A primer on QSAR/QSPR modeling: fundamental concepts. Springer International Publishing, pp 37–59. <https://www.springer.com/gp/book/978331917280>
48. Golbraikh A, Tropsha A (2018) QSAR/QSPR revisited. In: Engel T, Gasteiger J (eds) Chemoinformatics: basic concepts and methods. Wiley, Weinheim, pp 465–495
49. KAshinhou Tool for Ecotoxicity (KATE) is an ecotoxicity prediction system that consists of QSAR models and was researched and developed under contract with the Ministry of the Environment, Government of Japan from fiscal year 2004 to fiscal year 2018 by the Center for Health and Environmental Risk Research of the National Institute for Environmental Studies. <https://kate.nies.go.jp/>. Accessed 29 Apr 2019
50. Golbamaki A, Cassano A, Lombardo A, Moggio Y, Colafranceschi M, Benfenati E (2014) Comparison of in silico models for prediction of *Daphnia magna* acute toxicity. SAR QSAR Environ Res 25:673–694
51. Hsieh S-H, Hsu C-H, Tsai D-Y, Chen C-Y (2006) Quantitative structure-activity relationships for toxicity of nonpolar narcotic chemicals to *Pseudokirchneriella subcapitata*. Environ Toxicol Chem 25:2920–2926
52. Tsai K-P, Chen C-Y (2007) An algal toxicity database of organic toxicants derived by a closed-system technique. Environ Toxicol Chem 26:1931–1939
53. Aruoja V, Moosus M, Kahru A, Sihtmae M, Maran U (2014) Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*. Chemosphere 96:23–32
54. Fu L, Huang T, Wang S, Wang X, Su L, Li C, Zhao Y (2017) Toxicity of 13 different antibiotics towards freshwater green algae *Pseudokirchneriella subcapitata* and their modes of action. Chemosphere 168:217–222
55. U.S. Environmental Protection Agency KOW-WINTM: Estimates the log octanol-water partition coefficient, log K_{ow}, of chemicals using an atom/fragment contribution method. <https://www.epa.gov/tsca-screening-tools/epi-suitem-estimation-program-interface>. Accessed 7 May 2019
56. OECD (2014) OECD series on testing and assessment, No. 194, Guidance on grouping of chemicals, 2nd edn. OECD, Paris
57. Von der Ohe PC, Kühne R, Ebert RU, Altenburger R, Liess M, Schüürmann G (2005) Structural alerts – a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. Chem Res Toxicol 18:536–555



Chapter 25

QSAR Approaches and Ecotoxicological Risk Assessment

Mabrouk Hamadache, Othmane Benkortbi, Abdeltif Amrane, and Salah Hanini

Abstract

Hundreds of thousands of chemicals that can affect human health or the quality of aquatic and terrestrial ecosystems are introduced directly or indirectly into the air, water, or soil. Therefore, the awareness of the serious and harmful effects caused by these chemical compounds has revealed the absolute and compelling need to resort to the evaluation of potential risks incurred as a result of exposure to these compounds. In the aim to provide a high level of protection for human, animal, and environmental health, many regulatory agencies have established strict legislation for both toxicological and ecotoxicological risk assessments of existing and new chemical compounds. To limit the *in vivo* experiments which are a tedious and costly practice and generate a large sacrifice of animals, the REACH regulation recommends the use of *in silico* methods, such as quantitative structure–activity relationship (QSAR) models.

Key words Ecosystems, Pollutants, Ecotoxicity, QSAR models, Adverse effects

1 Introduction

As a result of human activity over many decades, hundreds of thousands of chemicals that can affect human health or the quality of aquatic and terrestrial ecosystems are introduced directly or indirectly into the air, water or soil. In this regard, in recent decades, the number of research dedicated to pollution and the incidence of ecotoxic chemical compounds (perfumes, cosmetics, pharmaceuticals, ionic liquids, food and preservative colors, detergents, varnishes, paints, nanoparticles, illicit drugs, surfactants, plasticizers, pesticides, metals, solvents, etc.) on human health, living species and ecosystems is constantly increasing [1–7]. These chemical compounds used daily, in addition to contributing to the well-being and comfort of the human being, are nevertheless responsible for the decline in biodiversity [8] and the extinction of certain species [9–11].

The serious nuisances caused by these chemical compounds pose a serious threat to both humans and the ecosystem as a

whole. Therefore, the awareness of the serious and harmful effects of these chemicals has revealed the absolute and the compelling need to resort to the evaluation of potential risks incurred as a result of exposure to these compounds. In this regard, the interest in assessing the risks inherent in their production, use and fate in the environment has increased to such an extent that has become a growing area of scientific research [12]. As a result, many regulatory agencies have established strict legislation for both toxicological and ecotoxicological risk assessments of existing and new chemical compounds in the aim to provide a high level of protection for human, animal, and environmental health [13, 14].

The experimental evaluation of the toxicological and ecotoxicological risks of chemical substances takes place according to three types of approaches: in vivo experiments on animals, in vitro experiments using tissue culture cells, and in silico experiments refer to the use of the computer tool. In vivo and in vitro tests often require a high cost, relatively, a long time and a large number of laboratory animals. In addition, tests on animals are now considered ethically unacceptable by animal rights organizations. For all these reasons, and taking into consideration the significant number of chemicals entering the market daily, the need for fast, accurate and cost-effective assessments is more than a requirement. Also, the current trend is to use the computer tool (in silico approach) as a viable alternative to the classical methods of animal experiments. In fact, several organizations advocate and encourage the use of modeling for risk assessment. These organizations include the United States Environmental Protection Agency (US EPA), the REACH regulation (Registration, Evaluation, Authorization and Restriction of Chemicals) in Europe whose legislation came into effect in 2007, the European Center for Validation of Alternative Methods (ECVAM) of the European Union and the European Union Commission Scientific Committee on Toxicity, Ecotoxicity, and Environment (CSTEE).

The QSAR/QSPR (quantitative structure–activity relationship/quantitative structure–property relationship) approach is one of those many modeling methods that aim to assess the toxicological and ecotoxicological risks of chemical substances. This is a quantitative approach that aims to develop a QSAR/QSPR model to link quantified structural characteristics into a set of molecular descriptors or physicochemical properties of compounds such as toxicity property. Combined with genetic algorithms and statistical learning methods (e.g., artificial neural networks, support vector machines), this approach has proved effective and promising. For this, a large set of known structure and properties of interest of compound determined experimentally is necessary to develop the model. Once established, this model is subjected to several internal and external validation tests to evaluate its robustness and its power of prediction. For an effective mastery of the QSAR/QSPR tool,

reference books dealing with fundamental concepts of QSAR modeling and their basic concepts for applications in risk assessment are currently available in the literature [15–17].

This chapter deals with two aspects: the first is devoted to the generalized pollution of all environmental compartments and the proven or suspected adverse effects of chemical substances. The second aspect is devoted to the various QSAR/QSPR studies and results inherent in modeling applied to toxicology and ecotoxicology.

2 State of the Art on Pollution and Its Negative Impacts

Agricultural and industrial activities, waste disposal, oil spills, and acid rain are all factors that produce thousands of pollutants (chemicals and toxic gases, plastic waste, electronic waste, heavy metals, paints, rubber, waste oil, batteries, etc.). They represent not only a real threat that could pose serious health risks to all living systems (humans, bird populations, mammals, fish, aquatic invertebrates, and other species) but also contamination of all ecosystems (atmosphere, hydrosphere, lithosphere, and biosphere) [18–27]. The systematic review of the recently compiled scientific literature has edified us on the extent and importance of the studies inherent to this pollution (Fig. 1) and its environmental risk (Fig. 2). This part of the chapter is an overview of the general pollution of the environment and the harmful impacts of this pollution.

2.1 Ecosystem Pollution

The terrestrial ecosystem is an essential component of our environment. In addition to being a source of production of the majority of food for humans and animals, it is a source of raw materials and a reservoir of genes and species that ensure biodiversity. According to various studies around the world, it has been noticed that various types of soils, including cultivated fields, vegetable fields, and forest lands, are contaminated by a multitude of pollutants [28, 29]. For example, soil contamination by heavy metals, pesticides, and various sewage and other pollutants were the focus of the recent years. Thus, several recent studies have highlighted this contamination in the world, particularly in Italy [30, 31], Turkey [32, 33], Iran [34], Macedonia [35], Mexico [36], Russia [37], Laos [38], China [39–45], Kyrgyzstan [46], Saudi Arabia [47, 48], South Korea [49], Spain [50], Algeria [51–54], India [55, 56], and Germany [57].

The aquatic ecosystem as a whole (groundwater, surface water, and seawater) is a vital natural resource for humans, animals, plants, and aquatic organisms. Unfortunately, this ecosystem is drowned with all kinds of toxic pollutants such as chemicals, wastes, and radioactive substances. Because of the dangerous risks of this kind of pollution, it has become a source of major concern for the scientific community, which is constantly reporting the degradation

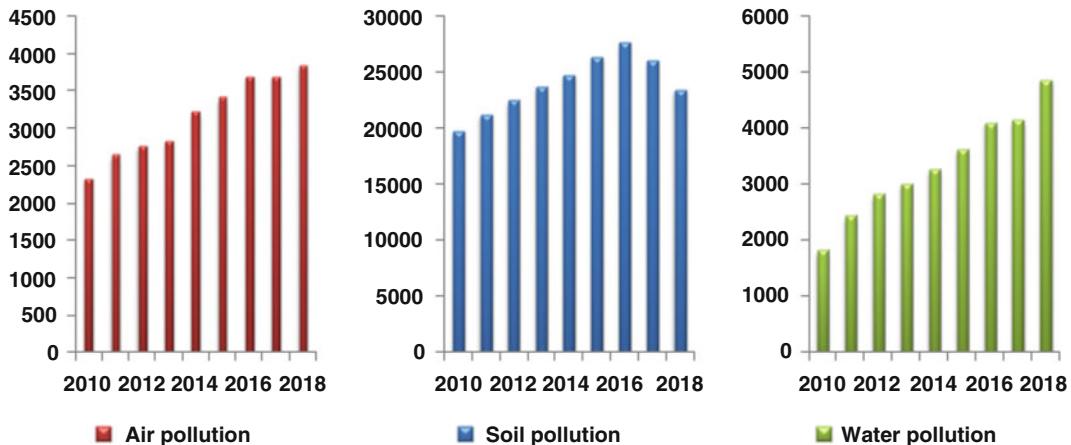


Fig. 1 Number of publications dealing with air, soil, and water pollution from 2010 to 2018. (PubMed accessed on March 17, 2019)

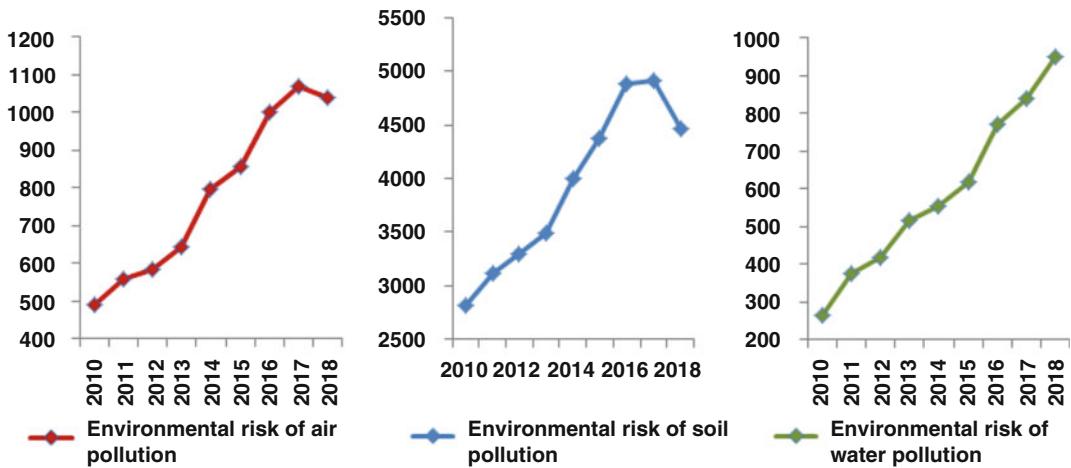


Fig. 2 Number of publications dedicated to the environmental risk of air, soil, and water pollution from 2010 to 2018. (Access to PubMed on March 17, 2019)

of the aquatic environment. Examples of heavy metal pollution are reported by numerous authors [58–71]. In addition, other recent works have been dedicated to water pollution by pesticides, especially in Ecuador [72], China [73, 74], Egypt [75], England [76], Vietnam [77], Brazil [78], France [79], Europe and the USA [80], Guadeloupe [81], Spain [82], and Greece [83]. The literature on the pollution of aquatic systems by various sewage and other pollutants is also abundant. Among the most recent studies, we can mention the cases of water pollution in Tunisia [84], Sri Lanka [73], Pakistan [85], Kenya [86], and Brazil [87]. The risk of pollution by plastic microphones, ionic liquids, and pharmaceuticals has also been widely reported [88–93].

In addition to soil and water pollution, air pollution is also a global and growing problem [94–97]. According to DuPont [98], 92% of the world's population suffers from poor air quality, while in sub-Saharan Africa nearly half a million people die each year [99]. This pollution is the consequence of the presence of complex mixture of gaseous components, as well as solid and liquid particles suspended in the air. These pollutants include particles from the combustion of coal, carbon monoxide, sulfur dioxide, nitrogen dioxide, and heavy metals such as cadmium (Cd), lead (Pb), and mercury (Hg) [100–102].

2.2 Adverse Effects of Pollutants

In recent years, several studies have shown that ecosystem pollution is a recognized causal factor in many diseases and adverse effects on humans, animals, and plants for food [64]. According to Basu [96], pollution is responsible for more deaths than smoking, hunger, and natural disasters. In addition, the bibliographic review has edified us on the proven or suspected effects of the different pollutants. It has been reported that exposure to heavy metals has a significant impact on the development of health problems such as cancer [95, 103, 104], cardiovascular diseases [105], Parkinson's disease [106], chronic kidney disease [107–109], hypertension, gastrointestinal disorders [42, 110], bone degeneration, and lesions of the liver and lungs [64, 111]. Moreover, the effects of certain metals on plant processes such as the decline of seed germination and stunted root growth have been reported [112].

The harmful effects of pesticides are also similar to those of heavy metals. According to Yadav [22], about 200,000 people die and about three million are poisoned every year by pesticides around the world. Many studies have reported that exposure to certain pesticides is linked to the risk of various diseases, including Parkinson's disease [113, 114], Alzheimer's disease [115], congenital anomalies [116], hormonal disorders, cancers [117, 118], diabetes, and obesity [119]. In addition, other studies have been conducted on the impacts of pesticides on pollinators [10, 120, 121], birds [122], and aquatic organisms [123].

Due to their toxicity and their persistent and bioaccumulative nature, dust and coal residues cause adverse effects such as hypertension, headaches, irritability, abdominal pain, nerve damage, skeletal problems, pulmonary, hepatic and renal problems, anemia, intellectual impairments, fatal cardiac arrest, and carcinogenesis [102].

3 Review of Literature on QSAR Ecotoxicity Modeling

As mentioned above, it is of utmost importance to evaluate the toxicity of pollutants of all kinds. This assessment must be focused jointly on the risks for humans but also on the undesirable

implications of these pollutants for different species of the ecosystem. On the other hand, as noted above, it is almost impossible to test all these pollutants at the laboratory level. Also, among the computer techniques used in ecotoxicology, QSAR methodologies are an encouraging alternative in that they allow, from a limited number of available data for compounds, to predict the toxicity of several other compounds without resorting to animal testing, on the sole condition that they fall within the same field of applicability. In this part of the chapter, only those ecotoxicology QSAR models that have been published in the last decade are considered. These models are classified according to the nature of the pollutants.

3.1 QSAR Models for Ecotoxicity Prediction of Pesticides

Given their presence in all components of the environment, pesticides are among the pollutants for which risk assessment is a requirement and a top priority. Thus, a large number of QSAR models have been developed for the purpose of predicting the toxicity of pesticides to all living species in ecosystems. In this bibliographic review, the most recent ones are quoted.

Can et al. [124] proposed a quantitative structure-toxicity relationship (QSTR) model for estimating the acute oral toxicity of herbicides to male rats. The LD₅₀ (in mmol/kg) of 20 herbicides from the formation set and 7 others reserved for external validation were described using 4 descriptors (lipophilic character, polarity, molecular geometry, and a quantum chemistry descriptor). The development of the model to predict the toxicity of sulfonylurea and phenylurea herbicides was performed using multiple linear regressions. The statistical parameters of the model such as the MSE and the coefficient R² were, respectively, equal to 0.041 and 0.93. The authors estimated that lipophilic character, dipole moment, molar refractivity, and molecular weight are effective parameters that describe the toxicity of these substances. Also, any new herbicide of this family must have the following characteristics to be the least toxic possible: it should be highly polar and soluble in water and have a low refractivity and also a low molar mass.

In another study, *Hamadache et al.* [114] have developed a validated QSAR model to predict acute oral toxicity of 329 pesticides to rats. This QSAR model was based on 17 molecular descriptors; it was shown to be robust, externally predictive and characterized by a good applicability domain. The best results were obtained with a 17/9/1 artificial neural network model trained with the quasi-Newton back propagation (BFGS) algorithm. The prediction accuracy for the external validation set was estimated by the Q²_{ext} and the root mean square (RMS) error, which were equal to 0.948 and 0.201, respectively. 98.6% of external validation set was correctly predicted, and the model proved to be superior to models previously published. A validated QSAR model to predict contact acute toxicity (LD₅₀) of 111 pesticides

to bees was developed by *Hamadache et al.* [10]. The QSAR model was assessed according to the OECD principles for the validation of QSAR models. The calculated values for the internal and external validation statistic parameters (Q^2 and r^2_m) were greater than 0.85. In addition to this validation, a mathematical equation derived from the ANN model was used to predict the LD₅₀ of 20 other pesticides. A good correlation between predicted and experimental values was found ($R^2 = 0.97$ and RMSE = 0.14). As a result, this equation could be a means of predicting the toxicity of new pesticides.

In a study undertaken by *Basant et al.* [125], tree-based multi-species QSAR models were constructed for predicting the avian toxicity. A set of 4768 pesticides and a set of 9 descriptors derived directly from the chemical structures and following the OECD guidelines were used. The QSAR models (SDT, DTF, and DTB) were externally validated using the toxicity data in four other test species (mallard duck, ring-necked pheasant, Japanese quail, house sparrow). The external predictive power of the QSAR models was tested through rigorous validation deriving a wide series of statistical checks. The S36 and MW were the most influential descriptors identified by DTF and DTB models. The DTF and DTB performed better than the SDT model and yielded a correlation (R^2) of 0.945 and 0.966 between the measured and predicted toxicity values in test data array. The same authors [126] established local and global QSTR and ISC QSAAR (interspecies correlation quantitative structure activity–activity relationship) models for predicting the toxicities of 3767 pesticides in multiple aquatic test species using the toxicity data in crustacean (*Daphnia magna*, *Americanopsis bahia*, *Gammarus fasciatus*, and *Penaeus duorarum*) and fish (*Oncorhynchus mykiss* and *Lepomis macrochirus*) species in accordance with the OECD guidelines. Furthermore, the chemical applicability domains of these QSTR/QSAAR models were determined using the leverage and standardization approaches. The constructed local, global, and interspecies QSAAR models yielded high correlations (R^2) of >0.941, >0.943, and >0.826, respectively, between the measured and model predicted endpoint toxicity values in the test data. The authors concluded that the developed QSTR/QSAAR models are appropriate since they reliably predict the aquatic toxicity of structurally diverse pesticides in multiple test species and can be used for the screening and prioritization of new pesticides.

Recently, by using nine molecular fingerprints to describe pesticides, binary and ternary classification models were constructed by Sun et al. [127] to predict aquatic toxicity of pesticides via six machine learning methods, namely, naïve Bayes (NB), artificial neural network (ANN), k-nearest neighbor (kNN), classification tree (CT), Random Forest (RF), and Support Vector Machine (SVM). For the binary models, local models were obtained with

829 pesticides on rainbow trout (RT) and 151 pesticides on *Lepomis* (LP), and global models were constructed on the basis of 1258 diverse pesticides on RT, LP, and 278 other fish species. The 1258 pesticides were also used to build global ternary models. The best local binary models were Maccs_ANN for RT and Maccs_SVM for LP, which both exhibited an accuracy of 0.90. For global binary models, the best one was Graph_SVM with an accuracy of 0.89. Accuracy of the best global ternary model Graph_SVM was 0.81, which was slightly lower than that of the best global binary model. The authors suggested that this study provides a useful tool for an early evaluation of pesticides aquatic toxicity in environmental risk assessment. In the same year, *Toropov* [128] used optimal descriptors to establish QSAR models to predict acute toxicity of pesticides toward rainbow trout. A heterogeneous set of pesticides ($n = 116$) was considered, taken from the EFSA's Chemical Hazards Database: OpenFoodTox. The statistical characteristics of these models were the following: (1) for training set, correlation coefficient (R^2) was in the range 0.72–0.81, and root mean square error (RMSE) ranges 0.54–1.25; and (2) for external (validation) set, the ranges of values were 0.74–0.84 for R^2 and 0.64–0.75 for RMSE. Computational experiments have shown that the presence of chlorine, fluorine, sulfurs, and aromatic fragments induce an increase of the toxicity.

More recently, *Qin et al.* [129] developed a QSAR model for the toxicities (half-effect concentration, EC50) of 45 binary and multicomponent mixtures composed of 2 antibiotics and 4 pesticides. The acute toxicities of single compound and mixtures toward *Aliivibrio fischeri* were tested. A genetic algorithm was used to obtain the optimized model with three theoretical descriptors. Various internal and external validation techniques led to a coefficient of determination of 0.9366 and a root mean square error of 0.1345; the QSAR model predicted that 45 mixture toxicities presented additive, synergistic, and antagonistic effects. Compared with the traditional models, the QSAR model exhibited an advantage in predicting mixture toxicity. Thus, the presented approach may be able to fill the gaps in predicting nonadditive toxicities of binary and multicomponent mixtures. More recently, QSAR models for *Daphnia magna* toxicities of different classes of agrochemicals (fungicides, herbicides, insecticides, and microbiocides) individually as well as for the combined set with the application of Organization for Economic Cooperation and Development (OECD) recommended guidelines were suggested by *Khan et al.* [130]. The models were generated employing only simple and interpretable two-dimensional descriptors and subsequently strictly validated using test set compounds. All the individual models of different classes of agrochemicals as well as the global set of agrochemicals showed encouraging statistical quality and prediction ability. The general observations suggest that the toxicity increases

with lipophilicity and decreases with polarity. According to the authors, the generated models should be applicable for data gap filling for new or untested agrochemical compounds falling within the applicability domain of the developed models. Note that other recent QSAR models not detailed in this review are reported in the literature [14, 131, 132].

3.2 QSAR Models for Ecotoxicity Prediction of Ionic Liquids

Although they have interesting properties that meet industrial requirements, ionic liquids are characterized by their toxic properties throughout the living ecosystem. In recent years, the use of QSAR approaches for the prediction of the toxicity of these potential pollutants has been the subject of scientific publications.

Roy and Das [133] have developed predictive classification and regression models correlating the structurally derived chemical information of a group of 62 diverse ionic liquids (ILs) with their toxicity toward *Daphnia magna* and their interpretation. They have principally used the extended topochemical atom (ETA) indices along with various topological non-ETA and thermodynamic parameters as independent variables. The developed models have been subjected to multiple validation strategies. According to the results obtained, the lipophilicity, branching pattern, electronegativity, and chain length of the cationic substituents play a major role in ecotoxicity of ionic liquids toward *D. magna*. The authors concluded that this information can be successfully used to design better ionic liquid analogues acquiring the qualities of a true eco-friendly green chemical. In another study, Roy *et al.* [134] have suggested statistical models for toxicity of a set of ionic liquids (ILs) to *Daphnia magna* using computed lipophilicity, atom-type fragment, quantum topological molecular similarity (QTMS), and extended topochemical atom (ETA) descriptors. The models have been developed and validated in accordance with the Organization for Economic Cooperation and Development (OECD) guidelines for quantitative structure–activity relationships (QSARs). The best partial least squares (PLS) model outperforms the previously reported multiple linear regression (MLR) model [143] in statistical quality and predictive ability ($R^2 = 0.955$, $Q^2 = 0.917$, $R^2_{\text{pred}} = 0.848$). In addition to the importance of lipophilicity, the best model clearly shows the importance of aromaticity in ecotoxicity of ionic liquids toward *D. magna*. These results suggest that ILs with less toxicity may be designed by avoiding aromaticity, nitrogen atoms, and increasing branching in the cationic structure.

The quantitative structure–activity relationship (QSAR) models, including genetic function approximation (GFA) and least squares support vector machine (LSSVM), were developed by Ma *et al.* [135] for predicting the ecotoxicity of 69 ILs toward the marine bacterium *Vibrio fischeri*. Five descriptors were selected by GFA and used to develop the linear model. In order to capture the nonlinear nature, the LSSVM model was also built for more

accurately predicting the ecotoxicity. The GFA and LSSVM models were performed for rigorous internal and external validation, further verifying these models with excellent robustness and predictive ability. From the used descriptors, the cation structure was the main factor involved in the toxicity, which mainly depended on the size, lipophilic, and 3D molecular structure of cations. *Das et al.* [136] have developed predictive QSAR models using topological and quantum chemical descriptors models for *V. fischeri* toxicity using the largest available set of ionic liquids ($n = 305$) with the experimental toxicity data using Microtox®. The whole study has been performed in consonance with the OECD guidelines in terms of dataset selection, model development, applicability domain determination, model validation, and mechanistic interpretation of the diagnosed chemical attributes. In order to experimentally validate the models, a set of IL compounds with low predicted toxicity values was designed and subsequently synthesized, and their toxicity against *V. fischeri* was then experimentally tested. The authors point out that it was the first attempt to perform both true external validation and experimental validation of QSPR models for toxicity of ionic liquids with regard to *V. fischeri*. In addition, the designed ionic liquids were experimentally confirmed to be harmless or practically harmless as defined in the acute toxicity determination criteria by the European Commission.

Das et al. [137] have done a study to explore the chemical attributes of diverse ionic liquids responsible for their cytotoxicity in a rat leukemia cell line (IPC-81) by developing predictive classification as well as regression-based mathematical models. This study using simple and interpretable descriptors derived from a two-dimensional representation of the chemical structures along with quantum topological molecular similarity indices meets the guidelines of the Organization for Economic Cooperation and Development (OECD) for QSAR modeling. The models were subjected to rigorous validation tests proving their predictive potential. After analyzing the results, the authors proposed that the cytotoxicity of ILs could be reduced by making suitable structural changes including reduced cationic surfactant behavior by the use of short-length side chains and decreased cationic lipophilicity, and the employment of nonaromatic cations whenever possible (considering the desired application), avoiding dialkylamino substituent at 4-position of the pyridinium nucleus, and using anions of limited size. Recently, σ -profile descriptors were used by *Ghanem et al.* [138] to build linear and nonlinear QSAR models to predict the ecotoxicities of a wide variety of 111 ionic liquids toward bacterium *Vibrio fischeri*. Linear model was constructed using five descriptors resulting in high accuracy prediction of 0.906. The model performance and stability were ascertained using k-fold cross-validation method. The MLR model clearly emphasized the proportional relation between the length of the alkyl chain and the

increase on the toxicity of ILs. The selected descriptors set from the linear model were then used in multilayer perceptron (MLP) technique to develop the nonlinear model. The accuracy of this MLP model was further enhanced achieving high correlation coefficient of training, validation, and testing sets as high as 0.979 with a highest mean square error of 0.157. The proposed QSAR models can be used as a primary step for screening and designing inherently safer ILs. Note that other recent QSAR models not detailed in this review are reported in the literature [139, 140].

3.3 QSAR Models for Ecotoxicity Prediction of Pharmaceuticals

Scientific reports on the development of QSAR models focused on pharmaceutical products (especially from hospital effluents, wastewater from pharmaceutical industries, household waste, and human and animal excreta) due to their toxicity and adverse effects.

In 2010, *Kar and Roy* [141] have developed interspecies toxicity correlation between *Daphnia magna* (zooplankton) and fish (species according to OECD guidelines) assessing the ecotoxicological hazard potential of diverse 77 pharmaceuticals. Developed models were also used to predict fish toxicities of 59 pharmaceuticals (for which *Daphnia* toxicities are present) and *Daphnia* toxicities of 30 pharmaceuticals (for which fish toxicities are present). According to the authors, this study should allow a better and comprehensive risk assessment of pharmaceuticals for which toxicity data is missing for a particular endpoint. *Das et al.* [142] have used a dataset of 194 compounds with reported rodent, fish, daphnia, and algae toxicity data to develop interspecies models using molecular descriptors. Rigorous validation of all the developed models was performed using multiple validation strategies. Acceptable results were obtained in both cases of direct and interspecies extrapolation quantitative structure–activity relationship models.

The environmental risk assessment of 26 pharmaceuticals and personal care products (PPCPs) of relevant consumption and occurrence in the aquatic environment in Spain was accomplished by *De Garcia et al.* [143] based on the ecotoxicity values obtained by bioluminescence and respirometry assays. The real risk of impact of these compounds in wastewater treatment plants (WWTPs) and in the aquatic environment was predicted. The experimental ecotoxicity results showed that 65.4% of PPCPs under study were at least harmful to aquatic organisms according to the GHS classification based on two different ecotoxicity tests. Acetaminophen, ciprofloxacin, clarithromycin, clofibrate, ibuprofen, omeprazole, triclosan, parabens, and 1, 4-benzoquinone showed some type of risk for the aquatic environments. *Sangion and Gramatica* [144] have developed predictive quantitative activity–activity relationship (QAAR) models to investigate the relationship between toxicities in different species. QAAR models implemented by theoretical molecular descriptors to improve the quality and predictivity of the interspecies relationships were developed using QSARINS software

and validated for their external predictivity. The authors concluded that the models based on *Daphnia* toxicity values can help in reducing more complex and expensive tests on fish, reducing also the tested animals.

In another study, new externally validated QSAR models, specific to predict acute toxicity of a large set of more than 1000 active pharmaceutical ingredients (APIs) in algae, *Daphnia*, and 2 species of fish, were developed by *Sangion and Gramatica* [145]. The models were based on theoretical molecular descriptors calculated by free PaDEL-Descriptor software and selected by genetic algorithm. The models are statistically robust, externally predictive (CCC_{ext} range 85–95% and Q^2_{Fn} range 70–90%) and characterized by a wide structural applicability domain. The accuracy of the models on training and different external sets was compared with the accuracy of the commonly used ECOSAR software, and the authors concluded that their models showed better performances.

In a very recent work, quantitative structure–activity relationship (QSAR) models have been developed by *Khan et al.* [146] to predict the toxicity of a large dataset of approximately 9300 drug-like molecules of pharmaceuticals on 4 different aquatic species, namely, *Pseudokirchneriella subcapitata*, *Daphnia magna*, *Oncorhynchus mykiss*, and *Pimephales promelas*, using genetic algorithm (GA) for feature selection followed by partial least squares regression technique according to the Organization for Economic Cooperation and Development (OECD) guidelines. Only 2D descriptors were used for capturing chemical information and model building, whereas validation of the models was performed by considering various stringent internal and external validation metrics. The applicability domain study was performed in order to set a pre-defined chemical zone of applicability for the obtained QSAR models. In order to prove the robustness and the predictability of the obtained models, an additional comparison was made with ECOSAR, an online expert system for toxicity prediction of organic pollutants. As suggested by *Sangion and Gramatica* [144, 145], the authors confirmed the positive contribution of hydrophobicity and the negative contribution of polar bonds such as hydrogen bonds.

3.4 QSAR Models for Ecotoxicity Prediction of Other Pollutants

Several reliable QSAR models for ecotoxicity prediction studies of different pollutants such as solvents, organic and organometallic compounds, and metals have been developed in recent years.

The methods based on decision tree boost (DTB) and decision tree forest (DTF) were used by *Singh et al.* [147] to develop QSAR models for algae (*P. subcapitata*) experimental ecotoxicity data of chemicals. These models have been developed and validated on the basis of OECD principles for QSAR acceptance and regulation; they are characterized by high external predictivity and wide applicability domain. The QSAR models were successfully applied to

predict toxicities of wide groups of chemicals in other test species including algae (*S. obliquus*), *Daphnia*, fish, and bacteria. The DTB-QSAR models yielded an R^2 of 0.793 for the test set (*P. subcapitata*) and 0.575–0.672 for the other four external validation species, while the DTF-QSAR models yielded an R^2 of 0.753 for the test set and 0.605–0.689 for the other four species. To test the influence of structural parameters on ecotoxicity of a series of 60 phosphonates, a QSAR model was performed by Petrescu *et al.* [148]. The obtained models showed that the toxicity of phosphonates was influenced by steric and molecular geometry which cause inhibition of cholinesterase activity.

Levet *et al.* [149] suggested reliable QSAR models in order to model both invertebrate and algae EC₅₀ for organic solvents by using multiple linear regression using the ordinary least squares method. The chemically heterogeneous organic solvents ($n = 122$) were described by physicochemical descriptors and quantum theoretical parameters. The four-parameter QSAR developed for invertebrate pEC₅₀ prediction included LogP, surface tension, dielectric constant, and the minimal atomic charge. A two-parameter QSAR involving LogP and LUMO energy allowed well-predicting algae pEC₅₀ for all solvents other than amines. To evaluate robustness and predictive performance of the QSARs developed, several strategies were considered, and external validation techniques as required by the REACH regulation and according to the OECD guidelines were performed. In view of the results obtained, the authors deduced that these models constitute a major tool for a reliable assessment of environmental risk related to organic solvents. During the same year, Basant *et al.* [150] developed probabilistic neural network (PNN) and generalized regression neural network (GRNN) models, which were constructed using neurotoxicity data of 47 structurally diverse organic solvents in rats and mice following OECD guideline principles for model development. The prediction and generalization abilities of these models were evaluated. Several statistical validation tests were performed which revealed a high predictivity for the qualitative and quantitative models and rendered high statistical confidence. The results of the applicability domain analysis using the leverage method revealed a single compound (in mouse) as the response outlier and thus confirmed the applicability of the constructed QSTR models over a wide chemical space. For the authors, this study is useful in cost and effort reduction toward the neurotoxicity evaluation of new chemicals.

QSAR models of amine oxide (pure linear C8, C10, C12, C14, and C16 amine oxide surfactants) toxicity were developed by Belanger *et al.* [151] on alga (*Desmodesmus subspicatus*), an invertebrate (*Daphnia magna*), and a fish (*Danio rerio*) using the appropriate array of OECD Test Guidelines. The R^2 of these models were in the range 0.920–0.980. Local and global QSAR models for

predicting the mutagenic activity of various chemicals (aldehydes, aliphatic amines, aromatic amines, benzylamines, hydrazines, anilines, dinitrobenzenes, esters, imidazoles, neutral organics, phenols, benzyl alcohols, vinyl allyl ethers, halides, ketones, and alcohols) against *Salmonella typhimurium* (TA) bacterial strains (TA98 and TA100) were suggested by Basant *et al.* [152]. Relevant structural features that were responsible and influence the mutagenic activity were identified. The applicability domains of the developed models were defined. Accuracies greater than 96%, as well as test set root mean square error (RMSE) and mean absolute error (MAE) values emphasized the usefulness of the developed models for predicting new compounds. According to the authors, the developed models can be used as tools for screening new chemicals for their mutagenicity assessment for regulatory purpose. In the same year, Basant and Gupta [153] established a multi-target QSTR (mt-QSTR) model using four different experimental toxicity datasets of metal oxide nanoparticles (MeONPs) in *E. coli* and HaCaT cells. The optimal validated model yielded high correlation coefficients (R^2 between 0.828 and 0.956) between the experimental and simultaneously predicted endpoint toxicity values in test arrays for all the four systems, revealing high predictivity. No nanoparticles were X-outlier and out of the applicability domain of mt-QSTR model. The analysis of the results obtained showed that oxygen percent, LogS, and Mulliken's electronegativity have direct relationships with the accepted toxicity mechanisms in *E. coli* as well as in HaCaT cells. In conclusion, the authors emphasized that the proposed approach would not only help in reducing the efforts, time, and computational costs but can also provide useful guidance for a new design of oxide nanoparticles.

During this year, three studies caught our attention. De Morais Silva *et al.* [154] developed a predictive model of ecotoxicity for 993 organic micropollutants (OMPs) identified in different sources of water in Brazil to a freshwater crustacean (*Daphnia magna*) and a fish (fathead minnow—*Pimephales promelas*), both commonly used in acute toxicity studies of the investigated agents. Results obtained for *D. magna* showed a lower prediction validation value, however a higher accuracy regarding predicted values for 20 OMPs. According to these authors, the models developed in this study could be used as virtual screening tools for the prediction of aquatic toxicity of organic compounds against both organisms. Ha *et al.* [155] have built a QSAR model and predicted the correlation between the ecotoxicity value and the $\log K_{ow}$ value of eight kinds of polycyclic aromatic hydrocarbons (PAHs: benzene, toluene, naphthalene, 2-methylnaphthalene, fluorene, dibenzothiophene, phenanthrene, pyrene). The root mean square error (RMSE) values of *Daphnia magna* and *Hyalella azteca* were 6.0049 and 5.9980, respectively, when the QSAR model was constructed using the toxicity data for PAHs. The QSAR model was compared with the

ECOSAR data to confirm its validity, showing a correlation with the ECOSAR results. The R^2 value of correlation between the predicted and observed results was 0.9356. Therefore, the authors estimated that the QSAR model developed in this study can accurately provide data for predicting the toxicity of PAHs and may be helpful for use in the toxicity prediction for other kinds of PAHs.

Six machine learning methods to develop highly predictive local and global binary models using different species for saltwater crustacean (*Mysidae* data for local model and *Mysidae*, *Palaemonidae*, and *Penaeidae* data for global binary models) were developed by *Liu et al.* [156] to predict aquatic toxicity of diverse organic chemicals, including pesticides and industrial chemicals. After the clustering of the toxic molecules in the training set, the number of compounds in the training set was reduced to 192 for local models and 261 for global models. Ten descriptors with higher scores were used to develop local models, while 11 descriptors with higher scores were used to develop the global models. The applicability domains of the six better local models and the six better global models were further analyzed. The AUC (area under the receiver operating characteristic curve) values of the better local and global models were around 0.8 and 0.9 for the test sets, respectively. Several chemicals with selective toxicity on different species were identified, and the relationship between chemical aquatic toxicity and the molecular descriptors was explored. According to the authors, the results of this study would be helpful to predict chemical aquatic toxicity. Note that other recent QSAR models not detailed in this review are reported in the literature [157–162].

4 Conclusions

Following the analysis of the abundant literature, it is clear that all components of the environment are polluted by different types of chemical compounds. However, regardless of their benefits and importance in almost all areas, these pollutants seriously affect not only the ecosystem but also various living species because of their toxic properties. In this respect, many researches around the world point out the link between these compounds and the appearance of harmful effects on humans, fauna, and flora. Therefore, one of the most important tasks of the ecotoxicological risk assessment of these products is the experimental evaluation of their toxicity.

On the one hand, this evaluation is a tedious and expensive practice and generates a great sacrifice of animals. On the other hand, given the considerable number (hundreds of thousands) of chemicals, it is almost impossible to test all these pollutants at the laboratory level. To overcome these imponderables and limit in vivo experiments, the REACH regulation recommends the use of in silico methods, such as quantitative QSAR models.

A detailed analysis of the QSAR models devoted to ecotoxicology modeling led us to divide them into two categories. The first consists of models developed from a congeneric series of compounds and proposed for a long time. These, though interesting, suffer from a number of shortcomings. They are obtained from limited datasets for structurally similar compounds and have been validated only on the basis of one or two statistical parameters. The second category includes QSAR models developed on the basis of OECD principles and benefit from the development of the computer tool.

From our point of view, the QSAR models developed over the past decade are promising in that they are more efficient in predicting the toxicity of various compounds. In addition, it is tried, at best, through these models, to meet the challenge faced by modeling in ecotoxicology: the presence of a toxic compound in contact with several species and the contact of a species with several toxic products. Some of the relevant advances made by recent QSAR models include the following:

1. The use of high-quality and increasingly large databases with structurally different compounds.
2. Calculation of a range of descriptors and improvement of the most relevant selection techniques.
3. Validation of the predictive performance of the models through a large number of statistical parameters.
4. Identification of the applicability domain of the model.
5. The interspecies quantitative structure-toxicity-toxicity relationship (QSTTR) and the interspecies quantitative structure-toxicity relationship (i-QSTR) are useful to compensate for the lack of toxicity data for some species.

In conclusion, the satisfaction of all the criteria listed above is targeted at the development of predictive models capable of serving as an alternative to animal experimentation but also to provide useful indications for a design of chemical compounds that are friendly to the environment.

References

1. Ceriani L, Papa E, Kovarich S, Boethling R, Gramatica P (2015) Modeling ready biodegradability of fragrance materials. Environ Toxicol Chem 34(6):1224–1231
2. Gajewicz A (2017) What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. Nanoscale 9(24):8435–8448
3. Kleandrova VV, Luan F, González-Díaz H, Russo JM, Speck-Planche A, Cordeiro MN (2014) Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. Environ Sci Technol 48(24):14686–14694

4. Aubakirova B, Beisenova R, Boxall AB (2017) Prioritization of pharmaceuticals based on risks to aquatic environments in Kazakhstan. *Integr Environ Assess Manag* 13(5):832–839
5. Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P (2017) Computational methodologies for the risk assessment of pesticides in the European Union. *J Agric Food Chem* 65(10):2017. <https://doi.org/10.1021/acs.jafc.7b00516>
6. Riva F, Zuccato E, Davoli E, Fattore E, Castiglioni S (2019) Risk assessment of a mixture of emerging contaminants in surface water in a highly urbanized area in Italy. *J Hazard Mater* 361:103–110
7. Raitano G, Goi D, Pieri V, Passoni A, Mattiussi M, Lutman A, Romeo I, Manganaro A, Marzo M, Porta N (2018) (Eco) toxicological maps: a new risk assessment method integrating traditional and in silico tools and its application in the Ledra River (Italy). *Environ Int* 119:275–286
8. Van den Brink PJ, Boxall AB, Maltby L, Brooks BW, Rudd MA, Backhaus T, Spurgeon D, Veroustraete V, Ajao C, Ankley GT (2018) Toward sustainable environmental quality: priority research questions for Europe. *Environ Toxicol Chem* 37(9):2281–2295
9. Musee N (2011) Nanotechnology risk assessment from a waste management perspective: are the current tools adequate? *Hum Exp Toxicol* 30(8):820–835
10. Hamadache M, Benkortbi O, Hanini S, Amrane A (2018) QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (*Apis mellifera* L.). *Environ Sci Pollut Res* 25(1):896–907
11. Ortiz-Santiestra ME, Maia JP, Egea-Serrano A, Lopes I (2018) Validity of fish, birds and mammals as surrogates for amphibians and reptiles in pesticide toxicity assessment. *Ecotoxicology* 27(7):819–833
12. Grech A, Brochot C, Dorne J-L, Quignot N, Bois FY, Beaudouin R (2017) Toxicokinetic models and related tools in environmental risk assessment of chemicals. *Sci Total Environ* 578:1–15
13. Papa E, van der Wal L, Arnot JA, Gramatica P (2014) Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis. *Sci Total Environ* 470:1040–1046
14. Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P (2018) Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Sci Total Environ* 634:1530–1539
15. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic, London, pp 1–46
16. Roy K, Kar S, Das RN (2015) A primer on QSAR/QSPR modeling: fundamental concepts. Springer, Cham, pp 1–35
17. Devillers J. (2008) Artificial Neural Network Modeling in Environmental Toxicology. In: Livingstone D.J. (eds) Artificial Neural Networks. Methods in Molecular Biology™, vol 458. Humana Press, Switzerland, pp 59–77
18. Ihedioha J, Ukoja P, Ekere N (2017) Ecological and human health risk assessment of heavy metal contamination in soil of a municipal solid waste dump in Uyo, Nigeria. *Environ Geochem Health* 39(3):497–515
19. Ma L, Sun J, Yang Z, Wang L (2015) Heavy metal contamination of agricultural soils affected by mining activities around the Ganxi River in Chenzhou, Southern China. *Environ Monit Assess* 187(12):731. <https://doi.org/10.1007/s10661-015-4966-8>
20. Weissmannová HD, Pavlovský J (2017) Indices of soil contamination by heavy metals—methodology of calculation for pollution assessment (minireview). *Environ Monit Assess* 189(12):616. <https://doi.org/10.1007/s10661-017-6340-5>
21. Sapančin A, Čakal M, Imamović B, Salihović M, Pehlje E, Jacimović Z, Jancan G (2016) Herbicide and pesticide occurrence in the soils of children's playgrounds in Sarajevo, Bosnia and Herzegovina. *Environ Monit Assess* 188(8):450. <https://doi.org/10.1007/s10661-016-5463-4>
22. Yadav IC, Devi NL, Syed JH, Cheng Z, Li J, Zhang G, Jones KC (2015) Current status of persistent organic pesticides residues in air, water, and soil, and their possible effect on neighboring countries: a comprehensive review of India. *Sci Total Environ* 511:123–137
23. Hamadache M, Khaouane L, Benkortbi O, Si Moussa C, Hanini S, Amrane A (2014) Prediction of acute herbicide toxicity in rats from quantitative structure–activity relationship modeling. *Environ Eng Sci* 31(5):243–252
24. Chowdhary P, Raj A, Bharagava RN (2018) Environmental pollution and health hazards from distillery wastewater and treatment approaches to combat the environmental threats: a review. *Chemosphere* 194:229–246
25. Chaza C, Sopheak N, Mariam H, David D, Baghdad O, Moomen B (2018) Assessment

- of pesticide contamination in Akkar groundwater, northern Lebanon. Environ Sci Pollut Res 25(15):14302–14312
26. Pattnaik P, Dangayach G, Bhardwaj AK (2018) A review on the sustainability of textile industries wastewater with and without treatment methodologies. Rev Environ Health 33 (2):163–203
 27. Shen Z, Zhang Y, Jin F, Alessi DS, Zhang Y, Wang F, McMillan O, Al-Tabbaa A (2018) Comparison of nickel adsorption on biochars produced from mixed softwood and Miscanthus straw. Environ Sci Pollut Res 25 (15):14626–14635
 28. Ahmed DA, Slima DF (2018) Heavy metal accumulation by *Corchorus olitorius* L. irrigated with wastewater. Environ Sci Pollut Res 25(15):14996–15005
 29. Pereira R, Cachada A, Sousa JP, Niemeyer J, Markwiese J, Andersen CP (2018) Ecotoxicological effects and risk assessment of pollutants. In: Soil pollution. Academic Press, Elsevier, Cambridge, Massachusetts, pp 191–216
 30. Borgese L, Federici S, Zacco A, Gianoncelli A, Rizzo L, Smith D, Donna F, Lucchini R, Depero L, Bontempi E (2013) Metal fractionation in soils and assessment of environmental contamination in Valcamonica, Italy. Environ Sci Pollut Res 20(7):5067–5075
 31. Ferrante M, Fiore M, Ledda C, Cicciù F, Alonzo E, Fallico R, Platania F, Di RM, Valenti L, Sciacca S (2013) Monitoring of heavy metals and trace elements in the air, fruits and vegetables and soil in the province of Catania (Italy). Ig Sanita Pubbl 69 (1):47–54
 32. Özkul C (2016) Heavy metal contamination in soils around the Tunçbilek thermal power plant (Kütahya, Turkey). Environ Monit Assess 188(5):284. <https://doi.org/10.1007/s10661-016-5295-2>
 33. Uzen N, Cetin O, Unlu M (2016) Effects of domestic wastewater treated by anaerobic stabilization on soil pollution, plant nutrition, and cotton crop yield. Environ Monit Assess 188(12):664. <https://doi.org/10.1007/s10661-016-5680-x>
 34. Mirzaei M, Marofi S, Solgi E, Abbasi M, Karimi R, Bakhtyari HRR (2019) Ecological and health risks of soil and grape heavy metals in long-term fertilized vineyards (Chaharmahal and Bakhtiari province of Iran). Environ Geochem Health 1–17. <https://doi.org/10.1007/s10653-019-00242-5>
 35. Popov BB, Hristova VK, Ahmad MA, Petrovská M (2014) Monitoring of heavy metals and trace elements contamination in the soil and vegetables and air pollution in the Republic of Macedonia. Int J Enhanced Res Sci Technol Eng 3(1):205–214
 36. Arcega-Cabrera F, Fargher L, Quesadas-Rojas M, Moo-Puc R, Oceguera-Vargas I, Noreña-Barroso E, Yáñez-Estrada L, Alvarado J, González L, Pérez-Herrera N (2018) Environmental exposure of children to toxic trace elements (Hg, Cr, As) in an urban area of Yucatan, Mexico: water, blood, and urine levels. Bull Environ Contam Toxicol 100(5):620–626
 37. Stepanova N, Fomina S, Valeeva E, Ziyatdinova A (2018) Heavy metals as criteria of health and ecological well-being of the urban environment. J Trace Elem Med Biol 50:646–651
 38. Vongdala N, Tran H-D, Xuan T, Teschke R, Khanh T (2019) Heavy metal accumulation in water, soil, and plants of municipal solid waste landfill in Vientiane, Laos. Int J Environ Res Public Health 16(1):22. <https://doi.org/10.3390/ijerph16010022>
 39. Bai H, Hu B, Wang C, Bao S, Sai G, Xu X, Zhang S, Li Y (2017) Assessment of radioactive materials and heavy metals in the surface soil around the Bayanwula prospective uranium mining area in China. Environ Res Public Health 14(3):300. <https://doi.org/10.3390/ijerph14030300>
 40. He B, Zhao X, Li P, Liang J, Fan Q, Ma X, Zheng G, Qiu J (2019) Lead isotopic fingerprinting as a tracer to identify the pollution sources of heavy metals in the southeastern zone of Baiyin, China. Sci Total Environ 660:348–357
 41. Kong X, Liu T, Yu Z, Chen Z, Lei D, Wang Z, Zhang H, Li Q, Zhang S (2018) Heavy metal bioaccumulation in rice from a high geological background area in Guizhou Province, China. Environ Res Public Health 15 (10):2281. <https://doi.org/10.3390/ijerph15102281>
 42. Lu Y, Song S, Wang R, Liu Z, Meng J, Sweetman AJ, Jenkins A, Ferrier RC, Li H, Luo W (2015) Impacts of soil and water pollution on food safety and health risks in China. Environ Int 77:5–15
 43. Tang Z, Chai M, Cheng J, Jin J, Yang Y, Nie Z, Huang Q, Li Y (2017) Contamination and health risks of heavy metals in street dust from a coal-mining city in eastern China. Eco-toxicol Environ Saf 138:83–91
 44. Tang Z, Zhang L, Huang Q, Yang Y, Nie Z, Cheng J, Yang J, Wang Y, Chai M (2015) Contamination and risk of heavy metals in soils and sediments from a typical plastic

- waste recycling area in North China. *Ecotoxicol Environ Saf* 122:343–351
45. Xiao R, Wang S, Li R, Wang JJ, Zhang Z (2017) Soil heavy metal contamination and health risks associated with artisanal gold mining in Tongguan, Shaanxi, China. *Eco-toxicol Environ Saf* 141:17–24
46. Toichuev RM, Zhilova LV, Makambaeva GB, Payzildaev TR, Pronk W, Bouwknegt M, Weber R (2018) Assessment and review of organochlorine pesticide pollution in Kyrgyzstan. *Environ Sci Pollut Res* 25 (32):31836–31847
47. El-Saeid M, Al-Turki A, Al-Wable M, Abdell-Nasser G (2011) Evaluation of pesticide residues in Saudi Arabia ground water. *Res J Environ Sci* 5(2):171–178
48. Al-Wabel M, El-Saeid M, Al-Turki A, Abdell-Nasser G (2011) Monitoring of pesticide residues in Saudi Arabia agricultural soils. *Res J Environ Sci* 5(3):269–278
49. Jung Min Ahn SK-SK (2019) Selection of priority management of rivers by assessing heavy metal pollution and ecological risk of surface sediments. *Environ Geochem Health.* <https://doi.org/10.1007/s10653-019-00284-9>
50. Ruiz-Guerra I, Molina-Moreno V, Cortés-García FJ, Núñez-Cacho P (2019) Prediction of the impact on air quality of the cities receiving cruise tourism: the case of the Port of Barcelona. *Heliyon* 5(3):e01280. <https://doi.org/10.1016/j.heliyon.2019.e01280>
51. Drif F, Abdennour C, Cigerci İH, Ali MM, Mansouri O, Messarah M (2019) Preliminary assessment of stress and genotoxicity biomarkers in bivalve molluscs from the Gulf of Annaba, Algeria. *Bull Environ Contam Toxicol* 102:1–5. <https://doi.org/10.1007/s00128-019-02583-4>
52. Rebhi A, Lounici H, Lahrech M, Morel J (2018) Response of Artemisia herba alba to hexavalent chromium pollution under arid and semi-arid conditions. *Int J Phytoremediation* 21:1–6. <https://doi.org/10.1080/15226514.2018.1524841>
53. Bouaroudj S, Menad A, Bounamous A, Ali-Khodja H, Gherib A, Weigel DE, Chenchouni H (2019) Assessment of water quality at the largest dam in Algeria (Beni Haroun Dam) and effects of irrigation on soil characteristics of agricultural lands. *Chemosphere* 219:76–88
54. Rabhi L, Lemou A, Cecinato A, Balducci C, Cherifi N, Ladji R, Yassaa N (2018) Polycyclic aromatic hydrocarbons, phthalates, parabens and other environmental contaminants in dust and suspended particulates of Algiers, Algeria. *Environ Sci Pollut Res* 25 (24):24253–24265
55. Narasimha A, Qian H, Wang H (2019) Assessment of heavy metal (HM) contamination in agricultural soil lands in northern Telangana, India: an approach of spatial distribution and multivariate statistical analysis. *Environ Monit Assess* 191:246. <https://doi.org/10.1007/s10661-019-7408-1>
56. Rather MY, Tilwani YM, Dey A (2019) Assessment of heavy metal contamination in two edible fish species Carassius carassius and Tripholysa kashmirensis of Dal Lake, Srinagar, Kashmir, India. *Environ Monit Assess* 191(4):242. <https://doi.org/10.1007/s10661-019-7382-7>
57. Rinklebe J, Antoniadis V, Shaheen SM, Rosche O, Altermann M (2019) Health risk assessment of potentially toxic elements in soils along the Central Elbe River, Germany. *Environ Int* 126:76–88
58. Al-Omari A, Farhan I, Kandakji T (2019) Zarqa River pollution: impact on its quality. *Environ Monit Assess* 191(3):166. <https://doi.org/10.1007/s10661-019-7283-9>
59. Bolisetty S, Peydayesh M, Mezzenga R (2019) Sustainable technologies for water purification from heavy metals: review and analysis. *Chem Soc Rev* 48(2):463–487
60. Eid EM, Shaltout KH, Moghami FS, Youssef MS, El-Mohsny E, Haroun SA (2019) Bioaccumulation and translocation of nine heavy metals by Eichhornia crassipes in Nile Delta, Egypt: perspectives for phytoremediation. *Int J Phytoremediation* 21:1–10. <https://doi.org/10.1080/15226514.2019.1566885>
61. Rahman Z, Singh VP (2018) Assessment of heavy metal contamination and Hg-resistant bacteria in surface water from different regions of Delhi, India. *Saudi J Biol Sci* 25 (8):1687–1695
62. Saddik M, Fadili A, Makan A (2019) Assessment of heavy metal contamination in surface sediments along the Mediterranean coast of Morocco. *Environ Monit Assess* 191(3):197. <https://doi.org/10.1007/s10661-019-7332-4>
63. Patel M, Kumar R, Kishor K, Mlsna T, Pittman CU Jr, Mohan D (2019) Pharmaceuticals of emerging concern in aquatic systems: chemistry, occurrence, effects, and removal methods. *Chem Rev* 119:3510. <https://doi.org/10.1021/acs.chemrev.8b00299>
64. Sarma GK, Gupta SS, Bhattacharyya KG (2019) Nanomaterials as versatile adsorbents

- for heavy metal ions in water: a review. Environ Sci Pollut Res 26:1–34. <https://doi.org/10.1007/s11356-018-04093-y>
65. Xia F, Qu L, Wang T, Luo L, Chen H, Dahlgren RA, Zhang M, Mei K, Huang H (2018) Distribution and source analysis of heavy metal pollutants in sediments of a rapid developing urban river system. Chemosphere 207:218–228
 66. Soleimanifar H, Deng Y, Barrett K, Feng H, Li X, Sarkar D (2019) Water treatment residual-coated wood mulch for addressing urban stormwater pollution. Water Environ Res 91:523. <https://doi.org/10.1002/wer.1055>
 67. Mohanakavitha T, Divahar R, Meenambal T, Shankar K, Rawat VS, Haile TD, Gadafa C (2019) Dataset on the assessment of water quality of surface water in Kalingarayan Canal for heavy metal pollution, Tamil Nadu. Data Brief 22:878–884
 68. Siddiqui E, Pandey J (2019) Assessment of heavy metal pollution in water and surface sediment and evaluation of ecological risks associated with sediment contamination in the Ganga River: a basin-scale study. Environ Sci Pollut Res 26(11):10926–10940
 69. Sabarathinam C, Bhandary H, Al-Khalid A (2019) A geochemical analogy between the metal sources in Kuwait Bay and territorial sea water of Kuwait. Environ Monit Assess 191(3):142
 70. Deng T, Wu L, Gao J-M, Zhou B, Zhang Y-L, Wu W-N, Tang Z-H, Jiang W-C, Huang W-L (2018) Occurrence and health risk assessment of organotins in waterworks and the source water of the Three Gorges Reservoir Region, China. Environ Sci Pollut Res 25(15):15019–15028
 71. Mahdavinia GR (2018) Polyvinyl alcohol-based nanocomposite hydrogels containing magnetic laponite RD to remove cadmium. Environ Sci Pollut Res 25(15):14977–14988
 72. Deknock A, De Troyer N, Houbraken M, Dominguez-Granda L, Nolivos I, Van Echelpoel W, Forio MAE, Spanoghe P, Goethals P (2019) Distribution of agricultural pesticides in the freshwater environment of the Guayas river basin (Ecuador). Sci Total Environ 646:996–1008
 73. Gunawardena A, Wijeratne E, White B, Hailu A, Pandit R (2017) Industrial pollution and the management of river water quality: a model of Kelani River, Sri Lanka. Environ Monit Assess 189(9):457. <https://doi.org/10.1007/s10661-017-6172-3>
 74. Tang X-Y, Yang Y, Tam NF-Y, Tao R, Dai Y-N (2019) Pesticides in three rural rivers in Guangzhou, China: spatiotemporal distribution and ecological risk. Environ Sci Pollut Res 26(4):3569–3577
 75. Megahed AM, Dahshan H, Abd-El-Kader MA, Abd-Elall AMM, Elbana MH, Nabawy E, Mahmoud HA (2015) Polychlorinated biphenyls water pollution along the River Nile, Egypt. Sci World J 2015:1
 76. Ibrahim IM, Gilfoyle L, Reynolds R, Voulvoulis N (2019) Integrated catchment management for reducing pesticide levels in water: engaging with stakeholders in East Anglia to tackle metaldehyde. Sci Total Environ 656:1436–1447
 77. Nguyen LD, Gassara S, Bui MQ, Zaviska F, Sistat P, Deratani A (2019) Desalination and removal of pesticides from surface water in Mekong Delta by coupling electrodialysis and nanofiltration. Environ Sci Pollut Res <https://doi.org/10.1007/s11356-018-3918-6>
 78. Rocha O, Neto AJG, dos Santos Lima JC, Freitas EC, Miguel M, da Silva MA, Moreira RA, Daam MA (2018) Sensitivities of three tropical indigenous freshwater invertebrates to single and mixture exposures of diuron and carbofuran and their commercial formulations. Ecotoxicology 27(7):834–844
 79. Gaullier C, Dousset S, Billet D, Baran N (2018) Is pesticide sorption by constructed wetland sediments governed by water level and water dynamics? Environ Sci Pollut Res 25(15):14324–14335
 80. Schreiner VC, Szöcs E, Bhowmik AK, Vijver MG, Schäfer RB (2016) Pesticide mixtures in streams of several European countries and the USA. Sci Total Environ 573:680–689
 81. Dromard CR, Guéné M, Bouchon-Navaro Y, Lemoine S, Cordonnier S, Bouchon C (2018) Contamination of marine fauna by chlordcone in Guadeloupe: evidence of a seaward decreasing gradient. Environ Sci Pollut Res 25(15):14294–14301
 82. dos Santos CF, da Costa SN, Santos RFB, Meneses JO, do Couto MVS, de Almeida FTC, de Sena Filho JG, Carneiro PCF, Maria AN, Fujimoto RY (2018) Deltamethrin-induced nuclear erythrocyte alteration and damage to the gills and liver of Collossoma macropomum. Environ Sci Pollut Res 25(15):15102–15110
 83. Tsaboula A, Papadakis E-N, Vryzas Z, Kotopoulou A, Kintzikoglou K, Papadopoulou-Mourkidou E (2019)

- Assessment and management of pesticide pollution at a river basin level part I: aquatic ecotoxicological quality indices. *Sci Total Environ* 653:1597–1611
84. El Zrelli R, Rabaoui L, Alaya MB, Daghbouj N, Castet S, Besson P, Michel S, Bejaoui N, Courjault-Radé P (2018) Seawater quality assessment and identification of pollution sources along the central coastal area of Gabes Gulf (SE Tunisia): evidence of industrial impact and implications for marine environment protection. *Mar Pollut Bull* 127:445–452
 85. Hussain B, Sultana T, Sultana S, Al-Mulhim N, Mahboob S (2018) Pollutant fate and spatio-temporal variation and degree of sedimentation of industrial-and municipal wastes in Chakbandi drain and River Chenab. *Saudi J Biol Sci* 25(7):1326–1331
 86. Njuguna SM, Yan X, Gituru RW, Wang Q, Wang J (2017) Assessment of macrophyte, heavy metal, and nutrient concentrations in the water of the Nairobi River, Kenya. *Environ Monit Assess* 189(9):454. <https://doi.org/10.1007/s10661-017-6159-0>
 87. Medeiros AC, Faial KRF, Faial KCF, da Silva Lopes ID, de Oliveira LM, Guimarães RM, Mendonça NM (2017) Quality index of the surface water of Amazonian rivers in industrial areas in Pará, Brazil. *Mar Pollut Bull* 123 (1–2):156–164
 88. Hahladakis JN, Velis CA, Weber R, Iacovidou E, Purnell P (2018) An overview of chemical additives present in plastics: migration, release, fate and environmental impact during their use, disposal and recycling. *J Hazard Mater* 344:179–199
 89. Anbumani S, Kakkar P (2018) Ecotoxicological effects of microplastics on biota: a review. *Environ Sci Pollut Res Int* 25 (15):14373–14396
 90. Azuma T, Otomo K, Kunitou M, Shimizu M, Hosomaru K, Mikata S, Mino Y, Hayashi T (2018) Performance and efficiency of removal of pharmaceutical compounds from hospital wastewater by lab-scale biological treatment system. *Environ Sci Pollut Res* 25 (15):14647–14655
 91. Herbst AF, Sturm MT, Schuhem K (2018) A new approach for the agglomeration and subsequent removal of polyethylene, polypropylene, and mixtures of both from freshwater systems—a case study. *Environ Sci Pollut Res* 25(15):15226–15234
 92. Le Guet T, Hsini I, Labanowski J, Mondamert L (2018) Sorption of selected pharmaceuticals by a river sediment: role and mechanisms of sediment or Aldrich humic substances. *Environ Sci Pollut Res* 25 (15):14532–14543
 93. Pereira BV, Matus GN, Costa MJ, Dos Santos ACA, Silva-Zacarin EC, do Carmo JB, Nunes B (2018) Assessment of biochemical alterations in the neotropical fish species *Phallichthys harpagos* after acute and chronic exposure to the drugs paracetamol and propranolol. *Environ Sci Pollut Res* 25 (15):14899–14910
 94. Landrigan PJ, Fuller R, Acosta NJ, Adeyi O, Arnold R, Baldé AB, Bertollini R, Bose-O'Reilly S, Boufford JI, Breysse PN (2018) The Lancet Commission on pollution and health. *Lancet* 391(10119):462–512
 95. Alias C, Benassi L, Bertazzi L, Sorlini S, Volta M, Gelatti U (2019) Environmental exposure and health effects in a highly polluted area of Northern Italy: a narrative review. *Environ Sci Pollut Res* 26 (5):4555–4569
 96. Basu N, Lanphear BP (2019) The challenge of pollution and health in Canada. *Can J Public Health* 110(2):159–164
 97. Xing L, Wang L, Zhang R (2018) Characteristics and health risk assessment of volatile organic compounds emitted from interior materials in vehicles: a case study from Nanjing, China. *Environ Sci Pollut Res* 25 (15):14789–14798
 98. DuPont A (2018) Improving and monitoring air quality. *Environ Sci Pollut Res* 25 (15):15253–15263
 99. Hanif I (2018) Impact of economic growth, nonrenewable and renewable energy consumption, and urbanization on carbon emissions in Sub-Saharan Africa. *Environ Sci Pollut Res* 25(15):15057–15067
 100. Afsar B, Elsurer Afzar R, Kanbay A, Covic A, Ortiz A, Kanbay M (2018) Air pollution and kidney disease: review of current evidence. *Clin Kidney J* 12(1):19–32
 101. de Luna MDG, Laciste MT, Tolosa NC, Lu M-C (2018) Effect of catalyst calcination temperature in the visible light photocatalytic oxidation of gaseous formaldehyde by multi-element doped titanium dioxide. *Environ Sci Pollut Res* 25(15):15216–15225
 102. Ishtiaq M, Jehan N, Khan SA, Muhammad S, Saddique U, Iftikhar B (2018) Potential harmful elements in coal dust and human health risk assessment near the mining areas in Cherat, Pakistan. *Environ Sci Pollut Res* 25 (15):14666–14673
 103. Fei X, Lou Z, Christakos G, Ren Z, Liu Q, Lv X (2018) The association between heavy metal soil pollution and stomach cancer: a

- case study in Hangzhou city, China. *Environ Geochem Health* 40(6):2481–2490
104. Yajima I, Zou C, Li X, Nakano C, Omata Y, Kumasaka M (2015) Analysis of heavy-metal-mediated disease and development of a novel remediation system based on fieldwork and experimental research. *Nihon Eiseigaku Zasshi* 70(2):105–109
 105. Lin W-W, Chen Z-X, Kong M-L, Xie Y-Q, Zeng X-W (2017) Air pollution and children's health in Chinese. In: *Ambient air pollution and health impact in China*. Springer, Singapore, pp 153–180
 106. Sun H (2018) Association of soil selenium, strontium, and magnesium concentrations with Parkinson's disease mortality rates in the USA. *Environ Geochem Health* 40 (1):349–357
 107. Tsai CC, Wu CL, Kor CT, Lian IB, Chang CH, Chang TH, Chang CC, Chiu PF (2018) Prospective associations between environmental heavy metal exposure and renal outcomes in adults with chronic kidney disease. *Nephrology* 23(9):830–836
 108. Pratush A, Kumar A, Hu Z (2018) Adverse effect of heavy metals (As, Pb, Hg, and Cr) on health and their bioremediation strategies: a review. *Int Microbiol* 21(3):97–106
 109. Cui X, Cheng H, Liu X, Giubilato E, Critto A, Sun H, Zhang L (2018) Cadmium exposure and early renal effects in the children and adults living in a tungsten-molybdenum mining areas of South China. *Environ Sci Pollut Res* 25(15):15089–15101
 110. Dada OA, Adekola FA, Odebunmi EO (2016) Kinetics and equilibrium models for sorption of Cu (II) onto a novel manganese nano-adsorbent. *J Dispers Sci Technol* 37 (1):119–133
 111. Eklund B, Watermann B (2018) Persistence of TBT and copper in excess on leisure boat hulls around the Baltic Sea. *Environ Sci Pollut Res* 25(15):14595–14605
 112. Kohli SK, Handa N, Sharma A, Gautam V, Arora S, Bhardwaj R, Wijaya L, Alyemeni MN, Ahmad P (2018) Interaction of 24-epibrassinolide and salicylic acid regulates pigment contents, antioxidative defense responses, and gene expression in *Brassica juncea* L. seedlings under Pb stress. *Environ Sci Pollut Res* 25(15):15159–15173
 113. Brouwer M, Huss A, van der Mark M, Nijssen PC, Mulleners WM, Sas AM, Van Laar T, de Snoo GR, Kromhout H, Vermeulen RC (2017) Environmental exposure to pesticides and the risk of Parkinson's disease in the Netherlands. *Environ Int* 107:100–110
 114. Hamadache M, Benkortbi O, Hanini S, Amrane A, Khaouane L, Moussa CS (2016) A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. *J Hazard Mater* 303:28–40
 115. Hamadache M, Hanini S, Benkortbi O, Amrane A, Khaouane L, Moussa CS (2016) Artificial neural network-based equation to predict the toxicity of herbicides on rats. *Chemom Intell Lab Syst* 154:7–15
 116. Kim J, Swartz M, Langlois P, Romitti P, Weyer P, Mitchell L, Luben T, Ramakrishnan A, Malik S, Lupo P (2017) Estimated maternal pesticide exposure from drinking water and heart defects in offspring. *Environ Res Public Health* 14(8):889
 117. Guibal R, Lissalde S, Leblanc J, Cleries K, Charriau A, Poulier G, Mazzella N, Rebillard J-P, Brizard Y, Guibaud G (2017) Two sampling strategies for an overview of pesticide contamination in an agriculture-extensive headwater stream. *Environ Sci Pollut Res* 25 (15):14280–14293
 118. Costa R, Pereira JL, Santos MA, Pacheco M, Guilherme S (2018) The role of contamination history and gender on the genotoxic responses of the crayfish *Procambarus clarkii* to a penoxsulam-based herbicide. *Ecotoxicology* 27(7):908–918
 119. Harmouche-Karaki M, Matta J, Helou K, Mahfouz Y, Fakhoury-Sayegh N, Narbonne J-F (2018) Serum concentrations of selected organochlorine pesticides in a Lebanese population and their associations to sociodemographic, anthropometric and dietary factors: ENASB study. *Environ Sci Pollut Res* 25 (15):14350–14360
 120. Boyle NK, Sheppard WS (2017) A scientific note on seasonal levels of pesticide residues in honey bee worker tissues. *Apidologie* 48 (1):128–130
 121. Heard MS, Baas J, Dorne J-L, Lahive E, Robinson AG, Rortais A, Spurgeon DJ, Svendsen C, Hesketh H (2017) Comparative toxicity of pesticides and environmental contaminants in bees: are honey bees a useful proxy for wild bee species? *Sci Total Environ* 578:357–365
 122. Hallmann CA, Foppen RP, van Turnhout CA, de Kroon H, Jongejans E (2014) Declines in insectivorous birds are associated with high neonicotinoid concentrations. *Nature* 511 (7509):341
 123. Sánchez-Bayo F, Goka K, Hayasaka D (2016) Contamination of the aquatic environment with neonicotinoids and its implication for

- ecosystems. *Front Environ Sci* 4:71. <https://doi.org/10.3389/fenvs.2016.00071>
124. Can A, Yildiz I, Guvendik G (2013) The determination of toxicities of sulphonylurea and phenylurea herbicides with quantitative structure-toxicity relationship (QSTR) studies. *Environ Toxicol Pharmacol* 35 (3):369–379
 125. Basant N, Gupta S, Singh KP (2015) Predicting toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. *J Chem Inf Model* 55(7):1337–1348
 126. Basant N, Gupta S, Singh KP (2016) Modeling the toxicity of chemical pesticides in multiple test species using local and global QSTR approaches. *Toxicol Res* 5(1):340–353
 127. Sun L, Zhang C, Chen Y, Li X, Zhuang S, Li W, Liu G, Lee PW, Tang Y (2015) In silico prediction of chemical aquatic toxicity with chemical category approaches and substructural alerts. *Toxicol Res* 4(2):452–463
 128. Toropov AA, Toropova AP, Marzo M, Dorne JL, Georgiadis N, Benfenati E (2017) QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environ Toxicol Pharmacol* 53:158–163
 129. Qin L-T, Chen Y-H, Zhang X, Mo L-Y, Zeng H-H, Liang Y-P (2018) QSAR prediction of additive and non-additive mixture toxicities of antibiotics and pesticide. *Chemosphere* 198:122–129
 130. Khan PM, Roy K, Benfenati E (2019) Chemometric modeling of *Daphnia magna* toxicity of agrochemicals. *Chemosphere* 224:470. <https://doi.org/10.1016/j.chemosphere.2019.02.147>
 131. Villaverde JJ, Sevilla-Morán B, López-Gotí C, Calvo L, Alonso-Prados JL, Sandín-España P (2018) Photolysis of clethodim herbicide and a formulation in aquatic environments: fate and ecotoxicity assessment of photoproducts by QSAR models. *Sci Total Environ* 615:643–651
 132. Como F, Carnesecchi E, Volani S, Dorne J, Richardson J, Bassan A, Pavan M, Benfenati E (2017) Predicting acute contact toxicity of pesticides in honeybees (*Apis mellifera*) through a k-nearest neighbor model. *Chemosphere* 166:438–444
 133. Roy K, Das RN (2013) QSTR with extended topochemical atom (ETA) indices. 16. Development of predictive classification and regression models for toxicity of ionic liquids towards *Daphnia magna*. *J Hazard Mater* 254:166–178
 134. Roy K, Das RN, Popelier PL (2014) Quantitative structure–activity relationship for toxicity of ionic liquids to *Daphnia magna*: aromaticity vs. lipophilicity. *Chemosphere* 112:120–127
 135. Ma S, Lv M, Deng F, Zhang X, Zhai H, Lv W (2015) Predicting the ecotoxicity of ionic liquids towards *Vibrio fischeri* using genetic function approximation and least squares support vector machine. *J Hazard Mater* 283:591–598
 136. Das RN, Sintra TE, Coutinho JA, Ventura SP, Roy K, Popelier PL (2016) Development of predictive QSAR models for *Vibrio fischeri* toxicity of ionic liquids and their true external and experimental validation tests. *Toxicol Res* 5(5):1388–1399
 137. Das RN, Roy K, Popelier PL (2015) Interspecies quantitative structure–toxicity–toxicity (QSTTR) relationship modeling of ionic liquids. Toxicity of ionic liquids to *V. fischeri*, *D. magna* and *S. vacuolatus*. *Eco-toxicol Environ Saf* 122:497–520
 138. Ghанem OB, Mutualib MA, Lévéque J-M, El-Harawi M (2017) Development of QSAR model to predict the ecotoxicity of *Vibrio fischeri* using COSMO-RS descriptors. *Chemosphere* 170:242–250
 139. He W, Yan F, Jia Q, Xia S, Wang Q (2018) QSAR models for describing the toxicological effects of ILs against *Staphylococcus aureus* based on norm indexes. *Chemosphere* 195:831–838
 140. Khan MI, Zaini D, Shariff AM, Moniruzzaman M (2018) Probabilistic ecotoxicological risk assessment of imidazolium ionic liquids with amino acid and halide anions. *J Mech Eng Sci* 12(3):3798–3810
 141. Kar S, Roy K (2010) First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals. *Chemosphere* 81 (6):738–747
 142. Das RN, Sanderson H, Mwambo AE, Roy K (2013) Preliminary studies on model development for rodent toxicity and its interspecies correlation with aquatic toxicities of pharmaceuticals. *Bull Environ Contam Toxicol* 90 (3):375–381
 143. De García SAO, Pinto GP, García-Encina PA, Irusta-Mata R (2014) Ecotoxicity and environmental risk assessment of pharmaceuticals and personal care products in aquatic environments and wastewater treatment plants. *Ecotoxicology* 23(8):1517–1533
 144. Sangion A, Gramatica P (2016) Ecotoxicity interspecies QAAR models from *Daphnia* toxicity of pharmaceuticals and personal care

- products. *SAR QSAR Environ Res* 27(10):781–798
145. Sangion A, Gramatica P (2016) Hazard of pharmaceuticals for aquatic environment: prioritization by structural approaches and prediction of ecotoxicity. *Environ Int* 95:131–143
 146. Khan K, Benfenati E, Roy K (2019) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotoxicol Environ Saf* 168:287–297
 147. Singh KP, Gupta S, Kumar A, Mohan D (2014) Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. *Chem Res Toxicol* 27(5):741–753
 148. Petrescu A-M, Putz MV, Ilia G (2015) Quantitative structure–activity/ecotoxicity relationships (QSAR/QEcoSAR) of a series of phosphonates. *Environ Toxicol Pharmacol* 40(3):800–824
 149. Levet A, Bordes C, Clément Y, Mignon P, Morell C, Chermette H, Marote P, Lantéri P (2016) Acute aquatic toxicity of organic solvents modeled by QSARs. *J Mol Model* 22(12):288
 150. Basant N, Gupta S, Singh KP (2016) Predicting the acute neurotoxicity of diverse organic solvents using probabilistic neural networks based QSTR modeling approaches. *Neurotoxicology* 53:45–52
 151. Belanger SE, Brill JL, Rawlings JM, McDonough KM, Zoller AC, Wehmeyer KR (2016) Aquatic toxicity structure-activity relationships for the zwitterionic surfactant alkyl dimethyl amine oxide to several aquatic species and a resulting species sensitivity distribution. *Ecotoxicol Environ Saf* 134:95–105
 152. Basant N, Gupta S (2017) QSAR modeling for predicting mutagenic toxicity of diverse chemicals for regulatory purposes. *Environ Sci Pollut Res* 24(16):14430–14444
 153. Basant N, Gupta S (2017) Multi-target QSTR modeling for simultaneous prediction of multiple toxicity endpoints of nano-metal oxides. *Nanotoxicology* 11(3):339–350
 154. de Moraes e Silva L, Lorenzo VP, Lopes WS, Scotti L, Scotti MT (2019) Predictive computational tools for assessment of ecotoxicological activity of organic micropollutants in various water sources in Brazil. *Mol Inf*. <https://doi.org/10.1002/minf.201800156>
 155. Ha H, Park K, Kang G, Lee S (2019) QSAR study using acute toxicity of *Daphnia magna* and *Hyalella azteca* through exposure to polycyclic aromatic hydrocarbons (PAHs). *Eco-toxicology* 28(3):333–342
 156. Liu L, Yang H, Cai Y, Cao Q, Sun L, Wang Z, Li W, Liu G, Lee PW, Tang Y (2019) In silico prediction of chemical aquatic toxicity for marine crustaceans via machine learning. *Toxicol Res* 8:341. <https://doi.org/10.1039/c8tx00331a>
 157. Stoyanova-Slavova IB, Slavov SH, Pearce B, Buzatu DA, Beger RD, Wilkes JG (2014) Partial least square and k-nearest neighbor algorithms for improved 3D quantitative spectral data–activity relationship consensus modeling of acute toxicity. *Environ Toxicol Chem* 33(6):1271–1282
 158. Perales E, García JI, Pires E, Aldea L, Lomba L, Giner B (2017) Ecotoxicity and QSAR studies of glycerol ethers in *Daphnia magna*. *Chemosphere* 183:277–285
 159. Su Q, Lu W, Du D, Chen F, Niu B, Chou K-C (2017) Prediction of the aquatic toxicity of aromatic compounds to *tetrahymena pyriformis* through support vector regression. *Oncotarget* 8(30):49359
 160. Tugcu G, Saçan MT (2018) A multipronged QSAR approach to predict algal low-toxic-effect concentrations of substituted phenols and anilines. *J Hazard Mater* 344:893–901
 161. Mu Y, Wang Z, Wu F, Zhong B, Yang M, Sun F, Feng C, Jin X, Leung KM, Giesy JP (2018) Model for predicting toxicities of metals and metalloids in coastal marine environments worldwide. *Environ Sci Technol* 52(7):4199–4206
 162. de Silva LDM, Alves MF, Scotti L, Lopes WS, Scotti MT (2018) Predictive ecotoxicity of MoA 1 of organic chemicals using in silico approaches. *Ecotoxicol Environ Saf* 153:151–159



Chapter 26

Multi-scale QSAR Approach for Simultaneous Modeling of Ecotoxic Effects of Pesticides

Alejandro Speck-Planche

Abstract

Pesticides are chemical or biological agents, whose ultimate purpose is to eradicate pests, thus preventing crop losses by protecting the plants from multiple diseases. Despite the importance of their use, pesticides constitute a focus of serious concern because of their harmful effects on the environment. In silico approaches have played a key role in diminishing time and financial resources when assessing the ecotoxicity of the pesticides. While many models based on quantitative structure-activity relationships (QSARs) have been reported to predict specific ecotoxicological endpoints, to date, there is no model capable of simultaneously predicting the ecotoxicological profiles of the pesticides under a wide spectrum of experimental conditions. This book chapter introduces for the first time a multi-scale QSAR model able to assess the ecotoxicity of the pesticides by considering different measures of ecotoxic effects, many bioindicator species, several different assay guidelines, and the multiple times during which the bioindicator species have been exposed to the pesticides. The multi-scale QSAR model correctly classified/predicted more than 75% of the data in both training and test sets. By interpreting different molecular descriptors in the models, this work offers the first view regarding the physicochemical properties and structural features that are common for the appearance of multiple ecotoxic effects in any chemical used as a pesticide. Finally, several molecular fragments are suggested as substructural features that can positively contribute to the diminution of the ecotoxic potential of pesticides.

Key words Artificial neural network, Ecotoxic, Fragment, Multi-scale, Pesticides, QSAR

1 Introduction

Pests are living organisms that are invasive or troublesome to plants or animals, causing massive crop losses, and detrimental to humans, livestock, and forestry. In this sense, with the advances of science and technology, pesticides have been created to minimize such pest-related damages that cause a negative impact on the economy. Although they may also have a biological origin, pesticides are referred to as substances or mixture of substances whose purpose

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-0716-0150-1_26) contains supplementary material, which is available to authorized users.

is to prevent, repel, mitigate, or eradicate any pests [1]. Nevertheless, currently, serious concerns have emerged regarding the harmful effects of the pesticides on the ecosystems [2, 3]. This situation has its foundations in the high toxicity of the pesticides, which is worsened by the indiscriminate use of these chemical products. Consequently, assessing the ecotoxic potential of the pesticides constitutes one of the prime goals in agricultural and environmental sciences.

Nowadays, disciplines such as chemoinformatics have propelled the application of computational approaches [4], where quantitative structure-activity relationships (QSAR) have become an integral part of many scientific projects in areas including (but not limited to) drug discovery [5], toxicology [6], nanotechnology [7], and many others fields of research involving complex molecular, biological, and ecological systems [8–11]. In the context of pesticide management, several seminal works reported in the last 5 years have been devoted to predicting the ecotoxicity of pesticides [12–17]. Unfortunately, these QSAR models reported to date have at least one of the following disadvantages. First, they have focused on only one measure of ecotoxicity. Second, the studies have been carried out against only one bioindicator species. Last, the QSAR models have not been capable of giving a sufficiently clear interpretation regarding the physicochemical properties and/or structural features that are required to diminish the ecotoxic effects of pesticides. It should be noted that each bioindicator species has a specific sensitivity to each pesticide. The development of an advanced QSAR model able to predict multiple ecotoxic effects against different bioindicators and under multiple experimental conditions would provide deeper insights regarding the extent and intensity of the toxic potential of the pesticides on the environment.

Recently, several research groups have emphasized the need to develop multi-scale QSAR (ms-QSAR) models. Such models use mathematical operators to characterize the deviations (perturbations) of a query chemical with respect to the average (expected) values of all the chemicals experimentally tested against the same measure of the biological effect (activity, toxicity, pharmacokinetic property, reaction yield, and others), and/or the same target (microorganism, cell line, animal, etc.), and/or the same assay protocol. In general, the multi-scale modeling philosophy and related approaches have found successful applications in diverse scientific fields of research such as organic chemistry [9, 18, 19], materials science and nanotechnology [7, 20–26], neuroscience [27–31], cancer research [32–38], immunology and immunotoxicity [39–41], and infectious diseases [42–48].

Taking into consideration all these ideas, this book chapter reports the first ms-QSAR model focused on the simultaneous prediction of multiple ecotoxic effects of pesticides under dissimilar experimental conditions. The present study provides a fragment-

based interpretation of the molecular descriptors in the ms-QSAR model, allowing the extraction of key structural and physicochemical aspects that can positively contribute to the diminution of the ecotoxicological profile of a pesticide.

2 Materials and Methods

2.1 Construction of the Dataset and Calculation of the Molecular Descriptors

The procedure underpinning the development of the type of multi-scale QSAR models presented in this work has been recently explained in detail in the literature [49]. Initially, 568 unique pesticides involving the three main groups (fungicides, insecticides, and herbicides) were retrieved from the literature [50–53]. The names of these pesticides were matched with their corresponding ecotoxicity records present in the OPP Pesticide Ecotoxicity Database [54]. It should be noted that the process of curation of this database had a series of challenges. From one side, the OPP Pesticide Ecotoxicity Database was focused not only on ecotoxicity data derived from the active ingredients, but also a huge amount of the data is based on formulations whose compositions are very complex. On the other hand, most of the pesticides were tested more than one time under the same experimental conditions (duplicates). In addition, in most cases, pesticides were not reported with the exact ecotoxicity values; instead, they were reported above or below certain cutoffs. Finally, all the ecotoxicity values reported in the aforementioned database were expressed in units of mass per volume (or bodyweight). Bearing in mind all these factors, all the data based on formulations were removed. For the case of the duplicates, as it was not possible to average the ecotoxicity endpoints for a defined pesticide, only the value indicating the highest toxicity (the lowest ecotoxicity value) was chosen. All the ecotoxicity endpoints were converted to values expressed in units of amount of substance (mole) per volume (or bodyweight), which guaranteed a correct comparison between the ecotoxic potential of the pesticides tested under the same assay conditions.

The dataset used in this study contained 259 pesticides, which were assayed by considering at least 1 out of 8 measures of ecotoxicity (m_e), against 1 out of 28 bioindicator species (b_s), involving at least 1 out of 8 assay guidelines (a_g), where at least 1 out of 8 exposure times (e_p) was reported. Notice that the combination of the elements m_e , b_s , a_g , and e_p defines a unique experimental condition (c_j), which can be viewed as an ontology with the form $c_j \rightarrow (m_e, b_s, a_g, e_p)$. Most of the pesticides present in the dataset had not been experimentally tested by considering all the possible combinations of the aforementioned elements. At the end, the dataset ended up containing 3610 statistical cases. It should be emphasized that the dataset presented here is characterized by a great dispersion of the data. In any case, the great success of the

multi-scale QSAR models mentioned in the previous section is based on their ability to handle dispersed and heterogeneous data.

Each statistical case present in the dataset was annotated as non-ecotoxic [$TE_i(c_j) = 1$] or ecotoxic [$TE_i(c_j) = -1$], $TE_i(c_j)$ being a categorical variable that characterized the ecotoxic effect of the i th case under a defined experimental condition c_j . The assignments of positive and negative cases according to the different ecotoxic effects were realized by considering different cutoff values that are depicted in Table 1. It should be highlighted that these cutoffs comply with two aspects. From one side, the cutoff values were selected to be as stringent as possible, enabling with this, a more rigorous prediction of chemicals that may be used as pesticides without posing threat to the environment. On the other hand, the cutoff values were chosen in such a way that for each measure of the ecotoxic effect reported in the dataset, they prevented the excessive imbalance between the number of pesticides annotated as non-ecotoxic and those assigned as ecotoxic.

The SMILES of the 3610 cases were stored in a *.txt file. After manually converting the *.txt file to *.smi, a subsequent conversion from *.smi to *.sdf was made by the software OpenBabel

Table 1
Different cutoff values of the ecotoxic effects used in this study

Measure of ecotoxic effect (m_e)	Cutoffs ^a	Concept
EC ₂₅ (mmol/ac)_TP	≥ 143.54	Concentration (expressed in millimole per acre) required to cause a toxic effect in 25% of the terrestrial plants tested
EC ₅₀ (nM)_AP	≥ 916.25	Concentration (expressed in nanomolar) required to cause a toxic effect in 50% of the aquatic plants tested
EC ₅₀ (nM)_C	≥ 10632.57	Concentration (expressed in nanomolar) required to cause a toxic effect in 50% of the crustaceans tested
EC ₅₀ (nM)_M	≥ 3001.11	Concentration (expressed in nanomolar) required to cause a toxic effect in 50% of the molluscs tested
LC ₅₀ (nM)_Av	≥ 10363739.85	Lethal concentration (expressed in nanomolar) required to kill 50% of the birds tested according a dietary toxicity assay
LC ₅₀ (nM)_C	≥ 1085.87	Lethal concentration (expressed in nanomolar) required to kill 50% of the crustaceans tested
LC ₅₀ (nM)_F	≥ 7149.9	Lethal concentration (expressed in nanomolar) required to kill 50% of the fishes tested
LD ₅₀ (μ mol/kg)_Av	≥ 4147.49	Lethal dose (expressed in micromoles per kilograms of body weight) required to kill 50% of the birds tested via an acute oral toxicity assay

^aValues from which a pesticide was annotated as non-ecotoxic chemical

v2.4.1 [55]. The purpose of such conversion was to get information regarding the 2D connectivity of the molecules, where the option of removing hydrogen atoms was used. Following this, the software QUBILs-MAS v1.0 [56] was employed for the calculation of the molecular descriptors known as the total and local atom-based quadratic indices from the stochastic adjacency matrix. These topological indices have been widely reported in different research areas related to medicinal chemistry and early drug discovery [57–60]; they can be calculated in the following way:

$$TssAq_k(x) = \sum_{i=1}^n \sum_{j=1}^n {}^k a_{ij} \cdot x_i \cdot x_j \quad (1)$$

In Eq. 1, $TssAq_k(x)$ represents the total stochastic quadratic index of order k , which considers each atom i and its chemical environment (j th neighbors) at the topological distance $d = k$. Also, the symbol x defines the different atomic physicochemical properties, namely, hydrophobicity (*HYD*), electronegativity (*E*), atomic weight (*AW*), polar surface area (*PSA*), and polarizability (*POL*). In addition, the element ${}^k a_{ij}$ is used to describe the presence (or absence) of adjacency between any two atoms in a molecule. The local counterparts [$LssAq_k(x)Z$] of the total quadratic indices can be calculated in a similar manner:

$$LssAq_k(x)Z = \sum_{i=1}^n \sum_{j=1}^n {}^k a_{ijZ} \cdot x_i \cdot x_j \quad (2)$$

Here, all the symbols have the same meanings as in the case of Eq. 1. The only difference is that Eq. 2 characterizes the presence of specific fragments based on certain types of atoms (Z) such as hydrogen bond donors, hydrogen bond acceptors, and different types of carbon atoms, among others. In total, 252 molecular descriptors were calculated by the software QUBILs-MAS v1.0. For such purpose, some configurations were used [algebraic form, quadratic; constrains, atom-based; matrix form, stochastic; maximum order, 6; cutoff, keep all; groups, total and local (hydrogen bond donors, hydrogen bond acceptors, aliphatic carbons, aromatic carbons, halogens, carbons in methyl groups, and heteroatoms); properties, already mentioned in the previous paragraph; aggregation, Manhattan distance].

Notice that the descriptors calculated in Eqs. 1 and 2 cannot differentiate the chemical structure of a pesticide when this is tested by varying any of the elements of the experimental condition c_j . In this sense, several works have established that Box-Jenkins operators can be used to solve this problem [44, 61]. When applied to the QSAR philosophy based on multi-scale modeling, the Box-Jenkins operators are used to generate new molecular descriptors able to consider both the

chemical structure and a specific element of the experimental condition c_j , which were mentioned above:

$$\text{avg}QI(c_j) = \frac{1}{n(c_j)} \times \sum_{a=1}^{n(c_j)} QI_a \quad (3)$$

In Eq. 3, QI_a is a general symbol that represents any total (or local) stochastic atom-based quadratic index according to (see Eqs. 1 and 2). As commented before, each experimental condition c_j depends on the elements m_e , b_s , a_g , and e_p . Therefore, Eq. 3 is applied to each of these elements. For instance, for the case of the element b_s , $\text{avg}QI(c_j)$ is the average value of any total quadratic index calculated of all the pesticides annotated as positive, which have been experimentally assayed against the same bioindicator species. At the same time, for $n(c_j)$, the same rule is used; $n(c_j)$ is the number of pesticides considered as positive while tested against the bioindicator species. Similar deductions can be made from Eq. 3 when this is applied to the other elements of c_j . In the second step, the following mathematical formalism is used:

$$DQI_a(c_j) = \frac{QI_a - \text{avg}QI(c_j)}{(QI_{MX} - QI_{MN}) \times p(c_j)} \quad (4)$$

In Eq. 4, $DQI_a(c_j)$ is a multi-scale descriptor, and it measures how much a pesticide structurally deviates from a set of pesticides annotated as positive and assayed by considering the same element of the experimental condition c_j (m_e , b_s , a_g , or e_p). At the same time, QI_{MX} and QI_{MN} are the maximum and minimum values of each quadratic index calculated in this work, respectively. Last, $p(c_j)$ reflects an a priori probability of finding a chemical tested under certain assay condition:

$$p[(m_e)l_c] = \frac{n_T(m_e)}{N_T(l_c)} \quad (5)$$

$$p[(b_s)t_m] = \frac{n_T(b_s)}{N_T(t_m)} \quad (6)$$

$$p[a_g] = \frac{n_T(a_g)}{N_T} \quad (7)$$

$$p[(e_p)t_c] = \frac{n_T(e_p)}{N_T(t_c)} \quad (8)$$

In Eqs. 5, 6, 7, and 8, the meanings of $n_T(m_e)$, $n_T(b_s)$, $n_T(a_g)$, and $n_T(e_p)$ have been already exemplified when Eq. 3 was explained by using the symbol $n(c_j)$. In Eq. 5, $p[(m_e)l_c]$ is the a priori probability of a pesticide being assayed by considering a certain measure of ecotoxic effect with respect to the total number of pesticides assayed by using all the measures of ecotoxicity that exhibit the

same degree of lethality [$N_T(l_c)$]. Thus, $N_T(l_c)$ can be the sum of all the pesticides tested by using all the measures of nonlethal ecotoxic effects (e.g., the sum of all the EC₂₅s and EC₅₀s) or the same sum but for the lethal counterparts (sum of all the LC₅₀s and LD₅₀s). In Eq. 6, $p[(b_s)t_m]$ is the a priori probability of a pesticide being assayed against a specific bioindicator with respect to the total number of pesticides tested against bioindicators that belong to the same general class. Therefore, $N_T(t_m)$ is the total number of pesticides tested against all the terrestrial plants, or all the aquatic plants, or all crustaceans, or all the aves (birds), and so forth. In Eq. 7, $p[a_g]$ represents the a priori probability of a pesticide being tested by using a specific assay guideline with respect to the total number of pesticides N_T . Last, In Eq. 8, $p[(e_p)t_c]$ is the a priori probability of a pesticide being assayed by considering a defined exposure time with respect to the total number of pesticides tested by considering all the exposure times that belong to the same classification [$N_T(t_c)$] in terms of duration. In this study, the exposure times have been divided into three labels: short (48–120 h), medium (7–14 days), and large (21–28 days). So, $N_T(t_c)$ refers to the total number of pesticides assayed by considering any of these time intervals. Taking into account all this information, it should be emphasized that Eqs. 3, 4, 5, 6, 7, and 8 were derived from the training set; in the end, the total amount of descriptors of the type $DQI_a(c_j)$ was 252×4 (four elements, m_e , b_s , a_g , and e_p) = 1008.

2.2 Development of the ms-QSAR Model

The model was created by following different steps (Fig. 1). In this sense, the dataset formed by the 3610 cases was randomly split into two series: training and test sets. The training set was used to search for the best model, and it was formed by 1500 cases assigned as non-ecotoxic and 1220 annotated as ecotoxic, taking 2720 (75.35% of the dataset) compounds in the training set. The purpose of the test set was to validate the model by assessing its predictive power; this set was formed by 890 cases (24.65%), 492 annotated as non-ecotoxic and 398 assigned as ecotoxic.

The software IMMAN was employed to select the molecular descriptors with the highest discriminant power according to their values of differential Shannon entropies [62]. When extracting the most appropriate descriptors, the correlations between them were analyzed. In this sense, the interval $-0.7 < PCC < 0.7$ was used as a cutoff to indicate the lack of correlation, with PCC being Pearson's correlation coefficient [63]. Then, the chosen descriptors were used as inputs by the Intelligent Problem Solver of the artificial neural networks' (ANN) package of the program STATISTICA v6.0 [64]; the purpose was to generate the best ms-QSAR-ANN model. Here, a first run was performed to determine the most appropriate ANN architectures; four different architectures were considered, namely, linear neural networks (LNN), radial basis function (RBF), and multilayer perceptron (MLP). Subsequent

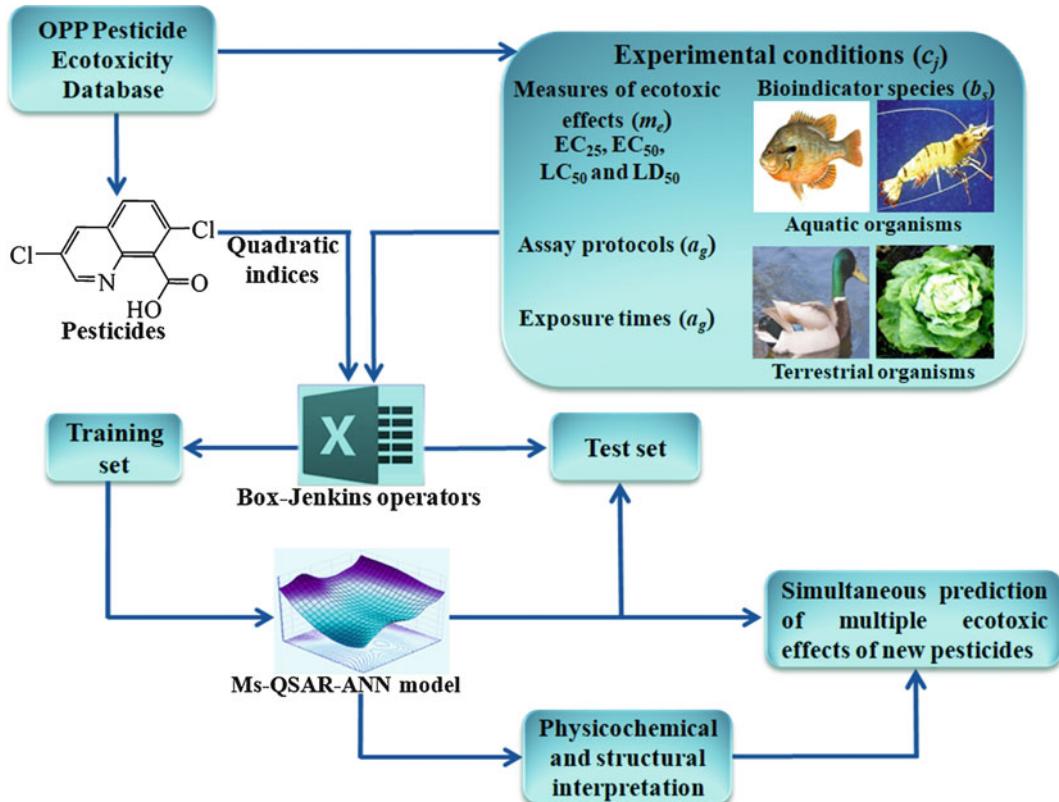


Fig. 1 Steps involved in the development of the ms-QSAR-ANN model

runs served to determine the optimum number of neurons by analyzing the training and test errors, as well as the values of the statistical indices known as sensitivity [$Sn(%)$], and specificity [$Sp(%)$], accuracy [$Ac(%)$], and the Matthews correlation coefficient (MCC) [65] in both training and test sets. In addition, these statistical indices were also used to assess the internal quality (training set) and the predictive power (test set) of the ms-QSAR-ANN model. Last, a sensitivity analysis was performed to rank the descriptors previously selected by IMMAN according to their significances in the ms-QSAR-ANN model.

3 Results and Discussion

3.1 The ms-QSAR-ANN Model

After examining the different ANN architectures, the best ms-QSAR-ANN model found in this work had the profile RBF 9:9-525-1:1. From this, it can be inferred that the ms-QSAR-ANN model reported here is based on a radial basis function that uses 9 molecular descriptors as input nodes (input layer), 525 neurons in the hidden layer, and an output node (output layer) describing the

Table 2
Symbols and definitions of the molecular descriptors used to develop the ms-QSAR-ANN model

Descriptors	Definition
$D[TssAq_2(HYD)]m_e$	Deviation of the total stochastic atom-based quadratic index of order 2, weighted by the hydrophobicity. This descriptor depends on the chemical structure and the measure of the ecotoxic effect
$D[LssAq_6(POL)C]m_e$	Deviation of the local stochastic atom-based quadratic index of order 6, weighted by the polarizability, and focused on the aliphatic carbon atoms. This descriptor depends on the chemical structure and the measure of the ecotoxic effect
$D[LssAq_5(PSA)D]m_e$	Deviation of the local stochastic atom-based quadratic index of order 5, weighted by the polar surface area, and focused on the atoms able to act as hydrogen bond donors. This descriptor depends on the chemical structure and the measure of the ecotoxic effect
$D[LssAq_2(AW)G]m_e$	Deviation of the local stochastic atom-based quadratic index of order 2, weighted by the atomic weight, and focused on the halogen atoms. This descriptor depends on the chemical structure and the measure of the ecotoxic effect
$D[LssAq_5(PSA)D]b_s$	Deviation of the local stochastic atom-based quadratic index of order 5, weighted by the polar surface area, and focused on the atoms able to act as hydrogen bond donors. This descriptor depends on the chemical structure and the specific bioindicator used in the assay
$D[LssAq_5(POL)G]b_s$	Deviation of the local stochastic atom-based quadratic index of order 5, weighted by the polarizability, and focused on the halogen atoms. This descriptor depends on the chemical structure and the specific bioindicator used in the assay
$D[LssAq_0(PSA)A]a_g$	Deviation of the local stochastic atom-based quadratic index of order 0, weighted by the polar surface area, and focused on the atoms able to act as hydrogen bond acceptors. This descriptor depends on the chemical structure and the specific assay guideline
$D[LssAq_2(AW)X]a_g$	Deviation of the local stochastic atom-based quadratic index of order 2, weighted by the atomic weight, and focused on the heteroatoms. This descriptor depends on the chemical structure and the specific assay guideline
$D[LssAq_5(HYD)C]e_p$	Deviation of the local stochastic atom-based quadratic index of order 5, weighted by the hydrophobicity, and focused on the aliphatic carbon atoms. This descriptor depends on the chemical structure and the exposure time

response (predicted) variable of the ecotoxic effect [$Pred_TE_i(c_j)$]. The different molecular descriptors used to build the ms-QSAR-ANN model are represented in Table 2. The file containing the network of the model can be provided upon request to the author.

The ms-QSAR-ANN model exhibits good performance for the classification/prediction of different ecotoxicity endpoints by considering multiple experimental conditions. In this sense, this model correctly classified 2246 out of 2720 cases in the training set, which is equivalent to an $Ac(\%)$ value of 82.57%. In the test set, 4650 out of 890 cases were rightly predicted, with $Ac(\%) = 76.4\%$. Such a

Table 3
Statistical indices used to assess the performance of the ms-QSAR-ANN model

Symbols ^a	Training set	Test set
$N_{\text{non-ecotoxic}}$	1500	492
$CC_{\text{non-ecotoxic}}$	1233	381
$Sn(\%)$	82.20%	77.44%
N_{ecotoxic}	1220	398
CC_{ecotoxic}	1013	299
$Sp(\%)$	83.03%	75.13
MCC	0.65	0.524

^a $N_{\text{non-ecotoxic}}$ number of pesticides annotated as non-ecotoxic, $CC_{\text{non-ecotoxic}}$ pesticides correctly classified as non-ecotoxic, $Sn(\%)$ sensitivity (percentage of pesticides correctly classified as non-ecotoxic), N_{ecotoxic} number of pesticides assigned as ecotoxic, CC_{ecotoxic} pesticides correctly classified as ecotoxic, $Sp(\%)$ specificity (percentage of pesticides correctly classified as ecotoxic), MCC Matthews correlation coefficient

performance was also confirmed by the statistical indices $Sn(\%)$ and $Sp(\%)$, which exhibited values higher than 75% in both training and test sets (Table 3). In addition, the statistical index MCC is equal to 0.65 and 0.524 for training and test sets, respectively. Observe that MCC can take values from +1 (ideal performance) to -1 (the poorest quality) with zero representing a random predictor. Considering that the MCC values obtained for the ms-QSAR-ANN model are closer to +1, it can be concluded that there is a strong correlation between the observed and predicted (categorical) values of the variable $TE_i(c_j)$, which characterizes the different ecotoxic effects. All the chemical and biological data used to create the ms-QSAR-ANN model are available in the *Electronic Supplementary Material 1* that accompanies this book chapter. The results of the classifications/predictions are stored in *Electronic Supplementary Material 2*.

It should be emphasized that the ms-QSAR-ANN model classifies/predicts different ecotoxic effects (m_e) by considering many bioindicators (b_s), diverse assay protocols (a_g), and multiple exposure times (e_p). Although the values for the statistical indices $Sn(\%)$ and $Sp(\%)$ indicate good internal quality and predictive power, they offer information regarding the global performance of the model. Considering this fact, the local counterparts of these indices were calculated for the aforementioned elements of the experimental condition, namely, $[Sn(\%)]m_e$, $[Sp(\%)]m_e$, $[Sn(\%)]b_s$, $[Sp(\%)]b_s$, $[Sn(\%)]a_g$, $[Sp(\%)]a_g$, $[Sn(\%)]e_p$, and $[Sp(\%)]e_p$. All the values of these local sensitivities and specificities are available in *Electronic Supplementary Material 3*. In this sense, for the training set, $[Sn(\%)]m_e$ and $[Sp(\%)]m_e$ were higher than 74%. For the test set, these local statistical indices exhibited values higher than 70%, except for the

measures of ecotoxic effects EC₅₀ (nM)_C with [Sn(%)]_{m_e} = 57.69%, as well as EC₅₀ (nM)_M and LC₅₀ (nM)_F with [Sp(%)]_{m_e} equal to 58.82% and 68.89%, respectively. On the other hand, the values of [Sn(%)]_{b_s} or [Sp(%)]_{b_s} (never both) inferior to 60% were reported in 1 out of 28 and 7 out of 28 bioindicators in the training and test sets, respectively. The only exception was *Pimephales promelas* (fathead minnow) whose [Sn(%)]_{b_s} and [Sp(%)]_{b_s} values were in the interval 50–55%. For the other bioindicators, [Sn(%)]_{b_s} and [Sp(%)]_{b_s} exhibited values in the interval 61–100%; particularly, 75% of all the bioindicators reported in this study presented [Sn(%)]_{b_s} and [Sp(%)]_{b_s} values \geq 60% in both training and test sets.

Regarding the assay protocols, [Sn(%)]_{a_g} and [Sp(%)]_{a_g} yielded values \geq 74% in the training set, while in the test set, the values for this statistical indices were above 70%, except for [Sn(%)]_{a_g} in the assay 72-2 against crustaceans (57.14%) and [Sp(%)]_{a_g} in the assay 72-1 against fishes (66.67%). Last, for the case of the different exposure times, [Sn(%)]_{e_p} and [Sp(%)]_{e_p} behaved in the interval 75–100% in both training and test sets. Once exception was exposure time of 5 days whose [Sn(%)]_{e_p} and [Sp(%)]_{e_p} values were around 65% in both training and test sets. The other two exceptions occurred in the test set, where [Sn(%)]_{e_p} had a value of 58.62% for the time 48 h and [Sp(%)]_{e_p} was equal to 68.31% for the exposure time of 96 h. Altogether, the statistical analysis presented here demonstrates that the ms-QSBER model developed in this work has good quality and predictive power.

3.2 Applicability Domain

Currently, most of the in silico models are used with the sole purpose of filtering the chemical space by virtually screening large and heterogeneous databases. Consequently, estimating the reliability of the predictions performed by any model is an aspect of paramount importance. In this sense, the applicability domain is a well-established concept in predictive modeling; it focuses on defining the regions of the chemical and/or chemico-biological space that may be reliably predicted by a model. A series of applicability domain approaches have been reported in the literature [66], but until now, there is no consensus regarding the superiority of one approach over the others.

The applicability domain assessed in this study was defined by considering the descriptor space approach as reported in a recent work [67]. This means that the applicability domain assessed here was derived from the subset of pesticides correctly classified in the training set. In this sense, for each molecular descriptor present in the ms-QSAR-ANN model, the maximum and minimum values are determined. Then, if for a defined descriptor, a compound had a descriptor value falling beyond the boundaries established by the maximum and minimum, a local score was generated, being equal to zero, otherwise, the local score took the value of one. As

mentioned above, the ms-QSAR-ANN model was constructed from nine molecular descriptors. Therefore, nine local scores were generated. The sum of the local scores should ideally yield a total score equal to nine, which indicates that the chemical completely falls within the applicability domain of the model. The results of the applicability domain are represented in *Electronic Supplementary Material 4*.

3.3 Interpretation of the Molecular Descriptors

Nowadays, as the predictive models are used to prioritize huge amount of chemicals, the molecular descriptors are considered as mere numerical tools that encode some structural information (often believed to be unclear). As a result, the importance of providing an insightful physicochemical and structural interpretation of a model by means of the molecular descriptors is usually neglected and underestimated [49, 67]. The ms-QSAR-ANN model developed in this work has an additional characteristic, which is commonly considered detrimental to the interpretation; this model is based on ANNs with the RBF architecture, and therefore, it is nonlinear. Such a characteristic is the reason for which the models based on ANNs or other machine learning methods are treated as black boxes. Here, the molecular descriptors will be interpreted according to an approach reported by Speck-Planche and co-workers [67–69], which offers a solution to gather information from the molecular descriptors in nonlinear models. Such an approach considers three distinctive elements.

First, while interpreting the molecular descriptors, the approach will rely on analyzing the sensitivity values (*SV*) of the molecular descriptors; the *SV* reflects the relative importance of each molecular descriptor in the model (Fig. 2). The larger the *SV* of a molecular descriptor, the more influential that descriptor will be.

Second, the approach focuses on the calculation of the class-based mean values for each molecular descriptor present in the ms-QSAR-ANN model (Table 4). In this sense, for the subset of pesticides correctly classified in the training set, two mean values were calculated for each descriptor: one for the pesticides annotated as non-ecotoxic and the other for the pesticides assigned as ecotoxic. Then, by comparing the two mean values for each molecular descriptor, it is possible to estimate a tendency of variation, i.e., how the molecular descriptor should vary in order to diminish all the ecotoxic effects.

Last, this approach suggested by Speck-Planche and co-workers benefits from the fragment-based information contained within each topological (graph-based) descriptor. Notice that it is well-established that each graph-theoretical descriptor can always be represented as a linear combination of the number of times in which different molecular fragments (both connected and disconnected) appear in a molecule [70]. Therefore, when

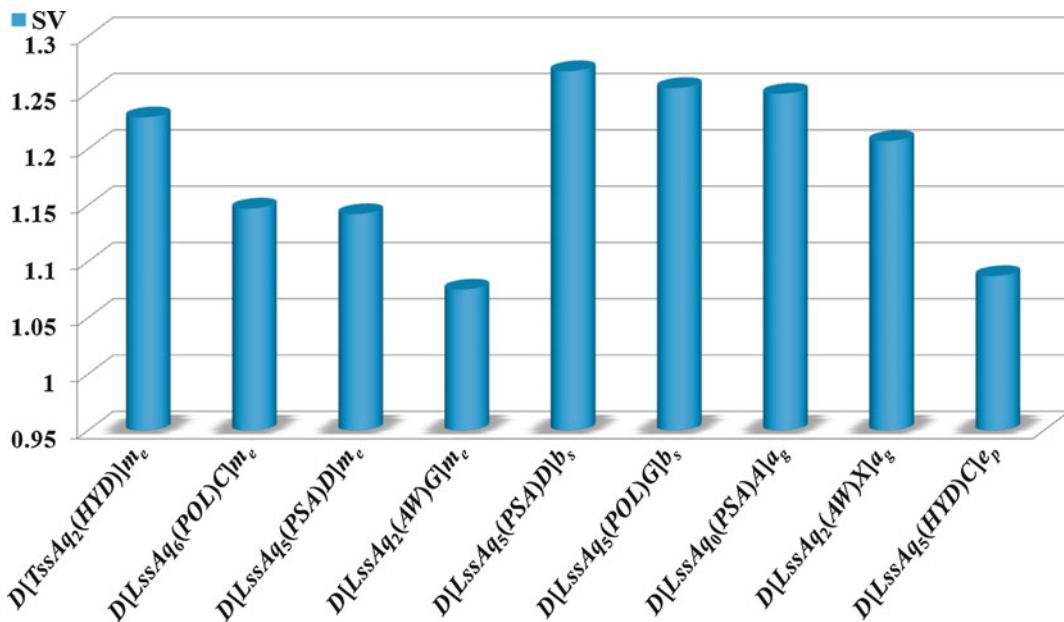


Fig. 2 Relative influences of the molecular descriptors in the ms-QSAR-ANN model

Table 4
Tendency of variation of the different molecular descriptors in the ms-QSAR-ANN model by considering the class-based means

Symbol	Non-ecotoxic	Ecotoxic	Tendency ^a
$D[TssAq_2(HYD)]m_e$	-0.010	0.074	Decrease
$D[LssAq_6(POL)C]m_e$	-0.005	0.060	Decrease
$D[LssAq_5(PSA)D]m_e$	0.018	-0.057	Increase
$D[LssAq_2(AW)G]m_e$	-0.010	0.074	Decrease
$D[LssAq_5(PSA)D]b_s$	0.018	-0.061	Increase
$D[LssAq_5(POL)G]b_s$	-0.024	0.147	Decrease
$D[LssAq_0(PSA)A]a_g$	0.029	0.003	Increase
$D[LssAq_2(AW)X]a_g$	-0.039	0.362	Decrease
$D[LssAq_5(HYD)C]e_p$	0.005	-0.008	Increase

^aTendency, referred to the potential variation (increase or diminution) of a molecular descriptor in order to decrease the ecotoxic effects

interpreting each molecular (graph-based) descriptor present in the ms-QSAR-ANN model, several molecular fragments will be mentioned as examples of substructures that can vary the value of each descriptor, causing an improvement in the safety profiles of any chemical intended to be used as a pesticide. Recently, this last step

has been suggested to help designing new chemicals with desired properties [69, 71].

Before explaining the meanings of the quadratic indices in the ms-QSAR-ANN model, it should be pointed out that they consider the distribution of the different physicochemical properties at the topological distance $d = k$, with k being the order of the descriptor. In addition, the term “topological distance” is used to refer to the number of bonds (without considering bond multiplicity) that exist between any two atoms.

First, we have $D[TssAq_2(HYD)]m_e$, which characterizes the decrease of the joint hydrophobic contribution of any two atoms placed at a topological distance equal to 2. This is the fourth most significant in the ms-QSAR-ANN model. It should be pointed out that the term “joint hydrophobic contribution” is referred to as the multiplication of the hydrophobicity values of any two atoms. In this sense, according to the approach reported by Ghose and Crippen [72–74], each atom in the molecule is described by its neighboring atoms. For instance, hydrogen and halogen atoms are classified by the hybridization and oxidation states of the carbon atom to which they are attached, and for hydrogen atoms, heteroatoms attached to a carbon atom in α position are further considered. At the same time, carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms [72–74]. The Ghose-Crippen approach is based on the use of an atomic hydrophobicity scale derived from each atom type [72–75]. Heteroaromatic rings and mono-halogenated benzenes can be beneficial to the diminution of $D[TssAq_2(HYD)]m_e$, therefore, decreasing the ecotoxic effect of a chemical. On the other hand, non-substituted benzenes as well as the aliphatic portions containing three or more carbon atoms can enhance the harmful effect of the pesticides.

Continuing with the hydrophobic factors, we have the descriptor $D[LssAq_5(HYD)C]e_p$, which indicates the increase of the joint hydrophobic contribution of any two atoms placed at a topological distance equal to 5. Here, one of the two atoms must be an aliphatic carbon. Therefore, $D[LssAq_5(HYD)C]e_p$ (the eighth most influential) focuses on the presence of large hydrophobic regions (cyclic and acyclic) based on aliphatic carbons. Nevertheless, in terms of hydrophobic contributions, it should be pointed out that if aliphatic atoms are placed at the topological distance of 5 with respect to non-substituted aromatic carbons, this could increase $D[LssAq_5(HYD)C]e_p$, diminishing the environmental impact of the pesticides (Fig. 3).

In the ms-QSAR-ANN model, 4 out of 9 descriptors contain information regarding the influence of steric factors. Interestingly, these four descriptors see the size (of course in different ways) as a structural aspect whose diminution can favor the safety profile of a pesticide (Fig. 3). One of them is $D[LssAq_6(POL)C]m_e$ (the sixth

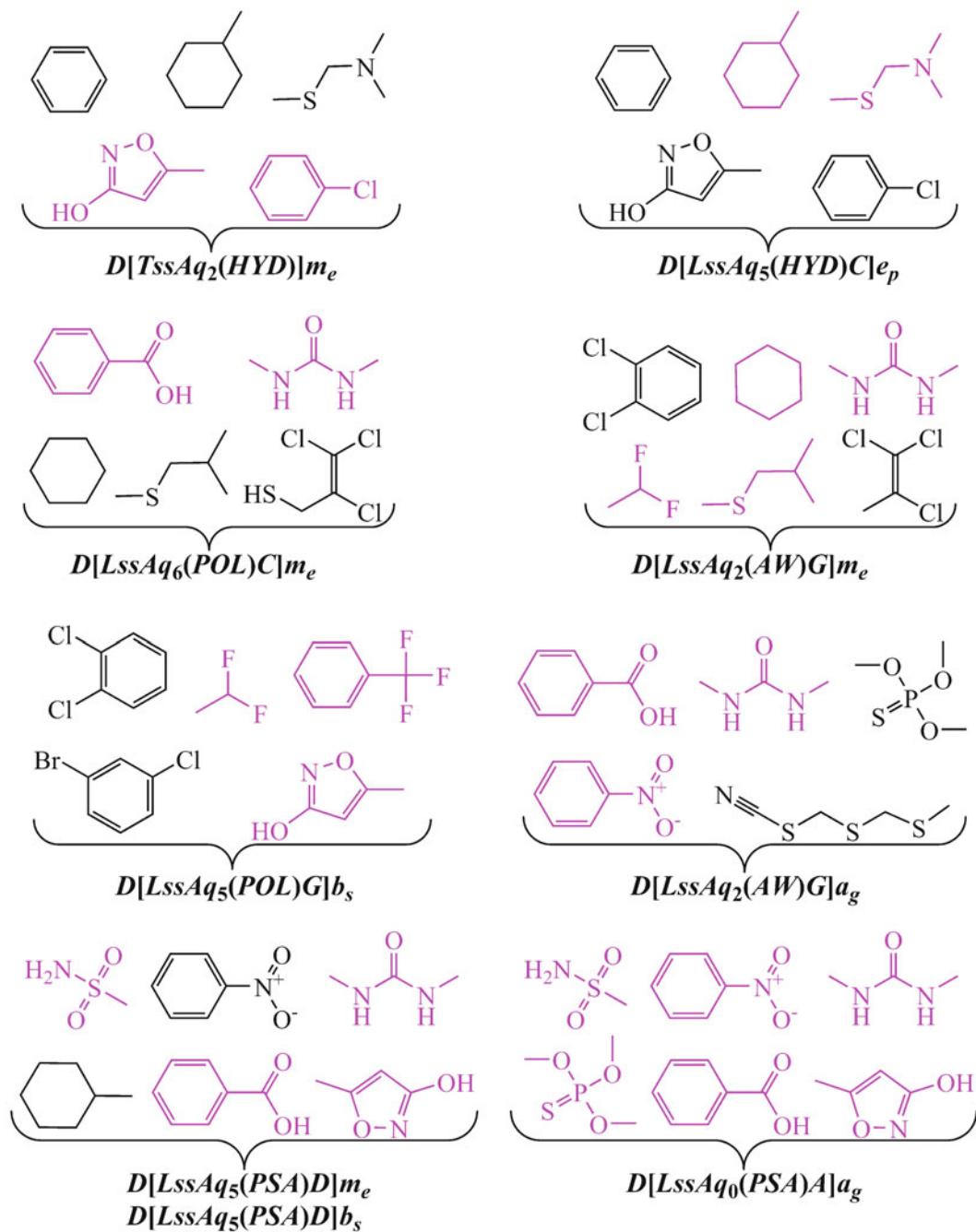


Fig. 3 Suitable and unfavorable fragments generated according to the physicochemical and structural interpretations of the molecular descriptors

most important), which based on the diminution of the polarizability in those regions where any two atoms are placed at a topological distance equal to 6, one of these atoms being an aliphatic carbon. Therefore, this means that in contrast to the information

previously provided by $D[LssAq_5(HYD)C]e_p$, the descriptor $D[LssAq_6(POL)C]m_e$ constrains the presence of aliphatic portions. Thus, when present in a pesticide, the aliphatic carbons must appear in the periphery of the molecules, and they should be separated at the topological distance equal to 6 with respect to nitrogen and oxygen atoms, particular those able to act as hydrogen bond donors. Another descriptor is $D[LssAq_2(AW)G]m_e$, which involves the diminution of the number of halogens and/or the decrease of the atomic weight in regions where any two atoms (one of them being a halogen) are placed at a topological distance equal to 2. Consequently, fragments containing fluorine atoms (or not containing halogens at all) can decrease the value of $D[LssAq_2(AW)G]m_e$ (least influential descriptor). If chlorine, bromine, or iodine atoms are present in a pesticide, it should appear in the periphery of the molecules surrounded by only one carbon atom (if possible). Similarly, at the structural level, the descriptor $D[LssAq_2(AW)X]a_g$ offers information regarding the diminution of the atomic weight in the same regions as $D[LssAq_2(AW)G]m_e$. However, $D[LssAq_2(AW)X]a_g$ (the fifth most important) focuses on preventing the presence of heavy heteroatoms such as sulfur and phosphorus. If such a heavy heteroatom is present in a pesticide, it should be in the periphery of a molecule connected to just one carbon atom (if possible).

From sensitivity analysis reported in Fig. 2, it can be inferred that $D[LssAq_5(POL)G]b_s$ is the second most significant descriptor in the ms-QSAR-ANN model, accounting for the diminution of the polarizability in those regions where any two atoms (one of them being a halogen) are placed at a topological distance equal to 5. Once again, regions containing fluorine atoms can diminish $D[LssAq_5(POL)G]b_s$, positively contributing to the attenuation of the ecotoxic impact of a pesticide. The regions where the halogens (other than fluorine) are placed at the topological distance of 5 with respect to heavy atoms (sulfur or phosphorus) must be avoided.

Finally, we have three molecular descriptors that include information regarding the effect of the hydrophilicity of the pesticides. In this sense, all of the three indicates the increase of the hydrophilicity as a factor that can diminish the ecotoxic effect of the pesticides (Fig. 3). From one side, $D[LssAq_5(PSA)D]m_e$ and $D[LssAq_5(PSA)D]b_s$ describe the augmentation of the polar surface area in regions where any two atoms acting as hydrogen bond donors (or one hydrogen bond donor and one hydrogen bond acceptor) are placed at the topological distance equal to 5. Consequently, at the aforementioned topological distance, all the fragments containing nitrogens belonging to primary or secondary amines, non-substituted and N-substituted amides, alcohols, phenols, and carboxylic acids favor the increment of $D[LssAq_5(PSA)D]m_e$ and $D[LssAq_5(PSA)D]b_s$, thus decreasing the negative

environmental impact of the pesticides. Notice that at the structural level, $D[LsAq_5(PSA)D]m_e$ and $D[LsAq_5(PSA)D]b_s$ characterize similar information. However, while $D[LsAq_5(PSA)D]m_e$ (having the seventh-highest influence) depends on the measure of the ecotoxic effects, $D[LsAq_5(PSA)D]b_s$ focuses on the different bioindicator species used in the assays. It should be highlighted that $D[LsAq_5(PSA)D]b_s$ is the most influential descriptor in the ms-QSAR-ANN model. The information provided by the two previous hydrophilicity-based descriptors is convergent with that present in $D[LsAq_0(PSA)A]a_g$, which indicates the increase of the global polar surface area of the molecule based on atoms able to act as hydrogen bond acceptors. This descriptor is the third most significant in the ms-QSAR-ANN model, and in addition to the functional groups considered by $D[LsAq_5(PSA)D]m_e$ and $D[LsAq_5(PSA)D]b_s$, the descriptor also includes tertiary amines, N, N-substituted amides, ethers, and esters.

It should be emphasized that several fragments considered contributing to improving the safety profiles could be present in highly ecotoxic pesticides. The opposite is also valid; fragments selected as negative could be present in pesticides with relatively low ecotoxic impact. This demonstrates that the presence of a specific fragment is not enough to assess a pesticide as ecotoxic or non-ecotoxic. However, the intrinsic physicochemical properties of the different fragments and the way in which they are connected to other fragments will determine if the pesticide is ecotoxic or not. Thus, the selection of fragments in Fig. 3 has been made on the basis that if all the fragments could be appropriately connected, those highlighted in purple color in the aforementioned figure would better contribute to diminishing the ecotoxicity of a pesticide because of their intrinsic physicochemical properties. All this highlights the importance of interpreting the molecular descriptors.

4 Conclusions

Assessing the ecotoxicity of pesticides constitutes a goal as well as a challenge in environmental sciences. The QSAR models have become pillars in the quest to establish rigorous guidelines for the regulation of pesticides. Nowadays, the QSAR models and the other computational tools should focus on predicting the ecotoxicity of the aforementioned chemicals by considering a wider variety of experimental conditions. This can provide deeper insights regarding the environmental impact of any pesticides. The ms-QSAR-ANN developed here represents the first attempt to concurrently predict many multiple ecotoxic effects of pesticides by changing different factors such as the measures of ecotoxicity, the number, diversity, and complexity of the bioindicator species, the

assay protocols, and the exposure times. The interpretations of the molecular descriptors have provided general guidelines regarding the physicochemical and structural requirements that should be considered for the simultaneous decrease of the different ecotoxic effects of the pesticides. This work opens new horizons toward the application of QSAR modeling in the field of pesticide control and management.

Acknowledgments

Speck-Planche acknowledges the financial support provided by the I.M. Sechenov First Moscow State Medical University under the agreement № Y-187.

References

- Plimmer JR, Gammon DW, Ragsdale NN (2003) Encyclopedia of agrochemicals. Hoboken, Wiley
- Monteiro HR, Pestana JLT, Novais SC, Soares A, Lemos MFL (2019) Toxicity of the insecticides spinosad and indoxacarb to the non-target aquatic midge Chironomus riparius. *Sci Total Environ* 666:1283–1291
- He J, He H, Yan Z, Gao F, Zheng X, Fan J, Wang Y (2019) Comparative analysis of freshwater species sensitivity distributions and ecotoxicity for priority pesticides: implications for water quality criteria. *Ecotoxicol Environ Saf* 176:119–124
- Bunin BA, Bajorath J, Siesel B, Morales G (2007) Chemoinformatics: theory, practice and products. Springer, Dordrecht
- Oprea T (2005) Chemoinformatics in drug discovery. Weinheim, Wiley-VCH Verlag GmbH & Co. KGaA
- Cruz-Monteagudo M, Ancede-Gallardo E, Jorge M, Cordeiro MNDS (2013) Chemoinformatics profiling of ionic liquids – automatic and chemically interpretable cytotoxicity profiling, virtual screening, and cytotoxicophore identification. *Toxicol Sci* 136:548–565
- Gonzalez-Durruthy M, Alberici LC, Curti C, Naal Z, Atique-Sawasaki DT, Vazquez-Naya JM, Gonzalez-Diaz H, Munteanu CR (2017) Experimental-computational study of carbon nanotube effects on mitochondrial respiration: in silico nano-QSPR machine learning models based on New Raman spectra transform with Markov-Shannon entropy invariants. *J Chem Inf Model* 57:1029–1044
- Duardo-Sanchez A, Munteanu CR, Riera-Fernandez P, Lopez-Diaz A, Pazos A, Gonzalez-Diaz H (2013) Modeling complex metabolic reactions, ecological systems, and financial and legal networks with MIANN models based on Markov-Wiener node descriptors. *J Chem Inf Model* 54:16–29
- Gonzalez-Diaz H, Arrasate S, Gomez-SanJuan A, Sotomayor N, Lete E, Besada-Porto L, Russo JM (2013) General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Curr Top Med Chem* 13:1713–1741
- Gonzalez-Diaz H, Riera-Fernandez P, Pazos A, Munteanu CR (2013) The Rucker-Markov invariants of complex bio-systems: applications in parasitology and neuroinformatics. *Biosystems* 111:199–207
- Gonzalez-Diaz H, Arrasate S, Juan AG, Sotomayor N, Lete E, Speck-Planche A, Russo JM, Luan F, Cordeiro MNDS (2014) Matrix trace operators: from spectral moments of molecular graphs and complex networks to perturbations in synthetic reactions, micelle nanoparticles, and drug ADME processes. *Curr Drug Metab* 15:470–488
- He L, Xiao K, Zhou C, Li G, Yang H, Li Z, Cheng J (2019) Insights into pesticide toxicity against aquatic organism: QSTR models on *Daphnia Magna*. *Ecotoxicol Environ Saf* 173:285–292
- Toropov AA, Toropova AP, Marzo M, Dorne JL, Georgiadis N, Benfenati E (2017) QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environ Toxicol Pharmacol* 53:158–163

14. Basant N, Gupta S, Singh KP (2016) Modeling the toxicity of chemical pesticides in multiple test species using local and global QSTR approaches. *Toxicol Res (Camb)* 5:340–353
15. Basant N, Gupta S, Singh KP (2015) Predicting aquatic toxicities of chemical pesticides in multiple test species using nonlinear QSTR modeling approaches. *Chemosphere* 139:246–255
16. Basant N, Gupta S, Singh KP (2015) Predicting toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. *J Chem Inf Model* 55:1337–1348
17. Hamadache M, Benkortbi O, Hanini S, Amrane A, Khaouane L, Si Moussa C (2016) A Quantitative Structure Activity Relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. *J Hazard Mater* 303:28–40
18. Simon-Vidal L, Garcia-Calvo O, Oteo U, Arrasate S, Lete E, Sotomayor N, Gonzalez-Diaz H (2018) Perturbation-Theory and Machine Learning (PTML) model for high-throughput screening of parham reactions: experimental and theoretical studies. *J Chem Inf Model* 58:1384–1396
19. Aranzamendi E, Arrasate S, Sotomayor N, Gonzalez-Diaz H, Lete E (2016) Chiral bronsted acid-catalyzed enantioselective alpha-amidoalkylation reactions: a Joint Experimental and Predictive Study. *ChemistryOpen* 5:540–549
20. Blay V, Yokoi T, Gonzalez-Diaz H (2018) Perturbation theory-machine learning study of zeolite materials desilication. *J Chem Inf Model* 58:2414–2419
21. Gonzalez-Durruthy M, Werhli AV, Seus V, Machado KS, Pazos A, Munteanu CR, Gonzalez-Diaz H, Monserrat JM (2017) Decrypting strong and weak single-walled carbon nanotubes interactions with mitochondrial voltage-dependent anion channels using molecular docking and perturbation theory. *Sci Rep* 7:13271
22. Concu R, Kleandrova VV, Speck-Planche A, Cordeiro M (2017) Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology* 11:891–906
23. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2015) Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine (Lond)* 10:193–204
24. Luan F, Kleandrova VV, Gonzalez-Diaz H, Ruso JM, Melo A, Speck-Planche A, Cordeiro MNDS (2014) Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* 6:10623–10630
25. Kleandrova VV, Luan F, Gonzalez-Diaz H, Ruso JM, Speck-Planche A, Cordeiro MNDS (2014) Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ Sci Technol* 48:14686–14694
26. Kleandrova VV, Luan F, Gonzalez-Diaz H, Ruso JM, Melo A, Speck-Planche A, Cordeiro MNDS (2014) Computational ecotoxicology: simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ Int* 73C:288–294
27. Ferreira da Costa J, Silva D, Caamano O, Brea JM, Loza MI, Munteanu CR, Pazos A, Garcia-Mera X, Gonzalez-Diaz H (2018) Perturbation theory/machine learning model of ChEMBL data for dopamine targets: docking, synthesis, and assay of new l-prolyl-l-leucyl-glycinamide peptidomimetics. *ACS Chem Neurosci* 9:2572–2587
28. Abeijon P, Garcia-Mera X, Caamano O, Yanez M, Lopez-Castro E, Romero-Duran FJ, Gonzalez-Diaz H (2017) Multi-target mining of Alzheimer disease proteome with Hansch's QSBR-perturbation theory and experimental-theoretic study of new thiophene isosters of rasagiline. *Curr Drug Targets* 18:511–521
29. Romero-Duran FJ, Alonso N, Yanez M, Caamano O, Garcia-Mera X, Gonzalez-Diaz H (2016) Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* 103:270–278
30. Speck-Planche A, Luan F, Cordeiro MNDS (2012) Role of ligand-based drug design methodologies toward the discovery of new anti-Alzheimer agents: futures perspectives in Fragment-Based Ligand Design. *Curr Med Chem* 19:1635–1645
31. Molina E, Sobarzo-Sanchez E, Speck-Planche A, Matos MJ, Uriarte E, Santana L, Yanez M, Orallo F (2012) Monoamino oxidase a: an interesting pharmacological target for the development of multi-target QSAR. *Mini Rev Med Chem* 12:947–958
32. Bediaga H, Arrasate S, Gonzalez-Diaz H (2018) PTML combinatorial model of

- ChEMBL compounds assays for multiple types of cancer. *ACS Comb Sci* 20:621–632
33. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2013) Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med Chem* 13:791–800
 34. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2012) Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anticancer Agents Med Chem* 12:678–685
 35. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2012) Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur J Pharm Sci* 47:273–279
 36. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2012) Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg Med Chem* 20:4848–4855
 37. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2011) Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg Med Chem* 19:6239–6244
 38. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2011) Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur J Med Chem* 46:5910–5916
 39. Martinez-Arzate SG, Tenorio-Borroto E, Barbabosa Pliego A, Diaz-Albiter HM, Vazquez-Chagoyan JC, Gonzalez-Diaz H (2017) PTML model for proteome mining of B-cell epitopes and theoretical-experimental study of Bm86 protein sequences from Colima, Mexico. *J Proteome Res* 16:4093–4103
 40. Tenorio-Borroto E, Ramirez FR, Speck-Planche A, Cordeiro MNDS, Luan F, Gonzalez-Diaz H (2014) QSPR and flow cytometry analysis (QSPR-FCA): review and new findings on parallel study of multiple interactions of chemical compounds with immune cellular and molecular targets. *Curr Drug Metab* 15:414–428
 41. Tenorio-Borroto E, Penuelas-Rivas CG, Vasquez-Chagoyan JC, Castanedo N, Prado-Prado FJ, Garcia-Mera X, Gonzalez-Diaz H (2014) Model for high-throughput screening of drug immunotoxicity – Study of the antimicrobial G1 over peritoneal macrophages using flow cytometry. *Eur J Med Chem* 72:206–220
 42. Herrera-Ibata DM, Pazos A, Orbegozo-Medina RA, Romero-Duran FJ, Gonzalez-Diaz H (2015) Mapping chemical structure-activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of U.S. counties. *Biosystems* 132–133:20–34
 43. Herrera-Ibata DM, Orbegozo-Medina RA, Gonzalez-Diaz H (2015) Multiscale mapping of AIDS in U.S. countries vs anti-HIV drugs activity with complex networks and information indices. *Curr Bioinform* 10:639–657
 44. Gonzalez-Diaz H, Herrera-Ibata DM, Duardo-Sanchez A, Munteanu CR, Orbegozo-Medina RA, Pazos A (2014) ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J Chem Inf Model* 54:744–755
 45. Speck-Planche A, Cordeiro MNDS (2014) Review of current chemoinformatic tools for modeling important aspects of CYPs-mediated drug metabolism. Integrating metabolism data with other biological profiles to enhance drug discovery. *Curr Drug Metab* 15:429–440
 46. Speck-Planche A, Kleandrova VV, Cordeiro MNDS (2013) New insights toward the discovery of antibacterial agents: multi-tasking QSAR model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *Eur J Pharm Sci* 48:812–818
 47. Speck-Planche A, Cordeiro MNDS (2014) Simultaneous virtual prediction of anti-*Escherichia coli* activities and ADMET profiles: a chemoinformatic complementary approach for high-throughput screening. *ACS Comb Sci* 16:78–84
 48. Speck-Planche A, Kleandrova VV, Russo JM, Cordeiro MNDS (2016) First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple Gram-positive pathogens. *J Chem Inf Model* 56:588–598
 49. Speck-Planche A, Cordeiro MNDS (2017) Speeding up early drug discovery in antiviral research: a fragment-based in silico approach for the design of virtual anti-hepatitis C leads. *ACS Comb Sci* 19:501–512
 50. Speck-Planche A, Cordeiro MNDS, Guilarte-Montero L, Yera-Bueno R (2011) Current computational approaches towards the rational design of new insecticidal agents. *Curr Comput Aided Drug Des* 7:304–314
 51. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS (2012) Predicting multiple eco-toxicological profiles in agrochemical

- fungicides: a multi-species chemoinformatic approach. *Ecotoxicol Environ Saf* 80:308–313
52. Speck-Planche A, Kleandrova VV, Scotti MT (2012) Fragment-based approach for the in silico discovery of multi-target insecticides. *Chemom Intel Lab Syst* 111:39–45
 53. Perez Gonzalez M, Gonzalez Diaz H, Molina Ruiz R, Cabrera MA, Ramos de Armas R (2003) TOPS-MODE based QSARs derived from heterogeneous series of compounds Applications to the design of new herbicides. *J Chem Inf Comput Sci* 43:1192–1199
 54. EPA. OPP pesticide ecotoxicity database. Access Date: 28 Feb 2019. Available from: www.ipmcenters.org/ecotox/
 55. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3:33
 56. Valdes-Martini JR, Marrero-Ponce Y, Garcia-Jacas CR, Martinez-Mayorga K, Barigye SJ, Vaz d'Almeida YS, Pham-The H, Perez-Gimenez F, Morell CA (2017) QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *J Cheminform* 9:35
 57. Medina Marrero R, Marrero-Ponce Y, Barigye SJ, Echeverria Diaz Y, Acevedo-Barrios R, Casanola-Martin GM, Garcia Bernal M, Torrens F, Perez-Gimenez F (2015) QuBiLS-MAS method in early drug discovery and rational drug identification of antifungal agents. *SAR QSAR Environ Res* 26:943–958
 58. Marrero-Ponce Y, Siverio-Mota D, Galvez-Llompart M, Recio MC, Giner RM, Garcia-Domenech R, Torrens F, Aran VJ, Cordero-Maldonado ML, Esguera CV, de Witte PA, Crawford AD (2011) Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and in vivo screening: the nitroindazolinone chemotype. *Eur J Med Chem* 46:5736–5753
 59. Montero-Torres A, Garcia-Sanchez RN, Marrero-Ponce Y, Machado-Tugores Y, Nogal-Ruiz JJ, Martinez-Fernandez AR, Aran VJ, Ochoa C, Meneses-Marcel A, Torrens F (2006) Non-stochastic quadratic fingerprints and LDA-based QSAR models in hit and lead generation through virtual screening: theoretical and experimental assessment of a promising method for the discovery of new antimalarial compounds. *Eur J Med Chem* 41:483–493
 60. Marrero-Ponce Y, Medina-Marrero R, Torrens F, Martinez Y, Romero-Zaldivar V, Castro EA (2005) Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg Med Chem* 13:2881–2899
 61. Speck-Planche A, Cordeiro MNDS (2014) Chemoinformatics for medicinal chemistry: in silico model to enable the discovery of potent and safer anti-coccidi agents. *Future Med Chem* 6:2013–2028
 62. Urias RW, Barigye SJ, Marrero-Ponce Y, Garcia-Jacas CR, Valdes-Martini JR, Perez-Gimenez F (2015) IMMAN: free software for information theory-based chemometric analysis. *Mol Divers* 19:305–319
 63. Pearson K (1895) Notes on regression and inheritance in the case of two parents. *Proc R Soc Lond* 58:240–242
 64. Statsoft-Team (2001) STATISTICA. Data analysis software system. v6.0. Tulsa
 65. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405:442–451
 66. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17:4791–4810
 67. Speck-Planche A (2018) Combining ensemble learning with a fragment-based topological approach to generate new molecular diversity in drug discovery: in silico design of Hsp90 inhibitors. *ACS Omega* 3:14704–14716
 68. Speck-Planche A, Kleandrova VV (2012) QSAR and molecular docking techniques for the discovery of potent monoamine oxidase B inhibitors: computer-aided generation of new rasagiline bioisosteres. *Curr Top Med Chem* 12:1734–1747
 69. Speck-Planche A (2019) Multicellular target QSAR model for simultaneous prediction and design of anti-pancreatic cancer agents. *ACS Omega* 4:3122–3132
 70. Baskin II, Skvortsova MI, Stankevich IV, Zefirov NS (1995) On the basis of invariants of labeled molecular graphs. *J Chem Inf Comput Sci* 35:527–531
 71. Speck-Planche A, Cordeiro MNDS (2017) De novo computational design of compounds virtually displaying potent antibacterial activity and desirable in vitro ADMET profiles. *Med Chem Res* 26:2345–2356
 72. Ghose AK, Crippen GM (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. *J Comput Chem* 7:565–577

73. Ghose AK, Crippen GM (1987) Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J Chem Inf Comput Sci* 27:21–35
74. Ghose AK, Pritchett A, Crippen GM (1988) Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: modeling hydrophobic interactions. *J Comput Chem* 9:80–90
75. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK (1989) Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J Chem Inf Comput Sci* 29:163–172



Chapter 27

Quantitative Structure-Toxicity Relationship Models Based on Hydrophobicity and Electrophilicity

Gourhari Jana, Ranita Pal, Shamik Sural, and Pratim Kumar Chattaraj

Abstract

In pharmaceutical research, particularly in the preclinical stages of drug discovery, quantitative structure-activity relationship (QSAR) is being increasingly utilized to avoid costly experimentation and tedious extraction of relevant information from big chemical databases. QSAR modelling is also used in modelling environmental toxicity of chemicals. In the current study, toxicity ($pLC_{50}/pIGC_{50}$) to *Pimephales promelas* and *Tetrahymena pyriformis* has been investigated by using electrophilicity index, its square and cubic terms. Hydrophobicity is known as one of the important predictors, and accordingly it has also been employed to improve the models. The widely used multiple linear regression (MLR) method has been implemented to determine regression coefficients indicating the predictive power of the descriptors used.

Key words QSTR, Global electronic descriptor, Hydrophobicity, Multiple linear regression (MLR), *Pimephales promelas*, *Tetrahymena pyriformis*

1 Introduction

Quantitative structure-activity relationship (QSAR) attempts to correlate structural properties of a series of molecules to their biological or ecotoxicological activities, creating models to be used in evaluating activities of new compounds. Molecular properties such as electronic, hydrophobic, steric, etc. act as descriptors for generating mathematical/computational models capable of predicting their activities with remarkable accuracy. Drug discovery, evaluation of toxicity, etc. have become a lot easier and environment-friendly over the years with the increasing development in computational chemistry, quantum simulations, and statistical techniques, since these have resulted in a drastic reduction in animal testing and time-consuming experimental procedures. The idea is to identify the factors or structural features responsible for a certain biological activity and to successfully predict that activity using computational and statistical techniques bypassing animal experiments. The process of drug design through computational

techniques begin with an understanding of molecular structures and interactions with the subject, followed by geometrical and energy calculations, and then relating the structural and chemical features with the activity. The mathematical form of QSAR is usually represented as follows:

$$\text{Activity} = f(\text{physicochemical properties})$$

i.e., in the linear form,

$$\text{Activity} = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots$$

where the descriptors denoted by x_n either are experimentally obtained or are computationally derived and the coefficients a_n are calculated using various statistical methods.

Mapping molecular features and physicochemical properties to biological activity began in the mid-nineteenth century. In 1863, Cros [1] found out a correlation between the solubility of alcohol in water and their toxic effects on mammals. Half a decade later, Crum-Brown and Fraser [2] generalized Cros' findings by expressing the physiological action of a substance as a function of its chemical composition. In the next few years, while Richardson [3] discovered a relation between water solubility of ethers and alcohols and their toxic activity, Mills [4] quite accurately predicted the melting and boiling points in a homologous series using a QSPR model. Later in 1893, Richet [5] put forward his discovery stating that there exists an inverse relation between cytotoxicities of simple organic compounds and their corresponding water solubilities, following which Meyer [6] and Overton [7] independently employed olive oil/water partition coefficients of a series of organic molecules as descriptors to describe their narcotic effects. In the 1930s, the relations between rate constants or equilibrium constants of reactions and molecular structures were of particular interest among scientists which led to Hammett's invention of the electronic substituent (σ) and reaction (ρ) constants used for characterization of different types of electronic effects on reaction mechanisms [8, 9]. Ferguson's thermodynamic approach toward correlating the relative saturation of volatile compounds used in vehicles to their depressant action [10]; Albert, Bell, and Robin's study [11–13] showing the significance of ionization of bases and weak acids in bacteriostatic activity; and Taft's separation of polar, steric, and resonance effects and introduction of the first steric parameter E_S [14, 15] were some of the important contributions toward the development of QSAR. Meanwhile, Hansch et al. [16, 17] used Hammett's sigma constant in combination with *n*-octanol/water partition coefficient to construct QSAR models which are now famously known as the linear Hansch equations. This contribution of Hansch is considered to mark the beginning of the modern QSAR.

Quantum mechanics has come a long way in describing both global and local reactivity indices within the scope of conceptual density functional theory (CDFT) [18–25]. Electronic structure principles such as hard-soft acid-base (HSAB) principle [26], maximum hardness principle (MHP) [27], minimum polarizability principle (MPP) [28], minimum electrophilicity principle (MEP) [29–31], and generalized philicity provide theoretical basis for the reactivity descriptors. Molecular descriptors being the foremost components of QSAR/QSPR/QSTR, in the prediction analysis, are used to represent mathematical correlation models in terms of quantitative numbers. These descriptors are also developed for encoding significant chemical information in molecules. The prediction quality depends mainly on the extraction of chemical features rather than the statistics of modelling. From the very beginning, a significant number of descriptors or variables have been introduced to explore QSAR model analysis, which has become of paramount interest in the growing field of research. Various interesting aspects of electronic information, molecular topology, and bonding interactions in different environments have been addressed extensively in previous studies on structure-activity/property/toxicity paradigm [32–37]. Quantum chemical parameters like orbital energies (E_{LUMO} or E_{HOMO}), dipole moment (D), polarizability (α), chemical hardness (η), chemical softness (S), chemical potential (μ), electronegativity (χ), electrophilicity index (ω), etc. have been efficiently acting as global reactivity descriptors and atomic charges, Fukui function (FF), and local philicity as local reactivity descriptors in many QSAR models. In this chapter, we have focused on analyzing the predictive ability of computationally obtained electrophilicity index in comparison to the most commonly used *n*-octanol/water partition coefficient, using simple MLR technique.

2 Theory

Rationalizing toxic effects of chemical compounds requires the knowledge of the compound's mechanism of action toward its toxic effect, using which the toxicity of related compounds can be predicted. Toxicological studies involve receptor-mediated and non-receptor-mediated mechanisms among which the latter can be further divided into covalent and non-covalent categories. Accumulation of chemicals within cell membrane resulting in narcosis is an important form of aqueous toxicity following non-covalent mechanisms. While the narcotic action of nonpolar chemicals can be described in terms of hydrophobicity parameter alone, effective modelling of polar narcosis requires the inclusion of electronic terms to account for the polarization effect of an electronegative center in the molecule. In case of covalent mechanisms, a bond is

formed between the drug and the protein (DNA), and modelling this type of interaction cannot be done using hydrophobicity. This is because toxicity of a compound via covalent mechanism requires it to be electrophilic in nature, i.e., it must be susceptible to attack from electron-rich amino acid side chains [38]. The covalent mechanism can be well described by the HSAB principle “Among potential partners of a given electronegativity, hard likes hard and soft likes soft” [39]. Clearly, the reactivity of an electrophile-nucleophile interaction goes parallelly with the extent of the compound’s toxic response. Keeping this in mind, the assessment of electronic state of a chemical becomes extremely useful when it comes to prediction of its biological/ecotoxicological activities.

Conceptual density functional theory (CDFT) defines chemical concepts like chemical hardness (η) [40, 41] and electronegativity (χ) [42, 43] for a system consisting of N -electrons. Electronegativity, in a way, quantifies the chemical reactivity of the system, and its definition has undergone gradual modifications over time. Pauling’s electronegativity scale [42, 44], Mulliken’s formulation including ionization potential and electron affinity [45], Allred-Rochow’s introduction of force into the electronegativity theory [46], etc. have enriched the definition toward the various complexities of the concept [47–54]. Parr et al. [53] finally established the link between electronegativity and quantum chemistry by defining electronegativity as the negative of the chemical potential (μ) within the scope of DFT [18, 55, 56]. Combining this definition with that provided by Iczkowski and Margrave [49] stating χ as the negative of the variation in energy with that of the number of electrons, we get the following equation:

$$\chi = -\mu = -\left(\frac{\partial E}{\partial N}\right)_{v(\vec{r})} \quad (1)$$

where E is the total energy and $v(\vec{r})$ is the external potential.

Chemical hardness (η) for an N -electron system is described by Parr and Pearson [57] as the second-order derivative of energy with respect to number of electrons (N), i.e., first derivative of the chemical potential with respect to N (from Eq. 1):

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(\vec{r})} = \left(\frac{\partial \mu}{\partial N}\right)_{v(\vec{r})} \quad (2)$$

It is to be noted that earlier the definition of η had a factor of $1/2$ in it to make it symmetrical to the definition of μ . However, nowadays the convention without this factor is more commonly used [58, 59]. This equation is in keeping with the HSAB principle.

Finite difference approximation leads to the formulation of μ and η as [18]:

$$\mu = -\frac{(\text{IP} + \text{EA})}{2} \quad (3)$$

$$\eta = \text{IP} - \text{EA} \quad (4)$$

where IP and EA are the vertical ionization potential and electron affinity of the system calculated at constant external potential, i.e., at a fixed nuclear position. Now in order to bypass the time-consuming and computationally costly procedure of calculating E_N , E_{N+1} , and E_{N-1} for the evaluation of IP and EA, Koopmans' theorem for closed-shell molecules is employed which defines IP and EA as the negative of highest occupied (E_{HOMO}) and lowest unoccupied molecular orbital energies (E_{LUMO}), respectively. Hence μ and η become

$$\mu = \frac{E_{\text{LUMO}} + E_{\text{HOMO}}}{2} \quad (5)$$

$$\eta = E_{\text{LUMO}} - E_{\text{HOMO}} \quad (6)$$

In the context of an earlier proposal by Maynard et al. [60], Parr et al. [61] quantified electrophilicity index (ω) as the ground state stabilization energy of atoms/molecules on acceptance of electron(s) from a donor. The measure of ω , i.e., electrophilic power, can be considered analogous to the electrostatic power in classical physics which is formulated as:

$$\text{Power} = \frac{V^2}{R} \quad (7)$$

$$\omega = \frac{\mu^2}{2\eta} = \frac{\chi^2}{2\eta} \quad (8)$$

where V and R are potential and resistance, analogous to μ and η in Eq. 8.

Electrophilicity index can be successfully used in developing QSAR models where the electronic environment of the compounds being studied is more or less similar. The prediction of biological activity [62] or toxicity [63] of several pollutants including polychlorinated biphenyls and benzidine [64–67] has been successfully displayed using the electrophilicity index. Now coming to hydrophobicity, its contribution depends on the type of receptor site and mechanism of action. Very slight or no dependence on hydrophobic parameter is observed in case the receptor site is polar in nature or the reaction occurs in an aqueous environment [68–73]. However, hydrophobicity becomes a very important descriptor when it comes to the receptor site being nonpolar or the reactions occurring in a lipid environment [74–77]. In QSAR studies, the logarithm of the partition coefficient of the molecule in *n*-octanol and water is usually used as a measure of its hydrophobicity.

3 Method

3.1 Computational Details

Geometries of all the compounds considered in this study are optimized at a particular level of theory based on the type of elements present in the compounds. For a single study, all the compounds must be optimized at the same level to ensure comparability. Frequency analysis is done at the same level to check the absence of any imaginary frequency which would otherwise mean that the optimized structures do not lie on the minima in their respective potential energy surfaces. From the optimized geometry, quantities like chemical potential (μ), hardness (η), and global electrophilicity index (ω) are calculated following Koopmans' theorem using Eqs. 5, 6, and 8.

We have used HF/6-311G* level of theory in Gaussian 03 package [78] for the study against *Tetrahymena pyriformis* and B3LYP/6-31G(d) level in Gaussian 09 program package [79] for the toxicity study against *Pimephales promelas*.

3.2 Regression Analysis

The present chapter has been introduced to provide a clear overview of the computational methodologies such as the simplest and the most commonly used multiple linear regression method (MLR) to build QSAR models. The prediction performance has been analyzed on the basis of regression coefficient (R^2) and standard deviation (SD). In this method, the relative importance of each descriptor to the QSTR activity is indicated by the magnitudes of descriptors and the sign coming along with it. In accordance with the information of sign of the coefficients coming together with the magnitude of molecular descriptors, we can suggest whether the descriptors contribute negatively or positively to that specific activity.

The regression analysis requires the division of the entire dataset into two sets, namely, training and test sets. Construction of the regression model is carried out by considering experimental toxicity values (pLC_{50} or $pIGC_{50}$) as the dependent variable and the computed descriptors (in this case, ω , ω^2 and ω^3) as independent variables for the training set. The developed model is then employed to calculate the toxicity of compounds in the test set. The correlation coefficient (denoted by R^2) between these computed toxicity values and their respective experimental values define the efficiency of the model constructed. To remove any bias, a threefold cross-validation study is performed by dividing the dataset into three groups (sets A, B, and C) containing an equal number of molecules, among which two are taken as training sets and the other is taken as the test set. Simple QSAR model assumes a linear relationship between the physiochemical properties of a set of compounds (i.e., descriptors, denoted by x_n) and a certain biological or ecotoxicological activity (denoted by y).

Table 1
Data set for *Pimephales promelas*

Sl. No.	Set division	Compounds	Experimental pLC ₅₀
1	Set A	Hexachlorobenzene	6.38
2		1,2-Dichlorobenzene	4.40
3		Chlorobenzene	3.77
4		1,3-Dichlorobenzene	4.30
5		2-Xylene	3.48
6	Set B	1,2,4,5-Tetrachlorobenzene	5.85
7		1,4-Dichlorobenzene	4.56
8		4-Chlorotoluene	4.33
9		3-Xylene	3.82
10		Benzene	3.40
11	Set C	1,2,4-Trichlorobenzene	5.00
12		1,2,3-Trichlorobenzene	4.89
13		Bromobenzene	3.89
14		4-Xylene	4.21
15		Toluene	3.32

Experimental values are taken from ref. 81

MLR, being one of the earliest conventional and most commonly used techniques for the construction of QSAR/QSPR/QSTR models, is still used to date due to its some specific advantages like easy interpretability and simplistic form over several other approaches like partial least squares (PLS) analysis, principal component regression (PCR), etc., which are more abstract and difficult to interpret. However, the use of MLR technique is limited only to linear QSAR models containing molecular descriptors that are mathematically independent of one another. Thus, the efficiency of the models constructed is subject to the accuracy of the assumption that the relation between the activity and respective descriptors is linear in nature.

In the present chapter, we have established a QSTR analysis by constructing models proceeding with two different data sets (1) toxicity (96-h LC₅₀) of 15 benzene derivatives toward fathead minnow (*Pimephales promelas*) [80] (see Table 1) and (2) toxicity (pIGC₅₀) of 169 aliphatic compounds encompassing different groups like saturated alcohols, carboxylic acids, etc. [82] (see Table 2). Considerations have been made to avoid any large-scale computation or experiment by fetching all possible combinations of hydrophobic ($\log P$, $\{\log P\}^2$) and electronic (ω , ω^2 , ω^3) parameters to investigate their ability in predicting the toxicity.

3.2.1 *Pimephales promelas*

QSTR analysis has been done by initially dividing the dataset (15 molecules) into three equal sets (A, B, and C) containing five

Table 2
Data set for *Tetrahymena pyriformis*

Sl. No.	Compounds	Experimental pIC ₅₀
Saturated alcohols (<i>N</i> = 32)	Methyl alcohol	-2.6656
	Ethyl alcohol	-1.9912
	1-Propanol	-1.7464
	2-Propanol	-1.8819
	1-Butanol	-1.4306
	(+/-)-2-Butanol	-1.5420
	2-Methyl-1-propanol	-1.3724
	2-Pentanol	-1.1596
	3-Pentanol	-1.2437
	3-Methyl-2-butanol	-0.9959
	tert-Amyl alcohol	-1.1729
	2-Methyl-1-butanol	-0.9528
	3-Methyl-1-butanol	-1.0359
	2,2-Dimethyl-1-propanol	-0.8702
	2-Methyl-2-propanol	-1.7911
	1-Hexanol	-0.3789
	3,3-Dimethyl-1-butanol	-0.7368
	4-Methyl-1-pentanol	-0.6372
	1-Heptanol	0.1050
	2,4-Dimethyl-3-pentanol	-0.7052
	1-Octanol	0.5827
	2-Octanol	0.0011
	3-Octanol	0.0309
	1-Nonanol	0.8551
	2-Nonanol	0.6183
	3-Ethyl-2,2-dimethyl-3-pentanol	-0.1691
	1-Decanol	1.3354
	(+/-)-4-Decanol	0.8499
	3,7-Dimethyl-3-octanol	0.3404
	1-Undecanol	1.9547
	1-Dodecanol	2.1612
	1-Tridecanol	2.4497
Carboxylic acids (<i>N</i> = 28)	Propionic acid	-0.5123
	Butyric acid	-0.5720
	Valeric acid	-0.2674
	Hexanoic acid	-0.2083
	Heptanoic acid	-0.1126
	Octanoic acid	0.0807
	Nonanoic acid	0.3509
	Decanoic acid	0.5063
	Undecanoic acid	0.8983
	Isobutyric acid	-0.3334
	Isovaleric acid	-0.3415
	Trimethylacetic acid	-0.2543
	3-Methylvaleric acid	-0.2331
	4-Methylvaleric acid	-0.2724
	2-Ethylbutyric acid	-0.1523
	2-Propylpentanoic acid	0.0258

(continued)

Table 2
(continued)

Sl. No.	Compounds	Experimental pIGC ₅₀
	2-Ethylhexanoic acid	0.0756
	Succinic acid	-0.9395
	Glutaric acid	-0.6387
	Adipic acid	-0.6060
	Pimelic acid	-0.5845
	3,3-Dimethylglutaric acid	-0.6643
	Suberic acid	-0.5116
	Sebacic acid	-0.2676
	1,10-Decanedicarboxylic acid	-0.0863
	Crotonic acid	-0.5448
	trans-2-Pentenoic acid	-0.2774
	trans-2-Hexenoic acid	-0.1279
Monoesters (<i>N</i> = 31)	Ethyl acetate	-1.2968
	Propyl acetate	-1.2382
	Isopropyl acetate	-1.5900
	Butyl acetate	-0.4864
	Amyl acetate	0.1625
	Hexyl acetate	-0.0087
	Octyl acetate	1.0570
	Decyl acetate	1.8794
	Ethyl propionate	-0.9450
	Butyl propionate	0.1704
	Isobutyl propionate	-0.6935
	Propyl propionate	-0.8148
	tert-Butyl propionate	-0.4095
	Ethyl butyrate	-0.4903
	Ethyl isobutyrate	-1.2709
	Ethyl valerate	-0.3580
	Propyl butyrate	-0.4138
	Butyl butyrate	0.5157
	Propyl valerate	0.0094
	Amyl propionate	-0.0431
	Ethyl hexanoate	0.0637
	Methyl butyrate	-1.2463
	Methyl valerate	-0.8448
	Methyl hexanoate	-0.5611
	Methyl heptanoate	0.1039
	Methyl octanoate	0.5358
	Methyl nonanoate	1.0419
	Methyl decanoate	1.3778
	Methyl undecanoate	1.4248
	Methyl formate	-1.4982
	tert-Butyl formate	-1.3719
Diesters (<i>N</i> = 20)	Diethyl malonate	-0.9975
	Diethyl sebacate	1.3536
	Diethyl suberate	0.7018
	Diethyl succinate	-0.8511
	Dimethyl malonate	-1.2869

(continued)

Table 2
(continued)

Sl. No.	Compounds	Experimental pIGC_{50}
	Dibutyl adipate	0.7918
	Dimethyl succinate	-1.0573
	Diethyl adipate	-0.1265
	Dimethyl brassylate	1.6536
	Dimethyl sebacate	1.0106
	Dimethyl suberate	0.2962
	Diethyl pimelate	0.4069
	Dibutyl suberate	1.6556
	Diethyl butylmalonate	0.5566
	Diethyl ethylmalonate	-0.2422
	Diethyl-3-oxopimelate	-0.3778
	Diethyl-4-oxopimelate	-0.6378
	Diethyl methylmalonate	-0.5114
	Diethyl propylmalonate	0.1341
	Dibutyl succinate	0.5123
Ketones ($N = 15$)	Acetone	-2.2036
	2-Butanone	-1.7457
	2-Pentanone	-1.2224
	3-Pentanone	-1.4561
	4-Methyl-2-pentanone	-1.2085
	2-Heptanone	-0.4872
	5-Methyl-2-hexanone	-0.6459
	4-Heptanone	-0.6690
	2-Octanone	-0.1455
	2-Nonanone	0.6598
	2-Decanone	0.5822
	3-Decanone	0.6265
	2-Undecanone	1.5346
	2-Dodecanone	1.6696
	7-Tridecanone	1.5214
Amino alcohols ($N = 18$)	2-(Methylamino)ethanol	-1.8202
	4-Amino-1-butanol	-0.9752
	2-(Ethylamino)ethanol	-1.6491
	2-Propylaminoethanol	-1.6842
	DL-2-amino-1-pentanol	-0.6718
	3-Amino-2,2-dimethyl-1-propanol	-0.9246
	6-Amino-1-hexanol	-0.9580
	DL-2-amino-1-hexanol	-0.5848
	DL-2-amino-3-methyl-1-butanol	-0.5852
	2-Amino-3,3-dimethyl-butanol	-0.7178
	2-Amino-3-methyl-1-pentanol	-0.6594
	2-Amino-4-methyl-pentanol	-0.6191
	2-(Tert-butylamino)ethanol	-1.6730
	Diethanolamine	-1.7941
	1,3-Diamino-2-hydroxy-propane	-1.4275
	N-Methyldiethanol amine	-1.8338
	3-(Methylamino)-1,2-propanediol	-1.5341
	Triethanolamine	-1.7488

(continued)

Table 2
(continued)

Sl. No.	Compounds	Experimental pIGC_{50}
Unsaturated alcohols ($N = 25$)	2-Methyl-3-buten-2-ol 4-Pentyn-1-ol 2-Methyl-3-butyn-2-ol trans-3-Hexen-1-ol cis-3-Hexen-1-ol 5-Hexyn-1-ol 3-Methyl-1-pentyn-3-ol 4-Hexen-1-ol 5-Hexen-1-ol 4-Pentyn-2-ol 5-Hexyn-3-ol 3-Heptyn-1-ol 4-Heptyn-2-ol 3-Octyn-1-ol 3-Nonyn-1-ol 2-Propen-1-ol 2-Buten-1-ol $(+/-)$ -3-Buten-2-ol cis-2-Buten-1,4-diol cis-2-Penten-1-ol 3-Penten-2-ol trans-2-Hexen-1-ol 1-Hexen-3-ol cis-2-Hexen-1-ol trans-2-Octen-1-ol	-1.3889 -1.4204 -1.3114 -0.7772 -0.8091 -1.2948 -1.3226 -0.7540 -0.8411 -1.6324 -1.4043 -0.3231 -0.6160 0.0170 0.3401 -1.9178 -1.4719 -1.0529 -2.1495 -1.1052 -1.4010 -0.4718 -0.8113 -0.7767 0.3654

Experimental values are taken from ref. 83

Table 3
Combinations of sets used in the training and test sets

	Training set	Test set
Case 1	Set A + set B	Set C
Case 2	Set A + set C	Set B
Case 3	Set B + set C	Set A

Reprinted from Pal et al. [80] with permission

molecules each. Out of these three sets, two were considered as the training set, and the rest as test set (see Table 3).

Out of the studied four parameters, i.e., global electrophilicity index (ω), its square term (ω^2), hydrophobicity ($\log P$), and $(\log P)^2$, construction of models has been done by choosing one parameter at a time as input to bring out a comparative interpretation of the results. The result of our investigation using single parameter-based QSTR models on the training and the test set has been tabulated (Tables 4 and 5), and the effectiveness of the correlation

Table 4

Regression models on the training sets in case of *Pimephales promelas* using ω , ω^2 , $\log P$, and $(\log P)^2$ as descriptors

Sl. No.		Regression equations	R^2	R_{ad}^2	SD
1	Case 1	$pLC_{50} = 1.55332 + 1.3107 * \omega$	0.921	0.911	0.293
2		$pLC_{50} = 3.04853 + 0.26165 * \omega^2$	0.926	0.917	0.283
3		$pLC_{50} = 1.17553 + 0.92114 * (\log P)$	0.929	0.921	0.276
4		$pLC_{50} = 2.88654 + 0.11495 * (\log P)^2$	0.931	0.922	0.273
5	Case 2	$pLC_{50} = 1.77513 + 1.18538 * \omega$	0.870	0.854	0.343
6		$pLC_{50} = 3.08854 + 0.24448 * \omega^2$	0.907	0.895	0.290
7		$pLC_{50} = 1.04942 + 0.93738 * (\log P)$	0.939	0.932	0.234
8		$pLC_{50} = 2.86562 + 0.11287 * (\log P)^2$	0.926	0.917	0.258
9	Case 3	$pLC_{50} = 1.88129 + 1.1977 * \omega$	0.803	0.779	0.367
10		$pLC_{50} = 3.06405 + 0.2821 * \omega^2$	0.838	0.818	0.333
11		$pLC_{50} = 0.97142 + 0.98868 * (\log P)$	0.944	0.937	0.195
12		$pLC_{50} = 2.60524 + 0.14291 * (\log P)^2$	0.960	0.955	0.164

Reprinted from Pal et al. [80] with permission

Table 5

R^2 and SD values obtained from MLR analysis on the test sets of *Pimephales promelas* using ω , ω^2 , $\log P$, and $(\log P)^2$ as descriptors

Compounds	ω		ω^2		$\log P$		$(\log P)^2$		
	R^2	SD	R^2	SD	R^2	SD	R^2	SD	
Case 1	1,2,4-Trichlorobenzene	0.693	0.504	0.731	0.391	0.961	0.140	0.941	0.149
	1,2,3-Trichlorobenzene								
	Bromobenzene								
	4-Xylene								
	Toluene								
Case 2	1,2,4,5-Tetrachlorobenzene	0.916	0.261	0.926	0.221	0.950	0.226	0.976	0.132
	1,4-Dichlorobenzene								
	4-Chlorotoluene								
	3-Xylene								
	Benzene								
Case 3	Hexachlorobenzene	0.972	0.195	0.991	0.136	0.934	0.353	0.947	0.391
	1,2-Dichlorobenzene								
	Chlorobenzene								
	1,3-Dichlorobenzene								
	2-Xylene								

has been judged in terms of coefficient of determination (R^2). A representative regression graph is provided as reference to highlight the obtained result (see Fig. 1). It is quite transparent that easily computable electronic descriptor ω^2 and ω provide comparable result with that obtained from the widely used lipophilic descriptors, $\log P$ and $(\log P)^2$.

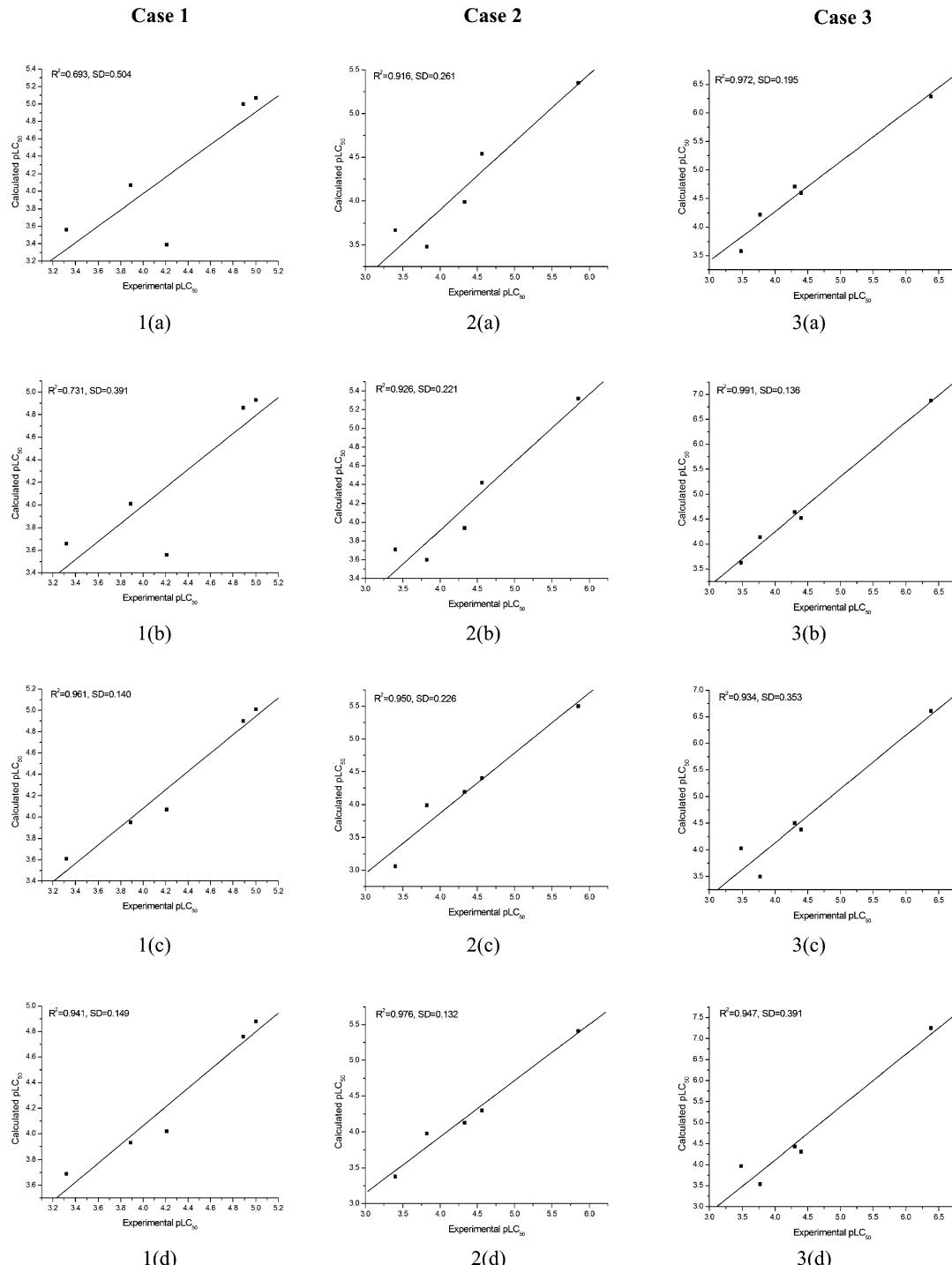


Fig. 1 Plots of experimental versus calculated values of pLC_{50} for the test set with models constructed using MLR. (a-d) Represent plots w.r.t. ω , ω^2 , $\log P$, and $(\log P)^2$, respectively, for cases 1–3. (Reprinted from Pal et al. [80] with permission)

3.2.2 Tetrahymena pyriformis

In order to develop linear prediction models, the investigation toward *Tetrahymena pyriformis* (a freshwater protozoan) has been performed by considering toxicity (pIGC_{50}) of 169 aliphatic compounds as dependent variable, whereas all possible combinations of electronic descriptors (ω , ω^2 , ω^3) and lipophilic descriptors $\{\log P, (\log P)^2\}$ used as independent variables in the constructed models [82]. Here the study includes two different approaches, (1) simple MLR analysis on each of the seven group of compounds (without dividing the complete set into training and test sets) and (2) by diving each of the studied seven group of compounds (i.e., saturated alcohols, carboxylic acid, monoesters, diesters, ketones, amino alcohols, and unsaturated alcohols) into three equal sets and taking two of them as training set and the third as the test set (Table 3). We have also found, among several available descriptors, which combination(s) is/are efficient to provide a good correlation coefficient (R^2). The results for the complete set study have been presented in Table 6 and that for the threefold cross-validation study in Table 7.

The electronic factors used for constructing single- and double-parameter QSARs provide satisfactory result for the set of compounds with similar electronic environment. When they are used along with hydrophobic descriptor, a substantial improvement in estimation power has been achieved.

Table 6

R^2 values obtained from MLR analysis on complete sets for *Tetrahymena pyriformis* using ω , ω^2 , ω^3 , $\log P$, $(\log P)^2$, and their combinations separately as descriptors

	R^2 values w.r.t.											
	ω	ω^2	ω^3	$\log P$	$(\log P)^2$	$\omega, \log P$	$\omega^2, \log P$	$\omega, (\log P)^2$	$\omega^2, (\log P)^2$	ω, ω^2	$\log P, (\log P)^2$	
Saturated alcohols	0.715	0.709	0.703	0.981	0.895	0.981	0.982	0.905	0.906	0.732	0.983	
Carboxylic acids	0.750	0.734	0.728	0.919	0.882	0.919	0.919	0.917	0.918	0.785	0.937	
Monoesters	0.756	0.758	0.760	0.930	0.889	0.932	0.932	0.889	0.900	0.763	0.933	
Diesters	0.739	0.733	0.725	0.910	0.814	0.958	0.957	0.911	0.912	0.748	0.912	
Ketones	0.779	0.771	0.762	0.975	0.881	0.975	0.975	0.959	0.959	0.876	0.975	
Amino alcohols	0.748	0.746	0.743	0.340	0.156	0.879	0.878	0.857	0.853	0.748	0.387	
Unsaturated alcohols	0.301	0.296	0.288	0.868	0.790	0.868	0.868	0.790	0.799	0.302	0.890	

Table 7

R^2 values obtained from MLR analysis on the cross-validated sets for *Tetrahymena pyriformis* using ω , ω^2 , ω^3 , $\log P$, $(\log P)^2$, and their combinations separately as descriptors

R^2 values w.r.t.													
		ω	ω^2	ω^3	$\log P$	$(\log P)^2$	ω ,	ω^2 ,	$(\log P)$	$(\log P)^2$	ω , ω^2	ω^2 , $(\log P)^2$	$\log P$, $(\log P)^2$
Saturated alcohols	Case 1	0.794	0.790	0.785	0.975	0.843	0.973	0.973	0.851	0.852	0.902	0.976	
	Case 2	0.497	0.503	0.509	0.980	0.896	0.985	0.985	0.894	0.895	0.474	0.982	
	Case 3	0.808	0.808	0.807	0.988	0.936	0.989	0.990	0.935	0.935	0.806	0.988	
Carboxylic acids	Case 1	0.711	0.700	0.688	0.874	0.934	0.874	0.874	0.942	0.942	0.771	0.917	
	Case 2	0.792	0.776	0.759	0.948	0.918	0.940	0.940	0.946	0.946	0.865	0.959	
	Case 3	0.719	0.710	0.701	0.939	0.910	0.937	0.934	0.943	0.943	0.754	0.941	
Monoesters	Case 1	0.846	0.849	0.852	0.955	0.916	0.957	0.957	0.916	0.916	0.847	0.954	
	Case 2	0.829	0.834	0.838	0.940	0.877	0.933	0.933	0.877	0.877	0.816	0.928	
	Case 3	0.632	0.627	0.621	0.925	0.899	0.928	0.929	0.895	0.897	0.539	0.927	
Diesters	Case 1	0.546	0.539	0.534	0.961	0.856	0.944	0.863	0.878	0.883	0.569	0.978	
	Case 2	0.759	0.749	0.737	0.931	0.945	0.935	0.931	0.906	0.903	0.766	0.885	
	Case 3	0.870	0.853	0.836	0.866	0.786	0.952	0.954	0.937	0.939	0.947	0.880	
Ketones	Case 1	0.802	0.799	0.796	0.981	0.970	0.981	0.981	0.988	0.989	0.844	0.978	
	Case 2	0.773	0.769	0.765	0.981	0.977	0.979	0.979	0.971	0.972	0.825	0.976	
	Case 3	0.864	0.854	0.845	0.979	0.804	0.979	0.979	0.925	0.926	0.873	0.952	
Amino alcohols	Case 1	0.844	0.841	0.836	0.252	0.108	0.855	0.863	0.841	0.843	0.844	0.260	
	Case 2	0.873	0.871	0.867	0.294	0.205	0.888	0.879	0.875	0.866	0.871	0.368	
	Case 3	0.610	0.611	0.820	0.622	0.306	0.888	0.894	0.858	0.855	0.607	0.695	
Unsaturated alcohols	Case 1	0.620	0.617	0.612	0.813	0.829	0.813	0.813	0.833	0.834	0.015	0.842	
	Case 2	0.417	0.439	0.454	0.919	0.893	0.764	0.761	0.840	0.850	0.353	0.925	
	Case 3	0.171	0.194	0.217	0.876	0.922	0.793	0.791	0.776	0.779	0.130	0.909	

4 Conclusion

This chapter has focused mainly on the predictive ability of electrophilicity in developing robust QSTR models. It also demonstrates a comparative study between electrophilicity and the widely used lipophilic parameter $\log P$ (*n*-octanol/water partition coefficient). The choice of very simple and most common multiple linear regression (MLR) technique has been made to make the computation relatively easier and to check its accuracy in a simple manner. In our approach, an easily computable global parameter, electrophilicity index (ω), and its square term (ω^2) have been used as predictor descriptors to generate structure-activity/property/toxicity models, and comparisons have been made for the confirmation of the consistency check of the obtained results by employing hydrophobicity, i.e., $\log P$ and its square term $(\log P)^2$. The simple

electrophilic parameters (ω , ω^2) can be satisfactorily used as basic descriptors toward toxicity prediction, upon which improvements can be made by incorporating additional descriptors like $\log P$, if required. This study may help in extracting meaningful judgment on structure-activity prediction models which can be used in more practical applications by using the knowledge of data set division.

Acknowledgments

PKC would like to thank the Volume Editor, Prof. Kunal Roy, for kindly inviting him to contribute a chapter entitled, “Quantitative Structure-Toxicity Relationship Models Based on Hydrophobicity and Electrophilicity” for the book *Ecotoxicological QSARs*. He also thanks DST, New Delhi, for the J. C. Bose National Fellowship. SS thanks CSE for the computational facilities. GJ and RP thank IIT, Kharagpur, and CSIR, respectively, for their fellowships.

References

- Cros A (1863) Action de l'alcool amylique sur l'organisme. Faculté de médecine de Strasbourg, France
- Brown AC, Fraser TR (1868) V.—On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Earth Environ Sci Trans R Soc Edinburgh* 25:151–203
- Richardson B (1869) Physiological research on alcohols. *Med Times Gazzette* 2:703–706
- Mills EJ (1884) On melting point and boiling point as related to composition. *Philos Mag* 17:173–187
- Richet C (1893) Comptes rendus des séances de la société de biologie et de ses filiales. *Soc Biol Ses Fil* 9:775–776
- Meyer H (1899) The theory of alcohol narcosis [Zur Theorie der Alkoholnarkose] arch. *Exp Pathol Pharmakol* 42:109–118
- Overton CE (1901) Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie. Fischer, Jena
- Hammett LP (1935) Some relations between reaction rates and equilibrium constants. *Chem Rev* 17:125–136
- Hammett LP (1937) The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc* 59:96–103
- Ferguson J (1939) The use of chemical potentials as indices of toxicity. *Proc R Soc Lond Ser B* 127:387–404
- Albert A, Rubbo S, Goldacre R, Davey M, Stone J (1945) The influence of chemical constitution on antibacterial activity. Part II: a general survey of the acridine series. *Br J Exp Pathol* 26:160
- Albert A (1985) Selective toxicity, 7th edn. Chapman & Hall, London, p 33
- Roblin RO Jr, Bell PH (1942) Structure and reactivity of sulphanilamide type compounds. *J Am Chem Soc* 64:2905–2917
- Taft RW Jr (1952) Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J Am Chem Soc* 74:3120–3128
- Taft R (1956) Separation of polar, steric and resonance effects in reactivity. In: Steric effects in organic chemistry. Wiley, New York, pp 556–675
- Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178
- Hansch C, Fujita T (1964) $p\text{-}\sigma\text{-}\pi$ analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 86:1616–1626
- Calais JL (1993) Density-functional theory of atoms and molecules. R.G. Parr and W. Yang,

- Oxford University Press, New York, Oxford, 1989. *Int J Quantum Chem* 47:101–101
19. Chermette H (1999) Chemical reactivity indexes in density functional theory. *J Comput Chem* 20:129–154
 20. Geerlings P, De Proft F, Langenaeker W (2003) Conceptual density functional theory. *Chem Rev* 103:1793–1874
 21. Chattaraj PK, Roy D, Giri S, Mukherjee S, Subramanian V, Parthasarathi R, Bultinck P, Van Damme S (2007) An atom counting and electrophilicity based QSTR approach. *J Chem Sci* 119:475–488
 22. Chattaraj PK, Parr RG (1993) Density functional theory of chemical hardness. In: *Chemical hardness*. Springer, Berlin/Heidelberg, pp 11–25
 23. Chattaraj PK, Poddar A, Maiti B (2002) Chemical reactivity and dynamics within a density-based quantum mechanical framework. In: *Reviews of modern quantum chemistry: a celebration of the contributions of Robert G Parr*, vol 2. World Scientific, River Edge, pp 871–935
 24. Chattaraj PK (2009) Chemical reactivity theory: a density functional view. CRC Press, Boca Raton
 25. Kohn W, Becke AD, Parr RG (1996) Density functional theory of electronic structure. *J Phys Chem* 100:12974–12980
 26. Pearson RG (1990) Hard and soft acids and bases—the evolution of a chemical concept. *Coord Chem Rev* 100:403–425
 27. Parr RG, Chattaraj PK (1991) Principle of maximum hardness. *J Am Chem Soc* 113:1854–1855
 28. Chattaraj PK, Sengupta S (1996) Popular electronic structure principles in a dynamical context. *J Phys Chem* 100:16126–16130
 29. Chamorro E, Chattaraj PK, Fuentealba P (2003) Variation of the electrophilicity index along the reaction path. *J Phys Chem A* 107:7068–7072
 30. Parthasarathi R, Elango M, Subramanian V, Chattaraj PK (2005) Variation of electrophilicity during molecular vibrations and internal rotations. *Theor Chem Acc* 113:257–266
 31. Noorizadeh S (2007) Is there a minimum electrophilicity principle in chemical reactions? *Chin J Chem* 25:1439–1444
 32. Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. *Curr Med Chem* 8:1573–1588
 33. Balaban AT (1995) Chemical graphs: looking back and glimpsing ahead. *J Chem Inf Comput Sci* 35:339–350
 34. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* 35:1039–1045
 35. Kim KH (1993) 3D-quantitative structure-activity relationships: describing hydrophobic interactions directly from 3D structures using a comparative molecular field analysis (CoMFA) approach. *Quant Struct-Act Relat* 12:232–238
 36. Raevsky O, Skvortsov V (2005) Quantifying hydrogen bonding in QSAR and molecular modeling. *SAR QSAR Environ Res* 16:287–300
 37. Kubinyi H (2001) Hydrogen bonding, the last mystery in drug design. In: *Pharmacokinetic optimization in drug research*. Wiley-VCH: Weinheim, Germany, pp 513–524
 38. Aptula AO, Roberts DW (2006) Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem Res Toxicol* 19:1097–1105
 39. Chattaraj PK, Lee H, Parr RG (1991) HSAB principle. *J Am Chem Soc* 113:1855–1856
 40. Pearson RG (1997) Chemical hardness. Wiley-VCH, Weinheim
 41. Chattaraj PK, Parr RG (1993) Density functional theory of chemical hardness. In: Sen KD, Mingos DMP (eds) *Chemical hardness, Structure and bonding*, vol 80. Springer, Berlin
 42. Pauling L (1960) *The nature of the chemical bond*, vol 260. Cornell University Press, Ithaca
 43. Sen K, Jorgenson C (1987) *Structure and bonding, Electronegativity*, vol 66. Springer, Berlin
 44. Pauling L (1932) The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. *J Am Chem Soc* 54:3570–3582
 45. Mulliken RS (1934) A new electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities. *J Chem Phys* 2:782–793
 46. Allred AL, Rochow EG (1958) A scale of electronegativity based on electrostatic force. *J Inorg Nucl Chem* 5:264–268
 47. Gordy W (1946) A relation between bond force constants, bond orders, bond lengths, and the electronegativities of the bonded atoms. *J Chem Phys* 14:305–320

48. Sanderson R (1988) Principles of electronegativity Part I. General nature. *J Chem Educ* 65:112
49. Iczkowski RP, Margrave JL (1961) Electronegativity. *J Am Chem Soc* 83:3547–3551
50. Klopman G (1968) Chemical reactivity and the concept of charge-and frontier-controlled reactions. *J Am Chem Soc* 90:223–234
51. Hinze J, Jaffe HH (1962) Electronegativity. I. Orbital electronegativity of neutral atoms. *J Am Chem Soc* 84:540–546
52. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity-equalization method for the calculation of atomic charges in molecules. *J Am Chem Soc* 108:4315–4320
53. Parr RG, Donnelly RA, Levy M, Palke WE (1978) Electronegativity: the density functional viewpoint. *J Chem Phys* 68:3801–3807
54. Parr RG, Weitao Y (1989) Density-functional theory of atoms and molecules. Oxford University Press, New York
55. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. *Phys Rev* 136:B864
56. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140:A1133
57. Parr RG, Pearson RG (1983) Absolute hardness: companion parameter to absolute electronegativity. *J Am Chem Soc* 105:7512–7516
58. Berkowitz M, Parr RG (1988) Molecular hardness and softness, local hardness and softness, hardness and softness kernels, and relations among these quantities. *J Chem Phys* 88:2554–2557
59. Ayers PW (2001) Strategies for computing chemical reactivity indices. *Theor Chem Acc* 106:271–279
60. Maynard A, Huang M, Rice W, Covell D (1998) Reactivity of the HIV-1 nucleocapsid protein p7 zinc finger domains from the perspective of density-functional theory. *Proc Natl Acad Sci U S A* 95:11578–11583
61. Parr RG, Szentpály L, Liu S (1999) Electrophilicity index. *J Am Chem Soc* 121:1922–1924
62. Roy D, Pal N, Mitra A, Bultinck P, Parthasarathi R, Subramanian V, Chattaraj PK (2007) An atom counting strategy towards analyzing the biological activity of sex hormones. *Eur J Med Chem* 42:1365–1369
63. Parthasarathi R, Subramanian V, Roy DR, Chattaraj PK (2004) Electrophilicity index as a possible descriptor of biological activity. *Bioorg Med Chem Lett* 12:5533–5543
64. Parthasarathi R, Padmanabhan J, Subramanian V, Maiti B, Chattaraj PK (2003) Chemical reactivity profiles of two selected polychlorinated biphenyls. *J Phys Chem A* 107:10346–10352
65. Parthasarathi R, Padmanabhan J, Subramanian V, Maiti B, Chattaraj PK (2004) Toxicity analysis of 33'44'5-pentachloro biphenyl through chemical reactivity and selectivity profiles. *Curr Sci* 86:535
66. Roy D, Parthasarathi R, Subramanian V, Chattaraj PK (2006) An electrophilicity based analysis of toxicity of aromatic compounds towards *Tetrahymena pyriformis*. *QSAR Comb Sci* 25:114–122
67. Padmanabhan J, Parthasarathi R, Subramanian V, Chattaraj PK (2006) Group philicity and electrophilicity as possible descriptors for modeling ecotoxicity applied to chlorophenols. *Chem Res Toxicol* 19:356–364
68. Hermens J, Busser F, Leeuwanch P, Musch A (1985) Quantitative correlation studies between the acute lethal toxicity of 15 organic halides to the guppy (*Poecilia Reticulata*) and chemical reactivity towards 4-nitrobenzylpyridine. *Toxicol Environ Chem* 9:219–236
69. Roberts DW, Schultz TW, Wolf EM, Aptula AO (2009) Experimental reactivity parameters for toxicity modeling: application to the acute aquatic toxicity of SN2 electrophiles to *Tetrahymena pyriformis*. *Chem Res Toxicol* 23:228–234
70. Schultz TW, Netzeva TI, Roberts DW, Cronin MT (2005) Structure – toxicity relationships for the effects to *Tetrahymena pyriformis* of aliphatic, carbonyl-containing, α , β -unsaturated chemicals. *Chem Res Toxicol* 18:330–341
71. Suter GW (1989) Aquatic toxicology and environmental fate: eleventh volume, vol 11. ASTM International Chem Biol Drug Des, Philadelphia
72. Hansch C, Kurup A, Garg R, Gao H (2001) Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem Rev* 101:619–672
73. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD (2002) Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem Rev* 102:783–812
74. Roberts D, Williams D (1982) The derivation of quantitative correlations between skin sensitisation and physio-chemical parameters for alkylating agents, and their application to experimental data for sulphones. *J Theor Biol* 99:807–825
75. Roberts D, Goodwin B, Williams D, Jones K, Johnson A, Alderson J (1983) Correlations

- between skin sensitization potential and chemical reactivity for p-nitrobenzyl compounds. *Food Chem Toxicol* 21:811–813
76. Roberts D, Basketter D (1990) A quantitative structure activity/dose response relationship for contact allergic potential of alkyl group transfer agents. *Contact Dermatitis* 23:331–335
77. Roberts DW, Aptula AO, Patlewicz G (2007) Electrophilic chemistry related to skin sensitization. Reaction mechanistic applicability domain classification for a published data set of 106 chemicals tested in the mouse local lymph node assay. *Chem Res Toxicol* 20:44–60
78. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr. JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision C.02, Wallingford, CT
79. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA et al. (2009) Gaussian 09, Revision D.01, Wallingford CT, vol 121, pp 150–166
80. Pal R, Jana G, Sural S, Chattaraj PK (2018) Hydrophobicity versus electrophilicity: a new protocol toward quantitative structure-toxicity relationship. *Chem Bio Drug Des* 93:1083–1095. <https://doi.org/10.1111/cbdd.13428>
81. Bertinetto C, Duce C, Solaro R, Héberger K (2013) Modeling of the acute toxicity of benzene derivatives by complementary QSAR methods. *MATCH Commun Math Comput Chem* 70:1005–1021
82. Jana G, Pal R, Sural S, Chattaraj PK (2019) Quantitative structure – toxicity relationship: an “in silico study” using electrophilicity and hydrophobicity as descriptors. *Int J Quantum Chem* (Provisionally accepted)
83. Schultz TW (1997) TETRATOX database. *Toxicol Methods* 7:289. <http://www.vet.utk.edu/TETRATOX/>



Chapter 28

Environmental Toxicity (Q)SARs for Polymers as an Emerging Class of Materials in Regulatory Frameworks, with a Focus on Challenges and Possibilities Regarding Cationic Polymers

Hans Sanderson, Kabiruddin Khan, Anna M. Brun Hansen, Kristin Connors, Monica W. Lam, Kunal Roy, and Scott Belanger

Abstract

Polymers are highly diverse and understudied materials from an environmental toxicity point of view. For the past decades, polymers have largely been out of scope regarding detailed safety assessment in most regulatory programs as they are assumed to not possess relevant toxicological properties due to their size. This regulatory exclusion is currently being reconsidered. This chapter discusses the available information about selected cationic polymers and outlines (Q)SAR ((Quantitative) Structure-Activity Relationship) approaches that could be used to develop new models to demonstrate potential aquatic toxicity of polymers. The amount of publicly available, high-quality environmental toxicity data on industrial polymers such as cationic polyquaterniums is extremely limited. Given the large size (dimension and molecular weight) of the materials, typical hydrophobicity-driven toxicity is not expected. Relevant descriptors for cationic polymers need to be identified. Molecular weight and charge density are well-known physical-chemical attributes that are suspected to be correlated with aquatic toxicity, but there might be other relevant descriptors as well.

We suggest models that predict polymer properties may be useful for estimating relevant properties regarding toxicity. Moreover, novel fragment-based 2D and 3D hologram (Q)SAR ($H(Q)$ SAR) may prove relevant in determining these properties that can be used to derive hypotheses about toxic mechanisms and guide experimental test designs. In a regulatory context, (Q)SARs have to be transparent and scientifically robust which extends to fragment-based models that may be useful in categorizing polymers. The toxicity of category members can then be experimentally explored, and read-across strategies developed within the category.

The authors of this chapter are pursuing polymer (Q)SAR strategies in the coming years via generation of novel experimental and computational data on polyquaterniums. We will also evaluate the potential for fragment-based (Q)SARs for polymers in REACH.

Key words Polymers, Chemometric tools, Descriptors, Environmental toxicity, Cationic, Polyquaterniums

1 Introduction

Polymers are large macromolecules consisting of repeating monomer units. Polymers are an exceptionally diverse group of compounds and are used in a large range of applications. Polymers may be described as linear, branched, or cross-linked. They may exist as homopolymers and have one repeated monomer, or they may be copolymers and contain two or more monomers combined in random or ordered approaches. While many polymers in commerce are synthetic, there are also natural polymers or biopolymers that are important building blocks of life, such as amino acids, proteins, and cellulose. Many polymers are soluble and dispersible in water. These are often used in consumer and personal care products, pharmaceuticals, water treatment, and wood preservation. Novel uses and applications in biomedical and nano-industries are expected to grow significantly in the coming years.

1.1 Current Regulatory View of Polymers

Historically, polymers have been subject to exemptions or reduced regulatory requirements in countries practicing chemical legislation. The assumption was that the high molecular weight and reduced reactivity of polymers in environmental compartments were viewed as lower risk to human health and the environment when compared to lower molecular weight substances. Most chemical legislations have adopted 500 Da as the highest molecular weight in scope, which is based on one component of Lipinski rules of bioavailability whereby substances that are >500 Da are considered less bioavailable. Since polymers are predominantly represented by higher molecular weight components, it has long been assumed that much of the polymer is not bioavailable and inert in the environment and the focus of chemical registrations and data needs has been more on lower molecular weight impurities and unreacted monomers as well as additives (non-intentionally added substances, NIAS, and intentionally added substances, IAS). The focus of most regulatory programs is generally on new polymers, not existing polymers already in commerce. However, through K-REACH, Korea is the first country to require registration of current polymers, with all existing chemistries greater than 1 metric ton requiring registration by 2030. In addition, Environment Canada has polymers included in their Chemical Management Program, and the agency recently published draft safety assessments for the polyamines (December 2016).

For new polymers requiring registration, the criteria used by many global regulatory agencies to identify “polymers of low concern” include molecular weight and levels of monomers, in addition to the presence of specific structural features or functional groups. The “polymers of low concern” concept is intended to guide prioritization of polymer review by regulators. While this

approach has been the practice of many global regulatory agencies, polymers have been exempted from chemical registrations by the European Chemicals Agency (ECHA).

Previous guidance from ECHA on polymer registration was done with the view that polymers would eventually require registration. Below is an excerpt from Article 138 Section 2 [1] (ECHA):

“The (European) Commission may present legislative proposals as soon as a practicable and cost-efficient way of selecting polymers for registration on the basis of sound technical and valid scientific criteria can be established, and after publishing a report on the following:

- (a) The risks posed by polymers in comparison with other substances;
- (b) The need, if any, to register certain types of polymer, taking account of competitiveness and innovation on the one hand and the protection of human health and the environment on the other.”

In recent years, the simplified view of the potential risks associated with polymers has received increased scrutiny, and with the polymer exemption under REACH being revisited, the expected outcome is starting approximately in the year 2023; polymers identified as “polymers requiring registration” (PRR) will come within the scope of REACH. This activity suggests the potential reapplication of current REACH methods, such as categorization and use of (Q)SAR, to characterize and estimate safety data and even to support grouping and read-across approaches for these materials. The use of (Q)SAR may be of special interest to regulators due to the limited publicly available data for polymers. It is likely many suppliers and downstream users of polymers have privately held data, but this information is often protected as confidential business information (CBI) due to the competitive environment of polymer innovation. Without unrestricted access to environmental safety data on polymers, the need for (Q)SAR development becomes more acute and relevant for polymer registration with the potential to use (Q)SAR to predict toxicity of polymers in lieu of testing.

As mentioned above, the “polymers of low concern” approach is based on the assumption that there is little toxicological concern expected for polymers due to their decreased bioavailability as a result of their significant molecular weight and their inability to cross biological membranes. Many (Q)SAR models used to estimate environmental fate and effects have not included large molecular weight chemistries in their training set; therefore, (Q)SARs are not intended to be useful predictors for large molecular weight and complex polymers. Most (Q)SARs used to estimate toxicity are based on log K_{ow} as a surrogate for hydrophobicity/hydrophilicity or contributions of certain functional groups or structural features. Log K_{ow} are generally estimated using fragment-based approaches, leading to a gross overestimation for high molecular weight

polymers. Furthermore, polymers are classic forms of UVCBs (Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials) which are difficult to devise (Q)SARs for as they are not discrete chemical entities. There is, however, a somewhat historical (Q)SAR for polymers in ECOSAR in the EPI Suite [2], which is also built into the OECD (Q)SAR toolbox. These (Q)SARs were developed using a set of data from polymers exclusively.

In the mid- to late 1990s, the USEPA conducted a review of more than 10,000 polymeric substances notified to the USEPA [3] for market access, and based on this exercise, a guidance document was developed along with (Q)SAR domains for these materials. Although this guidance received a modest update in the mid-1990s [3], it remains the main reference point for the document (USEPA) [4].

The ECOSAR models are, in essence, based on chemicals with specific mechanisms of action causing excess toxicity (greater toxicity than that predicted by baseline toxicity) and compounds without a specific mechanism of action. There are a couple dozen known mechanisms of toxicity (e.g., organophosphates and others). However, the majority of compounds are nonspecific and have what is known as a narcotic mechanism of action (typically more than 75% of all chemicals) [5].

The narcotic mechanism of action, also known as baseline toxicity, is based on the assumption of disruption of the cell membrane integrity. This means that the critical cell membrane function

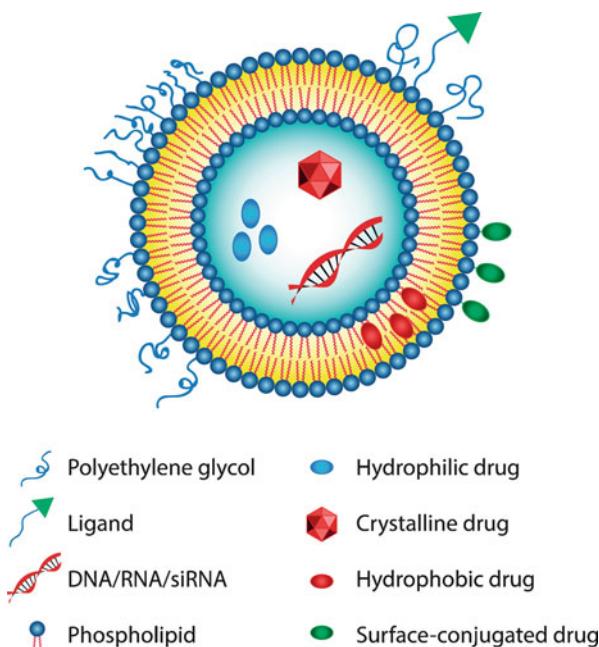


Fig. 1 Cell membrane (Colorbox)

necessary to sustain homeostasis in the chemical environment of the cell is impaired and the cell will die. The cell membrane (Fig. 1) is shared among all forms of life on Earth and is therefore a very good proxy for toxicity as the disruption of the cell membrane integrity will affect all life.

The mechanism of cell membrane disruption has not been entirely clarified, but it is usually characterized as a puncture or shift in fluidity of the lipid bilayer protecting the cell so that the functions needed to maintain homeostasis are impaired (e.g., efflux pumps or ligands are closed or lost). Some compounds can penetrate the membrane via pumps and receptor ligands in the membrane; these are typically the compounds with excess toxicity. Hence, an experimental proxy for the cell membrane was needed, and n-octanol/water partitioning coefficient was identified as a good model for the partitioning of chemicals between water and lipids such as the lipid bilayer. The log K_{ow} expresses the ability of the compound to disrupt the cell membrane, and, hence, the most significant acute environmental toxicity descriptor was defined [6]. Toxicity is derived from Greek with the original meaning “poisons arrow” and is defined as a compound’s ability to penetrate the cell membrane. The Paracelsus toxicity theorem that *dose makes the poison* we have used in toxicology for the past centuries is therefore enabled.

The aim of this chapter is to discuss the potential possibilities and challenges with the development and use of environmental toxicity (Q)SARs for polymers in a regulatory context for REACH. With current knowledge and available computational tools, we will explore how to build on the work by USEPA a quarter of a century ago and develop novel (Q)SARs for relevant polymers and used in anticipated REACH registrations. It is already clear that chemical assessments in REACH will not utilize (Q)SARs directly to satisfy registration requirements for specific endpoints (e.g., acute fish toxicity) but they may be exceptionally important in the establishment of chemical categories or groupings thereby lessening testing needs. The chapter will only take into consideration (Q)SAR and read-across as tools for risk assessment of polymers in a broad sense. Some details of the methodology involved in (Q)SAR and read-across will be discussed in Subheading 2 below after we have briefly reviewed the available data and provided a couple of examples.

2 Materials and Methods

Polymers as a group contain a wide variety of materials with differing structural attributes, functionalization, and physical/chemical properties. Polymers are composed of repeating monomer units, and copolymers are made up of more than one species of monomer.

2.1 Compounds: Polymers—A Brief Overview of Chemical Diversity and Available (Q)SARs

The copolymers are classified by how the units are arranged in the chain. The major groups include alternating, random, and block copolymers. Branched copolymers have a single main chain with one or more polymer side chains that are grafted or have branching that form other architectures. This complexity and diversity in polymeric species and structure present a significant challenge for their assessment and modeling. Polymers may contain structural alerts and/or specific functionalized properties (e.g., in pharmaceuticals and biocides) and may require specific toxicity analysis. Others may be completely toxicologically inert or have specific features and uses that warrant further assessment. It therefore makes sense to further define these materials.

According to USEPA definitions, compounds with a molecular weight greater than 1000 Daltons are too large to pass through the cellular membrane and are therefore unable to exert toxicity in a traditional manner. However, these compounds could cause mechanical effects (e.g., gill clogging) at high concentrations (e.g., typically observed at >1000 mg/L). Mechanical effects including binding to external and internal (e.g., gut) biological surfaces which are not “toxicity” in the traditional sense but for biological organisms may still be ecologically relevant. Polymer safety assessments may include separate considerations for the polymer, oligomers, and monomers depending on the polymer composition. USEPA [4] divides polymers into three categories based on the average molecular weight (M_{Wn}) and the amount of low molecular weight components (LMW):

Category 1: Polymers with low average molecular weight (M_{Wn} < 1000 Daltons). These can potentially be assessed as discrete structures in EPI Suite, within the normal limitation of the software, as long as the composition and structure of the polymer is known [4, 6].

Category 2: Polymers with high average molecular weight (M_{Wn} > 1000 Daltons) and large LMW material composition ($\geq 25\%$ with MW < 1000 Daltons; $\geq 10\%$ with MW < 500 Daltons). The environmental toxicity of these polymers can be assessed; however, oligomers may need separate assessment to account for any increased toxicity due to their lower molecular weight [4].

Category 3: Polymers with high average molecular weight (M_{Wn} > 1000 Daltons) and minimal LMW material (<25% with MW < 1000 Daltons; <10% with MW < 500 Daltons). These are generally assessed solely as the polymer (USEPA) [4].

The aquatic toxicity of polymers is influenced by solubility. Insoluble polymers are not expected to be toxic due to lack of bioavailability. Typical acute aquatic toxicity values for these polymers are >100 mg/L or > 10 mg/L for acute and chronic tests, respectively. However, physical or mechanical effects may occur if the insoluble polymer exists as a fine particle. Indeed, this is the case for microplastic particles [7].

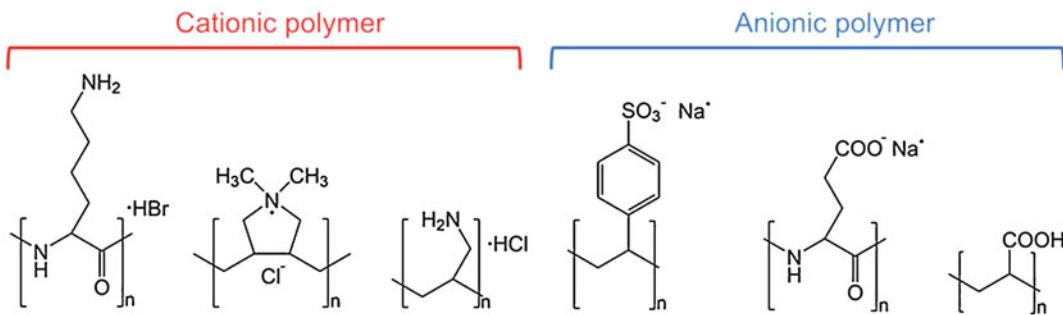


Fig. 2 Examples of cationic and anionic polymers

Polymer charge (neutral, anionic, cationic, amphoteric) can also modulate toxicity (Fig. 2). Nonionic polymers have very low water solubility and are generally believed to be of low hazard concern, unless they contain a significant amount of oligomer or if the polymer is used as a surfactant or dispersant. Anionic polymers are classified as poly(aromatic acids) or poly(aliphatic acids). Poly(aromatic sulfonate and carboxylate) polymers have moderate acute aquatic toxicities with fish, daphnids, and algae (LC50 1–100 mg/L). Poly(aliphatic acids) polymers have low toxicity to fish and daphnids (LC50 >100 mg/L), whereas algae seem to be more sensitive presumably due to chelating effects of nutrients. Due to chelation potential of many of these polymers, the mitigation potential of hard water further complicates study interpretation and design. The toxicity of both cationic and amphoteric polymers has been shown to increase with increasing cationic charge density [4]. As cationic polymers are believed to pose the greatest environmental hazard, the need for accurate aquatic acute toxicity (Q)SAR predictions is the greatest for these compounds. Other properties that may impact the toxicity of the pure polymer include physical form, particle size distribution, swellability, dispersibility, and of course in addition to these the presence and potentially weight fraction of reactive functional groups.

2.2 Cationic Polymers

Although cationic polymers are not limited to quaternary ammonium, phosphonium, and sulfonium functional groups, cationic polymers with quaternary ammonium groups are used in personal care and household cleaning products as conditioners or softeners and as flocculants in drinking water treatment plants. Therefore, there is a potential for the release into the aquatic environment. Sound environmental risk assessment, with a focus on the aquatic compartment, is of particular interest for cationic polymers, especially those with the quaternary ammonium functionality. At the time of the Boethling and Nabholz publication on polymer risk assessment, almost all of the cationic polymers reviewed contained a N-functionality [3].

Cationic polymers have a net positive charge at environmental pH and therefore have the potential to be highly sorptive and surface active. It has been suggested that cationic polymers will sorb to biological surfaces which are net negative, and previous studies have shown that cationic polymers have an impact on respiratory processes and may disrupt oxygen transfer (e.g., gill membranes of fish; Biesinger and Stokes) [8]. Total organic content (TOC) and dissolved organic carbon (DOC) have been shown to have a mitigating impact on the aquatic toxicity of cationic polymers, presumably due to sorption, but most of these investigations have been with highly charged polymers, and less mitigation may be presumed for lower cationic charged compounds. Traditional toxicity studies are conducted in clean media (e.g., standard OECD test media). The TOC and DOC levels in this media are not representative of environmental concentrations, and, therefore, the hazard values derived from these standardized studies may overestimate the environmental hazard of cationic polymers. Mitigation factors, specific to cationic polymers charge density and LMW composition, have been developed to adjust aquatic toxicity values to reflect environmental TOC levels. Based on confidential data from 53 cationic polymers, the USEPA described mitigation factors ranging from 7 to 290 [2–4]. However, these studies were conducted in the absence of analytical verification. Experimental TOC and DOC values may be important parameters to include in (Q)SAR modeling building exercises. These observations have resulted in some test conduct considerations as reflected in the OECD difficult test substance monograph (OECD) [9] and USEPA [10]. Because cationic polymers interact with anionically charged substances in general, we have observed toxicity mitigation as a function of water hardness in our laboratories (P&G, unpublished data). These are important, since toxicity often is linked to the positively charged polymers [11, 12].

In addition to DOC/TOC levels, other parameters to be considered in developing (Q)SARs for cationic polymers are physical-chemical properties that often serve as identity descriptors for the polymer. The cationic charge is typically found on a nitrogen group. For this reason, the % amine-nitrogen has been previously used as a descriptor in aquatic toxicity (Q)SARs. To further elaborate, from the historical work by Boethling and Nabholz [3], the cationic charge density based on %amine-nitrogen is because almost all the polymers submitted to the US TSCA office had their cationic charge based on nitrogen. The polymer backbone may also influence the toxicity. Cationic polymer backbone types can be carbon-based, silicon-based (e.g., Si-O), or natural (e.g., starch). The importance of backbone type and environmental hazard is not entirely clear. For fish, the toxicity silicon and carbon-based backbones are described using unique (Q)SAR equations. Natural polymer backbones are assumed to have equal or slightly less acute

toxicity than carbon backbone cationic polymers. However, daphnids have unique (Q)SAR equations for natural and carbon-based backbones, with silicon backbones having equal or slightly less acute toxicity than carbon-based backbones.

Fish and daphnid acute/chronic ratios range from 14 to 18, which suggest a narcotic mode of action [3]. The toxicity ranged from 0.006 mg/L towards algae for a carbon-based backbone polymer with a 7.8% amine-nitrogen charge density quaternary amine and 38% MW <500 to more than 1000 mg/L for a natural-based backbone with 0.07% amine-nitrogen charge density quaternary amine and 0% MW <500) [3].

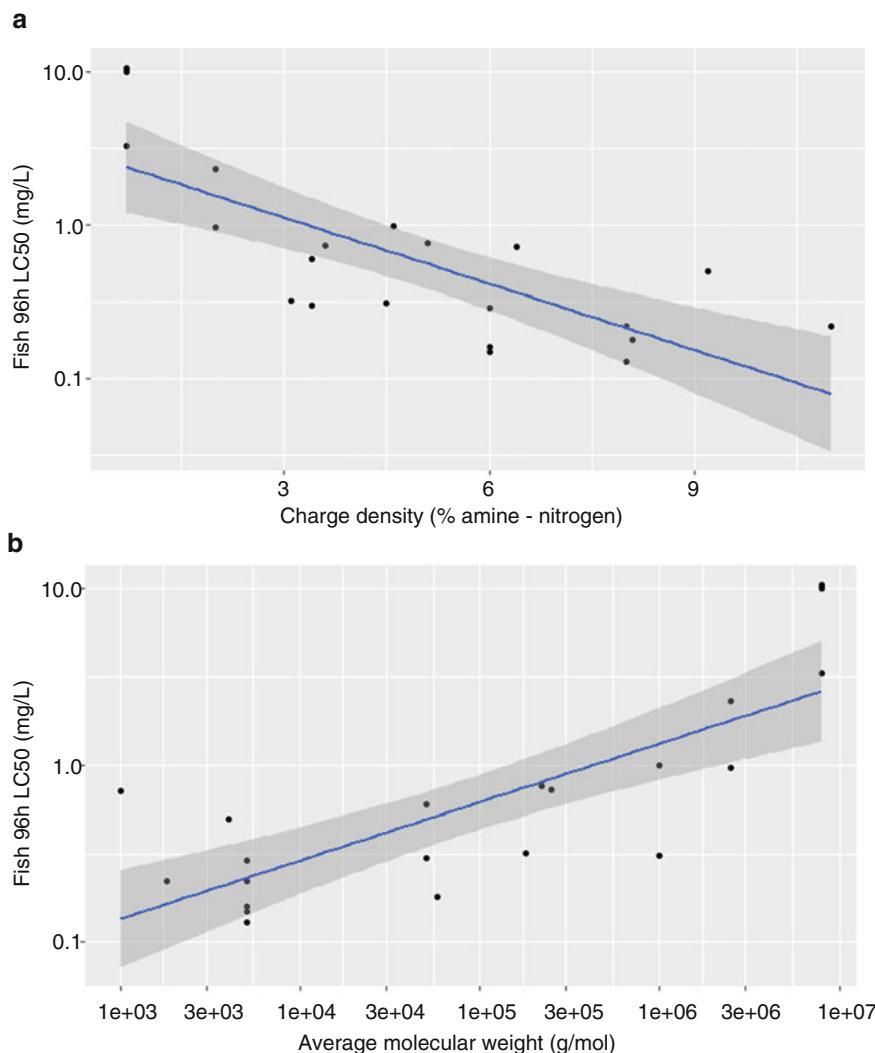


Fig. 3 Acute fish toxicity of quaternary amine cationic polymers (carbon-based backbone) as a function of (a) charge density (% amine-nitrogen) and (b) average molecular weight. Data obtained from Boethling and Nabholz (1996) [3]

It has been suggested by Boethling and Nabholz [3] that the aquatic toxicity of cationic polymers is influenced by charge density, molecular weight, and position of cation relative to the backbone. Figure 3 depicts the relationship between acute fish toxicity and (a) charge density (percent amine-nitrogen) or (b) average molecular weight in carbon-based backbone quaternary cationic polymers [3]. These plots hint towards increasing charge density leading to an increase in toxicity, whereas an increasing molecular weight corresponds with a decrease in toxicity—however much work is needed to develop reliable (Q)SARs.

In recognition of the impact of charge density of the environmental toxicity of cationic polymers, several regulatory agencies (e.g., Canada) have established a functional group equivalent weight (FGEW) cutoff of 5000 for the criteria of polymer of low concern (PLC). This concept has also been supported by the OECD review in 2009 of PLC criteria around the world. The FGEW cutoff concept can be a valuable tool in the prioritization of polymers to be selected for detailed regulatory reviews (e.g., REACH registration in EU).

2.3 Polyquaternium Cationic Polymers: A Complex Cationic Polymer Category

There is very limited data available on a specific class of polyquaternium that supports the observation that measured aquatic toxicity is influenced by charge density.

Polyquaternium cationic polymers represent a class of particular interest of cationic polymers due to their widespread use and releases to the aquatic environment. Polyquaterniums represent a very wide diversity of structures, and as of early 2019, there were approximately 40 registered active varieties with the Chemical Abstracts Service. Polyquaterniums are available as homopolymers or copolymers, and most are water soluble. Homopolymers vary in MW typically from <100,000 to 500,000 Daltons. All polyquaternium polymers contain a monomer with a quaternary ammonium functional group, such as diallyldimethylammonium chloride or trimethylammonium chloride. There is a diversity in monomer chemistries used as the copolymer for the quaternary ammonium monomer. A few examples of nonionic or anionic copolymers include vinylpyrrolidone, acrylic acid, polyvinyl alcohol, and acrylamide. Within each class of polyquaternium, the molecular weight will vary depending on the number of repeat units. While charge density remains constant for homopolymer polyquaterniums, the range in charge density or degree of substitution is dependent on the ratio of the monomers for copolymer polyquaterniums. The selection of monomers and fine-tuning of monomer ratios are necessary to give a range of physical-chemical properties and product benefits for diverse applications.

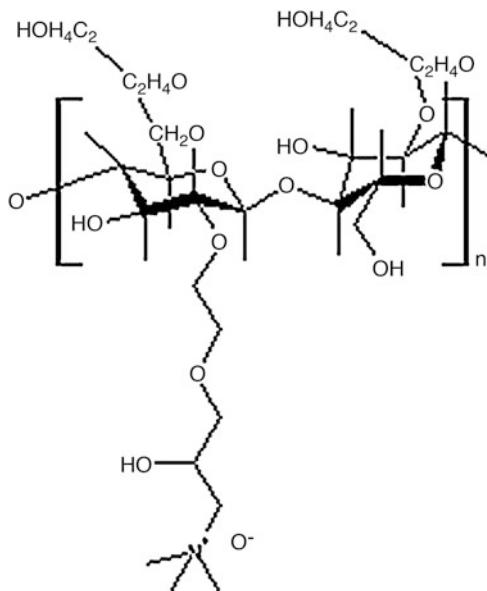
**Fig. 4** Representative structure of polyquaternium-10

Table 1
Polyquaternium-10 aquatic effects data

Variant	Charge density [13] (meq/g)	%N ²	Avg MW ^a [14] (kDa)	Viscosity (as 2% aq sol'n) (mPa/s) ²	96 h acute fish EC ₅₀ (<i>Gambusia holbrookii</i>) [15] (mg/L)	72 h EC _x algae (<i>Chlorella sp12</i>) [17] (mg/L)
UCARE JR125	High (0.9)	1.5–2.2	Low (250)	75–125	1.2	EC ₅₀ = 0.04
UCARE JR30M	High (1.0)	1.5–2.2	High (600)	30,000	1.5	EC ₁₀ = 0.002 EC ₅₀ = 0.05
UCARE JR400	High (1.2)	1.5–2.2	Low (400)	300–500	2.1	EC ₁₀ = 0.013 EC ₅₀ = 0.05
UCARE LK	Low (0.3)	0.4–0.6	Low (~400) ^b	300–500	100	Not available
UCARE LR30M	Low (0.4)	0.8–1.1	High (600)	30,000	66	Not available
UCARE LR400	Low (0.6)	0.8–1.1	Low (~400) ^b	300–500	64	Not available

^aSupplier information^bEstimated based on viscosity information [13–17]

Polyquaternium-10 (PQ10) is a cationic cellulose polymer with quaternary ammonium functionality, varying in charge density and MW (Fig. 4). The diversity in charge density is driven by the ratio of the monomer groups. A representative structure is illustrated in Fig. 4.

Table 1 below provides the measured and published aquatic toxicity of PQ10; the newest data is from 1991.

Based on the limited data available on aquatic effects, it could be proposed that charge density within a polymer class influences aquatic effects on fish and algae, while MW does not appear to have an impact. More information, from a well-structured toxicity investigation program, would be useful to determine the viability of the hypothesis. This observed trend supports the rationale to develop (Q)SAR to estimate aquatic effects when applicable. Since there is very limited publicly available data, it is not well understood whether a (Q)SAR developed for one polymer subclass could be leveraged by another subclass with some common structural features.

In a more recent example of research to understand whether physical properties of polymers can be used to estimate aquatic toxicity, Pereira et al. evaluated molecular weight, charge density, and integrative intrinsic viscosity of several cationic polyacrylamides to determine whether these structural features and variables could be used to predict the environmental effects [18]. The studied polyacrylamides were copolymers of acrylamide and acryloyloxyethyltrimethyl ammonium chloride with a cationic monomer content between 40 and 50% (w/w). The test species included in this study were bacteria, microalgae, macrophytes, and daphnids. While correlations were found between physical properties of the cationic polyacrylamides, the authors concluded that no clear ecotoxicity patterns correlating to physical properties were observed. While the observations may be valid for this particular group of polymers, the historical data from Boethling and Nabholz and Cumming et al. suggest there is a general relationship between certain structural features, such as charge density, and observed aquatic toxicity for cationic polymers, and in fact, (Q)SARs have been used for decades to estimate toxicity of cationic polymers by the USEPA [3, 15].

It is clear from the above that there is a strong need to explore (Q)SAR methodologies to describe the toxicity of polymers and cationic polymers in particular. The regulatory development of (Q)SARs for polymers has been advanced very little for the past decades, and publication of environmental toxicity data has also been sparse in that period. We will therefore in Subheading 3 section briefly describe possible options that may be applied in future elucidation of environmental toxicity (Q)SAR methods for polymers.

3 (Q)SAR Methods

Developing (Q)SARs based on curated PQ data is challenging as the data availability, transparency, and quality for the training set are limited and insufficient polymer descriptor information is available. The same is the case in a greater degree for cationic polymers in general [19]. And (Q)SARs are of course even more challenging for polymers in general based as they are much more diverse and data poor. Below is an outline of methods and approaches to consider.

3.1 Chemometric Tools in Ecotoxicological Evaluation of Polymers

In the recent decade, we have seen a notable rise in the use of alternative strategies in testing methods, including computational tools, for safety assessment of various organic/inorganic chemicals [20–22]. The *in silico* tools have demonstrated their successful application in detecting hazard potential of various chemicals belonging to several subclasses such as pharmaceuticals [23], agrochemicals, nanoparticles, and personal care products [23–27]. For emerging pollutants such as micro- and nano-sized particles [28] and polymers, such models are available to much a lower extent. There is a clear-cut deficit in the number of reports on application of *in silico* tools in toxicity (especially ecotoxicity) assessment of polymeric materials. The data scarcity on polymer ecotoxicity whether *in silico* or *in vitro* is evident mainly from availability of very few published studies in the literature. One possible reason is the high degree of proprietary nature for polymers and concerns with protecting confidential business information by disclosing identity descriptors for polymers in the public domain. While there are methods available that can estimate the effects of individual parent monomers [29, 30], the polymeric versions of the compounds are often left unevaluated (due to highly extensive computational requirement). Quantitative structure-activity/property relationship ((Q)SAR/QSPR) and quantitative read-across analysis (QRA) are widely accepted computational techniques, which are believed to be the most successful [2] two approaches that can be successfully implemented in identification of potent environmental pollutants among polymeric compounds (specifically, cationic polymers in view of their insufficient experimental data) using a very small amount of experimental results. It is also worth mentioning here that regardless of how statistically robust or significant a (Q)SAR/read-across model may be, it would be unavoidably associated with certain limitations [6]. These limitations are model specific, such as that a single (Q) SAR model may have its limited applicability owing to its restricted chemical domain which can be tackled by using intelligent consensus (Q)SAR approaches as proposed by Roy et al. [31]. Another major challenge in predictive toxicology is to effectively evaluate the reliability of obtained predictions of unknown/untested or not even synthesized chemicals. This limitation was also addressed recently with the introduction of prediction reliability indicator tool as proposed by Roy et al. [32]. Several commercially available tools for prediction of different endpoints for chemicals in general include TOPKAT software [33], CAESAR [34], ECOSAR [2], Toxicity Estimation Software Tool, etc.; however these tools generally do not include polymers in their training set which could be considered an important limitation [35]; hence we explore in the following sections alternative approaches.

It is well known that the most robust environmental toxicity tests are accompanied by confirmatory analytical verification of exposure. Analytical exposure determinations in aquatic toxicity tests are formally required, whenever feasible, under all typical OECD test guidelines for acute and chronic aquatic toxicity. However, limitations are also known for confirmation of exposures when polymers are tested. Indirect determinations can be useful in limited circumstances. These may include total organic carbon (sensitive down to perhaps 2 mg/L) or other alternatives such as measurement of an inorganic component such as silicon as was done in the 1990s during the programs addressing environmental safety of polydimethylsiloxane (PDMS) polymers [36]. (Q)SAR developments may be somewhat hampered by the lack of specific analytical verification of exposures until “high-end” analytical methods can be made routine and broadly available.

3.2 (Q)SAR Methodologies: Broad Classifications

The toxicity of whole polymeric structures or the structures in a monomeric form can be analyzed using (Q)SAR/QSPR methods, which can be classified as follows:

Regression-Based (Q)SAR This technique can be implemented to explore the quantitative correlation between toxicity of polymeric materials with the corresponding structural features. Multiple linear regression, partial least squares, and artificial neural networks are some of the examples of regression-based approaches. The use of regression approach for polymers is demonstrated in [37].

Classification-Based (Q)SAR For graded responses or where there is a lack of absolute quantitative toxicity data of polymers, classification-based techniques can be used to group the data into Boolean classes such as toxic, nontoxic, or moderately toxic classes. A classification-based technique like linear discriminant analysis (LDA) is also helpful in big data analysis [38].

3.3 Protocols for (Q)SAR Analysis in Polymers

(Q)SAR follows well-established protocols for developing statistically acceptable models for prediction of activity/property/toxicity chemical compounds [38, 39]. The Organization for Economic Cooperation and Development (OECD) has recommended five basic principles for (Q)SAR model development: (1) a defined endpoint, (2) an unambiguous algorithm, (3) a defined domain of applicability, (4) strict validation protocols, and (5) mechanistic interpretation, if possible [26]. The details of any (Q)SAR workflow are discussed below.

Collection of reported/generated biological data: For a (Q)SAR study involving polymeric compounds, the data collection should follow the prescribed guidelines of OECD [38, 39] which include uniform experimental conditions, uniform time of exposure for the desired effect, experiment with a standard species, analytical verification of the exposure concentrations, etc. The

data curation should be done effectively to check for duplicates/salts/ions, etc. Another important point in collecting homogenous ecotoxicity data for polymers includes ideality in experimental water conditions such as hardness, alkalinity organic carbon content (TOC and DOC), etc. that may affect the observed toxicity. In the case of algal testing and (Q)SARs, definition of specific anionic and cation components of media may also be important.

Descriptor calculation: For the descriptor calculation, in most of the cases, initially the monomeric or repeating unit is identified. The flanks of monomers are capped with hydrogen atom in order to satisfy the valence electron. Then the structure are subjected to descriptor calculating software such as Dragon [40], SiRMS [41], alvadesc [42], PaDEL-Descriptor [43], etc. to calculate molecular descriptors.

Division of the dataset: In order to obtain useful models, the collected data should be partitioned into training and test sets following unbiased methods. Some of the widely followed dataset division techniques include Kennard-Stone [44], Euclidean distance approach [45], k-medoids approach [46], and random sampling. These tools for dataset division are available, for example, at [http://teqip.jdvu.ac.in/\(Q\)SAR_Tools/](http://teqip.jdvu.ac.in/(Q)SAR_Tools/).

Feature selection: In feature selection, molecular descriptors important for the response values are identified. Some of the feature selection techniques include stepwise selection, genetic algorithm, double cross validation (DCV), and factor analysis [47]. The problems of small datasets (as in the case of polymer toxicity data, which is scarce) can be addressed to some extent using DCV. In DCV, the training set is split into calibration and validation sets, and these are utilized for model building and model selection, while the test set is exclusively used for model assessment. This process obviates the possibility of bias in descriptor selection. For ideal (Q) SAR models, the intercorrelation among the descriptors should be very less.

Modeling algorithms and chemometric tools used in (Q)SAR: The most commonly employed linear modeling algorithms include multiple linear regression (MLR) [48], univariate linear regression (ULR), ordinary least squares (OLS), partial least squares (PLS), principal component analysis (PCA) [27], principal component regression (PCR), etc.

Model validation metrics and mechanistic interpretation: Finally the developed model should be validated following internationally recognized guidelines. Some widely used validation metrics for regression models include leave-one-out (LOO) cross-validation R^2 (Q^2) and for training set evaluation and Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 , and concordance correlation coefficient (CCC) for test set evaluation. Some other stringent criteria for model validation include (1) mean absolute error (MAE) criteria proposed by Roy et al. [49] and (2) Golbraikh and Tropsha criteria for model

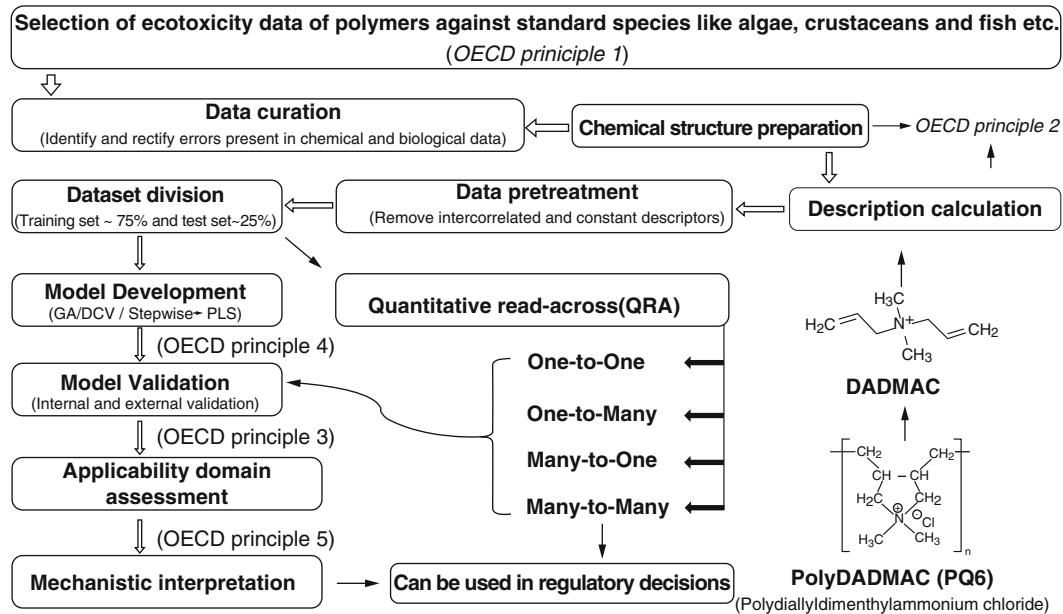


Fig. 5 Process involved in ecotoxicity study of polymers following in silico (Q)SAR and QRA approach

validation [46]. A mechanistic interpretation of a developed model is desired wherever possible. Figure 5 depicts the general outline of polymer toxicity modeling.

4 Applications of (Q)SAR to Polymers: A Literature Review—Applications of (Q)SAR in Ecotoxicity of Polymers

With the proprietary nature for many polymers, manufacturers and downstream formulators have generated aquatic effects data for stewardship reasons, but much of this data is privately held to protect confidential business information (CBI). However, there are some classes of polymers that have been studied with publicly available publications demonstrating potential toxicity of polymers with some aquatic species [19]. Several examples are presented below as case studies of (Q)SAR development for diverse polymer classes.

Acute algal toxicity: The very first and comprehensive (Q)SAR study on toxicity of polymers was conducted by Nolte et al. [19]. The data ($N = 43$) for growth rate inhibition (EC50) of algae were collected from the literature using Google Scholar and Web of Science. However, since the data was limited, the authors combined the data for two different times of exposure, i.e., 96-h and 72-h reflective of primarily USEPA and OECD algal test procedures, respectively. Three different models based on their charge separation (cationic ($N = 9$), anionic ($N = 16$), and

nonionic ($N = 17$) compounds) were developed using one theoretical descriptor following regression-based decision tree technique. More complex branched polymers, polymeric surfactants, and non-nitrogen cationic polymers were omitted from the study. The models predict that cellular adsorption, disruption of the cell wall, and photosynthesis could be the possible mechanisms of action for algal toxicity of cationic and nonionic polymers. The findings of the (Q)SAR results combined with molecular dynamics simulations proposed that nutrient depletion is likely the dominant mode of toxicity. (Q)SAR relationships for green algae growth inhibition, however with the low number of data for the generated (Q)SAR, were not statistically robust and do not comply with the quality criteria cited by Cherkasov et al. [6, 19].

4.1 Application of (Q)SAR in Toxicity Prediction of Polymers (Peptides)

Antimicrobial peptide toxicity: Langham and colleagues [51] developed (Q)SAR models to quantify and predict antimicrobial peptide toxicity against human host cells (epithelial and red blood cells) based on physicochemical properties like interaction energies and radius of gyration which were in turn calculated from molecular dynamics simulations of the peptides in aqueous solvent. For model development 60 peptides with experimentally determined toxicities were used. Langham and colleagues [51] proposed based on the findings of molecular modeling study that physicochemical properties of peptides and interactions in a solvent are responsible for their toxicity against human cells in their native state. The developed models were then employed in predicting several other protegrin-like peptides. The (Q)SAR model could correctly rank four out of five protegrin analogues newly synthesized and tested for toxicity in laboratory.

Although quantitative structure-toxicity relationship modeling reports involving polymers are scarce, there are several reports on (Q)SAR/QSPR modeling of their biological activity and property endpoints. We report here some of them to demonstrate that similar tools may be applied to develop models to predict toxicity of polymers.

4.2 Application of (Q)SAR to Biomedical Applications of Polymers

Cellular response and protein absorption: Khan and Roy [52] developed predictive (Q)SAR models for a cellular response (fetal rat lung fibroblast proliferation) and protein adsorption (fibrinogen adsorption (FA)) on the surface of tyrosine-derived polymers designed for the purpose of tissue engineering. These polymers were synthesized using a combinatorial approach which in turn is a decade long process used in tissue engineering applications; the process is briefed in the source paper [52]. The model consists of 66 data for cellular response and 40 data for protein adsorption on polymers. The models were developed using only selected 2D descriptors having definite physicochemical meaning. To enhance the biological domain of the model, multiple (Q)SAR models were

developed and then subjected to consensus modeling as proposed by Roy et al. [31]. The final consensus models were validated using strict OECD guidelines and accepted internal and external metrics.

Cellular response: Semiempirical QSPR models were developed to predict the cellular response to the surfaces of polymers designed for tissue engineering applications by Kholodovych and colleagues [53]. The findings of the models were then compared with experimental results which showed a high degree of accuracy proving its significance for biomedical applications. Partial least squares (PLS) regression technique was used for model development using 62 polyarylates and structure-based molecular descriptors.

Bioresponse modeling: Artificial neural networks (ANN) were applied to model bioresponse to the surfaces of polymers collected from combinatorial library [54]. For analysis, 22 structurally distinct polymers were modeled against human fibrinogen adsorption. Additionally, the developed models were used to model rat lung fibroblast and normal human fetal foreskin fibroblast proliferation in the presence of 24 and 44 different polymers. The root mean square was used for the error comparison with experimental finding, and it was lower than experimental results thus proving applicability of the developed models.

Protein adsorption: Smith et al. [55] proposed a surrogate model for the prediction of protein adsorption onto the surfaces of polymers designed for tissue engineering applications. The proposed surrogate model combines machine learning, molecular modeling, and an artificial neural network. The experimental errors were estimated using Monte Carlo technique. The dataset consists of 45 polymers with measurements of human fibrinogen adsorption. A total of 106 molecular descriptors were computed using the Molecular Operating Environment (MOE) software. The surrogate model was developed in two stages: firstly the three descriptors with highest correlation to the adsorption were identified, and then these three descriptors were used as input for the second stage, i.e., for artificial neural network (ANN) to predict fibrinogen adsorption. Here, a Monte Carlo approach enabled a direct assessment of the effect of the experimental uncertainty on the results. Only the training set (nearly 50%) was employed for ANN using random sampling followed by checking of experimental error using Monte Carlo analysis. The accuracy of ANN was then compared with experimental data for the remaining polymers (the validation set). The Pearson correlation coefficient was used as validation metric. In conclusion, the surrogate model was proposed to get accurate and unambiguous predictions of polymers to check for their range of fibrinogen absorption, an essential requirement for assessing polymers for regenerative tissue applications.

4.3 Applications of (Q)SAR in Property Estimation of Polymers

In other areas of (Q)SAR development, there are a number of publications that demonstrate that quantitative structure-property relationship (QSPR) models can be developed to predict certain physical properties of polymers. Though a number of studies in the available literature exist on modeling of various properties of polymers, we have reported here a few of the recent reports.

Refractive Index Khan et al. [56] proposed robust QSPR models to predict refractive indices (RIs) of a set of 221 diverse organic polymers employing simple 2D descriptors generated by using monomeric unit. The final model consists of six theoretical descriptors developed using partial least squares (PLS) regression technique. For feature selection, double cross-validation tool was used. Use of consensus modeling for predictions from multiple modeling was also demonstrated. Finally, four small virtual libraries were selected to predict their RIs values using obtained consensus model.

Glass Transition Temperature The glass transition property of 206 diverse polymers was studied by Khan and Roy [37] using the QSPR approach since it has a direct impact on polymer stability. Five individual QSPR models were obtained using six 2D molecular descriptors following partial least squares regression and DCV as the feature selection tool. The models were extensively validated, and Y-randomization (Y-scrambling) test was performed in order to prove nonrandom and robust nature of the developed models. At last, comparison with existing QSPR models was made to demonstrate the effectiveness of the novel models.

5 Discussion of Future Avenues: Application of Fragment-Based (Q)SAR and Read-Across in Ecotoxicity Predictions of Polymers

The area of (Q)SAR modeling for the evaluation of toxicity of polymers has remained largely unexplored, which could be used to motivate and inspire (Q)SAR modelers to contribute to this dynamic and vastly underdeveloped field. A major notable point here is many of the previous modeling studies [57, 58] on polymers involve computation of quantum-chemical descriptors which can be a time-consuming process. This problem can be solved effectively by using only 2D descriptors having simple more definite physicochemical meaning in order to avoid conformational analysis, computational complexity of energy minimization, and alignment problems.

Apart from the classical methods of (Q)SAR model development, one can also apply more novel and more appropriate methods as discussed below.

Fragment-based (Q)SAR: These use molecular substructures expressed in fingerprints as descriptors in the developed models.

Fragment (Q)SARs can be implemented in the ecotoxicological modeling of polymers when studying a part of a molecule or specific group in relation with the toxicity. A widely used group-based (Q) SAR is H(Q)SAR (Hologram-(Q)SAR) [38, 39].

H(Q)SAR: This is a modern 2D FB-(Q)SAR (fragment-based) technique which utilizes molecular substructures expressed in binary pattern also termed as fingerprints in model development as variables. The method does not involve calculation of any physicochemical chemical descriptor or 3D structure generation. The process follows three steps:

1. Fragment generation for each of the training set molecules
2. Representation of the fragments in the holograms
3. Finding correlation of the molecular holograms with the corresponding activity data using training set compounds employing the PLS technique

A number of parameters affect H(Q)SAR model generation such as hologram length, fragment size, and distinction [38]. H (Q)SAR encodes all possible fragments within the molecules along with sub-fragments; thus it is helpful in understanding the fragments responsible for the toxicity of polymers in reference species. The other possible applications of H(Q)SAR in ecotoxicity of polymers include exploring individual atomic contributions to the toxicity with a visual display of active centers in the compounds.

Read-across: The read-across approach is a practice based on the assumption that structurally similar compounds exhibit similar physicochemical, environmental fate, toxicological, and ecotoxicological properties. The process starts with the grouping of similar objects (here, structures), and then the response value of one or more chemicals can be used to predict the behavior of target chemicals. Four different strategies for read-across have been proposed so far, i.e., one-to-one, one-to-many, many-to-one, and many-to-many. As per the OECD guidelines [58], the QRA prediction can be performed in following one of the four ways:

1. Using similar chemicals for the endpoint to perform read-across
2. Using a mathematical scale to check the trend in experimental results using two or more similar chemicals (e.g., trend analysis)
3. Taking an average of endpoint values of two or more source chemicals
4. If sufficient data is available, using the most conservative value from the source chemicals in the whole category

A read-across strategy can be used to estimate the toxicity for a series of cationic or anionic polymers with acceptable levels of

uncertainty. Considering that the toxicity data are available for a limited number of polymers, read-across will be very helpful for bridging data gaps. However, efforts are needed to define how similar polymers should be grouped and what key physical-chemical properties should be used in the grouping scheme. Data anchors at the extremes of the biological attribute being used to develop the read-across are important to define. Previous groupings by ECHA or EPA may be too broad, and further work is needed to refine based on the diversity of polymers within classes or subclasses. In addition, it is possible that the grouping and read-across approach may need to be customized depending on polymer class or even route of exposure. The potential impact of polymers to human health and the environment may be estimated through developing (Q)SAR models and by enabling read-across to structural analogues and avoiding or minimizing the need to conduct safety studies. This would bring benefits to time, resources, and avoiding animal testing. (Q)SARs could also be leveraged in polymer innovation and providing guidance on the design space.

6 Conclusions

It is clear from the above that regulatory programs are increasingly starting to include polymers for environmental risk assessment, chiefly in REACH, and that there has been a paucity for a couple of decades in the development of aquatic toxicity (Q)SARs by the USEPA [4] for polymers. There is hence a need to develop models for this purpose. It is also clear that polymers are very diverse and this diversity needs to be reflected in the model development and domains [6]. It is also clear that key and necessary data that are needed to do assessments or generate regression-based (Q)SARs are currently largely missing [19] and the sparse available experimental data lacks insight on experimental exposure. Moreover, regression-based (Q)SARs still require identification of the most determinant toxicity descriptors of the polymer. It is highly questionable if this is hydrophobicity since the mechanism of action is either unknown or not narcotic since the molecules are too large to exert the narcotic mechanism we normally associate with narcosis. Cationic polymers are highlighted as an example in this chapter of a class of polymers of high and down-the-drain use, more specifically polyquaterniums. The toxicity of these materials is dependent upon charge density, molecular weight, %amine-nitrogen, solubility, and type of backbone. There may be other additional and currently uninvestigated descriptors that govern the toxicity of these and other cationic polymers. We have suggested a series of non-regression-based (Q)SAR approaches that may be applied to elucidate the potential descriptors. Figure 5 outlines a process for developing (Q)SARs which when combined with the learnings from

Cherkasov et al. [6] are important methods moving forward. Using polymer properties may be useful for estimating fate, effects, and even form in the environment. For example, the glass transition temperature (T_g) [37] may be used to estimate form. If a polymer is below T_g , then it has to be a solid. If it is above T_g , then it could be a solid or liquid depending on the melting temperature of the polymer, which would determine the bioavailability and toxicological availability of the material. 3D comparative molecular field analysis and other ANN or 2D H(Q)SAR may prove highly relevant—but in a regulatory setting, the models have to be transparent in which case the fragment-based models may initially be used to identify critical toxicity and availability descriptors which can then be used to cluster the polymers. The toxicity of these clusters can then be experimentally explored and recorded and subsequently develop read-across within these. The authors of this chapter are pursuing this in the coming years via generation of novel experimental and computational data on polyquaterniums, and we will also evaluate the potential for fragment-based (Q)SARs for polymers in REACH.

Acknowledgments

This work was supported by Cefic in the LRI project ECO46: iTAP (<http://cefic-lri.org/projects/eco-46-improved-aquatic-testing-and-assessment-of-cationic-polymers-itap/>). KK thanks Indian Council of Medical Research, New Delhi, for financial support in the form of a senior research fellowship.

References

1. ECHA, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Article 138 (2)
2. Mayo-Bean K, Moran K, Meylan B, Ranslow P (2012) Methodology document for the ECOlogical Structure-Activity Relationship Model (ECOSAR) class program. US-EPA, Washington D.C.
3. Boethling RS, Nabholz JV (1996) Environmental assessment of polymers under the US Toxic Substances Control Act. United States Environmental Protection Agency
4. USEPA (2013) Polymer guidance: <https://www.epa.gov/sites/production/files/2015-03/documents/polyguid.pdf>
5. Sanderson H, Thomsen M (2009) Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q) SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicol Lett* 187:84–93
6. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R (2014) (Q)SAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010

7. Connors KA, Dyer SD, Belanger SE (2017) Advancing the quality of environmental microplastic research. *Environ Toxicol Chem* 36 (7):1697–1703
8. Biesinger KE, Stokes GN (1986) Effects of synthetic polyelectrolytes on selected aquatic organisms. *J Water Pollut Control Fed* 58:207–213
9. OECD (2019) Guidance document on aqueous-phase aquatic toxicity testing of difficult test chemicals. Series on testing and assessment. No. 23 (second edition). Paris, 81p
10. USEPA (1996) Ecological effects test guidelines OPPTS 850.1085 fish acute toxicity mitigated by humic acid. EPA712-C-96-117. Washington D.C., p 10
11. de Rosemond SJ, Liber K (2004) Wastewater treatment polymers identified as the toxic component of a diamond mine effluent. *Environ Toxicol Chem* 23:2234–2242
12. Liber K, Weber L, Levesque C (2005) Sublethal toxicity of two wastewater treatment polymers to lake trout fry (*Salvelinus namaycush*). *Chemosphere* 61:1123–1133
13. Cumming J, Hawker D, Matthews C, Chapman H, Nugent K (2010) Analysis of polymeric quaternary ammonium salts as found in cosmetics by metachromatic polyelectrolyte titration. *Toxicol Environ Chem* 92:1595–1608
14. Siebert J, Luyt A, Ackermann C (1990) A new transmission electron microscopic (TEM) method to determine differences between cationic polymers in solution. *Int J Pharmaceut* 61:157–160
15. Cumming JL, Hawker DW, Nugent KW, Chapman HF (2008) Ecotoxicities of polyquaterniums and their associated polyelectrolyte-surfactant aggregates (PSA) to *Gambusia holbrooki*. *J Environ Sci Heal A* 43:113–117
16. Cumming J, Hawker D, Chapman H, Nugent K (2011) The fate of polymeric quaternary ammonium salts from cosmetics in wastewater treatment plants. *Water Air Soil Pollut* 216:441–450
17. Cumming JL (2008) Environmental fate, aquatic toxicology and risk assessment of polymeric quaternary ammonium salts from cosmetic uses. Griffith University, Mount Gravatt
18. Pereira JL, Vidal R, Goncalves FJM, Gabriel RG, Costa R, Rasteiro MG (2018) Is the aquatic toxicity of cationic polyelectrolytes predictable from selected physical properties? *Chemosphere* 202:145–153
19. Nolte TM, Peijnenburg WJ, Hendriks AJ, van de Meent D (2017) Quantitative structure-activity relationships for green algae growth inhibition by polymer particles. *Chemosphere* 179:49–56
20. Khan K, Baderna D, Cappelli C, Toma C, Lombardo A, Roy K, Benfenati E (2019) Ecotoxicological (Q)SAR modeling of organic compounds against fish: application of fragment based descriptors in feature analysis. *Aquat Toxicol* 212:162–174
21. Khan K, Khan PM, Lavado G, Valsecchi C, Pasqualini J, Baderna D, Marzo M, Lombardo A, Roy K, Benfenati E (2019) (Q)SAR modeling of *Daphnia magna* and fish toxicities of biocides using 2D descriptors. *Chemosphere* 229:8–17
22. Khan K, Roy K, Benfenati E (2019) Ecotoxicological (Q)SAR modeling of endocrine disruptor chemicals. *J Haz Mat* 369:707–718
23. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2018) Ecotoxicological assessment of pharmaceuticals using computational toxicology approaches: QSTR and interspecies QTTR modeling. In: Proceedings of MOL2-NET 2017, international conference on multidisciplinary sciences, 3rd edn. MDPI AG, Switzerland, Basel, p 1
24. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2019) Ecotoxicological modeling, ranking and prioritization of pharmaceuticals using QSTR and i-QSTTR approaches: application of 2D and fragment based descriptors. *Mol Inform*, 38, article 1800078, <http://dx.doi.org/10.1002/minf.201800078>
25. Khan K, Benfenati E, Roy K (2019) Consensus (Q)SAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotox Environ Safe* 168:287–297
26. Khan K, Roy K (2017) Ecotoxicological modelling of cosmetics for aquatic organisms: a QSTR approach. *SAR (Q)SAR Environ Res* 28:567–594
27. De P, Kar S, Roy K, Leszczynski J (2018) Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environ Sci Nano* 5:2742–2760
28. Braakhuis HM, Kloet SK, Kezic S, Kuper F, Park MV, Bellmann S, van der Zande M, Le Gac S, Krystek P, Peters RJ (2015) Progress and future of in vitro models to study translocation of nanoparticles. *Arch Toxicol* 89:1469–1495
29. ECHA European Chemicals Agency (2012) Guidance on registration. Version 2.0. Guidance for the implementation of REACH

30. Netzeva T, Pavan M, Worth A (2007) Review of data sources, (Q)SARs and integrated testing strategies for aquatic toxicity. European Communities, Luxembourg
31. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an “intelligent” use of multiple (Q)SAR/QSPR/QSTR models? *J Chemom* 32:e2992
32. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* 3:11392–11406
33. Enslein K, Gombar VK (1997) TOPKAT 5.0 and modulation of toxicity. *Mutat Res-Fund Mol M* 379:S14–S14
34. Plošnik A, Zupan J, Vračko M (2015) Evaluation of toxic endpoints for a set of cosmetic ingredients with CAESAR models. *Chemosphere* 120:492–499
35. De Vaugelade S, Nicol E, Vujošić S, Bourcier S, Pirnay S, Bouchonnet S (2018) Ultraviolet-visible phototransformation of dehydroacetic acid – structural characterization of photoproducts and global ecotoxicity. *Rapid Commun Mass Spectrom* 32:862–870
36. Fendinger NJ, McAvoy DC, Eckhoff WS, Price BB (1997) Environmental occurrence of poly-dimethylsiloxane. *Env Sci Technol* 31:1555–1563
37. Khan PM, Roy K (2018) QSPR modelling for prediction of glass transition temperature of diverse polymers. *SAR (Q)SAR Environ Res* 29:935–956
38. Roy K, Kar S, Das RN (2015) A primer on (Q) SAR/QSPR modeling: fundamental concepts. Springer, UK. <https://www.rsc.org/journals-books-databases/about-journals/environmental-science-nano/>
39. Roy K, Kar S, Das RN (2015) Statistical methods in (Q)SAR/QSPR. In: A primer on (Q) SAR/QSPR modeling. Springer, NY, USA, pp 37–59
40. Mauri A, Consonni V, Pavan M, Todeschini R, software D (2006) An easy approach to molecular descriptor calculations. *Match* 56:237–248
41. Kuz'min VE, Artyomenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) Hierarchic system of (Q)SAR models (1D–4D) on the base of simplex representation of molecular structure. *J Mol Model* 11:457–467
42. Alvadesc (2019) <https://www.alvascience.com/alvadesc/>
43. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466–1474
44. Kennard RW, Stone LA (1969) Computer aided design of experiments. *Technometrics* 11:137–148
45. Golmohammadi H, Dashtbozorgi Z, Acree WE Jr (2012) Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci* 47:421–429
46. Zhang Q, Coulouigner I (2005) A new and efficient k-medoid algorithm for spatial clustering. In: International conference on computational science and its applications. Springer, NY, USA, pp 181–189
47. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR (Q)SAR model development. *Chemom Intell Lab Syst* 159:108–126
48. De P, Aher RB, Roy K (2018) Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector Aedes aegypti: application of ETA indices. *RSC Adv* 8:4662–4670
49. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive (Q)SAR models. *Chemom Intell Lab Syst* 152:18–33
50. Langham AA, Khandelia H, Schuster B, Waring AJ, Lehrer RI, Kaznessis YN (2008) Correlation between simulated physicochemical properties and hemolytic of protegrin-like antimicrobial peptides: predicting experimental toxicity. *Peptides* 29:1085–1093
51. Khan PM, Roy K (2019) Consensus QSPR modelling for the prediction of cellular response and fibrinogen adsorption to the surface of polymeric biomaterials. *SAR (Q)SAR Environ Res* 30:363–382
52. Kholodovych V, Smith JR, Knight D, Abramson S, Kohn J, Welsh WJ (2004) Accurate predictions of cellular response using QSPR: a feasibility test of rational design of polymeric biomaterials. *Polymer* 45:7367–7379
53. J.R. Smith, D. Knight, J. Kohn, K. Rasheed, N. Weber, S. Abramson (2003) Using non-linear regression to predict bioreponse in a combinatorial library of biodegradable polymers, *MRS Online Proc Libr*, vol 804, Cambridge, UK
54. Smith JR, Knight D, Kohn J, Rasheed K, Weber N, Kholodovych V, Welsh WJ (2004) Using surrogate modeling in the prediction of fibrinogen adsorption onto polymer surfaces. *J Chem Inform Comput Sci* 44:1088–1097

55. Khan PM, Rasulev B, Roy K (2018) QSPR modeling of the refractive index for diverse polymers using 2D descriptors. *ACS Omega* 3:13374–13386
56. R Duchowicz P, C Comelli N, V Ortiz E, A Castro E (2012) (Q)SAR study for carcinogenicity in a large set of organic compounds. *Current Drug Saf* 7:282–288
57. Talevi A, L Bellera C, Di Ianni M, R Duchowicz P, E Bruno-Blanch L, A Castro E (2012) An integrated drug development approach applying topological descriptors. *Curr Comput Aided Drug Des* 8:172–181
58. Gajewicz A, Jagiello K, Cronin MTD, Leszczynski J, Puzyn T (2017) Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environ Sci Nano* 4:346–358

Part IV

Tools, Databases, and Web Servers



Chapter 29

Ecotoxicity Databases for QSAR Modeling

Shinjita Ghosh, Supratik Kar, and Jerzy Leszczynski

Abstract

Industrial chemicals, pharmaceuticals along with illicit drugs (IDs), as well as day-to-day personal care products (PCPs) are documented as contaminants of emerging concerns (CECs). They are environmental pollutants due to their substantial detrimental impacts on the environment through their frequent presence, persistence, and peril to the species living in aquatic, terrestrial, and soil compartments as well as to humans. Although the toxic effects and occurrence concentration of pharmaceuticals and chemicals have been studied and reported for the last three decades, PCPs and IDs are quite neglected substances along with the mixtures and transformation products (TPs) or metabolites of all CECs, in the context of their ecotoxicological risk evaluation. Among various compartments, the effects of CECs are largely documented for aquatic species where often very little information is available with regard to terrestrial and soil toxicity. This deficiency of knowledge has led to greater effort to create new methods and approaches which would measure their occurrence, metabolism, bioaccumulation, and biodegradability followed by the mechanism of action (MOA) behind toxicity to individual living species and ecosystem. This information is very important for risk assessment and risk management along with regulatory decision-making followed by *in silico* or computational modeling for future ecotoxicity. Thus, the obtained information needs to be documented under a system called “database” based on different categories including chemical class, toxicity testing, test species, environmental compartment-specific, toxicity MOA specific as well as country. Ecotoxicity has a number of open-access and commercial databases which are implemented for risk profiling, regulatory decision-making followed by ecotoxicity prediction of new and untested substances over the years. The present chapter deals with the most commonly used ecotoxicity databases followed by their detailed information so that one can use these databases efficiently in experimental as well as computational research.

Key words Database, Ecotoxicity, *In silico*, QSAR, Risk assessment, Risk management

1 Introduction

Micropollutants like industrial chemicals, pharmaceuticals and personal care products (PPCPs), agrochemicals, plastics, and polymers have been widely detected in day-to-day life with the developed state-of-the-art analytical methods [1]. The residues of these micropollutants are recurrently reported in surface and groundwater, influent and effluents of both the sewage treatment plants (STPs) and water treatment plants (WTPs), even in drinking

water due to their universal uncontrollable consumption, low biodegradability, and inappropriate disposal [2, 3]. Continuous bioaccumulation and bioconcentration of these compounds lead to a serious risk to each compartment of the environment as well as living species. Although multiple studies assessed the existence in terms of quantitative measures followed by toxicity evaluation, the amount of data considering their existence is still limited. In this perspective, computational or *in silico* models are helpful and proven alternative over the years to fill the ecotoxicity data gaps for new and/or untested chemicals [4].

The need of *in silico* techniques in predicting toxicological and hazardous properties of chemicals is taking the central stage of attention day by day among the scientific community, regulatory bodies, and the public in general for decision-making frameworks in safety assessments [5, 6]. *In silico* methods are capable of providing information about the physicochemical properties of organic chemicals, their environmental fate as well as their effects on human health. Thus, a great demand for large quantities of early information on toxicity, particularly at the designing stage, has arisen. The quantitative structure-activity relationship (QSAR) is one of the leading *in silico* methods increasingly being used for the prediction of different ecotoxicity properties (endpoints) of organic chemicals. Though a number of *in vitro* and high-throughput methods have been developed to provide an experimental evaluation of such properties, earliest possible identification of liability is desired, including even prior to synthesis. Thus, *in silico* approaches may be used for hypothetical compounds to guide synthetic efforts. Moreover, QSAR modeling may be done quickly, and a large number of compounds may be screened within a very short period. Additionally, the QSAR model will complement the 3Rs principle (replacement, refinement, and reduction of animals in research) minimizing animal testing. Around the world, extensive research work on predictive toxicology using QSAR is going on nowadays, and a great amount of applied research is still needed. It is essential to follow fundamental guidelines implemented by the Organization for Economic Co-operation and Development (OECD) to develop an acceptable and predictive QSAR model (<http://www.oecd.org/dataoecd/33/37/37849783.pdf>). One of the fundamental requirements to develop a reliable and predictive QSAR model is a precise and good quality of experimental data. Thus, generation and collection of experimental toxicity data for different endpoints and species are very much important [7]. A series of data for specific toxicity following similar or identical protocols is considered as database. In the present chapter, we have discussed the importance and application of major ecotoxicity databases which are utilized over the years by regulatory authorities and academicians. The future requirements for ideal ecotoxicity databases are also introspected with examples. The present chapter

is a rich source of information on ecotoxicity databases with their detailed data information for the experts along with the beginners who want to use these databases for QSAR modeling as well as high-throughput screening (HTS) of ecotoxicity and risk profiling of chemicals [8–10].

2 Ecotoxicity

The term ecotoxicity is used when the toxicity of a specific compound is tested in several organisms together, and the term environmental toxicity is related to hazards and risk associated with a chemical to the ecosystem or environment. However, in most of the literature, both terms are used interchangeably. In the present manuscript, we will use the term ecotoxicity to signify both terms. The ecotoxicity of a chemical can be defined in terms of the length and frequency of its exposure to the ecosystem followed by their hazardous effect or toxicity to the species living in the exposed environment [11, 12]. Industrial organic synthetic chemicals, pharmaceuticals, and personal care products are one of the CECs frequently detected in diverse compartments of the environment and are the leading substances responsible for ecotoxicity over the years. Based on the length and observed effects of toxicity, ecotoxicity can be categorized into two types, and they are [11]:

- (a) *Acute toxicity*: It is defined as harmful or toxic effects due to the exposure of a species/organism to a CEC hazard typically over a span of not more than 15 days. For the environmental toxicity assessment, the studied acute toxicity to fish, *Daphnia*, and algae are 96 h LC₅₀ in mg/l, 48 h EC₅₀ in mg/l, and 72–96 h EC₅₀ in mg/l, respectively.
- (b) *Chronic toxicity*: It is defined as harmful effects due to the long-term exposure (≥ 15 days to years) of a species/organism to the CEC hazard expressed as no observed effect concentration (NOEC) that is the concentration in water which below an unacceptable effect is unlikely to be observed. The study results of chronic toxicity for fish and algae are 28 days NOEC in mg/l and 21 days NOEC in mg/l, respectively.

Microorganism, phytoplankton, plants, amphipods, fish, and insects present in different compartments are directly or indirectly related to acute to chronic effects. Considering the whole cycle, higher class living systems including human are also affected enormously due to these ecotoxicities. The ecosystem is divided into many environmental compartments. Thus, ecotoxicity of CECs can be classified based on the toxicity occurrence to specific compartments. The major ones are the following:

Table 1
Species-specific OECD testing guidelines for aquatic toxicity

Species	Test guidelines as per OECD
Fish	Acute Toxicity Test (OECD TG 203), Early-life Stage Toxicity Test (OECD 210), Short-term Toxicity Test on Embryo and Sac-Fry Stages (OECD TG 212), Juvenile Growth Test (OECD TG 215)
<i>Daphnia</i>	Acute Immobilisation Test (OECD TG 202), <i>Daphnia magna</i> Reproduction Test (OECD TG 211)
Algae	Growth Inhibition Test (OECD TG 201)

- (a) *Aquatic toxicity:* The term aquatic itself suggests toxicity related to the aquatic environment. Aquatic toxicity can be in different forms like the surface, ground, drinking water, river, and ocean pollution [13–16]. The aquatic toxicity happens due to either specific toxic interactions or nonspecific mechanisms like necrosis. Most commonly considered reason for chemical toxicity to aquatic species is necrosis (either non-polar or polar) which affects the perturbation of cellular functions. The aquatic toxicity is typically tested on organism's representative of the three trophic levels, i.e., plants (algae), invertebrates (crustaceans), and vertebrates (fish). Acute and chronic aquatic toxicity data are vital for evaluating the environmental hazard categorization of a chemical under the Globally Harmonized System of Classification and Labelling of Chemicals (GHS). The commonly used species and testing guidelines [17] for aquatic toxicity are portrayed in Table 1.
 - (b) *Terrestrial toxicity:* Terrestrial toxicity can be defined as the effects of a chemical to terrestrial organisms and plants [18]. In most of the cases, agrochemicals go through their risk assessment for terrestrial toxicity before approval to the market which is significant as a protective measure to the ecosystem. But, PPCPs and most industrial chemicals are not subjected to ecotoxicity testing. Species considered under terrestrial toxicity testing comprise soil microorganisms, earthworms, birds, plants, and bees. The commonly used species and testing guidelines [17] for terrestrial toxicity are portrayed in Table 2. In a broader perspective, soil and sediment pollution, sewage sludge, as well as air pollution fall under terrestrial ecotoxicity.
- The fate of CECs and their transformation into transformed products (TPs) or metabolites [19] are directly or indirectly related to some form of hazards and toxicity. The most common ones are bioaccumulation, bioconcentration, and biodegradability of CECs.

Table 2
Species-specific OECD testing guidelines for terrestrial toxicity

Species	Test guidelines as per OECD
Earthworm	Acute Toxicity Tests (OECD 207), Enchytraeid Reproduction Test (OECD 220), Earthworm Reproduction Test (OECD 222)
Plants	Vegetative Vigour Test (OECD 227), Seedling Emergence and Seedling Growth Test (OECD 208)
Pollinators	Honeybees Acute Oral Toxicity Test (OECD 213), Honeybees, Acute Contact Toxicity Test (OECD 214), Honey Bee (<i>Apis Mellifera</i>) Larval Toxicity Test, Single Exposure (OECD 237)
Soil Microorganism	Nitrogen Transformation Test (OECD 216), Carbon Transformation Test (OECD 217)
Terrestrial vertebrates	Avian Acute Oral Toxicity Test (OECD 223), Avian Dietary Toxicity Test (OECD 205), OECD 206 Avian Reproduction Test
Other nontarget arthropods	Predatory mite reproduction test in soil (OECD 226), Determination of Developmental Toxicity of a Test Chemical to Dipteran Dung Flies (OECD 228), Collembolan Reproduction Test in Soil (OECD 232)

- (a) *Bioaccumulation:* It occurs in organisms or specific species when the rate of uptake of a chemical surpasses the rate of elimination by all possible routes (food, air, soil/sediment, and water) of exposure [20]. With the bioaccumulation, a chemical can persist in the living and/or ecosystem for a long time and can exert its hazardous effect on the environment. Therefore, bioaccumulation is a phenomenon which should be curtailed when designing eco-friendly chemicals.
- (b) *Bioconcentration:* It is a procedure leading to a higher concentration of a chemical in an organism than in environmental media to which it is exposed [21]. Bioconcentration can be described as a subset of bioaccumulation and refers to the uptake and concentration of chemicals from water into aquatic organisms. Thus, the bioconcentration factor (BCF) is the ratio between the concentration of the chemical in biota and the concentration in water at steady state. The BCF can be computed by the ratio of the first-order uptake and elimination rate constants, a method that does not require equilibrium conditions.
- (c) *Biodegradability:* To decrease the toxic effect of a chemical, it needs to be eliminated from the environment as quickly as possible. Thus, increased biodegradability maintaining the required effect of the chemical is integral to chemical design [22]. Chemicals which struggle biodegradation endure to exert toxic effects on the environment, and ones that are bioaccumulate are of even greater worry because their levels

may be attained in organisms that seem safe on the basis of a single daily exposure, but because the actual dose efficiently accumulates over time, the result may be unexpected toxic effects. Chemicals that are easily biodegraded are eliminated from the environment rapidly.

3 Role of the QSAR Model in Ecotoxicity Evaluation

A quantitative structure-activity relationship (QSAR) model defines a biological response/toxicity or property as a mathematical function of the molecular structure [23]. In the case of ecotoxicity modeling, environmental toxicity responses are employed to develop in silico models for assessing the risk and exposure of chemicals and PPCPs to the environment [5, 6, 24]. The principle objectives and significance of QSAR/QSTR analysis are [25]:

- Prediction of new analogues of the compound with lesser toxicity in respect to living system as well as the environment.
- Better understanding and exploration of the MOA for toxicity.
- Optimization of the lead compound with decreased toxicity.
- Reduction of wet laboratory experimentation and sacrifice of a large number of animals.
- Reduction of the cost, time, and manpower requirement by developing more effective compounds using a scientifically less exhaustive approach.
- The expert systems will also provide structural alerts to identify fragments mediating different toxicities.
- To combine a scientific and pragmatic approach to guide policy directions.
- To identify pollution prevention measures.
- To identify scientific data gaps.

Over the years, for the prediction of diverse toxicity endpoints, QSAR models have been developed as one of the alternative approaches for time-consuming and animal-dependent experiments. Zhao et al. [26] constructed QSAR models for predicting bioconcentration factor of 473 heterogeneous chemicals employing multiple linear regression (MLR), radial basis function neural network (RBFNN), and support vector machine (SVM) tools. QSAR models were developed by Xia et al. [27] for toxicity prediction of 91 aliphatic and aromatic chemicals using the linear (HM) and the nonlinear method radial basis function neural networks (RBFNN). A QSTR model was developed by Jalali-Heravi and Kyani [28] for a series of 268 substituted benzene derivatives using mechanistically interpretable descriptors employing

shuffling-adaptive neuro-fuzzy inference system (Shuffling-ANFIS) to select the important factors affecting the toxicity of substituted benzenes to *T. pyriformis*. First interspecies QSAR (i-QSAR) models were developed by Kar and Roy [29] to correlate the ecotoxicity of structurally diverse 77 pharmaceuticals to *Daphnia magna* and fish. Acute toxicity of 55 PPCPs toward the *Dugesia japonica* was modeled with the QSAR by Önlü and Saçan [30]. Khan et al. [31] developed ecotoxicological multiple QSAR models employing 260 pharmaceuticals on 3 trophic level species *Daphnia magna* (209), *Scenedesmus subspicatus* (134), and *Brachydanio rerio* (192) using the PLS approach and 2D descriptors for modeling. Sangion and Gramatica [32] developed QSAR models for 1267 pharmaceuticals collected from the ECOTOX database to predict acute toxicity toward four species *P. subcapitata*, *D. magna*, *O. mykiss*, and *P. promelas* spanning over three aquatic trophic levels. The endocrine disruption of perfluoroalkyl substances (PFASs) was modeled by classification and regression-based QSAR models followed by docking studies to interpret the significant structural features accountable for toxicity profiles by Kar et al. [33]. Kar et al. [34] reported statistically robust QSAR models employing single and mixture halogenated chemicals employing weighted descriptors approach for predicting developmental toxicity on *Danio rerio* embryos.

4 Toxicity and Ecotoxicity Databases

Assessment of potential aquatic toxicity, terrestrial toxicity, and fate and transformation of chemicals and PPCPs along with their bioaccumulation, bioconcentration, and biodegradation from compound's chemical structure information is extremely useful, and it defines collective goal of various academicians, industries, and government regulatory authorities. Although multiple techniques and criteria for toxicity and risk assessment are used, there is a necessity for reliable and open access to existing ecotoxicity data linked with chemical structure information. The databases are one of the starting points for computational or in silico modeling (e.g., QSAR, machine learning, HTS). Most commonly used ecotoxicity databases developed over the years are presented as a word cloud in Fig. 1.

4.1 Aggregated Computational Toxicology Online Resource (ACToR)

A publicly accessible database of industrial chemicals, pesticides, and drinking water contaminants is maintained by United States Environmental Protection Agency (US EPA) National Center for Computational Toxicology [35]. The database contains chemical structure and physicochemical information and offers in vitro and in vivo toxicology data for over 500,000 environmental chemicals. The ACToR is also a web applications warehouse for EPA's

**Fig. 1** Major ecotoxicity databases

Table 3
Summary statistics

Group	Total
Assays	506,534
Assay components	1,030,334
Assay results	44,420,380
Data collections	2701
Substances	3,221,191
Compounds	893,280
Generic chemicals	559,802
Generic chemicals with structure	456,918

computational toxicology information which provides chemical exposure, HTS, virtual tissues data, and sustainable chemistry which can be employed to explore and visualize multifaceted computational toxicology data. In ACToR, chemicals are systematized into three classes: substance (a substance is the article that was tested and provides a link to assay and other test data), compound (a compound holds chemical structure information), and generic chemical (a generic chemical aggregates a chemical structure plus all the corresponding substances. The common link is that all substances share the same CAS registry number). A brief statistic of the ACToR database resources is illustrated in Table 3.

Web Accessibility: <https://actor.epa.gov/actor/home.xhtml>

4.2 Birth Defects Systems Manager (BDSM)

The BDSM database, developed by the University of Louisville, deals with developmental toxicity [36]. The dataset comprises 232 microarrays of RNA samples by single or dual microarray platforms, human or mouse sequence information, and cDNA or oligonucleotide-based probes. The database is an open access and can be integrated with bioinformatics tools and materials to advance the stride of discovery in birth defects. Primary outcomes recognize system-level properties in the embryonic transcriptome as it responded to numerous developmental-teratological stimuli.

Web Accessibility: <http://systemsanalysis.louisville.edu/>

4.3 Carcinogenic Potency Database (CPDB)

The CPDB developed by the University of California, Berkeley, and the Lawrence Berkeley National Laboratory analyzes animal cancer tests used in support of cancer risk assessments for human [37]. It includes 6540 chronic, long-term animal cancer tests on 1547 chemicals from the literature as well as from the National Cancer Institute (NCI) and the National Toxicology Program (NTP). The carcinogenic potency is described in form of TD₅₀ which can be described as the daily dose rate in mg/kg/bodyweight/day for life to induce tumors in half of the test animals that would have remained tumor-free at zero dose. The TD₅₀ is an important standardized quantitative measure which can be employed for interpretation of diverse issues in carcinogenesis. The website is completely searchable by all required options and includes InChI codes, SMILES, and structures for all compounds. A collection of CPDB outcomes organized by target organ illustrates all chemicals that induce tumors in each of 35 target organs.

Web Accessibility: <http://potency.berkeley.edu/>

4.4 Chemical Carcinogenesis Research Information System (CCRIS)

Chemical carcinogenesis research information system (CCRIS), a government organization, formed by the National Cancer Institute (NCI) of the United States consists of mutagenicity, carcinogenicity, tumor inhibition, and tumor promotion test results for over 8000 chemicals [38]. Data are collected from current awareness tools, NCI reports, journals, books, etc. CCRIS offers information from the years 1985–2011. The only loophole of this database is no longer updated.

Web Accessibility: <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>

4.5 Danish (Q)SAR Database

The Danish (Q)SAR database (Fig. 2) is developed by the Technical University of Denmark with support from the Danish Environmental Protection Agency, National Food Institute, the Nordic Council of Ministers, and the European Chemicals Agency. This is a source for more than 200 QSARs from connected to physicochemical properties, environmental fate, ecotoxicity, and ADMET. Around 600,000 chemicals can be searched based on individual profile and chemical similarity [39].

Web Accessibility: <http://qsar.food.dtu.dk/>



Fig. 2 Screenshot of Danish (Q)SAR database

4.6 Developmental and Reproductive Toxicology Database (DART)

The DART covers teratology and reproductive and developmental toxicology of more than 400,000 journal references published since 1950 [40]. It's one the TOXNET database funded by the United States National Library of Medicine (NLM), the US EPA, the National Institute of Environmental Health Sciences, and the National Center for Toxicological Research of the Food and Drug Administration.

Web Accessibility: <https://toxnet.nlm.nih.gov/newtoxnet/dart.htm>

4.7 Developmental Toxicity (DevTox)

The DevTox database (Fig. 3) is planned to deliver a significant resource in the field of developmental toxicology for various strains of common laboratory animals. It represents the inclusive resources of images of developmental abnormalities. The DevTox Project was introduced by the German Federal Ministry of Food, Federal Ministry of the Environment, Nature Conservation and Nuclear Safety (BMU), and Agriculture and Consumer Protection (BMELV) under the sponsorships of the International Programme on Chemical Safety (IPCS) [41].

Three major parts which are accessible on this site of the project are the following:

- DevTox Background (supplementary information on the project)

Fig. 3 Screenshot of DevTox

- DevTox Nomenclature (contains nomenclature for developmental toxicology following the revised International Federation of Teratology Societies (IFTS) terminology with more than 2500 images for skeletal, external, maternal-fetal, and soft tissue anomalies)
- DevTox Data (experimental data on developmental toxicity in rabbits and rats in diverse laboratories in various strains of common laboratory animals)

The DevTox has three aims:

- To harmonize the nomenclature employed to define developmental variances in laboratory animals
- To support in the visual recognition of developmental anomalies with the aid of photographs
- To offer a historical control database of developmental effects in laboratory animals

Web Accessibility: <http://www.devtox.org>

4.8 Distributed Structure-Searchable Toxicity Database (DSSTox)

The DSSTox database prepared by the National Center for Computational Toxicology, US EPA, offers downloadable, structure-searchable, standardized chemical structure files connected with chemical toxicity data of environmental relevance [42]. One of the significant features of this database is a precise mapping of bioassay and physicochemical properties data related

with chemicals to their corresponding chemical structures. It also incorporates cheminformatics workflows with the chemical infrastructure for EPA's Safer Chemicals Research, including the Tox21 and ToxCast high-throughput toxicology efforts. Thus, this database promotes the implication of formalized and structure-annotated toxicity models, serving to interface these results with QSAR modelers.

Web Accessibility: <http://www.epa.gov/ncct/dsstox/index.html>

4.9 ECOTOXicology Knowledgebase (ECOTOX)

The ECOTOX (Fig. 4) consists of toxicity data for terrestrial plants, aquatic life, and wildlife of chemicals. ECOTOX was formed and is maintained by the US EPA's National Health and Environmental Effects Research Laboratory's (NHEERL's) Mid-Continent Ecology Division (MED) [43]. Three independent databases named AQUIRE, TERRETOX, and PHYTOTOX were integrated to form ECOTOX which derived toxicity data mostly from the peer-reviewed articles. The database considered the following requirements and inclusions:

- Single chemicals relevant to environmental exposure are included.

The screenshot shows the ECOTOX Knowledgebase homepage. At the top, there is a navigation bar with links for Environmental Topics, Laws & Regulations, About EPA, Home, Search, Explore, Help, and Contact Us. The main content area features a dark background with white text. On the left, a sidebar displays "Data last updated Mar 14, 2019" and a link to "See update totals". In the center, there is a table showing database statistics: 11,695 Chemicals, 12,713 Species, 48,464 References, and 930,643 Results. Below this, a green banner reads "WELCOME TO ECOTOX VERSION 5! Please click here to provide feedback so that we can continue to improve your experience." The bottom section contains three main boxes: "About ECOTOX" (describing the database as a comprehensive, publicly available knowledgebase), "Getting Started" (with a list of links for search, explore, user guides, and code appendix), and "New Data Visualizations!" (describing the EXPLORE feature). Each box includes small icons related to its content.

Fig. 4 Screenshot of ECOTOX

- Verifiable Chemical Abstract Services (CAS) number of chemical.
- Ecologically relevant species are considered only.
- Priority species are wild. Also, laboratory species and terrestrial domestic are employed to fill data gaps when required.
- Biological effect on live, whole organisms needs to be considered.
- Adverse effects are priority for inclusion.
- Concurrent environmental chemical concentration/dose reported as concentration, dose, or application rate.
- Sediment studies must have a water concentration reported to be included.
- Duration reports an associated concurrent with a biological effect.

Data not satisfying the following requirements are left out from the ECOTOX database:

- Mixtures
- Air pollution (CO₂, ozone)
- Human, monkey, bacteria, viral, and yeast under species conditions
- Inhalation studies route, sediment only concentration, lead shot, log values under concentration/dose criteria
- Reviews, full-text foreign language, abstract-only format under publication search

Web Accessibility: <http://cfpub.epa.gov/ecotox/>

4.10 European Chemical Substances Information System (ESIS)

The ESIS database includes classifications and labellings of chemicals or groups of chemicals according to the criteria in Directive 67/548/EEC for one or more endpoints. The ESIS is an IT system which illustrates information on chemicals related to the following [44]:

- Persistent, bioaccumulative, and toxic (PBT) and very persistent and very bioaccumulative (vPvB) data
- European Inventory of Existing Commercial Chemical Substances (EINECS) O.J.C 146A, 15.6.1990
- European List of Notified Chemical Substances (ELINCS) in support of Directive 92/32/EEC, the 7th amendment to Directive 67/548/EEC
- No-Longer Polymers (NLP)
- Biocidal Products Directive (BPD) active substances listed in Annex I or IA of Directive 98/8/EC or listed in the so-called list of non-inclusions

- Classification and Labelling (C&L), substances or preparations in accordance with Directive 67/548/EEC (substances) and 1999/45/EC (preparations)
- Export and Import of Dangerous Chemicals listed in Annex I of Regulation (EEC) No 304/2003
- High Production Volume Chemicals (HPVCs) and Low Production Volume Chemicals (LPVCs), including EU Producers/ Importers lists
- Priority lists, risk assessment process, and tracking system in relation to Council Regulation (EEC) 793/93 also known as Existing Substances Regulation (ESR)
- IUCLID Chemical Data Sheets, IUCLID Export Files, OECD-IUCLID Export Files, and EUSES Export Files

Data available in zipped XLS format which include following information of chemicals:

- Index No
- Chemical names
- Notes related to substances
- EC No
- CAS No
- Classification
- Labelling
- Concentration limits
- Notes related to preparations

Web Accessibility: <https://old.datahub.io/dataset/esis>

4.11 Extension TOxicology NETwork (EXTOXNET)

The EXTOXNET (Fig. 5) is a cooperative effort of Oregon State University, University of California-Davis, Cornell University, Michigan State University, and the University of Idaho. The first edition was published in 1989 and funded by the United States Department of Agriculture (USDA) and the USEPA [45]. The second edition was made conceivable through a grant from the National Agricultural Pesticide Impact Assessment Program (NAPIAP), a program of the United States Department of



The Extension TOxicology NETwork

Check out the EXTOXNET Frequently Asked Questions (FAQs)

You may go directly to the "EXTOXNET Global Search and Browse" page.

Fig. 5 Screenshot of EXTOXNET

Agriculture. The EXTOXNET primary files are maintained and archived at the Oregon State University. This database is mostly related to pesticide toxicology. Pesticide Information Profiles (PIPs) and Toxicology Information Briefs (TIBs) provide a summary of each pesticide's toxic effects and their probable actions in the environment.

The database contains the following information for each pesticide:

- *Toxicological effects*

Acute toxicity, chronic toxicity, reproductive toxicity, teratogenic effects, mutagenic effects, carcinogenic effects, organ toxicity, fate in human and animal

- *Ecological effects*

Effects on birds, effects on aquatic species, effects on other animals (nontarget species)

- *Environmental fate*

Breakdown of chemical in soil and groundwater, breakdown of chemical in surface water, breakdown of chemical in vegetation

- *Physical properties*

Appearance, stability, CAS number, molecular weight, water solubility, solubility in other solvents, melting point, vapor pressure, partition coefficient (octanol/water), adsorption coefficient

- *Exposure Guidelines*

ADI, HA, RfD, PEL/TLV

Web Accessibility: <http://extoxnet.orst.edu/ghindex.html>

4.12 eTox

A drug safety database developed in collaboration among 13 pharmaceutical industries, 11 academic institutions, and 6 small- and medium-sized enterprises (SMEs) comprises of toxicology data. The eTOX project (Fig. 6) was approved as one of the first Innovative Medicines Initiative (IMI) projects which started in 2010 and successfully ended in 2016 [46]. Along with toxicological data within the pharmaceutical industry, it created a series of predictive models to support toxicity prediction for the future. The eTOX contains 1947 compounds and 8047 study design records from 6971 reports along with 265,502 substances and 1,088,007 records from public sources like DrugMatrix, ChEMBL, and Open TG-GATES. A platform called eTOXsys integrated both data and models which is a powerful system to access the eTOX data and the predictive models. Five versions of the eTOXsys were launched, and the final version is 2016.3 Vitic release which contains 200 predictive models and the Human Outcomes Module. The Human Outcomes Module was designed to advance

Fig. 6 Screenshot of the eTOX database

translational research from preclinical to clinical research. Multiple open-access prediction tools were prepared for the scientific community under the name of Auto tools. The significant ones are eTOXsys, Human Outcomes Module, eTOXlab, LiMTox, etc.

Web Accessibility: www.etoxproject.eu/

4.13 Fraunhofer RepDose

The RepDose (repeated dose toxicity) database (Fig. 7) contains around 3100 studies on subacute to chronic toxicity for diverse routes of administration (oral or inhalation exposure) for about 930 chemicals carried out in mice, rats, and dogs [47]. The database is important for the investigation of the relationship between chemical functional groups and target organs in repeated dose studies. This publicly accessible database considered only predominantly peer-reviewed studies with the following sources: German MAK-Documentations, EU RAR, EHC, HPV-chemicals, Reports from German BG Chemie, and NTP reports. The database has been established by Fraunhofer ITEM with support from Cefic LRI to analyze and improve chemical risk assessment strategies. Nowadays the data of RepDose is circulated commercially through association with MN-AM (Molecular Networks—Altamira). Individual chemicals are characterized by 1–15 studies. However, the majority of them have undergone one to four studies (Table 4). L (N)OEL values (lowest (no) observed effects level) are given for each effect and each study.

Web Accessibility: <http://www.fraunhofer-repdose.de/>

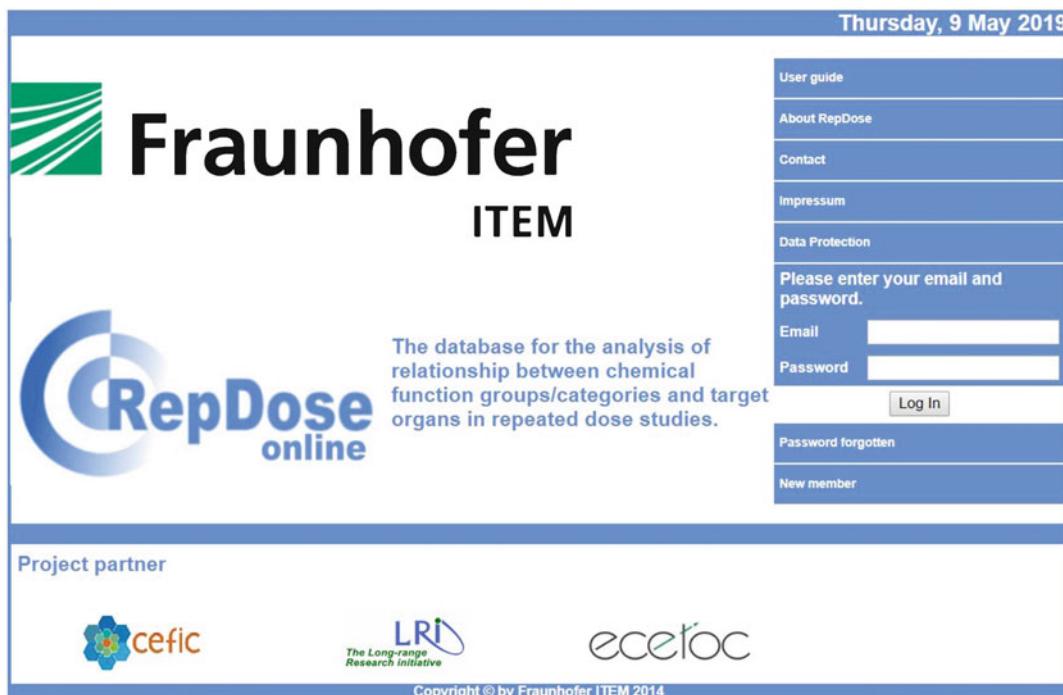


Fig. 7 Screenshot of Fraunhofer RepDose database

Table 4
Data quality parameter employed in RepDose database

Reliability	Description
A	Following OECD guidelines or comparable quality
B	Some deficits, but related for the evaluation
C	Quality cannot be assessed due to insufficient information
D	Special design for a certain endpoint

4.14 Genetic Alterations in Cancer (GAC)

GAC is a publicly available database that quantifies specific mutations found in cancers induced by environmental chemicals created by National Institute of Environmental Health Sciences (NIEHS) and US National Institutes of Health (NIH) [48]. The GAC web-based capable of evaluating data attained from peer-reviewed literature of genetic changes in tumors connected with exposure to physical, chemical, or biological agents, as well as natural tumors along with gene mutation data. The NIEHS mainly emphasizes on environmental reasons for cancer which helps in prevention strategies by recognizing adjustable risk factors and genetic factors that are involved in tumor development. Outcomes from human and rodent studies are incorporated and are systematized by strain,

species, tumor type and origin, target organ, and agent. Data mining features employed to inquiry the database integrate and summarize data from all experiments. The search tactics were applied to the peer-reviewed articles that encountered three critical conditions: (1) explanation of the tumor(s) and sign of which ones were connected with acquaintance to a specific agent and which were unprompted; (2) molecular investigation of the tumor sample for genetic modifications; and (3) recognition of the affected gene(s) and explanation of the gene change(s).

Web Accessibility: <http://www.niehs.nih.gov/research/resources/databases/gac/index.cfm>

4.15 Genetic Activity Profile (GAP)

The GAP has been developed to offer a matrix of data on the quantitative genotoxicity results of around 500 chemicals to support hazard classification of human carcinogens [49]. The complete procedure for the generation and assessment of GAPs has been settled in partnership with the International Agency for Research on Cancer (IARC) Monograph and US EPA. The database offers the overview of doses and test results data for individual chemicals followed by which either the highest ineffective dose (HID) or lowest effective dose (LED) is documented. The GAPs comprise of beneficial data for the generation of weight-of-evidence hazard ranking schemes and information of the likely genetic activity of multifaceted environmental mixtures.

Web Accessibility: <http://www.ils-inc.com/services/information-sciences>

4.16 Gene-Tox

The GENE-TOX, a TOXNET database, delivers mutagenicity test data from peer-reviewed scientific literature for more than 3000 chemicals from the US EPA [50]. The GENE-TOX was established to choose genetic toxicology assay data for evaluation, which review and recommend appropriate testing protocols. The database covers entries on chemicals and their likely negative effects on DNA. Users can search by compound and CAS. It covers the data from year 1991 to 1998, and it is no longer updated.

Web Accessibility: <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX>

4.17 Human and Environmental Risk Assessment (HERA)

The HERA is a volunteer industry risk assessment program on ingredients of household cleaning products (Fig. 8). It is an exclusive European partnership established in 1999 in Brussels, Belgium, between the chemical industry (Cefic) who provides the raw ingredients and the makers of household cleaning products (A.I.S.E.) [51]. Water-soluble linear polycarboxylates are major key ingredients used in dishwashing detergents, laundry detergents, hard surface cleaning formulations, and industrial and institutional cleaning processes and a variety of technical applications. Major polycarboxylates comprise two different types of polymer

Fig. 8 Screenshot of HERA database

families which can be differentiated based on their physicochemical properties and technical applications. In the HERA report, homopolymers of acrylic acid (P-AA) are described in **Part I**, and copolymers of acrylic/maleic acid (P-AA/MA) are described in **Part II**. This database is planned to support a risk-based method to chemicals legislation in the European Union and might aid as a pilot plan for the evaluation of safety data on the components used in these products in an active and translucent way.

Web Accessibility: [http://www.heraproject.com/
RiskAssessment.cfm](http://www.heraproject.com/RiskAssessment.cfm)

4.18 Hazard Evaluation Support System (HESS) Attached Database (HESS DB)

The repeated dose toxicity (RDT) is a significant endpoint in the risk assessment of hazardous chemicals. Due to the complexity of the endpoints connected with an entire body assessment, the development of the mechanistically interpretable structure-activity model is a challenging task. The structural alerts, read-across, and category approach built on mechanism evidence are effective tactics for data gap filling of RDT. Utilizing experimental RDT information, a toxicological category library is developed for 500 chemicals along with mechanistic information of the toxic effects of those chemicals on diverse organs. Here, 33 categories were well-defined for 14 kinds of toxicity, such as hemolytic anemia, hepatotoxicity, etc. This library was then incorporated in an integrated computational platform named HESS to offer mechanistically realistic predictions of RDT values for new and/or untested chemicals [52]. There is another attached database known as HESS DB (Fig. 9) which contains information from two databases. The first one is a toxicity knowledge database comprising data on RDT and toxicity mechanisms. The second one is a metabolism knowledge database containing rat metabolism maps and information on absorption, distribution, metabolism, and excretion (ADME) in humans and rats. The databases and platform are maintained by



Fig. 9 Screenshot of HESS database

the National Institute of Technology and Evaluation, Japan. For prediction purpose, the HESS is compatible with the OECD QSAR Toolbox.

Web Accessibility: <https://www.nite.go.jp/en/chem/qsar/hess-e.html>

4.19 Hazardous Substances Data Bank (HSDB)

The HSDB is a comprehensive toxicology database under TOX-NET, which contains toxicity records for about 5600 potentially hazardous chemicals (As of November 2014) [53]. It includes evidence on industrial hygiene, human exposure, environmental fate, emergency handling procedures, regulatory requirements, nanomaterials, and related areas. All data are summarized from government and regulatory documents, books, and peer-reviewed research articles. It can be accessed freely, and users can check the information by providing a chemical name, CAS registry number. Mixtures, radioactive materials, animal toxins, oil dispersants, and crude oil are included under the HSDB database.

Web Accessibility: <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>

4.20 International Agency for Research on Cancer (IARC) Monograph

The IARC is established by the World Health Organization (WHO) which classifies environmental factors which can upsurge the risk of human cancer. The monograph includes chemicals (phenobarbital), mixtures (air pollution), physical agents (UV radiation), occupational exposures (asbestos industry), biological agents (hepatitis B virus), as well as lifestyle factors (smoking) [54]. The complete principles, actions, and scientific process that direct the assessments are defined in the Preamble to the IARC Monographs (Fig. 10). Around 400 substances are identified as carcinogenic, probably carcinogenic, or possibly carcinogenic to humans from the study results of more than 1000 agents since 1971. It majorly identifies environmental factors which are carcinogenic to humans. The reports can help as scientific support for the national health agencies to avert exposure to latent



Fig. 10 Screenshot of IARC Monograph

Table 5
Classification of compounds in IARC monograph

Group	Classification	Number of agents
Group 1	Carcinogenic to humans	120
Group 2A	Probably carcinogenic to humans	82
Group 2B	Possibly carcinogenic to humans	311
Group 3	Not classifiable as to its carcinogenicity to humans	500
Group 4	The agent is probably not carcinogenic to humans	–

carcinogens. The IARC Monographs are supported with funds from the US National Cancer Institute, European Commission Directorate-General for Employment, Social Affairs and Inclusion, and US National Institute of Environmental Health Sciences. At the present moment, IARC Monographs have 123 volumes where agents are classified into 4 major groups based on carcinogenic effects of chemicals (see Table 5).

Web Accessibility: <http://monographs.iarc.fr/>

4.21 Integrated Risk Information System (IRIS)

The IRIS was created under the National Center for Environmental Assessment (NCEA), US EPA in 1985 to make available a database related to human health effects due to chemicals existing in the environment [55]. The database comprises data on 540 environmental chemicals and their possible hazardous effects on human. It is located in the Office of Research and Development (ORD) to make certain that IRIS can build up independent toxicity information to set national standards and clean up hazardous sites. The IRIS (Fig. 11) assessment covers a single chemical, group of chemicals, and complex mixture of chemicals.

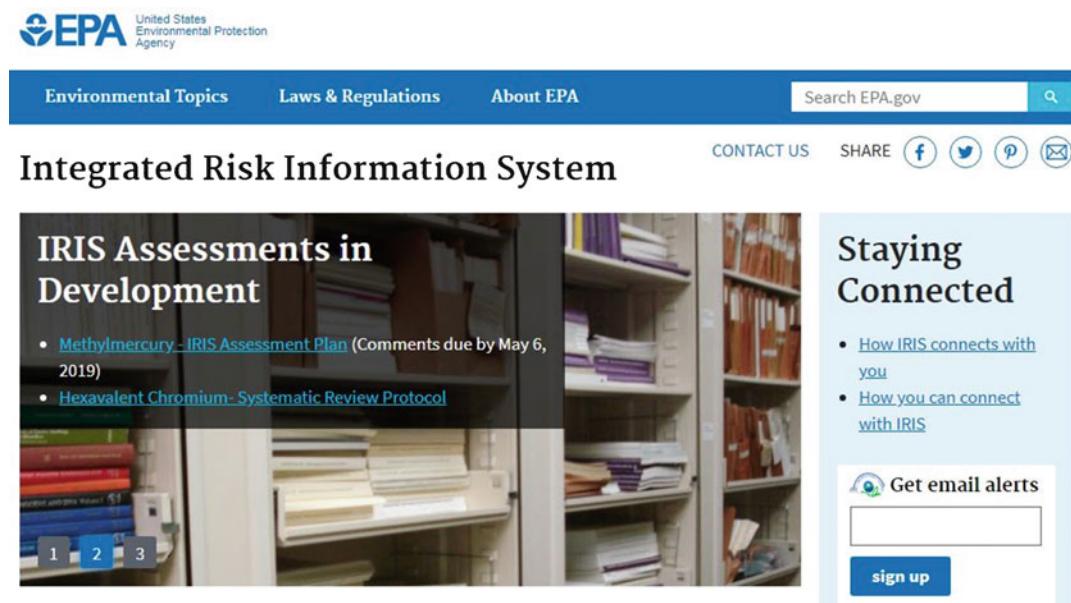


Fig. 11 Screenshot of IRIS

The IRIS database provides the following toxicity assessment values for health effects resulting from chronic exposure to chemicals:

- Reference dose (RfD)
- Reference concentration (RfC)
- Cancer descriptors characterize the chemical as:
 - (i) Carcinogenic to humans
 - (ii) Likely to be carcinogenic to humans
 - (iii) Suggestive evidence of carcinogenic potential
 - (iv) Inadequate information to assess carcinogenic potential
 - (v) Not likely to be carcinogenic to humans
- Inhalation unit risk (IUR)
- Oral slope factor (OSF)

The risk assessment is a four-step practice as mentioned by the National Research Council (NRC) in 1983:

- *Identification of hazard* associated with exposure to a chemical.
- *Dose-response assessment* which illustrates the relationship between each probable health hazard and chemical exposure in a quantitative manner, followed by an account for high to a low dose, animal to human, route to route, and other differences.

- *Exposure assessment* which identifies human exposure pathways and approximates the quantity of human exposure under diverse exposure.
- *Risk characterization* combines their exposure assessment with the hazard data and toxicity values from IRIS to illustrate prospective public health risks followed by *risk management*.

The IRIS follows below-mentioned steps for developing human health assessments:

- Step 1- Draft Development
- Step 2- Agency Review
- Step 3- Interagency Science Consultation
- Step 4- Public Comment and External Peer Review
- Step 5- Revise Assessment
- Step 6- Final Agency Review/Interagency Science Discussion
- Step 7- Final Assessment

Web Accessibility: <https://www.epa.gov/iris>

4.22 International Toxicity Estimates for Risk (ITER)

ITER is developed by TERA (Toxicity Excellence for Risk Assessment), and it consists of human health risk values and cancer classifications for over 680 chemicals of environmental concern [56]. It represents the key risk information from a series of organizations throughout the world in a side-by-side format explaining the variances in risk data evaluated by multiple organizations, and it has a direct access to individual organization's website and source documents for comprehensive information (Table 6). ITER (Fig. 12) is one of those organizations which includes risk assessment data from an independent organization whose risk values have undertaken autonomous peer review. It includes four groups of risk data: (1) cancer oral, (2) cancer inhalation, (3) noncancer inhalation, and (4) noncancer oral.

Web Accessibility: <http://www.tera.org/iter/>

4.23 Japan Existing Chemical Database (JECDB)

The JECDB is a toxicity database (Fig. 13) consisting of information related to hazard assessment and toxicity test reports of environmental and industrial chemicals maintained by the Japanese Ministry of Health, Labour and Welfare. The data are reviewed and obtained by scientists from the National Institute of Health Sciences and other institutes [57]. Toxicity data for single-dose toxicity test, a 28-day repeat dose toxicity test, a developmental/reproductive toxicity test, and mutagenicity tests are included for each chemical existing in the database. Reports comprise of the chemical's nomenclature and summarized data from the studies. Search operation can be performed by name, by CAS registry number, and by toxicity test.

Web Accessibility: http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp

Table 6
Type of organization and their data in ITR

Organization	Risk values and cancer classifications	Website
Agency for Toxic Substances and Disease Registry (ATSDR) –Toxicological Profiles	Chronic minimal risk levels	http://www.atsdr.cdc.gov/toxpro2.html
Health Canada –Priority Substances Assessment Reports	Tolerable daily intakes, tolerable concentrations, cancer potencies, and classifications	http://www.hc-sc.gc.ca/ewh-semt/pubs/contaminants/index_e.html
IARC Monographs	Cancer classifications	http://monographs.iarc.fr/
NSF International—Oral Risk Assessment Documents	Oral reference doses and cancer classifications	http://www.nsf.org
National Institute of Public Health and the Environment (RIVM), the Netherlands—Maximum Permissible Risk Level Reports	Maximum permissible risk levels with tolerable daily intakes, tolerable concentrations in air, cancer risk estimates	http://www.rivm.nl/bibliotheek/rapporten/711701025.pdf
US EPA IRIS	Reference doses and concentrations, cancer risk estimates and classifications	http://www.epa.gov/iris/index.html
“ITER PR” or “IPRV” Column—Government and private parties whose risk values have undergone independent peer review	Various values and information, depending on the source and chemical	http://www.tera.org/iter/about.htm

4.24 LeadsScope

LeadsScope is a commercial database (Fig. 14) of the National Institute for Occupational Safety and Health (NIOSH) Registry of Toxic Effects on Chemical Substances (RTECS) consisting of 400,000 acute, sub-chronic, genotoxicity, carcinogenicity, and reproductive toxicity data for around 180,150 chemicals. Every year, the updated database adds about 2000 new chemicals [58]. The RTECS database is improved by adapting the distributed ASCII data file into a XML format identified as ToXML (<http://www.xml.org>). Further, the obtained XML data is then uploaded into LeadsScope’s structure-searchable cheminformatics system which is available under LeadsScope’s data mining application or web services. A summary of the available toxicity information and endpoints ready for use in 2015 edition is outlined in Table 7.

Characteristics of the database:

- The toxicity data include:
 - The US Food and Drug Administration (US FDA) priority-based assessment of food additives (PAFA) database



(ITER) International Toxicity Estimates for Risk Assessment

ITER is an Internet database of human health risk values and cancer classifications for over 680 chemicals of environmental concern from multiple organizations worldwide. ITER is the only database that presents risk data in a tabular format for easy comparison, along with a synopsis explaining differences in data and a link to each organization for more information.

[Search ITER on TOXNET](#)

CHEMICAL FORMULA

Independent Non-profit Science
for public health protection.

Search

Connect with us on social media

Fig. 12 Screenshot of TERA

Japan Existing Chemical Database (JECDB)

[English | [Japanese](#)]

Background to the Publication of Toxicity Testing Reports for Environmental Chemicals and the Framework of the Reports, Safety Examination of Existing Chemicals and Safety Programmes in Japan.

Contents
[List of Chemical Substances Tested](#),
[List of Toxicity Test Reports](#).

General search

Select search item, and enter keyword.

Search Item CAS RN Name
Keyword
Toxicity test

Fig. 13 Screenshot of JECDB

- The RTECS database
- NTP Chronic Database
- DSSTox Carcinogenicity Potency Database (CPDB)
- Acute and sub-chronic hepatotoxicity, genetic toxicity, carcinogenicity, and reproductive and irritation toxicity with multiple dose studies

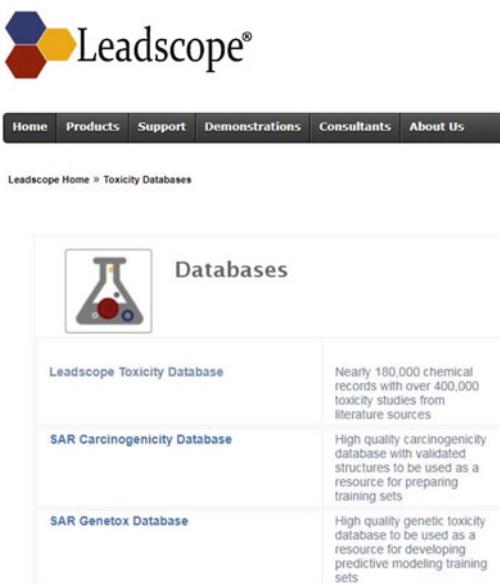


Fig. 14 Screenshot of Leadscape

- Chemical structure search option is present based on precise match, substructure, and similarity.
- Data mining is possible based on toxic effect, type of study, species, dosage, sex, duration, and route of exposure.

Web Accessibility: http://www.leadscape.com/toxicity_databases/

4.25 MDL

The MDL toxicity database is a commercially structure-searchable bioactivity database consisting of around 151,310 toxic chemicals from in vitro and in vivo studies. This database also includes data from the RTECS database by NIOSH [59]. It covers toxicity data from the year 1902 to till date from over 2950 literature and conference sources and is updated quarterly. This is an Oracle-based system reachable through MDL ISIS/Host covering mutagenicity, reproductive toxicity, carcinogenicity, eye irritation, skin toxicity, and multiple dose effects. The database comprises 65% compounds connected to drugs and drug-related substance, and remaining ones are industrial and petrochemicals, agrochemicals, flavors and fragrances, plant and animal extracts, organometallics, inorganic compounds, etc. Interestingly, MDL toxicity database can be employed concurrently with the MDL ISIS Toxicity Finder, the MDL Metabolite Database, and the MDL Metabolite Browser.

Web Accessibility: <http://www.iop.vast.ac.vn/theor/conferences/smp/1st/kaminuma/ChemDraw/toxicity.html>

Table 7
Toxicity information and endpoint types under LeadsScope database based on 2015 edition

<i>Toxicity study type</i>	<i>Study count</i>	<i>Structure count</i>
Acute	279,237	154,702
Irritation	8017	4880
Multiple dose	52,730	13,953
RTECS mutation	46,385	13,343
Reproductive	26,558	6851
Tumorigenic	10,517	3724
Totals	423,444	180,145
<i>Endpoint type</i>	<i>Structure count</i>	
Acute LC50	3618	
Acute LD50	177,082	
Irritation LOAEL—open test	607	
Irritation LOAEL—standard draize	5979	
Multiple dose TCLo	2240	
Multiple dose TDLo	19,075	
Multiple dose LOAEL—rat	141	
Reproductive TCLo	296	
Reproductive TDLo	10,151	
Tumorigenic TDLo	4781	
Tumorigenic TD	1062	
Total	225,032	

4.26 National Toxicology Program (NTP)

The NTP is an interagency program (Fig. 15) under the US Department of Health and Human Services to assess and identify the hazardous substances of public health concern by generating and using the modern toxicology and molecular biology studies [60]. The program conducts a number of studies enlisted in Table 8.

Web Accessibility: <http://ntp.niehs.nih.gov/>

4.27 Organization for Economic Cooperation and Development (OECD)

The OECD eChemPortal (Fig. 16) has access to information on physicochemical properties, environmental fate, and toxicity for a huge number of environmental chemicals [61]. The OECD is helping countries and industries in harmonizing toxicity testing guidelines for the testing of chemicals and good laboratory practice (GLP) to maintain the reliability and quality of test data to take

Fig. 15 Screenshot of NTP database

advantage from the OECD agreement on Mutual Acceptance of Data (MAD) and avoid duplicative testing. The OECD eChem-Portal is initiated in the year 2004 to provide toxicity information on chemical substances, in response to the request by the World Summit on Sustainable Development to improve the availability of hazard data of chemicals. The eChemPortal is also an outcome of the Strategic Approach to International Chemicals Management (SAICM). The current version of eChemPortal was made available on 12 June 2015. Countries like the United States, Japan, and Canada, the European Chemicals Agency, the European Commission, the Business and Industry Advisory Committee, the International Council of Chemical Industry Associations, the World Health Organization's International Programme on Chemical Safety, UNEP Chemicals, and Environmental NGO's are associated with the OECD Secretariat to develop this project. The foremost purposes of eChemPortal are:

1. To enable fast and effectual use of this information
2. To make this information on prevailing chemicals openly available and free of charge
3. To permit the efficient exchange of this information

Table 8
Toxicity study under NTP

Type	Description	Toxicity test
Toxicology and carcinogenicity	To address the gap in knowledge concerning the toxicity of substances To evaluate dose-response relationships between exposed and unexposed organisms To determine if a substance elicits toxic effects and/or causes cancer	There are two categories of toxicology/carcinogenicity studies: Short-term toxicity (14-day and 13-week) Long-term carcinogenicity (2-year) Species: rats, mice
Alternative toxicity models	NTP implements a variation of testing strategies to obtain data about possibly hazardous environmental and occupational substances followed by information to regulatory bodies	The purpose is the 3Rs approach Systems include: Computer-based predictive toxicology models In vitro cell- and tissue-based systems Transcriptomic profiling Microphysiological systems (“organs-on-a-chip”) Fish embryo models
Chemical disposition	Chemical disposition studies are done to evaluate what occurs to a chemical in an organism after that organism is exposed by measuring the absorption, distribution, metabolism, and excretion (ADME) of a chemical or substance	Studies are designed to determine the effect on each of these parameters of (1) species, (2) sex and age of animals, (3) dose level and frequency of dosing (e.g., single vs. repeated), and (4) route of exposure
Toxicokinetic (TK)	TK studies follow the change in concentration of parent substances and/or metabolite(s) with time in blood/plasma or other tissues of interest.	TK studies are designed to determine the effect of species, sex, and age of animals, dose level and frequency of dosing, and route of exposure on TK parameters and include an intravenous administration to determine the bioavailability following dosing via the route of interest
Developmental and reproductive toxicity	The prenatal developmental toxicity study is undertaken to recognize substances that may pose a risk to the developing fetus if pregnant women are exposed. Regulatory agencies use the results of well-conducted animal studies to help set human exposure guidelines. The experimental protocols are outlined by EPA Health Effects Test Guidelines OPPTS 870.3700 and ICH S5(R2)	Embryo-fetal developmental study, teratology study, or segment II study and testing regimes for evaluating the potentially toxic effects of exposure to environmental and occupational substances on the reproductive system
Genetic toxicology	Testing methods to evaluate the potential substances to damage DNA. These studies involve both in vivo (laboratory animals, human subjects) and in vitro (cells in culture) testing	Study: mutagenicity and cytogenetics Species: In vivo (rats, mice), <i>Salmonella typhimurium</i> , <i>Escherichia coli</i>

(continued)

Table 8
(continued)

Type	Description	Toxicity test
Immunotoxicity	NTP evaluates the potentially toxic effects of exposure to substances on the immune system which include: Food additives Natural products such as mycotoxins Pharmaceutical, agrochemicals, chemical, or consumer product industries	Immunotoxicity tests are designed to evaluate immune function and hypersensitivity. These tests are carried out using rodent models, cultured mammalian cells, and other <i>in vitro</i> methods
Neurotoxicity	Test to identify environmental chemicals which are responsible for neurodevelopmental (e.g., autism spectrum disorders) and neurodegenerative (e.g., Parkinson's, Alzheimer's) disorders. NTP has developed the Developmental NeuroToxicity Data Integration and Visualization Enabling Resource (DNT-DIVER) tool to help compare and visualize results across assays	In vivo neurotoxicity testing in rodents is conducted including the assessment of multiple components following developmental or adult exposures. Primary neurotoxicity assessments include: Clinical observations Developmental landmarks Motor activity Startle response Learning and memory Neurohistopathology
Toxicogenomics	Included testing program are microarrays, next-generation sequencing (NGS), proteomics, and metabolomics. Considered biological samples may be from animals or from cell culture studies. Study proposals are reviewed by the Toxicogenomics Faculty at NTP	Five-day rodent toxicogenomic studies are performed in rodents. Toxicogenomics investigates how the genome responds to environmental chemicals which can change the expression of genes, proteins, and metabolites in living cells. Determining genome-wide changes in affected tissues is beneficial for finding markers of toxicity or disease and for understanding how genetic variation among individuals can influence sensitivity to a substance

Databases currently participating in eChemPortal are the following:

- *ACToR*: US EPA Aggregated Computational Toxicology Resource
- *AGRITOX*: Base de données sur les substances actives phytopharmaceutiques
- *APVMA-CR*: The Australian Pesticides and Veterinary Medicines Authority database
- *CCR*: Canadian Categorization Results
- *CESAR*: Canada's Existing Substances Assessment Repository



Fig. 16 Screenshot of OECD database

- *Combined Exposures*: Collection of case studies on risk assessments of combined exposures to multiple chemicals
- *ECHA C&L inventory*: Public Classification and Labelling (C&L) Inventory according to the European Union (EU) CLP Regulation (EC) No 1272/2008
- *ECHA CHEM*: European Chemicals Agency's Dissemination portal with information on chemical substances registered under REACH
- *EFSA Open Food Tox*: Chemical Hazards Database of the European Food Safety Authority
- *EnviChem*: Data Bank of Environmental Properties of Chemicals
- *EPA HHBP*: EPA Human Health Benchmarks for Pesticides
- *EPA OPPALB*: EPA Office of Pesticide Programs' Aquatic Life Benchmarks
- *GDL*: Gefahrstoffdatenbank der Länder (Germany)
- *GHS-J*: GHS Classification Results by the Japanese Government
- *GSBL*: Joint substance data pool of the German Federal Government and the German Federal States
- *HPVIS*: High Production Volume Information System
- *HSDB*: Hazardous Substances Data Bank
- *HSNO CCID*: New Zealand Hazardous Substances and New Organisms Chemical Classification Information Database
- *IGS*: IGS-Public Informationssystem für gefährliche Stoffe (Germany)

- *INCHEM*: Chemical Safety Information from Intergovernmental Organizations
- *INERIS-PSC*: INERIS-Portail Substances Chimiques
- *IPCHEM*: Information Platform for Chemical Monitoring
- *J-CHECK*: Japan CHEmicals Collaborative Knowledge database
- *JECDB*: Japan Existing Chemical Data Base
- *NICNAS IMAP*: Australia's National Industrial Chemicals Notification and Assessment Scheme's (NICNAS) Inventory Multi-tiered Assessment and Prioritisation (IMAP) framework
- *NICNAS Other*: NICNAS assessments of existing chemicals other than Priority Existing Chemical assessments
- *NICNAS PEC*: NICNAS Priority Existing Chemical (PEC) Assessment Reports
- *OECD HPV*: OECD Existing Chemicals Database
- *OECD SIDS IUCLID*: OECD Existing Chemicals Screening Information Data Sets (SIDS) Database
- *SIDS UNEP*: OECD Initial Assessment Reports for HPV Chemicals including SIDS as maintained by United Nations Environment Programme (UNEP) Chemicals
- *SPIN*: Substances in Preparations In the Nordic countries
- *UK CCRMP Outputs*: UK Coordinated Chemicals Risk Management Programme Publications
- *US EPA IRIS*: The United States Environmental Protection Agency Integrated Risk Information System
- *US EPA SRS*: United States Environmental Protection Agency Substance Registry Services

Web Accessibility: <http://www.oecd.org/chemicalsafety/risk-assessment/echemportalglobalsportaltoinformationonchemicalsubstances.htm>

4.28 Optimized Strategies for Risk Assessment of Industrial Chemicals Through Integration of Non-test and Test Information (OSIRIS)

The OSIRIS database (Fig. 17) compiled aquatic toxicity data along with mutagenicity, carcinogenicity, and repeat dose toxicity data [62]. The database is a form of the report developed from a workshop under OSIRIS funded with the EU Commission within the Sixth Framework Programme held in Liverpool, UK. It deals with the potential of mode of action (MoA) information derived from non-testing and screening methodologies to support the risk hazard assessment. The aim of the OSIRIS project was to construct integrated testing strategies (ITS) suitable for the implication in the REACH system which would permit a substantial intensification in the use of non-testing information for decision-making of regulator authorities followed by minimizing animal testing. The ITS had



Fig. 17 Screenshot of OSIRIS database

been connected to a decision theory framework with alternate strategies like in vitro analysis, in vivo information on analogues, chemical and biological read-across, classification and regression-based QSAR models, and exposure-based waiving followed by thresholds of toxicological concern. OSIRIS accounts for more intelligent and compound-tailored approach rather than going for wide-ranging typical testing approaches.

The OSIRIS projects are organized in the five interlinked research pillars, and they are the following:

- Chemical domain
- Biological domain
- Exposure
- Integration strategies and tools
- Case studies

The major research of OSIRIS is directed to the following ITS:

- Skin sensitization
- Repeated dose toxicity
- Mutagenicity and carcinogenicity
- Bioconcentration factor
- Aquatic toxicity

Web Accessibility: www.osiris.ufz.de

4.29 Risk Assessment Information System (RAIS)

The RAIS, a web-based system (Fig. 18), deals with chemical-specific toxicity values useful in providing information for risk assessment [63]. Using structure-searchable and executable options, the RAIS provides necessary tools which can be tactfully employed to meet site-specific requirements. RAIS comprises of a series of chemical tools for the prediction of chemical toxicity and

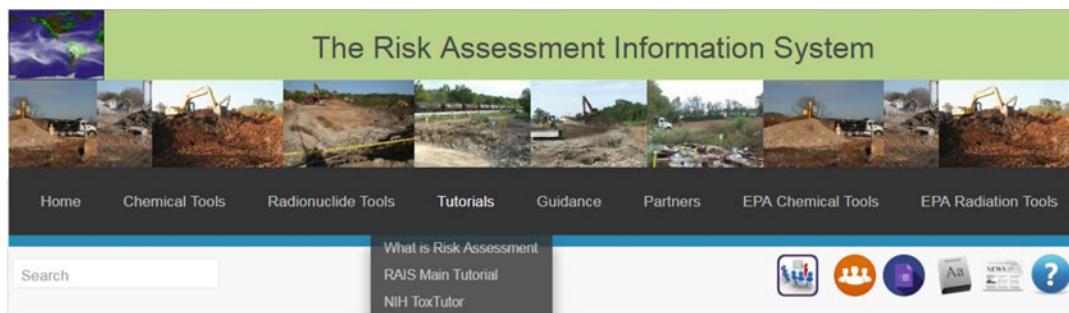


Fig. 18 Screenshot of RAIS database

parameters as well as seven tutorials which are designed to help the user in understanding and utilizing available RAIS tools to the risk assessment. The RAIS consists of seven modules under it, and they are the following:

- Module 1: Introduction and Content Map
- Module 2: Problem Identification
- Module 3: Designing Conceptual Site Models
- Module 4: Select COPCs
- Module 5: Toxicity Assessment
- Module 6: Risk Calculation
- Module 7: Documentation

The RAIS tutorial will assist in:

- Generating a fundamental site model
- Selection and categorization of probable toxic chemical of concern,
- Screening measures of toxicity
- Risk/hazards calculation species wise and environmental compartment wise
- Pull out information for a toxicity/risk assessment from all obtained toxicity data
- Documentation of the risk assessment report and management protocols

The RAIS is maintained by the US Department of Energy (DOE), Office of Environmental Management, and Oak Ridge Operations (ORO) Office through a contract between Bechtel Jacobs Company LLC and the University of Tennessee.

Web Accessibility: <http://rais.ornl.gov/>

The RiskIE is an open-access database (Fig. 19) which contains statements about an assortment of human health risks assessment



Fig. 19 Screenshot of RiskIE database

4.30 Risk Information Exchange (RiskIE)

projects like white papers, training module, and risk documents for both chemicals and non-chemicals [56]. RiskIE was formed in 2007 by TERA to notify about in-progress human health risk assessment work which tracks around 4000 risk assessment projects monitored by 35 organizations on behalf of 13 countries. RiskIE is available along with ITER under TERA project. The RiskIE delivers risk evaluators vital tools for straightforwardly categorizing and associating available risk data, for allocation in advancement assessments, and for enhancing interaction among risk assessment groups to lessening duplication of effort and to harmonize risk assessment procedures across organizations. RiskIE can also serve to bridge the communication gap among industry, government, environmental stakeholders, and academic (see Table 9).

Web Accessibility: <https://www.tera.org/Alliance%20for%20Risk/RiskIE.htm>

4.31 Registry of Industrial Toxicology Animal-Data (RITA)

The RITA (Fig. 20) represents a unique international cooperation between pharmaceutical and chemical industries and a nonprofit organization maintained by the Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM) Hannover for comparing and interpreting rodent carcinogenicity studies and tumor data [64]. In 2006 The Societies of Toxicologic Pathology from North America (STP), Japan (JSTP), and Europe (ESTP/BSTP) united in a common effort to review the terms and principles for a standard nomenclature of lesions in rodents and the International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice (INHAND) initiative. In this perspective, the RITA is helping scientifically and providing the tools for the online review collection of manuscripts and images. The RITA is gathering the data of control animals employed on rodent carcinogenicity studies from diverse laboratories in a constant manner. The accessibility of harmonized data from those studies in an inclusive database helps for the decisive interpretation of data. Most importantly, the cross-organizational review helps in optimizing the standards for robustness, reliability, and quality of the data. RITA is helping the standardization for the conduct and histopathological assessment of carcinogenicity studies (Table 10) which are mainly

Table 9
Organizations and projects included in RiskIE

Organization	Projects
Advisory Committee on Existing Chemicals (BUA) associated with the German Chemical Society	Health and environmental hazard assessment, Testing for health effects
American Conference of Industrial Hygienists (ACGIH)	Biological Exposure Indices (BEI), Identification of threshold limit value (TLV)
American Industrial Hygiene Association (AIHA)	Workplace Environmental Exposure Limit (WEEL), Emergency Response Planning Guideline (ERPG)
ATSDR	Assessment of minimal risk level (MRL) and toxicological profile (TP) of environmental hazards
California Environmental Protection Agency (CalEPA)	Chemical risk assessments, Public Health Goals (PHGs), Toxic Air Contaminants
Department of the Environment (UK): Environmental Hazard Assessment	Comprehensive health and environmental hazard estimation which employs a combination of inclusive exposure and effects data to reach conclusions about safety data of chemicals
Environment Canada	Canadian Soil Quality Guidelines, Canadian Water Quality Guidelines
European Union (EU)—European Chemicals Bureau (ECB)	Online European Risk Assessment Tracking System (ORATS), Occupational Exposure Limit (OEL)
Health Canada	Priority Substances List Assessment Reports, Risk Reduction Activities, Domestic Substance List (DSL) chemical assessments
International Programme for Chemical Safety (IPCS)	Concise International Chemical Assessment Documents (CICADS): Health and environmental hazard evaluation
Ministry of Health, Labour and Welfare (Japan)	Short-term testing for health effects, Long-term testing for health effects, Report of occupational health research
National Environmental Research Institute (Denmark)	Monitoring of environmental levels of chemicals or effects, including development
National Industrial Chemicals Notification and Assessment Scheme (Australia)	Initial health and/or environmental hazard evaluation which use a combination of data regarding toxic effects and limited exposure to reach an initial hazard assessment about a specific chemical
Organization for Economic Co-operation and Development (OECD)	SIDS project for OECD High Production Volume (HPV) chemicals, Initial health and/or environmental hazard evaluation which use a combination of data regarding toxic effects and limited exposure to reach an initial hazard assessment about a specific chemical

(continued)

Table 9
(continued)

Organization	Projects
Occupational Safety and Health Administration (OSHA)	Permissible Exposure Limit (PEL)
US EPA	Acute Exposure Guideline Limits (AEGLs), Toxicological reviews, Pesticide Reregistration (REDs and TREDs), IRIS cancer and noncancer risk values, IRIS Acute Exposure Duration, IRIS cancer effects, IRIS noncancer effects, IRIS chronic and less than lifetime exposure durations (CLLED)

**Fig. 20** Screenshot of RITA database**Table 10**
Current status of RITA (up to February 2019) [website reference]

Subject	Hamsters	Mice	Rats
Number of studies	5	123	191
Number of animals	500	10,882	19,773
Number of primary tumors	909	9719	30,395
Number of pre-neoplastic lesions	1636	5674	34,436
Total number of cases (including metastases)	2921	39,674	74,635

requested by regulatory authorities for the human risk assessment of pharmaceuticals or chemicals.

Web Accessibility: <https://reni.item.fraunhofer.de/reni/public/rita/>

4.32 Toxicology Testing in the 21st Century (Tox21)

The Tox21 is a collaboration program (Fig. 21) among different federal agencies of the United States like US EPA, NTP, NIEHS, US FDA, and National Center for Advancing Translational Sciences (NCATS) formed in the year 2008 [65]. Tox21 focused on creating methods to swiftly and efficiently assess the toxicity of



Environmental Topics

Laws & Regulations

About EPA

Search EPA.gov

Related Topics: [Safer Chemicals Research](#)

CONTACT US

SHARE



Toxicology Testing in the 21st Century (Tox21)

Toxicology in the 21st Century (Tox21) is a federal collaboration among EPA, NIH, including National Center for Advancing Translational Sciences and the National Toxicology Program at the National Institute of Environmental Health Sciences, and the Food and Drug Administration. Tox21 researchers aim to develop better toxicity assessment methods to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to negative health effects. One of EPA's contributions to Tox21 are the chemical screening results from the Toxicity Forecaster (ToxCast). [Learn more about the mission and goals of the Tox21 program.](#)



Fig. 21 Screenshot of Tox21 database

chemicals, pharmaceuticals, agrochemicals, food additives, and so on in the twenty-first century. The multi-collaborative research team has created and validated in vitro cell-based assays employing quantitative high-throughput screening. The Tox21 is currently screening over 10,000 chemicals and screened more than 70 assays. The aims of Tox21 are:

1. Classify mechanisms of chemically induced biological activity
2. Prioritize chemicals for more wide-ranging testing
3. Generate more pertinent and predictive models of in vivo toxicological responses

Tox21 data is publicly available through the EPA's Computational Toxicology Dashboard, the National Library of Medicine's PubChem, and NTP's Chemical Effects in Biological Systems. Detailed assay annotations, protocols, and performance statistics are publicly available on the EPA's Computational Toxicology website (www.epa.gov/comptox) and the NIH tripod website (<https://tripod.nih.gov/tox21>).

Web Accessibility: <http://www.epa.gov/ncct/Tox21/>

4.33 ToxCast

The ToxCast is a research program (Fig. 22) introduced within US EPA to advance the capability to predict toxicity data through bioactivity profiling (physical-chemical properties, biochemical properties based on throughput assays (HTS), genomic and metabolomic analyses of cells, cell-based phenotypic assays, and predicted biological responses from existing QSAR models), characterizing toxicity pathways followed by prioritizing chemicals for screening and testing to assist EPA programs for risk

The ToxCast Dashboard helps users examine high-throughput assay data to inform chemical safety decisions. To date, the ToxCast Dashboard has data on over 9,000 chemicals and information from more than 1,000 high-throughput assay endpoint components. Users of the ToxCast Dashboard can explore the data from a chemical or an assay viewpoint. Once the user selects the chemicals and assays of interest, they can then explore the biological activity for the chemical-assay combinations. Results from the selections are shown with tables, graphs and charts that can be downloaded by the user.



Fig. 22 Screenshot of the ToxCast database

Table 11
List of available ToxCast HTS assay

<i>Biochemical assays</i>	
Protein families	GPCR, NR, kinase, phosphatase, protease, other enzymes, ion channel, transporter
Assay formats	Co-activator recruitment, radioligand binding, enzyme activity
<i>Cellular assays</i>	
Primary cells	Human endothelial cells, human monocytes, human keratinocytes, human fibroblasts, human proximal tubule kidney cells, human small airway epithelial cells, rat hepatocytes, mouse embryonic stem cells (Sid Hunter)
Cell lines	HepG2 human hepatoblastoma, A549 human lung carcinoma, HEK 293 human embryonic kidney
Biotransformation competent cells	Primary human hepatocytes, primary rat hepatocytes
Assay formats	Cytotoxicity, reporter gene, gene expression, biomarker production, high-content imaging for cellular phenotype

management and regulation of CECs related to environment [66]. The data are generated with the collaboration of EPA, NTP, and the National Institutes of Health Chemical Genomics Center. At the present time, ToxCast has already evaluated over 2000 chemicals within over 700 HTS including 300 signalling pathways (Table 11).

Major goals of ToxCast are the following:

- Identify toxicity pathways
- Obtain HTS assays for pathways
- Screen a large chemical library
- Initially link HTS results to adverse effects—toxicity signatures
- Ultimately identify points of departure from HTS data—toxicity pathways

The most recent ToxCast data is available in the invitroDBv3.1 database (https://epa.figshare.com/articles/ToxCast_Database_invitroDB_/_6062623/2).

Web Accessibility: <https://www.epa.gov/chemical-research/toxcast-dashboard>

4.34 TOXMAP

TOXMAP combines an interactive and searchable US maps (Fig. 23) of EPA Toxics Release Inventory (TRI) and Superfund data. The TOXMAP overlays the US Census demographic data, income figures from the US Department of Commerce, and health data from the NCI SEER program [67]. The current version provides improved usability on mobile devices compared to previous versions.

The new TOXMAP has multiple updated datasets, and they are:

- NCI SEER cancer and disease mortality data (2011–2015)
- Canadian National Pollutant Release Inventory (NPRI) data (2016)
- Coal power plant data from the EPA Clean Air Markets Program (2017)
- US commercial nuclear power plants (2017)

Web Accessibility: <https://toxmap.nlm.nih.gov/toxmap/>

4.35 Toxicology Data Network (TOXNET)

The TOXNET represents a group of databases (see Fig. 24 and Table 12) under US National Library of Medicine of NIH that offers information related to chemicals and drug, environmental health, occupational safety, risk assessment and regulations, and toxicology. It is an open-access database and can be used by academicians, industries, as well as regulatory agencies [68].

Specialty databases under TOXNET can be accessed from <http://sis.nlm.nih.gov/enviro.html>:

- Dietary Supplement Label Database: Ingredients in supplements sold in the United States
- Pillbox: Rapid pill identification
- Drug Information Portal: Gateway to current and accurate drug information

Web Accessibility: <http://toxnet.nlm.nih.gov/>

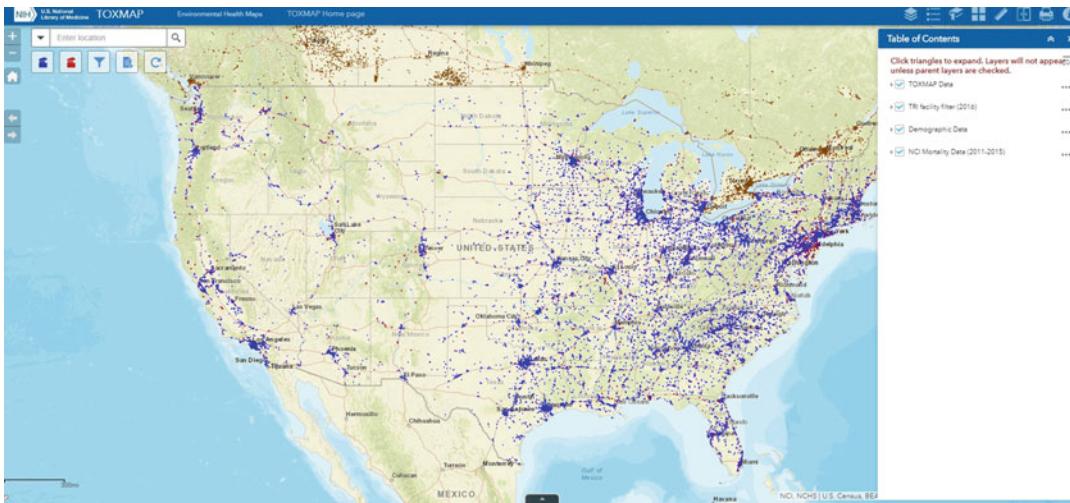


Fig. 23 Screenshot of TOXMAP database



Fig. 24 Screenshot of TOXNET database

4.36 Toxicity Reference Database (ToxRefDB)

The ToxRefDB captures information of in vivo toxicity study results including acute, (sub-)chronic, developmental, and reproductive endpoints for 474 chemicals [69]. The database offers complete study design, dosing, and observed treatment-related effects using standardized data. It enables connections with other public databases like ACToR and ToxCast databases which also provides detailed chemical toxicity data, public hazard, exposure,

Table 12
The major databases under TOXNET

Database name	Topics covered
ChemIDplus	Chemical names, formulas, structures
CCCRIS	Carcinogenicity, mutagenicity
CPDB	Cancer testing
GENE-TOX	Mutagenicity test data
IRIS	Human health risk assessment
ITER	Risk information
TOXLINE	Toxicology journal literature
DART	Reproductive toxicology journal literature
Haz-Map	Occupational health
Household Products Database	Products used in and around the home
HSDB	Health effects, toxicity, regulations
LactMed	Drugs and breastfeeding
TRI	Environmental releases of chemicals
TOXMAP	Interactive US maps of chemical releases

and risk resources. It consists of over 30 years and \$2 billion of animal testing results.

Web Accessibility: <http://www.epa.gov/comptox/toxrefdb/>

4.37 Toxic Substances Control Act Test Submissions (TSCATS)

The TSCATS is an online database of chemical testing results related to adverse effects of chemicals on health and ecological systems formed by the US Department of Commerce National Technical Information Service Alexandria, Virginia. The collection currently exceeds 29,000 titles of studies that are submitted to the US EPA by the US industry under several sections of the Toxic Substances Control Act (TSCA) [70]. The present version of the database is identified as TSCATS 2.0.

Web Accessibility: <https://catalog.data.gov/dataset/toxic-substances-control-act-test-submissions-2-0-tscats-2-0/resource/fbe133b5-d0bd-4c2c-a290-fd4deec4a5b9>

4.38 US FDA Chemical Evaluation and Risk Estimation System (CERES)

The US FDA CERES is a centralized data storage and management system (Fig. 25) that offers support in decision-making of pre- and post-market safety assessment for food-related ingredients and food contact substances [71]. The CERES is formed to address the technical challenges in food ingredient evaluation processes in the Office of Food Additive Safety (OFAS) by consolidating all data into one place and designed to be a knowledgebase of chemicals



Fig. 25 Screenshot of USFDA CERES database

regulated by the Center for Food Safety and Applied Nutrition (CFSAN). CERES is equipped with cheminformatics capabilities like:

- Predictive models (through collaborations with Altamira LLC and Molecular Networks GmbH)
- Cleft palate
- Bacterial reverse mutagenicity
- In vitro chromosome aberration
- In vivo micronucleus
- Mouse and rat tumor
- Skin sensitization hazard and potency
- Chemical structure similarity calculation

Web Accessibility: https://www.accessdata.fda.gov/scripts/fdatrack/view/track_project.cfm?program=cfsan&id=CFSAN-OFAS-Chemical-Evaluation-and-Risk-Estimation-System

4.39 VITIC

VITIC, a chemical structure-searchable database (Fig. 26), contains information about toxicological endpoints including mutagenicity, carcinogenicity, and hepatotoxicity from high-quality peer-reviewed sources. VITIC is capable of supporting the skin sensitization assessment of chemicals without animal testing. VITIC is developed by Lhasa Limited [72]. It contains updated and extensive coverage of chemical class and 430,000 toxicity data records for over 20,500 structures. The offered data help submissions to regulators required under ICH M7. Few of the major characteristics of the database are listed below:

- Assessment of the possible toxicity of new as well as existing chemicals through the implication of large chemical libraries.

The screenshot shows the Lhasa Limited homepage with a navigation bar at the top. The main content area displays the Vitic database product page, featuring a sidebar with product links like Derek Nexus, Meteor Nexus, Mirabilis, and Sarah Nexus. The main content area includes a brief description of Vitic as a structure searchable database and its benefits, followed by a large green Vitic logo.

Fig. 26 Screenshot of Vitic database

The screenshot shows the WikiPharma homepage with a navigation bar and a main content area. The main content area features a heading 'WELCOME TO THE WIKIPHARMA DATABASE!', a paragraph about the database's purpose, and a photograph of server racks. The URL <http://www.lhasalimited.org/products/vitic-nexus.htm> is provided below the image.

Fig. 27 Screenshot of WikiPharma database

- Decision-making regarding toxicity profile and approval of industrial chemicals considering ecotoxicity hazards when a complete experimental profile of the compound is absent.
- To save time, money and animal sacrifice to carry out toxicity experiments for a large number of chemicals.
- Sharing of toxicity data and knowledge.
- Modification of chemical structure by considering significant structural fragment or template responsible for toxicity employing in silico approaches and experimental knowledge.

Web Accessibility: <http://www.lhasalimited.org/products/vitic-nexus.htm>

4.40 WikiPharma

The WikiPharma, an open-access database (Fig. 27), covers eco-toxicity data for human pharmaceuticals available on the Swedish market. The database is updated continuously over a period to help researchers, industries, risk assessors, and regulators authorities worldwide [73]. The WikiPharma database is developed within the Swedish research program MistraPharma (www.mistrapharma.se) that carries basic data for around 831 active pharmaceutical ingredients (APIs) representing 35 different therapeutic classes. The effect data have been evaluated and counted in for

116 pharmaceuticals, and ecotoxicity test data have been extracted from 156 different sources. The MistraPharma functioned to classify human pharmaceuticals which are possible hazards to aquatic ecosystems and addressed the antibiotic resistance risk promotion in the environment along with the risk management strategies and wastewater treatment technologies. The operation of this research program occurred in the time frame of 2008–2015.

Web Accessibility: www.wikipharma.org

5 Application of Ecotoxicity Databases

Ecotoxicity databases consist of experimental data on toxic effects of a series of chemicals to diverse species living in the different environmental compartments. The most common application and importance of these databases are the following [8–10, 74, 75]:

- Qualitative and/or quantitative toxicity data along with experimental protocols, test species, and toxicity endpoints.
- The ecotoxicity databases are majorly employed for computational modeling purpose, for the future prediction of toxicity of new and/or untested chemicals.
- Although adverse drug reactions (ADRs) are mainly checked for drug discovery and market approval, in the present situation, many industries follow ecotoxicity protocol of new drug entity considering USEPA and REACH recommendations.
- Databases are frequently utilized for toxicity screening, ERA and ERM, safety evaluation, and regulatory decision-making along with the preparation of regulatory guidelines.
- Implementation of 3Rs principles for the reduction, replacement, and refinement of animal usage in experimental toxicity testing computational modeling is very much important, and a good database is the fundamental resource for acceptable and predictive model generation.
- Toxicity data gaps filling is another important role of the ecotoxicity database along with extrapolation of toxicity data employing interspecies experimental data available from inter-species computational models like i-QSAR or QTTR (quantitative toxicity-toxicity relationship).

6 Future Avenues and Conclusion

The available databases are the rich sources of information related to experimental protocols, assay techniques, and employed species for ecotoxicity testing. An access to comprehensive databases is significant for RA and RM, regulatory decision-making followed

by future toxicity prediction of new and untested chemicals employing the present data exist in a database. A process from the approval to rejection of a compound is highly dependent on a specific database. Thus, the importance of ecotoxicity databases is indispensable. Considering the present aspect, the existing databases need to be updated for future perspectives as many of them are lacking some fundamental aspects of ecotoxicity assessment as well as a good modeling source.

- (a) The ecotoxicity databases need to include experimental toxicity data of mixtures. As the majority of chemicals are identified as a mixture in the environment, thus to evaluate the hazards and risk associated with specific chemicals, one needs to additionally consider the mixture effect [76].
- (b) Most of the time, risk assessment is evaluated for the parent chemical, while a chemical undergoes multiple transformations into metabolites or TPs in diverse environmental compartments. In some cases, the metabolites are more toxic than the parent compound (e.g., prodrug concept). Therefore, complete life cycle and fate of specific chemical need to be studied along with toxicity testing of a parent as well as its major metabolites. The complete toxicity data need to be included in the databases for getting the complete scenario of toxicity for any chemical [77].
- (c) For the much faster and efficient prediction of ecotoxicity of new and/or untested chemicals, knowledge-based expert systems (KBES) need to be tied up with specific databases so that they can be competently employed for HTS. Additionally, databases for specific ecotoxicity need to be included in artificial intelligence techniques and machine learning tools for detailed and wide-ranging toxicity assessments.

Abovementioned implementation in future databases is necessary to make the scenario most competitive and predictive one in terms of risk profiling and toxicity prediction of any new chemicals in no time and economic way. We hope that the current chapter will be useful in evaluating existing databases and developing new ones that would be more efficient in holistic approach to toxicity of chemicals and their risk toward environment and leaving species.

Acknowledgments

S.G. wants to thank NSF-DMR-PREM, Grant#1826886 for financial assistance. S.K. and J.L. are thankful to the National Science Foundation (NSF/CREST HRD-1547754 and NSF/RISE HRD-1547836) for financial support.

Glossary

CEC	Contaminants of emerging concerns
HTS	High-throughput screening
ID	Illicit drug
MLR	Multiple linear regression
MOA	Mechanism of action
NCI	National Cancer Institute
NOEC	No observed effect concentration
NTP	National Toxicology Program
OECD	Organization for Economic Co-operation and Development
PCP	Personal care product
QSAR	Quantitative structure-activity relationship
RBFNN	Radial basis function neural networks
STP	Sewage treatment plant
SVM	Support vector machine
TP	Transformation product
USEPA	United States Environmental Protection Agency
WTP	Water treatment plants

References

- Wilkinson JL, Hooda PS, Barker J, Barton S, Swinden J (2016) Ecotoxic pharmaceuticals, personal care products, and other emerging contaminants: a review of environmental, receptor-mediated, developmental, and epigenetic toxicity with discussion of proposed toxicity to humans. *Crit Rev Environ Sci Technol* 46:336–381
- Evgenidou EN, Konstantinou IK, Lambropoulou DA (2015) Occurrence and removal of transformation products of PPCPs and illicit drugs in wastewaters: a review. *Sci Total Environ* 505:905–926
- Montesdeoca-Esponda S, Checchini L, Bubba MD, Sosa-Ferrera Z, Santana-Rodriguez JJ (2018) Analytical approaches for the determination of personal care products and evaluation of their occurrence in marine organisms. *Sci Total Environ* 633:405–425
- Cassani S, Gramatica P (2015) Identification of potential PBT behavior of personal care products by structural approaches. *Sustain Chem Pharm* 1:19–27
- Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, San Diego
- Roy K, Kar S, Das RN (2015) A primer on QSAR/QSPR modeling: fundamental concepts (SpringerBriefs in Molecular Science). Springer, New York
- Dearden JC (2016) The history and development of quantitative structure-activity relationships (QSARs). *IJQSPR* 1:1–44
- Kar S, Roy K (2012) Risk assessment for ecotoxicity of pharmaceuticals – an emerging issue. *Expert Opin Drug Saf* 11:235–274
- Roy K, Kar S (2016) In silico models for ecotoxicity of pharmaceuticals. In: Benfenati E (ed) In silico methods for predicting drug toxicity, methods in molecular biology, vol 1425. Springer, New York, pp 237–304
- Kar S, Roy K, Leszczynski J (2018) Impact of pharmaceuticals on the environment: risk assessment using QSAR modeling approach. In: Nicolotti E (ed) Computational toxicology. Springer, New York, pp 395–443
- Fent K, Weston AA, Caminda D (2006) Ecotoxicology of pharmaceuticals. *Aquat Toxicol* 76:122–159
- Peake BM, Braund R, Tong AYC, Tremblay LA (2016) Impact of pharmaceuticals on the environment. In: The life-cycle of pharmaceuticals in the environment. Woodhead Publishing, Amsterdam, pp 109–152
- Bebian MJ, Gonzalez-Rey M (2015) Ecotoxicological risk of personal care products and pharmaceuticals, chapter 16 in aquatic

- ecotoxicology. Academic Press, Amsterdam, pp 383–416
14. Santos LHMLM, Araújoa AN, Fachinia A, Pena A, Delerue-Matos C, Montenegro MC (2010) Ecotoxicological aspects related to the presence of pharmaceuticals in the aquatic environment. *J Hazard Mater* 175:45–95
 15. Park S, Choi K (2008) Hazard assessment of commonly used agricultural antibiotics on aquatic ecosystems. *Ecotoxicology* 17:526–538
 16. Cleuvers M (2003) Aquatic ecotoxicity of pharmaceuticals including the assessment of combination effects. *Toxicol Lett* 142:185–194
 17. OECD guidelines for the testing of chemicals, section 2. Access: https://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-2-effects-on-biotic-systems_20745761
 18. Fairbrother A, Hope B (2005) Terrestrial ecotoxicology. 2nd chapter NA. In: Wexler P (ed) Encyclopedia of toxicology. Elsevier Ireland Limited, Limerick, pp 138–142
 19. Yin L, Wang B, Yuan H, Deng S, Huang J, Wang Y, Yu G (2017) Pay special attention to the transformation products of PPCPs in environment. *Emerg Contam* 3:69–75
 20. Mackay D, Fraser A (2000) Bioaccumulation of persistent organic chemicals: mechanisms and models. *Environ Pollut* 110:375–391
 21. Arnot JA, Gobas FA (2006) A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environ Rev* 14:257–297
 22. Angelidaki I, Sanders W (2004) Assessment of the anaerobic biodegradability of macropollutants. *Rev Environ Sci Biotechnol* 3:117–129
 23. Tropsha A (2005) In: Oprea T (ed) Cheminformatics in drug discovery. Wiley-VCH, Weinheim
 24. Zheng S, Luo X, Chen G, Zhu W, Shen J, Chen K, Jiang H (2005) A new rapid and effective chemistry space filter in recognizing a druglike database. *J Chem Inf Comput Sci* 45:856
 25. Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ Health Perspect* 111:1391–1401
 26. Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73:1701–1707
 27. Xia B, Liu K, Gong Z, Zheng B, Zhang X, Fan B (2009) Rapid toxicity prediction of organic chemicals to *Chlorella vulgaris* using quantitative structure-activity relationships methods. *Ecotoxicol Environ Saf* 72:787–794
 28. Jalali-Heravi M, Kyani A (2008) Comparative structure-toxicity relationship study of substituted benzenes to *Tetrahymena pyriformis* using shuffling-adaptive neuro fuzzy inference system and artificial neural networks. *Chemosphere* 72:733–740
 29. Kar S, Roy K (2010) First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals. *Chemosphere* 81:738–747
 30. Önlü S, Saçan MT (2018) Toxicity of contaminants of emerging concern to *Dugesia japonica*: QSTR modeling and toxicity relationship with *Daphnia magna*. *J Hazard Mater* 351:20–28
 31. Khan K, Kar S, Sanderson H, Roy K, Leszczynski J (2018) Ecotoxicological modeling, ranking and prioritization of pharmaceuticals using QSTR and i-QSTTR approaches: application of 2D and fragment based descriptors. *Mol Inf* 37:1800078
 32. Sangion A, Gramatica P (2016) Hazard of pharmaceuticals for aquatic environment: prioritization by structural approaches and prediction of ecotoxicity. *Environ Int* 95:131–143
 33. Kar S, Sepúlveda MS, Roy K, Leszczynski J (2017) Endocrine-disrupting activity of per- and polyfluoroalkyl substances: exploring combined approaches of ligand and structure-based modeling. *Chemosphere* 184:514–523
 34. Kar S, Ghosh S, Leszczynski J (2018) Single or mixture halogenated chemicals? Risk assessment and developmental toxicity prediction on zebrafish embryos based on weighted descriptors approach. *Chemosphere* 210:588–596
 35. Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, Cathey T, Transue TR, Spencer R, Wolf M (2008) ACToR-aggregated computational toxicology resource. *Toxicol Appl Pharmacol* 233:7–13
 36. Singh AV, Knudsen KB, Knudsen TB (2005) Computational systems analysis of developmental toxicity: design, development and implementation of a Birth Defects Systems Manager (BDSM). *Reprod Toxicol* 19:421–439
 37. Fitzpatrick RB (2008) CPDB: carcinogenic potency database. *Med Ref Serv Q* 27:303–311
 38. United State National Library of Medicine, Toxicology Data Network (TOXNET), Chemical Carcinogenesis Research Information

- System (CCCRIS). <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCCRIS>. Accessed 15 May 2019
39. Wedebye EB, Dybdahl M, Reffstrup TK, Abildgaard Rosenberg S, Nikolov NG. New free Danish online (Q)SAR predictions database with >600,000 substances. Abstract from QSAR 2016 conference, Miami Beach
 40. Foster PM (2016) Influence of study design on developmental and reproductive toxicology study outcomes. *Toxicol Pathol* 45:107–113
 41. Solecki R, Heinrich V, Rauch M, Chahoud I, Grote K, Wölffel B, Buschmann J, Morawietz G, Kellner R, Lingk W (2010) The DevTox site: harmonized terminology and database. In: Charlene A (ed) McQueen, comprehensive toxicology, vol 12. Academic Press, Oxford, pp 339–346
 42. Richard AM, Williams CR (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res* 499:27–52
 43. ECOTOXicology Knowledgebase System User Guide-Version 5.0, 2018. <https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100UUBD.txt>. Accessed 15 May 2019
 44. Laamanen I, Verbeek J, Franco G, Lehtola M, Luotamo M (2008) Finding toxicological information: an approach for occupational health professionals. *J Occup Med Toxicol* 3:18
 45. Seyler LA. Extension toxicology network (EXTOXNET). Cornell University and Michigan State University, 1994. <http://extoxnet.orst.edu/index.htm>. Accessed 15 May 2019
 46. Briggs K, Barber C, Cases M, Marc P, Steger-Hartmann T (2014) Value of shared preclinical safety studies – the eTOX database. *Toxicol Rep* 2:210–221
 47. Tluczkiewicz I, Batke M, Kroese D, Buist H, Aldenberg T, Pauné E, Grimm H, Kühne R, Schüürmann G, Mangelsdorf I, Escher SE (2013) The OSIRIS Weight of Evidence approach: ITS for the endpoints repeated-dose toxicity (RepDose ITS). *Regul Toxicol Pharmacol* 67:157–169
 48. Jackson MA, Lea I, Rashid A, Peddada SD, Dunnick JK (2006) Genetic alterations in cancer knowledge system: analysis of gene mutations in mouse and human liver and lung tumors. *Toxicol Sci* 90:400–418
 49. Waters M, Stack H, Jackson M. Genetic Activity Profile (GAP) data base. U.S. Environmental Protection Agency, Washington, DC, EPA/600/D-91/049 (NTIS PB91177014)
 50. Cimino MC, Auletta AE (1993) Availability of the GENE-TOX database on the National Library of Medicine TOXNET system. *Mutat Res* 297:97–99
 51. HERA, Methodology document, 2002
 52. Sakuratani Y, Zhang HQ, Nishikawa S, Yamazaki K, Yamada T, Yamada J, Gerova K, Chankov G, Mekenyan O, Hayashi M (2013) Hazard Evaluation Support System (HESS) for predicting repeated dose toxicity using toxicological categories. *SAR QSAR Environ Res* 24 (5):351–363
 53. Fonger GC (1995) Hazardous substances data bank (HSDB) as a source of environmental fate information on chemicals. *Toxicology* 103:137–145
 54. Pearce N, Blair A, Vineis P et al (2015) IARC monographs: 40 years of evaluating carcinogenic hazards to humans. *Environ Health Perspect* 123:507–514
 55. Dourson ML (2018) Let the IRIS Bloom: regrowing the integrated risk information system (IRIS) of the U.S. Environmental Protection Agency. *Regul Toxicol Pharmacol* 97: A4–A5
 56. Wullenweber A, Kroner O, Kohrman M, Maier A, Dourson M, Rak A, Wexler P, Tomljanovic C (2008) Resources for global risk assessment: the International Toxicity Estimates for Risk (ITER) and Risk Information Exchange (RiskIE) databases. *Toxicol Appl Pharmacol* 233:45–53
 57. Matsumoto M, Kobayashi K, Takahashi M, Hirata-Koizumi M, Ono A, Hirose A (2015) Summary information of human health hazard assessment of existing chemical substances (I). Kokuritsu Iyakuhiin Shokuhin Eisei Kenkyusho Hokoku 133:42–47
 58. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE Jr (2000) LeadScope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 40:1302–1314
 59. Registry of Toxic Effects of Chemical Substances (RTECS) database produced by the National Institute of Occupational Safety and Health (NIOSH)
 60. Ring M, Eskofier BM (2015) Data mining in the U.S. National Toxicology Program (NTP) database reveals a potential bias regarding liver tumors in rodents irrespective of the test agent. *PLoS One* 10:e0116488
 61. Austin T, Denoyelle M, Chaudry A, Stradling S, Eadsforth C (2015) European Chemicals Agency dossier submissions as an experimental data source: refinement of a fish toxicity model for predicting acute LC50 values. *Environ Toxicol Chem* 34:369–378

62. Vonk JA, Benigni R, Hewitt M, Nendza M, Segner H, van de Meent D, Cronin MT (2009) The use of mechanisms and modes of toxic action in integrated testing strategies: the report and recommendations of a workshop held as part of the European Union OSIRIS Integrated Project. *Altern Lab Anim* 37:557–571
63. López-Roldán R, Rubalcaba A, Martín-Alonso J, González S, Martí V, Cortina JL (2016) Assessment of the water chemical quality improvement based on human health risk indexes: application to a drinking water treatment plant incorporating membrane technologies. *Sci Total Environ* 540:334–343
64. Nolte T, Rittinghausen S, Kellner R, Karbe E, Kittel B, Rinke M, Deschl U (2011) RITA—Registry of Industrial Toxicology Animal data: the application of historical control data for Leydig cell tumors in rats. *Exp Toxicol Pathol* 63:645–656
65. Thomas RT, Paules RS, Simeonov A, Fitzpatrick S, Crofton K, Casey W, Mendrick D (2018) The US federal Tox21 program: a strategic and operational plan for continued leadership. *ALTEX Altern Anim Exp* 35:163–168
66. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 95:5–12
67. Hochstein C, Szczur M (2006) TOXMAP: a GIS-based gateway to environmental health resources. *Med Ref Serv Q* 25:13–31
68. Wexler P (2001) TOXNET: an evolving web resource for toxicology and environmental health information. *Toxicology* 157:3–10
69. Plunkett LM, Kaplan AM, Becker RA (2015) Challenges in using the ToxRefDB as a resource for toxicity prediction modeling. *Regul Toxicol Pharmacol* 72:610–614
70. Toxic Substances Control Act (TSCA) test submissions database (TSCATS) – comprehensive update (on magnetic tape). Data file. United States
71. Hong H, Chen M, Ng HW, Tong W (2016) QSAR models at the US FDA/NCTR. *Methds Mol Biol* 1425:431–459
72. Elder DP, White A, Harvey J, Teasdale A, Williams R, Covey-Crump E (2015) Mutagenic impurities: precompetitive/competitive collaborative and data sharing initiatives. *Org Process Res Dev* 19:1486–1494
73. Molander L, Gerstrand M, Rudén C (2009) WikiPharma – a freely available, easily accessible, interactive and comprehensive database for environmental effect data for pharmaceuticals. *Regul Toxicol Pharmacol* 55:367–371
74. Cronin MTD (2017) (Q)SARs to predict environmental toxicities: current status and future needs. *Environ Sci: Processes Impacts* 19:213–220
75. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R et al (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010
76. Kar S, Leszczynski J (2019) Exploration of computational approaches to predict the toxicity of chemical mixtures. *Toxics* 7:15
77. Kar S, Leszczynski J (2017) Recent advances of computational modeling for predicting drug metabolism: a perspective. *Curr Drug Metab* 18:1106–1122



Chapter 30

VEGAHUB for Ecotoxicological QSAR Modeling

Emilio Benfenati and Anna Lombardo

Abstract

VEGAHUB is a freely available platform, which offers tens of QSAR models for many endpoints of environmental and ecotoxicological interest. In the last years, other tools have been added, for read across and prioritization. These tools can be used in an integrated way.

An interesting feature of VEGAHUB is the possibility to evaluate the reliability of the assessment, in particular for the QSAR models and for the software for prioritization.

Key words In silico models, QSAR, Read across, Prioritization, Screening, VEGAHUB

1 Introduction

There are several quantitative structure-activity relationships (QSAR) models, which have been developed in the last decades. We can group them within three families, regarding their availability: (1) commercial models, (2) free models available within the Internet, and (3) models published within journals but not easily available. VEGAHUB [1] belongs to the family of models implemented and made freely available through the Web. We did this choice in order to better disseminate the results of the activities on in silico models coordinated by the Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN), Milano, Italy.

In order to increase the use and application of the models for the assessment of the effects of chemical substances, VEGAHUB provides the results of the in silico models within a report. This report lists a number of issues useful to evaluate the reliability of the results, the reasons of possible uncertainty, and features, useful for reasoning on the factors involved in the ecotoxicological property of interest. All these elements have been introduced to reduce the barriers on the use of the in silico models, which may be unfamiliar to novel users.

VEGAHUB offers tools for the assessment of individual substances and for large sets of chemicals as well. Of course, the level of

details is different in these two cases, since the needs are different for these different applications. In case of the assessment of individual substances, the VEGAHub tools provide more details, while in case of screening of large sets of compounds, the user is typically interested in a system to filter the substances, to obtain subsets of chemical to be more carefully evaluated, for instance, the most toxic ones.

In order to increase the confidence and the reliability of the results, it is common that for the same endpoint, more than one model is available. This provides multiple lines of evidence and thus facilitates the decision on the property of interest. Indeed, the concordance between multiple values increases the confidence on the assessment; in case of conflicting results, this may be used to identify possible reasons for uncertainty, and thus, this will help the user in the final assessment.

The evolution of the VEGAHub tools followed the contributions of a quite numerous series of projects where IRFMN was active. However, VEGAHub benefits from contributions from many institutes, and the models it contains have been made available through the work done by many groups, as acknowledged at the VEGAHub web site [1].

One of the first projects that contributed to the models present in VEGAHub is DEMETRA. A book [2] describes in full details the results of this project. Within DEMETRA, five QSAR models for plant protection products (PPPs) have been developed:

- Rainbow trout
- Water flea
- Quail (both oral and dietary exposure)
- Honey bee

All the institutes and the PPPs companies of the consortium of the DEMETRA project worked to develop and check the models.

Of course, the PPPs need experimental data to be registered, but the main reason to develop these models was to apply them to related compounds, such as impurities, degradation products, metabolites, etc. There may be many of these substances for each parental PPP, and this requires a lot of work to generate the ecotoxicological data. Thus, the use of QSAR model may be of interest to generate data on substances related to the parental PPP.

Within the original proposal, we proposed not only to develop the QSAR models but also to make them freely available with a software system on the Internet. The reviewers of the proposal refused this hypothesis, because there were already programs to calculate the descriptors and it was judged not necessary to provide an easy way to get the predicted value. Thus, the funded DEMETRA project did not cover the implementation of the final models in an easy way. In our experience, in order to promote the use of the

QSAR models, it is much better to offer a single system. The use of multiple (usually complex) tools to calculate the molecular descriptors and then insert them into the algorithms at the basis of the models represents a barrier to the broad use of the models. Thus, even if DEMETRA was successful to generate models, its models were not widely used, and this shows the correctness of the strategy of VEGAHub. Later on, we implemented one of the DEMETRA models into VEGA.

It is interesting to notice that, even if DEMETRA models have been published in 2007, they are still quite advanced in their strategy, because they use the concept of hybrid models. In particular, each final model is based on more than one QSAR models, which are combined into the hybrid model. The combination is not through simple algebraic methods, but the output of the individual initial QSAR models is used as input for the final hybrid model [2].

What is important to know on the point of view of the user is that these models are dedicated to PPPs. The models are based on training sets that contain only PPPs. Thus, it may be expected that the models are “trained” to evaluate more difficult situations, like PPPs, which have chemical structures more complex than the typical “industrial” chemicals, and usually they show higher ecotoxicological values. Consequently, these models may be good for PPPs, probably for many biocides, but may overestimate the toxicity of simpler substances.

The project CAESAR [3] is an evolution of DEMETRA, since CAESAR made available the QSAR models requiring as input the structure of the chemical; the program calculates automatically the descriptors and uses them within the different QSAR models. This greatly facilitates the use of the models, and indeed CAESAR models have been and are still used. All the CAESAR models are now available within VEGAHub.

In order to further disseminate the results and use of the models, we also decided to have the description of the models published in open-access articles. Thus, the papers [4–9] describing in full details the models are also publicly available at the CAESAR web site [3].

CAESAR was on purpose dedicated to industrial chemical, and within this project, five models have been developed:

- Mutagenicity (Ames test)
- Carcinogenicity
- Developmental toxicity
- Skin sensitization
- Bioconcentration factor (BCF)

Thus, compared to DEMETRA, we developed here more models for toxicological properties, plus one for environmental

properties. Since these models have been developed for regulatory purposes, we gave particular attention to have conservative results. Therefore, in particular for certain endpoints where the number of chemicals in the training set was small, like in the case of skin sensitization, the user may have quite a number of false positives. This issue is reduced in the case of larger collections of values, like mutagenicity with Ames test, where many thousands of values were available.

Compared to DEMETRA, the training sets refer to “industrial” chemicals; thus these models may not be suitable for PPPs.

CAESAR also did a large improvement compared to DEMETRA, since it introduced the tool for the assessment of the reliability of the results, through an evaluation of the applicability domain. We will explain it in details later on. Anyhow, the use of the applicability domain tools should help informing the user that a certain model may be not adequate for a certain chemical.

It is important to know that in all cases, for the DEMETRA and CAESAR projects, and also for the following ones, we dedicated high attention to curate the values of the training sets, and for both projects, this activity required about 1 year. In particular, we checked the chemical orthography. In addition, we compared the experimental data from multiple sources, and in case of multiple values, we kept the substance only if the range of values was within a certain threshold. This also contributed to improve the quality of the results in prediction. This is also very useful when the user wants to use the software for read across, as we will describe later, since the quality of the experimental values is higher than in other cases.

As for DEMETRA, CAESAR took advantage of the work of the many partners in the consortium. Furthermore, we also collaborated with institutes out of the consortium. For instance, we built the model on developmental toxicity [4] with the US EPA, which implemented the model in the platform TEST [10].

Later on, a number of projects offered the opportunity to further evaluate the five models of CAESAR and to develop many others. We extended the perspective, to better address read across, and successively screening and prioritization. This prompted us to establish a new platform, VEGA-HUB, which merges all these modeling activities.

A series of other projects funded by the EC supported VEGA-HUB. These projects derive from projects funded by the Directorate-General (DG) for Research, DG Environment of the EC, and the European Food Safety Authority (EFSA). In addition, we acknowledge financial support from national authorities, such as the Italian Ministero della Sanità and Ministero dell’Ambiente e della Tutela del Territorio e del Mare, the German Umweltbundesamt (UBA), and the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety.

Within the different projects, we collaborated and often published with governmental agencies of many countries, in Europe and abroad, such as Portugal, France, the UK, Belgium, Austria, Germany, Italy, Denmark, the USA, Japan, and Canada. This is to document our attention to the regulatory applications and the possible use within legislative uses.

Furthermore, we thank the collaborations with industrial partners within the EC projects and the joint activities with research and academic groups. These last groups helped a lot in the development of the models. Today VEGAHub is in a certain way a community of groups willing to share and make available their models. The common interest is to reduce the (eco)toxicological impact of chemical substances. Finally, we also implemented some previously existing models, which are present within US EPA platforms, like EPI Suite [11], or from the Toxtree platform [12] of the Joint Research Center of the EC. We also acknowledge these valuable contributions.

We mention below some of the projects that contributed to VEGAHub.

ANTARES [13] addressed a number of existing QSAR models relevant for REACH, to identify which ones could be used at the beginning of REACH. Later one, other projects addressed REACH, in different years, covering different aspects: CALEIDOS [14], LIFE PROSIL [15] LIFE VERMEER [16], and LIFE CONCERT REACH [17]. The DG Environment funded ANTARES and the other projects. CALEIDOS checked the results of QSAR models compared with the experimental values of the substances registered in the first phase of REACH. Some difficulties appeared for the prediction of fish and daphnia acute toxicity. This prompted the development of more models, and this is one of the reasons why, now, there are several new models in VEGAHub developed for aquatic toxicity. Within CALEIDOS, we developed and implemented the software specific for read across called ToxRead [18].

VERMEER wants to integrate VEGAHub with models for exposure included in MERLIN-Expo; it also generated some new models, and, since it is ongoing, it will be addressed in the future development Subheading 4.2 below.

CONCERT REACH started very recently and will be mentioned below for the future development.

The German UBA and the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety funded three projects: PROMETHEUS, JANUS, and toDIVINE. The first two are related to screening and prioritization of chemicals based on the persistence, bioaccumulation, and toxicity (PBT) properties. In addition, JANUS includes other properties: carcinogenicity, mutagenicity, and reprotoxicity (CMR) and endocrine disruptors. The last one is on the integrated use of read across and QSAR models and is still ongoing; thus, it will be addressed

below for the future development. The first two refers to the possibility to identify certain groups of chemicals that deserve higher attention, or conversely, which may be less risky. Within these projects, we developed tools to integrate the results of different models, not only addressing the same property but also multiple properties. Indeed, the prioritization requires merging considerations related to different aspects, such as in the case of the screening for PBT. In this case, VEGAHub offers the possibility to run multiple models, covering different properties, in order to get the information of interest for PBT.

2 The Different Tools Available Within VEGAHub

VEGAHub provides different kinds of tools:

- QSAR models for specific endpoints
- A tool for the applicability domain measurement
- The read across system within VEGA
- The ToxRead tool for read across
- A system to integrate results from read across and QSAR models
- The tool for prioritization and screening
- A number of research tools for modeling, read across, etc.

2.1 The VEGA QSAR Models

The most classical tools are the QSAR models. In VEGA there are tens of different models, for many endpoints. These models are grouped within four families: models for physicochemical, environmental, ecotoxicological, and toxicological properties. Figure 1 shows some of the available models for the ecotoxicological endpoints.

It is common to find different models for the same endpoint. These models are based on different algorithms and approaches; thus they can offer orthogonal perspectives, which should be used within a weight-of-evidence strategy. EFSA in 2017 published a guidance document on weight of evidence [19]. Within this guidance, EFSA also made an example using VEGA and ToxRead. The strategy is to evaluate the relevance, reliability, and concordance of the different lines of evidence. Thus, it is quite useful to have multiple lines of evidence, and in the particular case, multiple in silico models. We also observe that not only EFSA but also ECHA and other authorities recommend using more than one model. This increases the confidence in the results.

However, the user should be aware of the difference that may occur between the different models. We already mentioned that VEGA offers models for PPPs or for industrial chemicals, for instance. Thus, the user should choose the model accordingly. Another example is that there are different models for fish acute



Fig. 1 The VEGA selection page, the ecotoxicological models

toxicity. They may be based on data related to different fish species; thus the results of a model based on experimental values obtained with rainbow trout may not be applicable to estimate the toxicity toward fathead minnow and vice versa.

In case of ecotoxicological properties, VEGA has models covering acute and chronic toxicity, for different trophic levels: fish, daphnia, and algae. There is also one model for honey bee.

The user should know that there are both models providing continuous values and other models that are classifiers. Both can be useful, and support each other, within a weight-of-evidence approach, as we said.

For the QSAR models, the user may introduce one single compound or collections of structures in batch. The most common way to enter the structures for VEGA is the simplified molecular-input line-entry system (SMILES) format, in order to facilitate the user. VEGA automatically checks if there are errors in the format and provides a warning. Furthermore, VEGA automatically standardizes the SMILES, in order to get more reproducible results. In this way, VEGA processes the same group always with the same format.

2.2 The Tool for Applicability Domain

In addition and integrated with the models for specific endpoints, VEGA automatically runs a tool to measure the applicability domain (AD). The results are provided as a quantitative value, with a score ranging from 1 to 0. If the value is high, the reliability is good. The threshold depends on the endpoint, because the different endpoints have different levels of uncertainty, mainly related to the experimental uncertainty and variability. Furthermore, if the number of substances in the training set is small, this also affects the overall reliability of the model.

The tool for AD is composed of several components, and the results of these components are then integrated to provide a single value, called Global AD Index (ADI). This ADI serves to evaluate the reliability of the results in prediction. The tool notifies which are the factors that deserve more attention in case there are issues.

The ADI components, some of them general other endpoint-specific, are:

- Similar molecules with known experimental values. It evaluates the presence of similar compounds in the training and test sets and allows the user to verify the performance in prediction through several aspects as explained below. The higher this parameter is, the higher the reliability of the prediction is. The availability of similar compounds means that the model “knows” the behavior of these similar compounds, and therefore it is more probable that its prediction is correct not only for them but also for the target. However, the absence of similar compounds in the training/test set does not automatically mean that the prediction is wrong. The meaning is that the user has to consider other lines of evidence (for instance, the results from other in silico models or for physicochemical parameters) to increase the confidence on the results. The similarity is not an absolute value of the substance. The similarity is always a comparison between two objects, and the result depends on which parameters are used for the comparison. There is no single way to measure the similarity between two chemicals. Multiple components compose the algorithm we use within VEGA, in order to balance different factors. We optimized it on four million compounds. In order to improve the dissemination and transparency, we published it in the open literature [20]. The algorithm for similarity works only on the chemical similarity, and it is the same for all models. We always recommend the user to visually evaluate the “similar” chemicals. In general, a similarity lower than 0.7 indicates a substance that is not similar. Another important point is how many similar compounds are present. VEGA calculates this parameter considering three or two most similar substances, depending on the model.
- Accuracy of prediction for similar molecules. It evaluates the agreement between the experimental and the predicted values

for the similar compounds. This strategy can be the best compromise to evaluate if the model is working for the chemical of interest. Of course, we assume that the experimental compound is unknown; thus we can only refer to something close to this substance. For this reason, we use the similar compounds to check if the specific VEGA model works for the target compound. The higher this parameter is, the higher the reliability of the prediction is. For the models that provide quantitative values, there is a further factor measured with the ADI tool, which is the maximum error in prediction among similar molecules. This indicates the largest error in prediction among the two or three (depending on the model) similar compounds. A large error reduces the ADI value; therefore in this case, the higher this parameter is, the lower the reliability of the prediction is.

- Concordance for similar molecules. It represents the concordance between the experimental values of the most similar compounds and the predicted value of the target compound. This parameter is different from the previous one. The previous one relates to the QSAR perspective: it accounts for the correctness of the predictions. The concordance factor relates to the read across perspective: it does not evaluate if the prediction is correct or not, but it evaluates if the related compounds have the same level of toxicity foreseen for the target one. This is close to the read across scheme. In case of conflicting value, the ADI is reduced. Therefore, the higher this parameter is, the higher the reliability of the prediction is.
- Model's descriptors range check. It verifies if the descriptors and the molecular weight of the target compound are within the range of the values of the chemicals of the training set. If the target is outside of this range, it means that the prediction could not be reliable; therefore the ADI is reduced. In this case, the output for this parameter is qualitative; true means that the target is in the model's descriptors range.
- Atom-centered fragments similarity check. VEGA verifies that there are fragments present in the target compound that are rare or not present at all in the training set. Since this may introduce uncertainty, an output of 1 means that no rare fragment is present.

Beyond the ADI components, VEGA also provides other pieces of information, depending on the property; for instance, for the bioconcentration factor (BCF), it verifies the logP values.

All these factors are expressed as quantitative values except the model's descriptors range check and are used to calculate the ADI score, which summarizes all these contributions. The strategy is to get a quantitative value, to keep into account of multiple factors,

and to evaluate more than one similar compound, weighting their contributions to the ADI based on their relative similarity. The most similar compounds will contribute more to the ADI value.

The ADI includes the evaluation of several parameters related from one side to the algorithm and the descriptors and to the other side to the structural similarity.

Thus, the VEGA tool is quite a complex one, compared to other programs, which typically only apply a series of statistical tools based on the structural information and associated descriptors; conversely, VEGA refers to the three conceptual components of any QSAR model: the property value, the chemical information, and the algorithm. Thus, for instance, the concordance and the accuracy of the predictions are pieces of information where the property value is also used and not only the chemical information.

For each chemical, in the summary page, there is a circle indicating the level of the effect, for instance, toxicity, with a color code indicating this (e.g., red toxic, orange moderately toxic, green not toxic). Then, there are up to three stars that represent the reliability of the prediction, which refers to the ADI score (i.e., one star out of AD, two stars could be out of AD, three stars in AD). This intuitively and immediately summarizes which is the prediction and if it may be reliable or not. A textual description reports the main issues, if identified, in the prediction.

The page with the ADI also reports symbols that declare if a certain parameter is correct or if there are levels of concern. The user should take particular attention evaluating the factors that are indicated as critical. The presence of critical aspects does not automatically mean that the prediction is wrong. The user may overrule these warning elements, but this should be done based on a solid argumentation. Indeed, the report of VEGA is intended for the user. The user should use it to prepare her/his report and address in particular the critical issues. The user takes the final decision.

At the VEGAHub web site, there is the explanation of the ADI, how to use it, and some examples. Furthermore, for each model VEGA provides guidelines, describing how the ADI is calculated, which varies for each model. In addition, for each model, there are the chemicals used for the training and test sets, with their property values. In some cases, there is also the QSAR model reporting format (QMRF) of the model.

2.3 The Use of VEGA for Read Across

Within VEGA, we provide a tool for read across, since VEGA shows the six most similar compounds. Thus, this information can be immediately used for all the endpoints as in VEGA. The read across can be done based on the similarity, as calculated by VEGA, based on the experimental values provided, and on the structural alerts/fragments which may be provided depending on the model. Thus, the read across can be done based on different features. With some models, for instance BCF, the software also provides a plot with the

experimental values of the three most similar compounds and the log Kow values. This can be also used for read across. In this case, the user does not need to look at the predicted values for the target, and the process is quite simple.

2.4 ToxRead

ToxRead further improved the way to evaluate chemicals within a read across perspective. Indeed, using VEGA to perform a read across, in case of multiple models for the same endpoint, the user should run all the models and perform the read across with each one. It means that in case of four models, the user should perform the read across four times. In addition, since the datasets used for each model may be different, the user may have different results. One of the advantages of the use of ToxRead is that all the dataset available are merged together. A second advantage is that the user can choose the number of similar compounds to consider. ToxRead provides the N most similar compounds as in VEGA, using exactly the same algorithm to calculate similarity. The user can decide the number of similar compounds (usually four or five are enough), which may be even tens of compounds. Figure 2 shows an example for BCF.

Another advantage of ToxRead is that it searches for rules/fragments/reasons of effects. Depending on the endpoint, there are structural alerts (in particular if the endpoint is defined as a

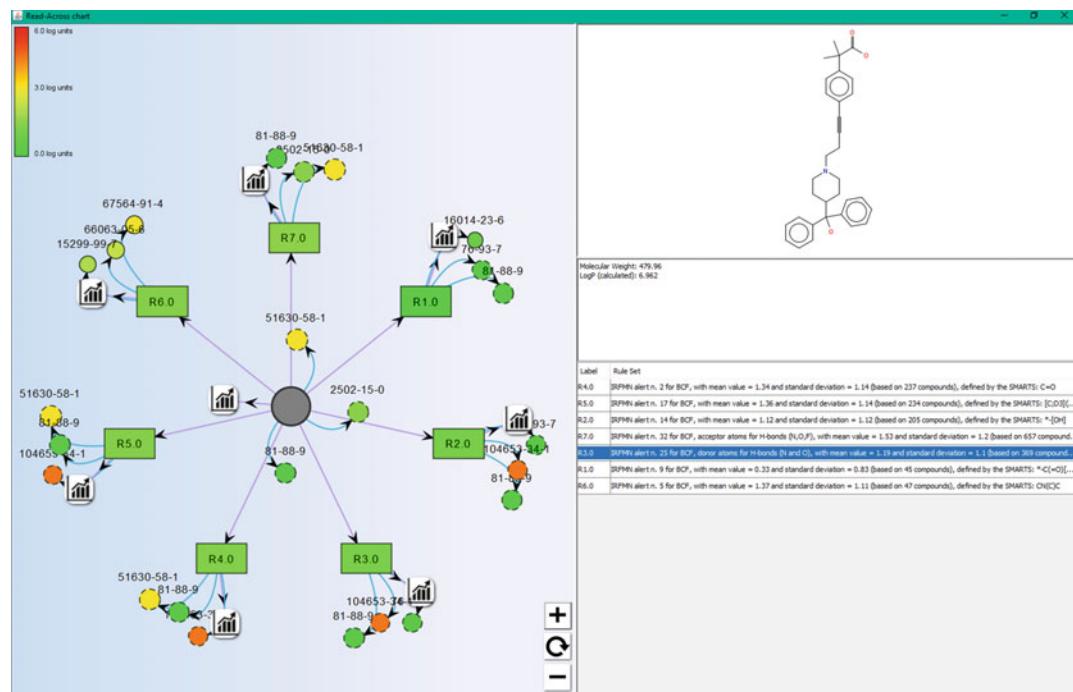


Fig. 2 A ToxRead example, the BCF (the number of similar compounds selected is 3). The molecule showed is the target selected

classifier: toxic or not) or also rules associated with a threshold in case of continuous values. In this way, the user can visualize which are the features associated with the effect, both increasing and decreasing the value, present in the molecule.

ToxRead shows the N most similar compounds, as chosen by the user, linked with arrows to the specific rule. The central circle represents the target compound. The other circles represent the similar chemicals, with different sizes depending on the similarity value (the higher the similarity is, the bigger the circle is). The color of the circles depends on the property value: red for toxic, green for not toxic, and yellow in between (the legend is reported in upper-left figure). Clicking on the circle, a window appears with the structure of the substance and the experimental values.

Triangles represent the rules used by ToxRead in case of structural alerts (i.e., toxic or not toxic), whereas rectangles represent the rules in case of continuous parameters (e.g., the molecule has a log Kow value associated with a certain property value). Clicking on the rules, the user can visualize the statistics related to the rules and the 100 chemicals most similar to the target compound that contain that rule.

In this way, the user has a scheme that on the same page shows all the most similar compounds, all the reasons for effect/lack of effect, and the most similar compounds that share a certain rule. This organized representation provides a standardized way to process the information for read across. It has been shown that ToxRead gives reproducible results in read across, while this is not always the case for other tools for read across [21].

2.5 ToxWeight

The user is typically interested in the assessment of the property value of the chemical substance, and for this purpose, he/she uses QSAR and/or read across. The ideal situation is when all these elements are combined within the same strategy. EFSA, in their guidance on weight of evidence, described this [19]. ToxWeight (available in the VEGAHub web site) is a first example of a program going in that direction. ToxWeight so far is working only for mutagenicity (Ames test). In the future, more endpoints will be added. In practice, when the user runs the new version of ToxRead, the results of the VEGA models are also presented through ToxWeight. It means that the user, through one single program, gets the results from the read across assessment, and from the five VEGA models on mutagenicity (four individual model and one which integrates the results of these four models), and then the system provides the overall assessment. All this is possible with one click, automatically.

2.6 PROMETHEUS and JANUS

The German UBA and the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety funded two software for prioritization and screening: PROMETHEUS and

JANUS. At the VEGAHub web site, it is possible to download the tool, the extended report of PROMETHEUS, and a short summary from Chemical Watch [22]. A scientific paper also summarized the project [23].

The strategy beyond both projects is to integrate a series of models and experimental data, in order to get an overall assessment regarding PBT (with PROMETHEUS) and additionally information on carcinogenicity, mutagenicity, and reprotoxicity (CMR) and effects as endocrine disruption (in the case of JANUS). JANUS is an evolution of PROMETHEUS, adding more properties, in particular for human toxicology. Thus, the overall assessment is done not on one single property value but on a large list of different endpoints, and the results are then integrated. The definitions for PBT and CMR are according to the European regulation. Thus, a chemical is PBT if it is at the same time persistent, bioaccumulative, and toxic. If it is not active for one of these properties, it is not PBT. Conversely, a chemical is persistent if it is persistent in at least one of the following compartment: water, sediment, or soil. As for the persistence, a chemical is classified as CMR if can be classified in at least one of the three categories (C, M, or R).

The thresholds applicable to define PBT and CMR are those according to the European legislation.

PBT and CMR are classifications of the substances. However, in order to prioritize the substances in a list, and rank them in a progressive order, we need continuous value. Thus, we used the continuous value for the toxicity properties, which are given as LD50 or LC50 (the dose or concentration that kills 50% of the animals). Thus for the ecotoxicological properties, which are addressed in the present chapter, the task is easy. More difficult is the situation for CMR. We will not address CMR here. Briefly, we used the potency value in this case.

In case of experimental values, the software uses them and gives a higher reliability, compared to the predicted values. In case of multiple values, the reliability is higher if there is consistency between the values. The software applies a series of checks: for instance, it keeps into account water solubility in case of aquatic toxicity, checking if the toxic value occurs at concentration higher than the water solubility.

The uncertainty of the assessment is provided, and it comes from the nature of the value (predicted or experimental), on the consistency between multiple values, and on the ADI in case of predicted values.

The software evaluates not only effects of the parental compound, but in addition it generates degradation products that may be produced into the environment. Then, the tool processes and prioritizes these degradation products.

Overall within JANUS there are 48 different QSAR models, running together, generating one single overall priority score.

However, the user can see the individual results for each property and the uncertainty associated with each value.

2.7 Research Tools for Modeling

VEGAHUB offers a series of tools, which can be used to develop other models, based on available list of chemicals, for any property of interest.

One example is SARpy [24]. This software has been developed by Politecnico di Milano and serves to build up classifier models, starting from lists of chemicals, simply using their SMILES structures. It progressively cuts the bonds generating fragments from each molecule. The prevalence of active/inactive substances with a certain fragment is used to build up collections of rules. SARpy has been used for several endpoints, for instance, ready biodegradability [25] and persistence [26].

An interesting advantage of this software is that it is quite easy, it can be used internally (for instance, by industries), and the model can be made public, without disclosing proprietary information on the chemical structures and toxicological data used to build up the model. Indeed, the model generated a series of structural alerts, which can be disclosed.

It is also interesting to notice that SARpy generates rules both for activity and lack of activity, which is different from the case of other expert systems which contain only rules associated with the adverse effect. In case of continuous values, the user can choose the threshold used to define toxicity. Different thresholds can be defined, and thus SARpy is not only a binary classifier.

QSARpy is an evolution of SARpy dedicated to continuous endpoints [27]. It breaks the molecules as SARpy into fragments, and then, for each fragment, it compares each couple of molecules of the training set that shares the same structure but different fragments. The fragment becomes a modulator associating to it a value (either positive or negative). In prediction, QSARpy compares the target molecule with the molecules of the training set. If it finds a similar molecule that is different only for one or few modulators (the target can be a superstructure, a substructure, or an intersection of the similar molecule), the predicted value is obtained combining the activity/property value of the similar and the value associated with each modulator. The list of modulators may be analyzed to verify the existence of a mechanistic explanation as for the rules extracted with SARpy.

CORAL is another useful software, which is available as a general tool to build up models based on the lists of chemicals available internally [28], and it has been also used to develop many QSAR models of ecotoxicological properties and others of environmental relevance [29–38].

3 Examples of Use of the VEGA Platform

We produced the following examples using with the VEGA version 1.1.5-b22. In this section, we describe the output of the models, the interpretation of the results of each model, and the overall assessment of the molecule.

3.1 Example 1: Glycolic Acid, Fish Acute Toxicity

VEGA has several models for fish acute toxicity. It is enough to enter the SMILES of the substance and choose the models of interest. This chemical is the hydroxyl derivative of acetic acid. Thus, it is a quite simple substance, quite polar. Its SMILES is C(C(=O)O)O. Figure 3 shows the target molecule and the three most similar chemicals of each model we used.

The first model we comment is Fish Acute (LC50) Toxicity classification (SarPy/IRFMN) 1.0.2 (named SARpy). This model, as we said, is a classifier. In this case, the threshold is 100 mg/l, which is the value of the limit test. Thus, the meaning of the result is that the chemical is not toxic, if predicted above 100. However, the output of the SARpy model can be more granular, and it also provides predictions given as between 10 and 100 mg/l, or between 1 and 10 mg/l, or less than 1 mg/l. The software provides a number of similar compounds, and the overall assessment done on these compounds is quite reassuring that the prediction may be correct. Indeed, the ADI is quite high (0.898), and all the

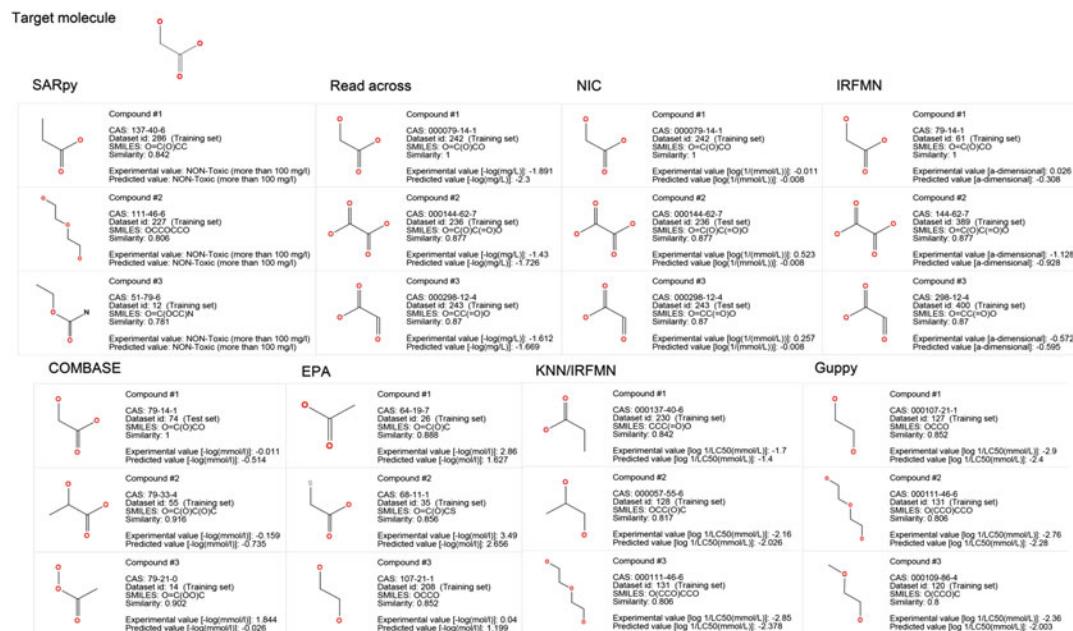


Fig. 3 The first example, the glycolic acid. The figure shows the target molecule and the three most similar compounds of each model used in the example

individual components of the ADI are correct, without any warning message. Anyhow, the user should manually verify all the values and have a look at the structures. The user, looking at the results on the similar compounds, should try to make considerations in the framework of a read across assessment. This is a second line of evidence, to be combined with the QSAR prediction. Read across regards the experimental data, the results based on the studies on similar compounds. VEGA helps in the selections of the similar compounds, but the user should also refer to background knowledge in the field. The most similar structure found by the system for this model is the propionic acid. It is not toxic (toxicity higher than 100 mg/l). Thus, this finding confirms the prediction done by the model. Propionic acid has a longer chain, compared to the target compound, and it is expected to be less polar. Thus, its toxicity should be higher. This is an advantage, because we have a similar compound that is expected to be more toxic. It is always preferable to identify which may be worse situations, to be used as boundaries.

The second most similar compound is 2,2'-dihydroxydiethyl ether. It is not toxic (>100 mg/l), but compared to the previous chemical, it may be less useful. Indeed, the acidic group is missing, and there is the ether moiety. The third similar compound is even less similar, with similarity value lower than 0.8. Thus, from this first model, we have the prediction, the read across (in particular on the first compound), and some reasoning about solubility that may contribute overall to come to the conclusion that the toxicity is not high, probably >100 mg/l.

The second model is the Fish Acute (LC50) Toxicity model (KNN/read across) 1.0.0 (named read across) model. It is based on the most similar compounds. In this case, the model contains the target compound in the training set. Thus, the system provides immediately the experimental value, which is reported 77.73 mg/l. This value may appear conflicting with the previous one, from the SARpy model. SARpy gave a predictive value, which may be questioned based on this experimental value, which is lower. However, one important consideration is the experimental value of propionic acid, which was also with the toxicity value >100 mg/l. Another important consideration is that the value 77.73 is not so different from the value >100 , even though these two belong to two toxicity classes. Indeed, in the literature it is possible to find fish acute toxicity values >100 for the target compound, as from the ECHA web site, for instance. We have always to keep in mind the experimental uncertainty and variability, and the difference between 78 and >100 mg/l is not so high. In the case that the VEGA model contains the experimental value, the ADI assessment refers to this, since this value should be used. Thus, we will not comment further the results from this model.

The third model is the Fish Acute (LC50) Toxicity model (NIC) 1.0.0 (named NIC) model, based on neural networks. In

this case, we also find the experimental value, and the predicted one is very close: 77.27 mg/l.

The fourth model is the Fish Acute (LC50) Toxicity model (IRFMN) 1.0.0 (named IRFMN) model, which again reports the experimental value as before, and the prediction is 55.4 mg/l.

The fifth model is the Fish Acute (LC50) Toxicity model (IRFMN/COMBASE) 1.0.0 (named COMBASE) model, with the experimental value as before, and the prediction is 247.44 mg/l.

The sixth model is the Fathead Minnow LC50 96 h (EPA) 1.0.7 (named US EPA) model. In this case, the system does not provide experimental value for the target compound (it is not contained in the training set of this model), and the prediction is 1614.88 mg/l. This value is surely different from all the others we have seen. In this case, the reliability of the prediction, as reported by VEGA, is low: only one star. The system reports a series of warnings. The ADI is 0.7, and the main issues indicated by the system refer to the accuracy of the predictions, the concordance for similar chemicals, and a large error in prediction observed in at least one case. The other components are not critical, and in particular, there are similar compounds. In this case, it is important to analyze the similar compounds manually and check what the system says. Indeed, the software makes predictions with an error of one log unit, for chemicals that are quite similar, like the acetic acid. Also for the other similar compounds, the errors in prediction are quite large. It means that the model is not so reliable for this kind of compounds. We may consider if the error is always in the same direction: over- or underpredictions. In this case, we observe that EPA model makes errors in different directions depending if the chemical is an acid or not. Thus, in principle we may consider this and apply an offset. Another point that should be considered is the fish species modeled. In the models we examined before, there is no specific fish species, whereas in the last case, the model refers to the fathead minnow. This species is not as sensitive as others (e.g., rainbow trout); therefore, a higher LC50 value is plausible.

The seventh model is the Fathead Minnow LC50 model (KNN/IRFMN) 1.1.0 one (named KNN/IRFMN). In this case, the predicted value is very high: 5422.77 mg/l, and there are two stars of reliability. Still some issues occurs. The ADI is 0.85. The main issue is about the concordance. Indeed, the predicted value is quite far from the experimental values of some similar substances. Here we can see which is the greater risk associated with the KNN models, in general. KNN models are based on the similarity of the substances, which are used to generate the prediction. In our case, the second most similar chemical is the 2-hydroxypropanol, with a similarity of 0.817 and an LC50 of more than 10 g/l. Thus, this chemical, which has two hydroxyl groups, is much less toxic than the other substances that we have seen so far, containing an acidic

moiety. The user should be aware of these critical issues, notified by VEGA. In particular, in the case of the KNN model, the concordance of the most similar substances becomes a critical aspect, because it immediately indicates that the property values of the similar compounds are not homogeneous; consequently, the reliability of this prediction is very low. The fact that the toxicity value of glycolic acid is similar to the toxicity value of the 2-hydroxypropanol is not realistic, on the basis of the evidence that other analogues of the target compound with the acidic moiety have toxicity values in the range of about 100 mg/l, and thus 100 times more toxic. Also in this case, the model refers to the fathead minnow species; therefore the considerations done previously are still valid.

The results of the last model, the Guppy LC50 model (KNN/IRFMN) 1.1.0 (named guppy), are also very different from those initially seen: indeed the toxicity is more than 60 g/l. The guppy model is also a KNN model. If we look at the most similar substance, used to feed the model, we can see they are all alcohols. This represents a bias that has been discussed above. Indeed, the ADI of this model is quite low: 0.511. For this model, however, the critical issues are the accuracy of the predictions, the maximum error in the prediction, and the presence of one or more rare fragment. This last point, in particular, refers to the fact that the substance has the hydroxyl-acid moiety, which is unknown to the model, and this is a serious limitation. In addition, this model refers to a specific fish species, which could be less sensitive than others (e.g., rainbow trout).

What we have done in the discussion of the above results follows the indications of the EFSA guidance on weight of evidence that we already introduced [19]. In particular, we gathered the different lines of evidence, which in our case are the different models. This is also in agreement with other authorities, like ECHA, which recommend using more than one QSAR model. Once we have identified the lines of evidence, all the models for fish acute toxicity, the associated experimental values, and the elements for reasoning, we have to evaluate each line of evidence individually. This is what we have done. The assessment should evaluate if the result is reliable. As we have seen, in this discussion, we used the predicted values, the experimental values, the data of the similar compounds (to be used for read across), and elements of reasoning, which derive from basic chemical knowledge, such as functional groups and expected polarity. Sometimes, in VEGA, the values of the similar compounds are given in log unit and in mmol/l, but to get the value in mg/l, if one does not want to do the conversion manually, it is sufficient to copy and paste the SMILES of the similar compound and run VEGA. At this point, VEGA will recognize that this substance is in the training set and thus will provide the experimental value, in mg/l.

Once the user has discussed the individual lines of evidence, the following step is to get an overall assessment, integrating the results, which have been characterized in their reliability.

In our example, we have a larger set of models that provide an experimental value of 77.73 mg/l (several models gave the same value, probably with the same original source). The classification model, in AD, predicted it as non toxic (>100 mg/l), whereas the three models on specific species gave values higher than 1 g/l, but they are not reliable (based on similar chemicals which are not relevant for our case, as we already discussed). We also have to remember that the regulation accepts experiments done on different fish species, and this variability increases the range of the expected values.

As a conclusion, we can accept that the toxicity value is about 100 mg/l, with good reliability, and the compound should be considered moderately toxic.

3.2 Example 2: 2-(4-Terbutylbenzyl) Propionaldehyde, Fish Acute Toxicity

The SMILES of this chemical is O=CC(C)Cc1ccc(cc1)C(C)(C)C. Figure 4 shows the target molecule and the three most similar chemicals of each model we used.

The SARpy model says that the toxicity is in the range 1–10 mg/l. Its ADI value is high, 0.933. There are several similar chemicals with the same range of toxicity, and one of them is an aldehyde. Overall, there are not major critical issues. The model

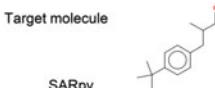
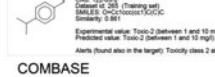
Target molecule				
SARpy	Read across	NIC	IRFMN	
 <p>Compound #1 CAS: 80-44-6 Dataset id: 481 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Experimental value: Toxic-2 (between 1 and 10 mg/l) Predicted value: Toxic-3 (between 10 and 100 mg/l) Alerts (found also in the target): Toxicity class 3 alert no. 24 Alerts (not found in the target): Toxicity class 3 alert no. 11</p> <p>Compound #2 CAS: 103-05-8 Dataset id: 147 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Experimental value: Toxic-2 (between 1 and 10 mg/l) Predicted value: Toxic-3 (between 10 and 100 mg/l) Alerts (found also in the target): Toxicity class 3 alert no. 24 Alerts (not found in the target): Toxicity class 3 alert no. 11</p> <p>Compound #3 CAS: 103-05-8 Dataset id: 295 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Experimental value: Toxic-2 (between 1 and 10 mg/l) Predicted value: Toxic-2 (between 1 and 10 mg/l) Alerts (found also in the target): Toxicity class 3 alert no. 24 Alerts (not found in the target): Toxicity class 3 alert no. 11</p>	 <p>Compound #1 CAS: 000140-66-9 Dataset id: 481 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.995</p> <p>Experimental value [-log(mmol/L)]: 0.446 Predicted value [-log(mmol/L)]: 0.175</p> <p>Compound #2 CAS: 000090-46-6 Dataset id: 476 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.89</p> <p>Experimental value [-log(mmol/L)]: 0.408 Predicted value [-log(mmol/L)]: -0.731</p> <p>Compound #3 CAS: 005510-99-6 Dataset id: 482 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.979</p> <p>Experimental value [-log(mmol/L)]: 0.826 Predicted value [-log(mmol/L)]: 0.687</p>	 <p>Compound #1 CAS: 000140-66-9 Dataset id: 481 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.995</p> <p>Experimental value [log (1/mmol/L)]: 2.758 Predicted value [log (1/mmol/L)]: 2.833</p> <p>Compound #2 CAS: 000090-46-6 Dataset id: 476 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.89</p> <p>Experimental value [log (1/mmol/L)]: 1.806 Predicted value [log (1/mmol/L)]: 1.814</p> <p>Compound #3 CAS: 005510-99-6 Dataset id: 482 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.879</p> <p>Experimental value [log (1/mmol/L)]: 3.138 Predicted value [log (1/mmol/L)]: 0.319</p>	 <p>Compound #1 CAS: 140-66-9 Dataset id: 378 (Test set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.895</p> <p>Experimental value [a-dimensional]: -4.57 Predicted value [a-dimensional]: -4.007</p> <p>Compound #2 CAS: 5510-99-6 Dataset id: 571 (Test set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.479</p> <p>Experimental value [a-dimensional]: -4.985 Predicted value [a-dimensional]: -4.714</p> <p>Compound #3 CAS: 99-71-8 Dataset id: 480 (Test set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.875</p> <p>Experimental value [a-dimensional]: -3.272 Predicted value [a-dimensional]: -3.033</p>	
 <p>Compound #1 CAS: 87564-51-4 Dataset id: 83 (Test set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.783</p> <p>Experimental value [-log(mmol/L)]: 2.037 Predicted value [-log(mmol/L)]: 2.912</p> <p>Compound #2 CAS: 100-51-6 Dataset id: 42 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.757</p> <p>Experimental value [-log(mmol/L)]: -0.631 Predicted value [-log(mmol/L)]: 0.541</p> <p>Compound #3 CAS: 104-56-2 Dataset id: 81 (Test set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.757</p> <p>Experimental value [-log(mmol/L)]: 1.692 Predicted value [-log(mmol/L)]: 1.868</p>	 <p>Compound #1 CAS: 000090-46-6 Dataset id: 483 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.89</p> <p>Experimental value [-log(mmol/L)]: 4.8 Predicted value [-log(mmol/L)]: 4.993</p> <p>Compound #2 CAS: 103-05-6 Dataset id: 177 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.864</p> <p>Experimental value [-log(mmol/L)]: 3.39 Predicted value [-log(mmol/L)]: 3.761</p> <p>Compound #3 CAS: 98-00-4 Dataset id: 148 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.853</p> <p>Experimental value [-log(mmol/L)]: 4.47 Predicted value [-log(mmol/L)]: 3.99</p>	 <p>Compound #1 CAS: 000090-46-6 Dataset id: 478 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.89</p> <p>Experimental value [log (1/C500/mmol/L)]: 1.8 Predicted value [log (1/C500/mmol/L)]: 1.918</p> <p>Compound #2 CAS: 00103-05-6 Dataset id: 437 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.864</p> <p>Experimental value [log (1/C500/mmol/L)]: 0.39 Predicted value [log (1/C500/mmol/L)]: 1.238</p> <p>Compound #3 CAS: 000133-35-3 Dataset id: 631 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.861</p> <p>Experimental value [log (1/C500/mmol/L)]: 1.35 Predicted value [log (1/C500/mmol/L)]: 0.877</p>	 <p>Compound #1 CAS: 000090-46-6 Dataset id: 478 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.89</p> <p>Experimental value [log (1/C500/mmol/L)]: 1.91 Predicted value [log (1/C500/mmol/L)]: 1.276</p> <p>Compound #2 CAS: 000098-54-4 Dataset id: 470 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.83</p> <p>Experimental value [log (1/C500/mmol/L)]: 1.46 Predicted value [log (1/C500/mmol/L)]: 1.443</p> <p>Compound #3 CAS: 014033-35-3 Dataset id: 472 (Training set) SMILES: O=C(C)Cc1ccc(cc1)C(C)(C)C Similarity: 0.85</p> <p>Experimental value [log (1/C500/mmol/L)]: 2.12 Predicted value [log (1/C500/mmol/L)]: 1.836</p>	

Fig. 4 The second example, the 2-(4-tert-butylbenzyl)propionaldehyde. The figure shows the target molecule and the three most similar compounds of each model used in the example

finds a structural alert (SA) associated with compounds with a toxicity between 1 and 10, the para-xylene. VEGA shows the three most similar compounds that share the same alert. In this case, the two most similar compounds have in common also other SAs. Only the third has one different SA, the ortho-xylene, but this SA is not used in the estimation because it indicates a lower toxicity (the classification is based on the worst case).

The KNN/read across model says 0.77 mg/l, which is lower than the previous result. The KNN models, as we said, reflect the toxicity values of the similar compounds. We observe that the similar compounds are phenols, which are a moiety not present in the target compound. Furthermore, there is an error in prediction quite high. This model should be considered less reliable than the previous one.

The NIC model predicts the toxicity at 1.77, with a very high reliability. Also in this case, the most similar compounds are phenols, but the situation is less critical than in the case of the KNN model, because the NIC model uses descriptors to make the predictions.

The IRFMN model gives a prediction in practice identical to what is seen for the NIC model; indeed the prediction is 1.78. However, in this case VEGA identifies a possible issue, related to the concordance values between one similar substance and the target compound. However, since the most similar substances are phenols with toxicity values that are also lower than 1, this issue should not be considered critical; it is sufficient to consider that these similar compounds may not be useful for the read across evaluation. Indeed, for this model, and for most of the previous ones, we have seen that the similar compounds are mainly phenols, and this is not very useful for the read across perspective.

The COMBASE model predicts the toxicity at 2.64 mg/l. There are some issues reported by VEGA, which indicates the ADI value = 0.392, thus quite low. The COMBASE model has been developed on a training set of biocides. This may have implications: the substances in the training set may be more toxic than the general, average population of the typical substance, which are assumed to be industrial chemicals. However, this fact does not seem to affect the prediction in this case, since the toxicity value is very close to the values seen for the previous models. The other major implication is that the training set is relatively small. This is probably the main reason for the low ADI value. Indeed, VEGA reports that there are not very similar compounds, and this is due to the smaller training set. Furthermore, the concordance is not high. Indeed, for instance, there is a similar compound with the toxicity value = 26 mg/l, thus ten times more the predicted value for the target compound. Furthermore, there are also one or more rare fragments present in the molecule. Again, these critical issues reflect the quite limited training set of the COMBASE model.

The US EPA model gives a prediction of 0.789, which is somehow lower than the values we have seen so far. The reliability is not high. This is mainly due to the concordance with similar substances, and there is also a large error in prediction for a similar compound.

The KNN/IRFMN model predicts the toxicity at 11.26. As we already commented, the KNN is reliable provided that the substances are quite homogeneous and related to the target compound. In this case, for instance, we notice that there is a similar compound, which is an aldehyde, with a toxicity value of about 6.6 mg/l. This substance, within a scheme for read across, may be useful. Thus, this model provides a kind of agreement with the previous models, in particular if we refer to the read across perspective, more than the QSAR prediction. We have also to remind that this model refers to a specific fish species, which could be less sensitive than the other species.

The Guppy model predicts 4.63 mg/l. There are critical issues reported by the model, in particular for rare fragments, and a large error in one case in prediction of a similar compound. This may be due to the small dataset used to build this model. We have also to remind that this model refers to a specific fish species, which could be less sensitive than the other species.

Overall, the results are quite consistent. In this example, we adopted a strategy somehow different from what we did for the first example. In the first example, we tried to analyze the different lines of evidence independently and then to integrate the results. An alternative approach starts from a model, and step by step, the user critically compares the results and sees how the new evidences may modify what has been established previously. The disadvantage is that in this case, there may be a bias, represented by the past evaluation of the previous results. However, it is reasonable to think that also in the other cases, it may happen that what has been already found may influence the successive results, since multiple concordant results will reinforce the confidence on a certain value. Thus, in all cases it is important to analyze critically all the novel elements. An advantage of the present approach is that we can immediately use the different findings already found to evaluate the individual results, and thus, for instance, we may refer to similar compounds found in a different model.

In this case, the overall toxicity seems to be a few mg/l; therefore, the compound should be considered toxic.

4 The Future Tools

VEGAHUB is a living platform, which continuously adds new models and tools, also thanks to a growing network of collaborations. We will describe below some important improvements that are planned within some projects.

4.1 The toDIVINE Project

The major improvements that are anticipated above are related to a new strategy for read across and weight of evidence, which better combines the read across with the results of the QSAR models. So far, we have discussed already how VEGA can be used also for read across. However, there are two limitations to the current approach: (1) the similarity is processed only on a structural point of view, and (2) certain features, like log Kow, etc., which we used in our discussion, rely on the expert's knowledge and thus are not used by the system.

The project toDIVINE, funded by the German UBA, wants to address this need to better integrate read across with QSAR, but using a deeper approach for read across. There are three partners within the toDIVINE project, which is coordinated by the IRFMN.

toDIVINE is using multiple features to identify substances that can be used for similarity. Beyond the structural similarity, as processed within VEGA, we will also use the physicochemical properties, the toxicological information, the pharmacokinetic information, and the information about degradation products that may be generated into the environment.

Of course, these properties and information are strictly associated with the endpoints of interest. toDIVINE is addressing fish, daphnia, and algae (both acute and chronic toxicity), persistence, bioaccumulation, and endocrine disruption.

Within toDIVINE, we decided to use parameters that can be calculated if the experimental value is not available to be able to process all chemicals of interest. For the physicochemical parameters, we will use log Kow, water solubility, Henry's constant, and molecular weight.

For the toxicological information, we will use the mode of action and a series of structural alerts specific for each endpoint. There are, indeed, some programs able to estimate the mode of action, for instance, for fish toxicity (as in TEST [10]), or for endocrine disruption, which is also available in VEGA. Unfortunately, this cannot be applied to all endpoints. For this reason, a series of fragments, which can be associated to the property value, will be also used, and these can be easily obtained with SARpy [24].

The pharmacokinetic information, which are also applicable to certain endpoints only, will be obtained from the kM for fish metabolism. It has been obtained from the EPISuite system [11]. Finally, we will also refer to the environmental biodegradation, because the overall critical impact into the environment may be due not to the parental compound but to the degradation product(s). The way to process the degradation products is done within the JANUS project.

Obviously, compared to the approach as in VEGA, this is a much more complex scheme, which also involves a series of phenomena which are usually addressed only in an implicit way with

the current approach or that are not addressed at all. There are some main advantages, which are (1) a larger number of similar compounds to be used for similarity, and thus it is easier to identify the relevant one—in principle it is possible to miss some relevant similar chemicals for read across, if we simply refer to the chemical structure; (2) a deeper evaluation of the similarity process, considering a number of features which are surely relevant to compare two substances; and (3) it is easier to reason about the causes of similarities and dissimilarities, since they are explicit.

We will summarize separately the results from the read across process and form the QSAR models into two values. Then, we will integrate these values to get the overall property value for the target compound. In all these processes, we will consider the reliability of the values, to weight the different contributions along the process and to report the uncertainty associated.

4.2 The *LIFE VERMEER Project*

LIFE VERMEER is a project funded by the EC and includes several partners in Europe, involving different stakeholders. The IRFMN coordinates the project.

The two main objectives of LIFE VERMEER are (1) to integrate the VEGA platform with the MERLIN-Expo platform and (2) to develop a software program to substitute risky substances.

MERLIN-Expo [39] is a platform developed within a previous EC project, to predict exposure. It addresses both environmental exposure and human internal exposure. There are multiple scenarios within MERLIN-Expo. The major advantage of the interaction of VEGA and MERLIN-Expo is that in this way, the user will have the possibility to get an overall prediction of the risk assessment using a unique software, LIFE SPHERA. This is a major improvement compared to any other existing tool.

Furthermore, within LIFE VERMEER, the aim is to develop a tool, named ToxEraser, to replace risky substances. Once identified as risky, based on the results of LIFE SPHERA, it will be possible to identify the reasons for the concern. We will base this assessment on an evolution of the ToxRead software. Indeed, ToxRead already has some important features: the possibility to identify reasons for the concern, and the knowledge on the neutral, or even counteracting factors that may reduce the adverse effect. Of course, currently within ToxRead this is limited to very few endpoints. Within LIFE VERMEER, we will fully exploit and extend this approach to all the relevant endpoints.

In order to integrate VEGA and MERLIN-Expo, we are developing a number of new models that will be available also within VEGA. We are working in particular on the endpoints useful to address the environmental fate and behavior, such as models for the Henry's law constant, partitioning between octanol and air, and octanol and carbon in soil.

A number of case studies will provide a sound basis to put in practice the outcomes of LIFE VERMEER. For instance, there will be case studies regarding (1) the human toxicological scenario, like the food contact materials, (2) both human and environmental scenarios, like the personal care products, and (3) case studies more directly related to the environmental scenarios, like green solvents, biocides, and dispersants (to be used in case of oil pollution). These case studies will surely address environmental impact, but the human toxicological aspects will be also covered. Indeed, the occupational risk is always present.

The challenge is to integrate the environmental with the human toxicological issues. In the future, these aspects have to be covered simultaneously. For instance, the persistence is very important also for the human toxicological endpoints, because if a substance is persistent, it will reach the population at a certain time. There are unfortunately several examples teaching us about this fact (e.g., the DDT).

The focus on the substitution makes this project particularly relevant for the European REACH regulation. Indeed, after the first phase, related to the registration of the substances present in the European market, now the new challenge related to REACH is the substitution of the substances of higher concern.

For this purpose, the new ToxEraser software will be particularly useful, assisting the industry to better adapt the lists of substances that can better meet the new safety requirements. In the past, industry was exploring new substances mainly through laboratory experiments. The information about the possible concern for certain toxicological and ecotoxicological effects was not available, and thus this issue was addressed in a later step, when a lot of efforts and work were already spent. The use of in silico tools, as in the case of the planned ToxEraser software, will be very beneficial for industries. Indeed, the industry will have available tool able to screen a large series of chemicals, before their preparation and synthesis.

4.3 The LIFE CONCERT REACH Project

The LIFE CONCERT REACH project started very recently, at the end of 2018. It follows a series of other LIFE projects addressing REACH: ANTARES, CALEIDOS, LIFE PROSIL, and LIFE VERMEER. The IRFMN coordinates the consortium of the project.

The main difference about LIFE CONCERT REACH and the previous LIFE projects is that finally the data on the registered substance became available. As we have seen previously, a number of models have been developed, based on historical collections of data, about a number of common endpoints, such as mutagenicity, fish toxicity, etc. However, while the speed of the improvements about the hardware and the software allowed in the last decades to introduce better and more efficient models, taking advantage of cheap and fast computers, and of a high number of molecular

descriptors and algorithm, the progress in the availability of experimental data is in general very slow, in particular, for data derived from in vivo experiments (in the case of in vitro data, the Tox21 and ToxCast initiatives are producing a large number of data). Thus, the REACH legislation is beneficial also regarding this important issue: to make data available. No data, no market was the motto. The unprecedented availability of data from the registrations represents a unique mine of data. As always, we have to be aware that the quality of the data is not homogeneous, and pruning will be necessary. Nevertheless, REACH offers, also to QSAR developer, a new perspective, starting from the data available within REACH.

We have to distinguish between the use of data to build up a QSAR models and for read across. In order to use data for read across, the user needs a letter of access, since the data are proprietary. Typically, the access requires a payment. Furthermore, it is not sufficient to have the value, for instance, the experimental aquatic toxicity value; for read across the registrant needs the full documentation with all the details about the experiment, and this is not available.

Conversely, to build up QSAR models, it is sufficient to have the list of values (better if with details on the test performed) and chemical structures. Thus, based on the REACH registration, now a vast amount of new data is available, for a larger set of endpoints. The LIFE CONCERT REACH wants to exploit these data in order to build up new models for endpoints that currently do not have models.

Furthermore, this project aims to establish a network between three of the most advanced and used systems for modeling: VEGA, the Danish QSAR database, and AMBIT. These systems will remain as independent platforms, but we will identify synergies, and we will address together the new models. We will direct and inform the user about the peculiar features of each of the three systems.

4.4 The OptiTox Project

EFSA funded the OptiTox project, which started at the end of 2018 and will last 4 years. The IRFMN coordinates the consortium of the project.

The OptiTox project wants to continue the work on the OpenFoodTox database of EFSA [40]. OpenFoodTox contains data on many thousands of substances of interest for EFSA, related to plant protection products, food contact materials, veterinary pharmaceuticals, contaminants, etc. The data regards tens of different taxa and contains data on mixtures. Thus, this is a quite large and rich database, with lot of details.

Compared to other databases, such as REACH, containing data provided by external sources, this database refers to the documents produced by the different units within EFSA, and thus the value refers to a curation process which is quite valuable.

Furthermore, starting from these data, we will develop new QSAR models within OptiTox, and we will make them available within VEGA.

Another interesting characteristic is that OptiTox will also exploit the data on the toxicokinetics available within EFSA. This will improve the predictive capability of the predictive models.

The issue of the toxicokinetics is also present within other projects we mentioned, like toDIVINE and LIFE VERMEER. Surely, it will become more and more important to have a deeper use of the available pieces of information. However, this will also require new efforts to cope with a more general evaluation of the risk assessment scheme and will require a new perspective on a regulatory point of view. Indeed, the toxicokinetics will bring a better evaluation about the internal dose and the behavior of the substance inside the body, either the human body or the animal one. However, the regulation currently refers only to the external dose, and does not identify any threshold/value for the internal dose. Thus, this aspect is fundamental on a scientific point of view, and is surely relevant for read across to identify commonalities between different substances, but in the case of QSAR will require more work, jointly done together with authorities.

4.5 Endocrine Disruption

Certain properties, like endocrine disruption, surely affect both the human population and the wild species. For this reason, we are extending the number of models able to address this phenomenon, and there will be several new models within VEGA in the near future. We will do it within a series of projects, like toDIVINE, LIFE VERMEER, and others not mentioned here, because of its relevance to human toxicology. These new models often derive from international collaboration, like the collaboration with the US EPA on the prediction of estrogen and androgen activity, within the CERAPP [41] and COMPARA [42] initiatives, respectively. In these projects, new models have been developed, for these two endpoints. We contributed with new models to be implemented within VEGA.

Another example of the collaborations related to endocrine disruption is with Prof. Kunal Roy, Jadavpur University, Kolkata, India. These joined studies allowed to address endocrine disruptors potentially affecting a number of species.

Furthermore, we established another collaboration with Prof. Wei Shi, Nanjing University, China. Her group developed a number of models for many nuclear receptors using SARpy, and we will implement these models within VEGA soon.

5 Conclusions

We introduced the platform VEGAHUB and its development. The main interest for VEGAHUB is to make models available and useful to better assess the chemical properties. This should improve the possibility to identify and thus avoid possible concern associated with the chemical structures.

During the years, thanks to a number of projects, mostly funded by the European Commission, VEGAHUB extended its targets and addressed not only QSAR but also read across, weight of evidence, exposure, and other targets which are under development. Offering more and more tools, VEGAHUB wants to provide a larger basis, not only to the prediction but also to the understanding and reasoning about the causes of the adverse effects. This is useful to increase the confidence of the overall assessment and to reduce the uncertainty.

VEGAHUB is not the only system offering predictions and models. The user should be aware of the limitations that still exist using VEGAHUB models, and that he/she can try other models.

We showed how VEGAHUB should be used to extract at the best all the different lines of evidence. The final decision is up to the expert. However, it would be a pity not to exploit all the data given by VEGAHUB, which is not only the predicted value but also a series of other elements, such as the similar chemicals and the rules for reasoning.

VEGAHUB is committed to improve the predictions and the way to inform the user about factors associated with the effect. This goes through a deeper integration of multiple tools, which today have to be run separately. Having multiple tools within the same platform will facilitate exploiting the results, establishing links between the different results.

Integration is a key word for the future improvements within VEGAHUB. Some examples are the integration between read across and QSAR, the integration between hazard and exposure, the integration between multiple properties, and the inclusion of toxicokinetics and toxicodynamic.

Eventually, VEGAHUB will be not only a prognostic and diagnostic system but also a therapeutic one, able to put remedies to the risky substance and identify safer, greener, candidates for substitution.

References

- Virtual models for property evaluation of chemicals within a global architecture (VEGA). www.vegahub.eu
- Benfenati E (ed) (2007) Quantitative Structure-Activity Relationships (QSAR) for pesticide regulatory purposes. Elsevier, Amsterdam
- CAESAR project. <http://www.caesar-project.eu/>

4. Cassano A, Manganaro A, Martin TM et al (2010) The CAESAR models for developmental toxicity. *Chem Cent J* 4(Suppl 1):S4
5. Fjodorova N, Vrachko M, Novich M et al (2010) New public QSAR model for carcinogenicity. *Chem Cent J* 4(Suppl 1):S3
6. Lombardo A, Roncaglioni A, Boriani E et al (2010) Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem Cent J* 4(Suppl 1):S1
7. Benfenati E (2010) The CAESAR project for in silico models for the REACH legislation. *Chem Cent J* 4(Suppl 1):I1
8. Chaudhry Q, Piclin N, Cotterill J et al (2010) Global QSAR models of skin sensitizers for regulatory purposes. *Chem Cent J* 4(Suppl 1):S5
9. Ferrari T, Gini G (2010) An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem Cent J* 4(Suppl 1):S2
10. Toxicity Estimation Software Tool (TEST). <https://www.epa.gov/chemical-research/forms/contact-us-about-toxicity-estimation-software-tool-test>
11. EPI Suite – Estimation Program Interface. <https://www.epa.gov/tsc-screening-tools/epi-suitetm-estimation-program-interface>
12. Toxtree. <http://toxtree.sourceforge.net/>
13. Alternative Non-Testing methods Assessed for REACH Substances (ANTARES), LIFE08 ENV/IT/000435. <http://www.antares-life.eu/>
14. Chemical Assessment according to Legislation Enhancing the In silico Documentation and Safe use (CALEIDOS), LIFE11 ENV/IT/000295. <http://www.life-caleidos.eu/>
15. Promoting the use of in silico methods in industry (LIFE PROSIL), LIFE12 ENV/IT/000154. <http://www.life-prosil.eu/>
16. Integrating VEGA, ToxRead, MERLIN-Expo and ERICA in a platform for risk assessment and substitution for risky substances (LIFE-VERMEER), LIFE16ENV/IT7000167. <https://www.life-vermeer.eu/>
17. Concerting experimental data and in silico models for REACH (LIFE CONCERT REACH). LIFE17 GIE/IT/000461. http://ec.europa.eu/environment/life/project/Projects/index.cfm?fuseaction=search.dspPage&n_proj_id=6791
18. Gini G, Franchi AM, Manganaro A et al (2014) ToxRead: a tool to assist in read across and its use to assess mutagenicity of chemicals. *SAR QSAR Environ Res* 25:999–1011
19. Hardy A, Benford D, Halldorsson T et al (2017) Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA J* 15:4971. <https://doi.org/10.2903/j.efsa.2017.4971>
20. Floris M, Manganaro A, Nicolotti O et al (2014) A generalizable definition of chemical similarity for read-across. *J Chem* 6:39
21. Benfenati E, Belli M, Borges T et al (2016) Results of a round-robin exercise on read-across. *SAR QSAR Environ Res* 27:371–384
22. PROMETHEUS. <https://www.vegahub.eu/portfolio-item/prometheus/>
23. Pizzo F, Lombardo A, Manganaro A et al (2016) Integrated in silico strategy for PBT assessment and prioritization under REACH. *Environ Res* 151:478–492
24. SARPy. <http://sarpy.sourceforge.net/>
25. Lombardo A, Pizzo F, Benfenati E et al (2014) A new in silico classification model for ready biodegradability, based on molecular fragments. *Chemosphere* 108:10–16
26. Pizzo F, Lombardo A, Brandt M et al (2016) A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH. *Environ Int* 88:50–260
27. Ferrari T, Lombardo A, Benfenati E (2018) QSARPy: a new flexible algorithm to generate QSAR models based on dissimilarities. The log Kow case study. *Sci Total Environ* 637–638:1158–1165
28. CORAL. <http://www.insilico.eu/coral/SOFTWARECORAL.html>
29. Toropov AA, Toropova AP, Marzo M et al (2017) QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environ Toxicol Pharmacol* 53:158–163
30. Toropov AA, Toropova AP, Cappelli CI, Benfenati E (2015) CORAL: model for octanol/water partition coefficient. *Fluid Phase Equilib* 397:44–49
31. Toropova AP, Toropov AA, Martyanov SE et al (2013) CORAL: Monte Carlo method as a tool for the prediction of the bioconcentration factor of industrial pollutants. *Mol Inform* 32:145–154
32. Toropov AA, Toropova AP, Benfenati E et al (2013) CORAL: QSPR model of water solubility based on local and global SMILES attributes. *Chemosphere* 90:877–880
33. Toropova AP, Toropov AA, Benfenati E et al (2012) The minimum number of “eccentric” substances: quantitative criterion to estimate the reliability of a QSPR. A case of water solubility. *Chem Phys Lett* 542:134–137

34. Toropov AA, Toropova AP, Lombardo A et al (2012) CORAL: the prediction of biodegradation of organic compounds with optimal SMILES-based descriptors. *Cent Eur J Chem* 10:1042–1048
35. Toropova AP, Toropov AA, Lombardo A et al (2012) CORAL: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas*). *J Comput Chem* 33:1218–1223
36. Toropova AP, Toropov AA, Benfenati E, Gini G (2012) QSAR models for toxicity of organic substances to *Daphnia magna* built up by using the CORAL freeware. *Chem Biol Drug Des* 79:332–338
37. Toropov AA, Toropova AP, Gonella Diaz R et al (2012) SMILES-based optimal descriptors: QSAR modeling of estrogen receptor binding affinity by correlation balance. *Struct Chem* 23:529–544
38. Toropova AP, Toropov AA, Martyanov SE et al (2012) CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*. *Chemosphere* 110:177–181
39. MERLIN-Expo. <https://merlin-expo.eu/>
40. OpenFoodTox. <https://www.efsa.europa.eu/en/microstrategy/openfoodtox>
41. Collaborative Estrogen Receptor Activity Prediction Project (CERAPP). <https://www.epa.gov/chemical-research/cerapp-collaborative-estrogen-receptor-activity-prediction-project-0>
42. Collaborative Modeling Project for Androgen Receptor (AR) Activity (CoMPARA). https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=341701&Lab=NCCT



Chapter 31

Enalos Cloud Platform: Nanoinformatics and Cheminformatics Tools

Dimitra-Danai Varsou, Andreas Tsoumanis, Antreas Afantitis, and Georgia Melagraki

Abstract

In this chapter, we present and discuss Enalos Cloud Platform designed and developed by NovaMechanics Ltd., as an easy-to-use portal to address a variety of challenges arising in the fields of cheminformatics and nanoinformatics. Enalos Cloud Platform also hosts predictive models as web services that can contribute to different aspects of material design and development, drug discovery, virtual screening of chemical substances, nanosafety, and the development of safe-by-design (nano)materials. All models included are developed and validated according to the OECD principles. The web services' interface is carefully designed with the aim of being simple and user-friendly, to allow also users with no informatics background to easily use the models and benefit from the produced predictions and results. At the end of the chapter, we aspire that readers will perceive the functionalities and the efficiency of the available web services and how these could be integrated in drug discovery or material design projects.

Key words Cheminformatics, Nanoinformatics, Enalos Cloud Platform, Predictive models, Virtual screening, Safe-by-design

1 Introduction

Drug discovery and (nano)material risk assessment need to be accelerated to pace with the vast number of compounds that can be developed and the new materials and structures that are included in everyday consumer products. Especially in the field of drug design, high-throughput screening techniques (HTS) produce significant amount of information that cannot be processed manually [1, 2]. At the same time, the ability to translate, with the aid of a computational software, the structural features into molecular descriptors that can be treated mathematically enables scientists to correlate the structure of molecules to their specific properties and thus to understand the mechanisms that control their biological behavior [3, 4].

The abundance of both experimental and calculated data actuated the field of cheminformatics that combines various computational strategies and deals with large datasets, to relate the structural characteristics of chemical compounds to their properties [5]. Quantitative structure-activity relationship (QSAR) models were developed especially to meet the needs of novel drug design; virtual screening studies can be performed using the developed predictive models in order to prioritize candidate compounds for experimental toxicity assessment [6–8]. In this way, experimental labor-intensive processes can be reduced, as well as resources and time constraints may be avoided, and the development of novel therapies for the treatment of various diseases can be accelerated.

The know-how acquired during the past decades in the field of cheminformatics can be applied in the emerging field of nanoinformatics, where similar to the risk assessment of chemical compounds, scientists are focused on the risk assessment of novel nanostructures that are produced and used in a wide range of industrial applications and consumer products [9, 10]. However, even after several years of advancing computational methods dedicated to assess the behavior of engineered nanomaterials (ENMs), the scarcity of experimental evaluation and property quantification impedes the development of robust models for the estimation of possible risks of ENMs for human health and the environment [11].

Due to the lack of sufficient experimental data and due to the ethics impediments considering the use of laboratory animals, the nanosafety community developed alternative *in silico* techniques for toxicity estimation of ENMs [12]. Read-across non-testing strategies for small ENM dataset hazard estimation were introduced recently by the European Chemicals Agency (ECHA) and are also supported by the Organisation for Economic Co-operation and Development (OECD) [13, 14]. These strategies include the organization of similar ENMs into groups and the toxicity (or query property estimation) within the group.

Beyond the development of a predictive model to solve emerging problems within the field of cheminformatics and nanoinformatics, it is equally important to disseminate the model produced to all interested parties, including academia, industry, and regulators, through a user-friendly environment that allows reliable predictions in minimum steps required and without the need of special computational skills [15–19].

In this chapter, Enalos Cloud Platform (<http://enaloscloud.novamechanics.com/>) is introduced, as an online toxicity and drug discovery platform that is publicly available for any interested user. Enalos Cloud Platform hosts a series of predictive models released as web services that aim to address the needs of acquiring fast and accurate predictions of the toxicity and properties of novel compounds and ENMs, reducing at the same time the cost and the time

spent in experimental procedures [18]. The integrated models are built on reliable open-source software, including KNIME Platform, WEKA, ImageJ, R, etc., as well as proprietary software (Enalos+ nodes) [20–24]. In this chapter, three web services hosted in Enalos Cloud Platform are presented; both the modeling development and the corresponding web service offered are discussed.

The web services presented address different problems spanning from reliable toxicity prediction for small molecules, cell association prediction for nanomaterials based on corona fingerprints, and a safe-by-design tool for carbon nanotubes (CNTs).

2 MouseTox

Toxicity assessment of the novel compounds is a very important step in the design of novel drugs as it is crucial to filter out toxic compound as early as possible within the drug discovery process. Especially when a large pool of compounds is available, resources and time can be spared by using computational tools that could indicate the most appropriate candidate compounds for the development of a specific drug. In this direction, the MouseTox quantitative structure-toxicity relationship (QSTR) model was integrated within Enalos Cloud Platform (<http://enaloscloud.novamechanics.com/EnalosWebApps/MouseTox/>), as an online tool that can be used for the estimation of the cytotoxicity of chemical compounds to NIH/3T3 (mouse embryonic fibroblast) cells [15].

The initial model was developed in the KNIME Platform using a dataset of 5416 compounds used in the toxicity assessment of potential drugs for the Chagas disease treatment [25, 26]. Each compound was labeled as “active” or “inactive” depending on the presence or the absence of cytotoxic effects to NIH/3T3 cells. For each compound, a set of molecular descriptors was calculated using Mold2 software incorporated in KNIME through Enalos+ nodes [22, 27]. In this way, the structural characteristics of the compounds were encoded into numerical values that could be used during modeling. The dataset was divided into training and test sets, and variable selection was performed in the training set in order to filter out noisy descriptors to the endpoint. Random forest modeling methodology was selected as the modeling methodology, given that it produced the most reliable predictions when applied to the external test set [28]. The QSTR model was fully validated before its public release via Enalos Cloud Platform (internal, validation, external validation, Y-randomization test), and the reported accuracy on the test set was up to 0.83 [15]. The web service is a ready-to-use application with the purpose to facilitate decision-making, as part of a safe-by-design framework for novel drugs.

MouseTox:Prediction of small molecules cytotoxic effect to NIH/3T3 cells through Enalos Cloud Platform



Fig. 1 MouseTox environment in Enalos Cloud Platform. At the left-handed side, the molecular drawing tool is found. At the top right-handed side, the SMILES input form can be seen followed by the option of importing an SDF file

2.1 Initiating the Analysis

In a virtual screening framework for the assessment of the cytotoxic effects of a range of novel compounds, users can initiate a prediction within the MouseTox web service (Fig. 1) by uploading the query structures to acquire toxicity predictions in minimum time required.

Different options are available in order to submit a structure for prediction that are briefly described below:

Compound Sketcher: *Users can design the chemical compound of interest (one structure at a time) using the provided chemical sketcher. The tool offers a variety of specific atoms, bonds, or substructures that can facilitate the design of the query compound.*

SMILES: *The query compounds can be uploaded using their SMILES notation. In case that the SMILES notation is not initially known, the aforementioned chemical sketcher gives the users the opportunity to draw the molecular structure and then copy the structure as SMILES within the corresponding field. This facilitates the generation of several structures, by allowing a multitude of modifications to be visualized and performed using the sketcher, so that a prediction for the whole set of produced structures is obtained.*

SDF File: *Users can select and import an SDF file with several structures that can be easily extracted from PubChem database or other repositories. This type of file contains molecular structure records, used as a standard exchange format for chemical information.*

After uploading one or several query compounds, toxicity predictions are produced by clicking on the *Execute* button, which can be found under the fields for data input.

Download files		
Row ID	Prediction (Class)	Prediction
"Row0"	"active"	"reliable"
"Row1"	"active"	"reliable"
"Row2"	"active"	"reliable"
"Row3"	"active"	"reliable"
"Row4"	"inactive"	"reliable"
"Row5"	"inactive"	"reliable"
"Row6"	"inactive"	"reliable"
"Row7"	"active"	"reliable"
"Row8"	"active"	"reliable"
"Row9"	"active"	"reliable"
"Row10"	"active"	"reliable"
"Row11"	"active"	"reliable"
"Row12"	"active"	"reliable"
"Row13"	"active"	"reliable"
"Row14"	"active"	"reliable"
"Row15"	"active"	"reliable"
"Row16"	"active"	"reliable"
"Row17"	"active"	"reliable"
"Row18"	"active"	"reliable"
"Row19"	"active"	"reliable"
"Row20"	"active"	"reliable"

Fig. 2 MouseTox-generated output page. The first column of the result table contains the compound's identification, the second column contains the prediction of each submitted compound, and the third column contains the reliability of each prediction based on the model's domain of applicability. This table can also be downloaded in CSV and HTML format by clicking on the corresponding button

2.2 Produced Results

For each set of submitted compounds, the results include the predicted cytotoxicity effect ("active"/"inactive") to NIH/3T3 cells and an indication of whether this prediction could be considered reliable based on the domain of applicability of the model (Fig. 2). Two options are available: the "reliable" option which indicates a prediction within the domain of applicability limits of the model and the "unreliable" option which is a warning for a prediction out of the model's domain of applicability.

By clicking on the *Download files* button, the table with the results can be downloaded in CSV and HTML formats.

3 A Safe-by-Design Tool for Functionalized Nanomaterials

Enalos Cloud Platform has recently incorporated the concept of read-across supported by both OECD and ECHA [29]. This concept is founded on the empirical knowledge that similar materials may exhibit similar behavior; therefore, the assessment of the properties of non-tested ENMs can be achieved using data within a group of similar tested ENMs [30, 31]. The *k*-nearest neighbors (*kNN*) algorithm belongs to the read-across approaches as through the training-testing procedures, small groups of *k* similar ENMs are

formed, and the prediction for each query ENM is performed within the corresponding group [32, 33].

A read-across-based tool integrated in the Enalos Cloud Platform is a safe-by-design workflow correlating molecular descriptors of the decorating molecules of functionalized multi-walled carbon nanotubes (MWCNTs) to their biological activity (protein binding of carbonic anhydrase) and toxicity (<http://enaloscloud.novamechanics.com/EnalosWebApps/CNT/>) [16]. The workflow was developed in the open-source KNIME including the Enalos+ nodes, using a dataset of 83 surface-modified MWCNTs [34]. Considering that all MWCNTs had an identical core, the assumption that the differences in their biological effects and toxicity were mostly due to the decorating molecules of their surface was made [35–37]. The chemical structure information of the ligands was quantified by calculating the necessary molecular descriptors using Mold2 software [27]. For each of the two endpoints, the variables that were the most critical for modeling purposes were selected, and the k NN method was employed with an optimal value of $k = 3$ for the CA-binding model and $k = 7$ for the toxicity model. The developed workflow was fully validated according to the OECD standards before it was released online via the Enalos Cloud Platform. A double-cross validation scheme was applied, and the reported predictive accuracy for the blank-external sets was over 0.8 for both models [16, 38]. The web service is a user-friendly application whose purpose is to facilitate decision-making, as part of a safe-by-design framework for novel MWCNTs.

3.1 Initiating the Analysis

For both CA-binding and toxicity profile estimation of a query-decorated MWCNT, users must provide through the platform one or several structures of compounds being considered as potential decorating molecules (Fig. 3) similar to MouseTox web service. After the structure input, descriptors are automatically calculated, and predictions are produced without any requirement of additional metadata.

Different options are available in order to submit a structure for prediction that are briefly described below:

Compound Sketcher: *Users can submit a potential decorator to the platform by drawing the molecular structure of interest using the provided chemical sketcher. The functionality enables the users to construct the decorating molecule using different tools that provide specific atoms, bonds, or substructures.*

SMILES: *Users can provide a list of the potential decorators using their SMILES notation.*

SDF File: *Users can upload the potential decorators as a list of structures in an SDF file that can be extracted from PubChem database or other repositories.*

Enalos Nanoinformatics Cloud Platform: A Safe-by-Design Tool for Functionalised Nanomaterials

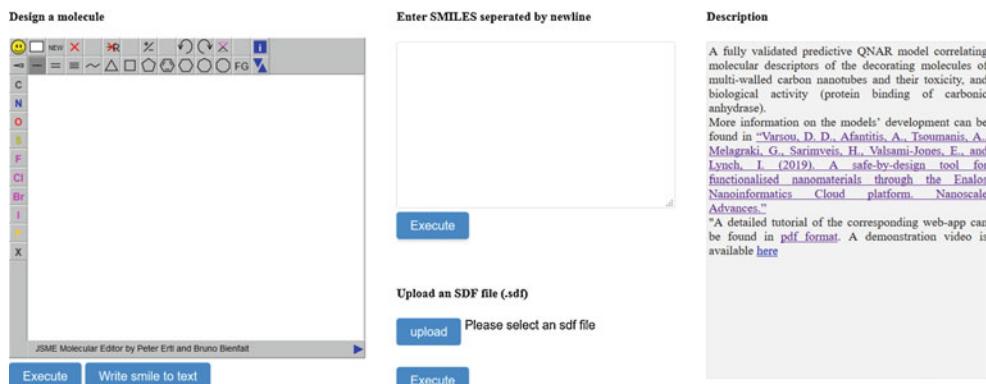


Fig. 3 Enalos Nanoinformatics Cloud Platform's user-friendly interface for MWCNT biological and toxicity assessment. At the left-handed side, the molecular drawing tool is found. At the top right-handed side, the SMILES input form can be seen followed by the option of importing an SDF file

Download files
[Knime report](#) powered by **Birt**

Safe-by-Design: Functionalised Nanomaterials

"Toxicity"	"Domain (Toxicity)"	"Activity"	"Domain (Activity)"
toxic	reliable	non-binder	reliable
non-toxic	reliable	binder	reliable
toxic	reliable	non-binder	reliable
toxic	unreliable	binder	reliable

Date: Apr 2, 2019 5:43 PM Author: NovaMechanics Ltd 1 of 1
www.knime.com

Fig. 4 Generated output page. The first column of the results table contains the toxicity prediction for each submitted ligand, and the second column contains the reliability of each prediction based on the model's domain of applicability. Similarly, the third and the fourth columns contain the CA-binding activity prediction and its reliability, respectively. This table can also be downloaded in CSV and HTML format by clicking on the corresponding button

By clicking on the *Execute* button, which corresponds each time to the used field for data input, predictions are performed and are displayed in a new page.

3.2 Produced Results

For each set of submitted decorating molecules, the results include the predicted CA-binding class ("binder"/"non-binder") to the MWCNTs and the toxicity class ("toxic"/"nontoxic") of the resultant-decorated MWCNTs and an indication of whether this prediction could be considered reliable based on the domain of applicability of the models (Fig. 4).

By clicking on the *Download files* button, the table can be downloaded, and in the downloaded files, the interested users can observe the neighbors of the training set used for the prediction of each one of the input compounds.

4 Protein Corona Fingerprints Tool for the Virtual Screening of Gold Nanoparticle Cellular Association

The composition of the protein corona that is formed in the surface of nanoparticles (NPs) that interact with biological media can be a source of valuable information concerning the biological mechanisms that are activated when the NPs are exposed to biological fluids, as well as the future interaction of the NPs with cells and organisms [39, 40]. As the so-called protein corona fingerprints contain more relevant information to biological endpoints than other NP physicochemical descriptors, they have already been used in the model development for the prediction of the interaction between cells and NPs [19, 41, 42]. In this section, a *k*NN model developed for the virtual screening of gold NP cellular association, which has also been included in Enalos Cloud Platform (<http://enalos.insilicotox.com/NanoProteinCorona/>), will be presented.

The initial model was developed using a validated dataset of 105 chemically diverse gold NPs with different surface coatings. For each NP, various physicochemical and biological descriptors were available (in total 805 parameters), as well as a cell association index (normalized log₂ values) measured when the gold NPs were incubated with A549 human lung epithelial carcinoma cells [41]. The initial dataset was filtered and randomly divided for validation purposes into training and test sets. The training set was used for variable selection and modeling purposes, and the test set was used in order to measure the robustness of the produced model. The *k*NN modeling strategy with an optimal value of *k* = 6 was used in order to correlate the selected variables to the cellular association. The proposed model was fully validated (reported external *R*² = 0.832), and the domain of applicability limits was also calculated [19].

4.1 Initiating the Analysis

For the virtual screening of the query gold NPs, users must provide the three physicochemical descriptors measured in serum (Z-average hydrodynamic diameter, zeta potential (mV), localized surface plasmon resonance index) and ten protein spectral counts (UNIPROT IDs: P01024, P02766, P08697, P19823, Q13103, Q9UK55, P02788, P02775, P14625, Q96KN2) as determined from the adsorbed corona from the human serum, and predictions of their cellular association will be produced in just a few clicks.

 Enalos Nano Protein Corona Platform

MNP Number	Z-AHD*	ZP**	LSPR***	P01024	P02766	P08697	P19823	Q13103	Q9UK55	P02788	P02775	P14625	Q96KN2
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													

* w/ serum Z-Average Hydrodynamic Diameter
** w/ serum Zeta Potential (mV)
*** w/ serum (AU) Localized Surface Plasmon Resonance (LSPR) index

Import a CSV file for High Throughput Virtual Screening (.csv)

No file selected.

Fig. 5 Enalos Cloud Platform protein corona model interface. At the top of the page the online form where the descriptor values can be filled in is presented, followed by the option of importing a CSV file

Two different options are available in order to submit a gold NP for prediction that are briefly described below:

Online Form: *Users can fill the provided form in the top of the interface (Fig. 5), by typing the corresponding numerical values for the necessary descriptors and protein spectral counts. The form can be used for up to 20 entries.*

CSV File: *In case that a large dataset of gold NPs (more than 20) is available for virtual screening, users are encouraged to provide the necessary information by uploading a CSV file containing this information.*

By clicking on the *Submit* button, which corresponds each time to the used field for data input, predictions are performed and are displayed in a new page.

4.2 Produced Results

For each one of the submitted gold NPs, the results include the predicted log₂ of the cellular association (in mL/µg (Mg)) and an indication of whether this prediction could be considered reliable based on the domain of applicability of the model (Fig. 6).

Cellular Interaction of Gold Nanoparticles Prediction Through Protein Corona Fingerprints

[Knime report](#) powered by [Birt](#)

"Id"	"log2 mL/ug(Mg)"	"Domain"
1	-2.374	unreliable
2	-2.507	unreliable
3	-5.353	unreliable
4	-5.916	unreliable
5	-2.09	unreliable

Date: Jul 18, 2019 6:13 PM

Author: NovaMechanics Ltd

1 of 1

www.knime.org

Fig. 6 Generated output page. The first column contains the IDs of the gold NPs, the second column contains the predicted normalized log2 cell association value, and the third column contains the reliability of each prediction based on the model's domain of applicability. This table can also be downloaded in CSV format

5 Conclusions

Current needs in both drug and material design render the development of robust and reliable in silico models inevitable. Models and tools developed can greatly underpin the efforts for assessing the risk of chemical compounds and ENMs, reducing the time and resources spent in experimental activities. However, while several predictive models have been built for assessing the biological activity and toxic side effects of small molecules and ENMs, these remain unexploited by the wider community as the developed predictive models have not been properly disseminated. All models developed should be integrated within a simple and user-friendly environment to reach all interested users and facilitate decision-making.

Enalos Cloud Platform addresses exactly this need for a user-friendly interface that can produce in few steps toxicity predictions and property calculations for chemical structures or ENMs. All web services presented (MouseTox, corona model, and MWCNTs' safe-by-design tool) are some of the tools and models offered as web services through Enalos Cloud Platform. Predictions are performed shortly after data input and are accompanied always by an indication of their reliability based on the results of the fully validated models running in the background. The produced results can be downloaded for further analysis and exploration contributing in this way, among others, in the understanding of activity mechanisms and read-across similarities.

Acknowledgments

This work was supported by the Cyprus Research Promotion Foundation, the Republic of Cyprus & the European Union under Grant agreement KOINA/ERASysAPP-ERA.NET/1113 and the European Union's Horizon 2020 research and innovation programme under grant agreements No 691095 (NANOGEN-TOOLS) & 731032 (NanoCommons).

References

1. Willett P (2002) Chemistry plans a structural overhaul The rising tide of data being generated by high-throughput. *Nature* 419:4–7
2. Melagraki G, Afantitis A, Sarimveis H et al (2006) A novel RBF neural network training methodology to predict toxicity to *Vibrio fischeri*. *Mol Divers* 10:213–221
3. Hong H, Xie Q, Ge W et al (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* 48:1337–1344
4. Mauri A, Consonni V, Pavan M et al (2006) Dragon software: an easy approach to molecular descriptor calculations. *Match* 56:237–248
5. Leach AR, Gillet VJ (2007) An introduction to chemoinformatics. Springer Netherlands, Dordrecht
6. Melagraki G, Afantitis A, Sarimveis H et al (2010) In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives. *Chem Biol Drug Des* 76:397
7. Tetko IV, Maran U, Tropsha A (2017) Public (Q)SAR services, integrated modeling environments, and model repositories on the web: state of the art and perspectives for future development. *Mol Inf* 36:1–14
8. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inf* 29:476–488
9. Gajewicz A, Rasulev B, Dinadayalan TC et al (2012) Advancing risk assessment of engineered nanomaterials: application of computational approaches. *Adv Drug Deliv Rev* 64:1663–1693
10. Winkler DA, Mombelli E, Pietrojasti A, et al (2013) Applying quantitative structure – activity relationship approaches to nanotoxicology: current status and future potential. <https://doi.org/10.1016/j.tox.2012.11.005>
11. Gajewicz A, Jagiello K, Cronin MTD et al (2017) Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environ Sci Nano* 4:346–358
12. Varsou D-D, Afantitis A, Melagraki G, et al (2019) Read-across predictions of nanoparticle hazard endpoints: a mathematical optimization approach. *Nanoscale Adv* 1:3485–3498
13. ECHA (2017) Appendix R. 6-1: recommendations for nanomaterials applicable to the guidance on QSARs and grouping 29
14. Schultz TW, Amcoff P, Berggren E et al (2015) A strategy for structuring and reporting a read-across prediction of toxicity. *Regul Toxicol Pharmacol* 72:586–601
15. Varsou D-D, Melagraki G, Sarimveis H et al (2017) MouseTox: an online toxicity assessment tool for small molecules through Enalos Cloud platform. *Food Chem Toxicol* 110:83–93
16. Varsou D-D, Afantitis A, Tsoumanis A et al (2019) A safe-by-design tool for functionalised nanomaterials through the Enalos Nanoinformatics Cloud platform. *Nanoscale Adv* 1:706
17. Braga RC, Alves VM, Muratov EN et al (2017) Pred-skin: a fast and reliable web application to assess skin sensitization effect of chemicals. *J Chem Inf Model* 57:1013–1017
18. Melagraki G, Afantitis A (2014) Enalos InSilicoNano platform: an online decision support tool for the design and virtual screening of nanoparticles. *RSC Adv* 4:50713–50725
19. Afantitis A, Melagraki G, Tsoumanis A et al (2018) A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints. *Nanotoxicology* 12:1148
20. KNIME KNIME Analytics Platform. <https://www.knime.org/knime-analytics-platform>
21. Abràmoff MD, Magalhães PJ, Ram SJ (2004) Image processing with ImageJ Second Edition. *Biophotonics Int* 11:36–42
22. Leonis G, Melagraki G, Afantitis A (2016) Open Source Chemoinformatics Software

- including KNIME Analytics Platform among a multitude. In: Leszczynski J (ed) *Handbook of computational chemistry*. Springer, Dordrecht
23. The University of Waikato Weka 3: machine learning software in Java. <https://www.cs.waikato.ac.nz/ml/weka/index.html>
 24. The R Project for statistical computing. <https://www.r-project.org/>
 25. National Center for Biotechnology Information PubChem BioAssay Database, AID=651744. <https://pubchem.ncbi.nlm.nih.gov/bioassay/651744>
 26. World Health Organisation WHO Chagas disease (American trypanosomiasis) Factsheet. <http://www.who.int/mediacentre/factsheets/fs340/en/>
 27. U.S. Food and Drug Administration, Mold2-Free software for fast-calculating descriptors from a two-dimensional chemical structure that is suitable for small and large datasets. <https://www.fda.gov/science-research/bioinformatics-tools/mold2>
 28. Witten IH, Frank E, Hall MA (2011) Data mining practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann Publishers, Burlington
 29. Mech A, Rasmussen K, Jantunen P et al (2019) Insights into possibilities for grouping and read-across for nanomaterials in EU chemicals legislation. *Nanotoxicology* 13:119–141
 30. Oomen AG, Bleeker EAJ, Bos PMJ et al (2015) Grouping and read-across approaches for risk assessment of nanomaterials. *Int J Environ Res Public Health* 12:13415–13434
 31. Lamon L, Aschberger K, Asturiol D et al (2019) Grouping of nanomaterials to read-across hazard endpoints: a review. *Nanotoxicology* 13:100–118
 32. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, United States
 33. Huluban R (2016) Practical guide-How to use and report (Q)SARs Practical Guide – How to use and report (Q)SARs, version 3.1. European Chemicals Agency, Helsinki
 34. Zhou H, Mu Q, Gao N et al (2008) A nano-combinatorial library strategy for the discovery of nanotubes with reduced protein-binding, cytotoxicity, and immune response. *Nano Lett* 8:859–865
 35. Chau YT, Yap CW (2012) Quantitative nanostructure-activity relationship modelling of nanoparticles. *RSC Adv* 2:8489–8496
 36. Toropov AA, Toropova AP, Puzyn T et al (2013) QSAR as a random event: modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* 92:31–37
 37. Kar S, Gajewicz A, Puzyn T et al (2014) Nano-quantitative structure-activity relationship modeling using easily computable and interpretable descriptors for uptake of magneto-fluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicol In Vitro* 28:600–606
 38. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intel Lab Syst*, Elsevier, 159:108
 39. Vilanova O, Mittag JJ, Kelly PM et al (2016) Understanding the kinetics of protein-nanoparticle corona formation. *ACS Nano* 10:10842–10850
 40. Cedervall T, Lynch I, Lindman S et al (2007) Understanding the nanoparticle–protein corona using methods to quantify exchange rates and affinities of proteins for nanoparticles. *Proc Natl Acad Sci* 104:2050–2055
 41. Walkey CD, Olsen JB, Song F et al (2014) Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* 8:2439–2455
 42. Varsou D-D, Tsiliki G, Nymark P et al (2018) toxFlow: a web-based application for read-across toxicity prediction using omics and physicochemical data. *J Chem Inf Model* 58:543–549



Chapter 32

alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints

Andrea Mauri

Abstract

In this chapter we will present alvaDesc, a software to calculate and analyze molecular descriptors and fingerprints.

Molecular descriptors and fingerprints play an essential role in quantitative structure-activity relationships (QSAR) as they are the mathematical representation of chemicals and they serve as the input for the data analysis methods used to build QSAR models.

The increasing number of newly proposed molecular descriptors and fingerprints and generally the attention paid by the scientific community to the development of novel methodologies to represent chemical structures are evidences of the relevance of these representations in the prediction of chemical properties.

Despite the complexity of dealing with a high number of variables, different types of molecular descriptors and fingerprints can highlight specific traits of molecular structures. These aspects, together with the increased availability of chemical data and methods for data analysis, are some of the challenges that researchers face in the development of QSAR models.

Key words Molecular descriptors, Molecular fingerprints, MACCS keys, Data analysis, Principal component analysis, Correlation analysis, Variable reduction, Software

1 Introduction

In silico methodologies, including quantitative structure-activity relationships (QSARs), have become essential in the analysis of chemicals. Every day new chemicals are synthesized, isolated, marketed, and even just imagined and drawn. New chemicals are constantly added to the CAS REGISTRY database (www.cas.org), and the number of chemical structures and chemical properties stored in online databases (like ChemSpider (www.chemspider.com), PubChem [1, 2], ChEMBL [3, 4], and ZINC database [5]) increases constantly.

This huge number of chemicals made the evaluation of experimental properties (physicochemical and ecotoxicological) an unfeasible task. Fortunately, the research on in silico molecule

representation, as well as the improvement in the methodology for data analysis and model building, provides researchers with the tools needed to implement QSAR models that can be used as an alternative to expensive experimental procedures.

Another incentive for the use of QSAR models is the introduction of regulations on the registration, evaluation, authorization, and restriction of chemicals. As an example, REACH in Europe [6] promotes the use of QSAR methodologies to predict unavailable experimental data.

Indeed, QSAR has been fruitfully applied to predict diverse chemical properties. QSAR models have been proposed for the determination of ecotoxicological properties of *Daphnia magna* [7–9], fish, and algae [10–12], as well as for chemical prioritization [13]. In a recent paper, QSAR has been applied to model toxicity data of endocrine disruptor chemicals (EDCs) on 14 different species [14].

The large amount of available data is essential, but once the data has been collected and curated, it is necessary to transform the structural information included in the chemical files to number values that can be used as the input for model building [15].

Those numbers, which are the results of mathematical manipulation of chemical structures, are the molecular descriptors and fingerprints. Even if a molecular fingerprint is not a simple value, in any case, it is a mathematical representation of a chemical structure [16, 17].

Many libraries, toolkits, and software are available for the calculation of molecular descriptors, among them Mordred [18], Padel-Descriptor [19], CDK [20, 21], RDKit [22], and Dragon [23]. Even if all these tools provide different functionalities and a variable number of molecular descriptors and fingerprints, all of them have the final goal of providing as many numerical representations of chemicals as possible.

The alvaDesc software [24] is one of the most recent tools for the calculation of molecular descriptors and fingerprints; it currently includes 5471 descriptors, 5305 of them are divided in 30 logical blocks (Table 1).

Together with the calculation of molecular descriptors, alvaDesc carries out the calculation of three different molecular fingerprints:

- MACCS166 fingerprint [25]
- Extended-connectivity fingerprints [26]
- Path fingerprints

Extended-connectivity fingerprints and path fingerprints can be tuned, not only with respect to the fingerprint size, fragment type, and dimensions, but even by defining atom and bond

Table 1
alvaDesc descriptor logical blocks with the number of descriptors included in each block

Constitutional indices	48	RDF descriptors	210
Ring descriptors	32	3D-MoRSE descriptors	224
Topological indices	79	WHIM descriptors	114
Walk and path counts	46	GETAWAY descriptors	273
Connectivity indices	37	Randić molecular profiles	41
Information indices	50	Functional group counts	154
2D matrix-based descriptors	607	Atom-centered fragments	115
2D autocorrelations	213	Atom-type E-state indices	172
Burden eigenvalues	96	Pharmacophore descriptors	165
P_VSA-like descriptors	55	2D atom pairs	1596
ETA indices	38	3D atom pairs	36
Edge adjacency indices	324	Charge descriptors	15
Geometrical descriptors	38	Molecular properties	20
3D matrix-based descriptors	99	Drug-like indices	28
3D autocorrelations	80	CATS 3D descriptors	300

parameters considered during fragment identifications (e.g., atom-type, aromaticity, the number of attached hydrogens, connectivity).

One of the most relevant features of alvaDesc is its capability to handle both full-connected and non-full-connected molecular structures, e.g., salts, mixtures, ionic liquids, and metal complexes. All the molecular descriptor calculation algorithms have been studied in order to provide different theoretical approaches for the calculation of molecular descriptors on such structures.

In addition to the calculation of descriptors and fingerprints, alvaDesc provides different tools to carry out a first exploration of chemical datasets:

- Molecule structure verification using PubChem services [2]
- Molecule structure visualization, charting, and filtering
- Principal component analysis (PCA) and correlation analysis

Due to its capability of calculating large numbers of molecular descriptors, alvaDesc provides variable reduction tools, including the fast V-WSP (variable reduction method adapted from space-filling designs) [27].

The software is available for all the three major operative systems (macOS, Linux, and Windows), and it is a multithreaded application that is provided both with a graphical and a command

line interface. Its command line interface can be easily integrated with KNIME [28] using the alvaDesc KNIME Plugin.

Additionally, an online version of alvaDesc is available as a service in the Online Chemical Modeling Environment (OCHEM) [29].

2 Molecular Structure Curation and Standardization

A molecule, or even a single fragment, can be represented in different ways; e.g., aromatic rings can be represented either in aromatic (i.e., conjugated bonds) or in Kekulé form (i.e., bond configuration double-single-double). Since the representation can be different, molecular descriptor values can be affected even if the chemical information is identical. Additionally, chemical structures, both retrieved from chemistry publications or from public and commercial databases, are not immune from errors [30, 31].

Molecular descriptors and fingerprints calculation is based on the assumption that the molecular structure on which the mathematical algorithms are applied to is correct, making molecular structure curation and standardization a fundamental step [15].

Since not all tools automatically standardize the molecules, it is the researcher's responsibility to verify it and eventually to carry out this step prior to the calculation of molecular descriptors and fingerprints.

In order to represent a molecule in the same way, independently from the original representation, alvaDesc performs its own standardization procedure which includes the nitro-group standardization, the addition of the implicit hydrogens, and the aromaticity detection. This standardization is performed to get the same internal molecular representation, and therefore the same descriptor values independently from the original representation, i.e., the same molecule represented in a Kekulé or aromatic form will be internally represented as the same molecule.

3 Molecular Descriptors

Molecular descriptors are the results of the application of mathematical functions on a well-defined representation of the chemical graph [17]. During the last decades, thousands of molecular descriptors have been proposed in literature [16] highlighting the interest of the scientific community in this field.

The increasing number of available molecular descriptors required the definition of few basic rules that molecular descriptors should comply with [32, 33]:

1. Be invariant to atom labelling and numbering
2. Be invariant to the molecule roto-translation
3. Be defined by an unambiguous algorithm
4. Have a well-defined applicability on molecular structures

These rules must be accomplished in order to define a molecular descriptor, but they do not guarantee that a molecular descriptor is useful to describe a defined property.

Molecular descriptors can be grouped in multiple ways; one of the most common is considering the information collected from the chemical structures from 0- to 3-dimensional descriptors.

The alvaDesc software calculates a variety of 0-dimensional, 1-dimensional, 2-dimensional, and 3-dimensional descriptors:

- 0-dimensional are those molecular descriptors obtained by a molecule representation that does not consider any information about the atom connections, e.g., molecular weight and atom-type counts.
- 1-dimensional descriptors consider a part and not the whole topology of the chemical structure, e.g., functional group counts, atom-centered fragments, and structural keys.
- 2-dimensional descriptors derive from the 2D representation of a chemical structure as a graph; they include the information about atomic composition and connectivity of atoms in the molecule, e.g., autocorrelation descriptors and topological indices.
- 3-dimensional descriptors are calculated using the 3D representation of the molecular graph, considering not only the connection between atoms but even their position in the 3-dimensional space, e.g., WHIMs (Weighted Holistic Invariant Molecular descriptors) [34], GETAWAYs (Geometry, Topology, and Atom-Weights Assembly descriptors) [35].

Most of the 0-dimensional descriptors, such as sum and average of atomic properties, atom counters, and the cyclomatic number (r), also known as circuit rank, are grouped in alvaDesc in the “Constitutional indices” block.

The generic formula for the sum of atomic properties is the following:

$$S_w = \sum_{i=0}^{|V|} w_i$$

where $|V|$ is the number of atoms included in the molecular structure and w_i the considered atomic property. Molecular weight is a

specific case of molecular descriptors as sum of atomic properties, where w_i is the atomic mass.

The cyclomatic number is defined as:

$$r = |E| - |V| + D$$

where $|E|$ is the number of bonds, $|V|$ the number of atoms, and D is the number of disconnected fragments in the molecular structure. The cyclomatic number r is the minimum number of edges to be removed from a molecular graph in order to remove all its cycles, making it into a forest, i.e., an acyclic graph. The cyclomatic number provides a basic description of ring systems since its value is the cardinality of the Smallest Set of Smallest Rings (SSSR).

Within the 2-dimensional group, there is a plethora of molecular descriptors, but basically, all of them are calculated considering a molecule as a topological graph. The molecular graph can be represented in different ways; typically a matrix is used. The adjacency matrix is the simplest matrix representation of a molecular graph, but also other matrices can be used to represent a graph in order to extract different information. These matrices are the starting point to derive molecular descriptors.

One of the most common topological descriptors is the Wiener index (W). The Wiener index was originally correlated with the boiling point of alkane molecules; it is a topological index defined as the half-sum of the lengths of the shortest paths between all pairs of vertices in the H-depleted chemical graph [36]:

$$W = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} d_{ij}$$

where $|V|$ is the number of atoms and d_{ij} is the topological distance between i -th and j -th atoms.

Analogously Harary [37] and Randić connectivity indices [38, 39] can be calculated as a generalization of the Wiener index formula, where the Harary index is calculated on the reciprocal distance matrix and the Randić connectivity index can be calculated from the χ matrix.

Topological indices, like the Wiener, Harary, and Randić connectivity index, can be derived applying mathematical operators to graph-theoretical matrices. Another approach for the definition of a topological index is the application of mathematical functions to local vertex invariants (LOVIs), where local vertex invariants are those numerical quantities of graph vertices that characterize specific properties of the molecule atoms [17].

Additional examples of 2-dimensional descriptors are the autocorrelation descriptors, like Moreau-Broto [40, 41] and Moran [42] autocorrelation descriptors. Autocorrelation descriptors can be defined using the following general formula:

$$D1_{(L;\alpha,\lambda,k)} = \alpha \cdot \sum_{i=1}^A \sum_{j=1}^A (\mathcal{L}_i \cdot \mathcal{L}_j)_{ij}^{\lambda} \cdot \delta(d_{ij};k)$$

where \mathcal{L} is a generic local vertex invariant; α and λ are a scaling and a power parameter, respectively; and $\delta(d_{ij};k)$ is a Kronecker delta function equal to one for pairs of substructure centers at topological distance $d_{ij} = k$ and zero otherwise; A is the number of substructure centers that typically are the molecule atoms.

The Moreau-Broto autocorrelation descriptors (*ATS*) can be derived from the general equation *D1* simply setting $\alpha = 1/2$ and $\lambda = 1$:

$$ATS_k = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} w_i \cdot w_j \cdot \delta(d_{ij};k)$$

Analogously, Moran autocorrelation descriptors (*MATS*) can be calculated using the following formula, derived from *D1* equation:

$$MATS_k = \frac{\frac{1}{\Delta_k} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} (w_i - \bar{w}) \cdot (w_j - \bar{w}) \cdot \delta(d_{ij};k)}{\frac{1}{|V|} \sum_{i=1}^{|V|} (w_i - \bar{w})^2}$$

where $|V|$ is the set of vertices, w_i and w_j are any atomic property, \bar{w} is the property mean considering the whole molecule, and Δ_k is the sum of the Kronecker deltas, that is, the number of atom pairs at distance equal to k . The $\delta(d_{ij};k)$ is the Kronecker delta as previously defined.

Atom-type autocorrelation (*ATAC*) descriptors are another case of autocorrelation descriptors. Conversely to Moreau-Broto and Moran descriptors, atom-type autocorrelation descriptors are discrete descriptors counting the occurrences of atom pairs at a predefined topological distance k .

Even the formula for the calculation of atom-type autocorrelation descriptors can be derived from the general equation *D1* as the following:

$$ATAC_k(u, v) = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \delta(i;u) \cdot \delta(j;v) \cdot \delta(d_{ij};k)$$

where u and v represent two different atom-types; $\delta(i;u)$, $\delta(j;v)$, and $\delta(d_{ij};k)$ are three Kronecker delta functions equal to one if atom i is of type u and atom j is of type v ; and the topological distance d_{ij} is equal to k and zero otherwise.

Atom-type autocorrelation descriptors can highlight different information simply by changing the atom-type

definition. An atom-type can be defined with the atom symbol (e.g., C, N, O...) or with one or more atomic properties or considering an atom-type as an atom-centered fragment (e.g., -COOH, -NH₃, -NO₂...).

Well-known atom-type autocorrelation descriptors are the CATS 2D (Chemically Advanced Template Search) descriptors [43, 44]. Atom-type definition for the calculation of CATS 2D descriptors is based on the concept of “potential pharmacophore points” (PPP), where a PPP is a generalized atom-type defined considering the atom as belonging to one of the following categories:

1. Hydrogen-bond donor (D)
2. Hydrogen-bond acceptor (A)
3. Positive (P)
4. Negative (N)
5. Lipophilic (L)

Any atom of the molecule can be assigned to none, one, or two atom-types, resulting in 15 possible atom pairs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL). CATS 2D are calculated on varying the topological distance from 0 to 9 obtaining a vector of 150 frequencies.

CATS 2D descriptors, as well as atom pair descriptors, are alignment-free and can be used for fast calculation of similarity even on large databases.

The last considered groups of molecular descriptors are those derived by using the 3-dimensional information of the molecular graph.

An example of a 3-dimensional descriptor is the 3D Wiener index (^{3D}W_H) which is a topographic index calculated by analogy with the Wiener index from the geometrical distance matrix as:

$${}^{3D}W_H = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} r_{ij}$$

where r_{ij} is the Euclidean distance between atoms i and j .

Two more examples of 3-dimensional descriptors are the well-known GETAWAYs (Geometry, Topology, and Atom-Weights AssemblY descriptors) [35] and WHIMs (Weighted Holistic Invariant Molecular descriptors) [34].

GETAWAY descriptors encode information about the role of each atom determining the whole molecule shape and evaluate the interactions among atoms with respect to their geometrical position. One of the matrices used for the calculation of GETAWAY descriptors is the molecular influence matrix \mathbf{H} , which is symmetric matrix, defined as:

$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^T \times \mathbf{M})^{-1} \times \mathbf{M}^T$$

where \mathbf{M} is the molecular matrix of the centered Cartesian coordinates (x, y, z) for a defined conformation. The fact that the Cartesian coordinates are centered grants that GETAWAY descriptors are independent from any alignment.

The GETAWAY descriptors, instead of providing a whole description of the molecular structure, have been developed in order to exploit local information based on the different contributions of atoms.

Conversely to GETAWAY descriptors, WHIM descriptors have been proposed in order to collect holistic information about the spatial distribution of molecule atoms, such as information on 3-dimensional molecular size, shape, symmetry, and atomic property distribution. WHIM descriptors are based on the calculation of eigenvalues and eigenvectors of a weighted covariance matrix of the centered Cartesian atomic coordinates. The WHIM descriptors do not consider the connections among atoms but only their position in the 3-dimensional space.

Moreover alvaDesc provides the calculation of several model-based physicochemical properties such as molar refractivity, topological polar surface area (TPSA) [45, 46], molecular volume estimations, two LogP models (Moriguchi [47] and Ghose-Chippen [48] octanol-water partition coefficient), and a significant list of drug-like and lead-like alerts including the well-known Lipinski alert index [49].

3.1 Molecular Descriptor Analysis and Interpretation

Molecular descriptors are not always easy to interpret. The majority of the descriptors present in literature have not been proposed in order to identify a specific chemical or physicochemical feature, such as the molecular weight or the presence of a well-defined chemical structure, but they are typically a mathematical manipulation of different representations of the chemical graph [16, 17]. Therefore, the resulting descriptors include chemical structure information, such as atom connections, bond orders, branching, and the aromaticity, but cannot necessarily be directly correlated to a chemical or toxicological aspect.

Due to the described complexity, a helpful approach in interpreting molecular descriptors is to study the molecular descriptor behavior within the considered chemical dataset.

The most used technique for the analysis of molecular descriptor values is the principal component analysis (PCA) [50] that can help in the identification of relationships among descriptors and to identify outliers or clusters of similar molecules. Together with principal component analysis, other tools can be used to evaluate relationships among chemicals within a dataset, including the well-known multidimensional scaling [51], and recently a new technique called t-SNE [52] has been proposed. While principal

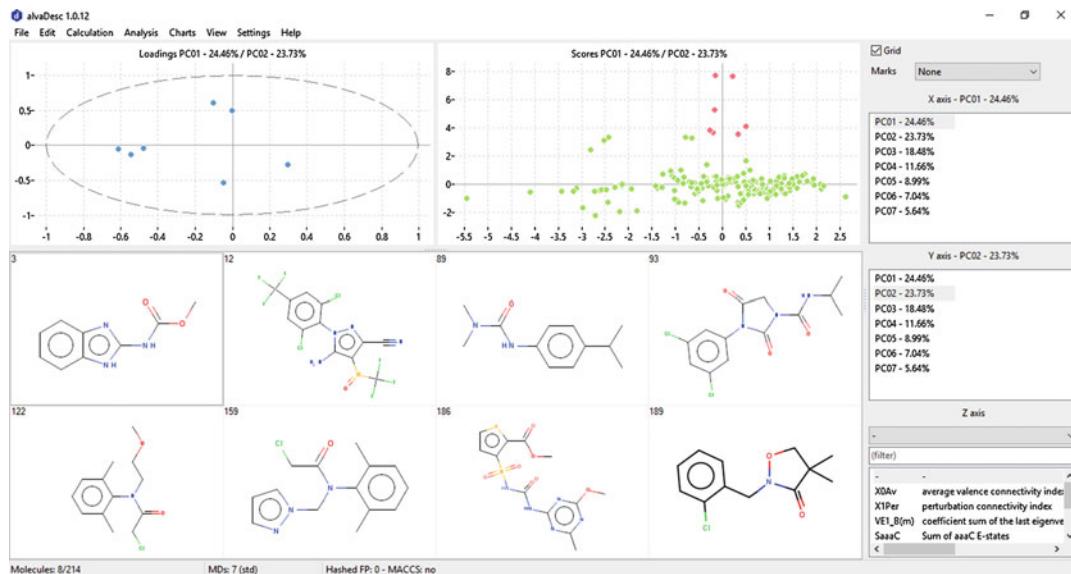


Fig. 1 Principal component analysis performed on BMF dataset [11] using *MlogP2*, *X0Av*, *X1Per*, *SaaaC*, *VE1_B(m)*, *B02[N-O]*, and *B03[N-Cl]* descriptors. Filtered molecules (red on score plot) have the highest *B02[N-O]* and *B03[N-Cl]* values

component analysis focuses on keeping the low-dimensional representation of dissimilar data points far apart, t-SNE tries to keep the low-dimensional representation of very similar data points close together (Fig. 1).

In order to evaluate the molecular descriptors behavior within a defined dataset, in a paper studying the laboratory-based fish bio-magnification factor (BMF) of organic chemicals, Grisoni et al. [11] proposed an ordinary least squares (OLS) model.

The proposed OLS model uses seven descriptors: *MlogP2*, *X0Av*, *X1Per*, *SaaaC*, *VE1_B(m)*, *B02[N-O]*, and *B03[N-Cl]*.

MlogP2 is the squared logarithm of the octanol-water partitioning coefficient (logKOW), as predicted by the Moriguchi model [47]; *B02[N-O]* and *B03[N-Cl]* are atom pair descriptors [53] where *B02[N-O]* is equal to 1 if there is at least 1 pair of N and O atoms separated by 2 bonds and 0 otherwise; and similarly *B03[N-Cl]* is equal to 1 if there is at least 1 pair N-Cl separated by 3 bonds and 0 otherwise.

SaaaC is a little more complex since it is an atom-type electro-topological state (E-state) index [54, 55] calculated as the sum of the E-states for all the carbon atoms in the molecule having three aromatic bonds. It is related to the molecular reactivity of such atom-types; *SaaaC* increases with the number of carbon atoms with three aromatic bonds as well as it increases with their reactivity.

All these three descriptors are quite easily interpretable but what about *X0Av*, *X1Per*, and *VE1_B(m)*?

$X0Av$ [56] is the average valence connectivity index of order 0, $X1Per$ [57] is a perturbation connectivity index, and $VE1_B(m)$ is a 2D matrix-based descriptor [17] obtained from the Burden [58] matrix weighted by mass ($B(m)$).

These three descriptors cannot easily be related to a chemical or toxicological feature, but while analyzing the dataset, the authors made some considerations.

They analyzed the behavior of $X0Av$ finding a relation among descriptor values, number of atoms with many valence electrons, and the number of aromatic and unsaturated bonds.

Similarly, $X1Per$ variability has been associated with the molecular shape and with the presence of heteroatoms and multiple bonds.

Finally, $VE1_B(m)$ variability has been associated with the molecular size and shape and with the presence of heavy heteroatoms and multiple bonds.

3.2 Variable Reduction

Thousands of molecular descriptors have been proposed in literature over the last few decades [16], and software implementing the calculation of molecular descriptors usually includes a number of molecular descriptors varying from hundreds to thousands of values for a single molecule.

Due to the huge amount of descriptors, the reduction of their number is necessary before proceeding with further steps to build a QSAR model (Figs. 2 and 3).

The alvaDesc software provides different unsupervised variable reduction methods that can be applied to decrease the number of molecular descriptors. Descriptors with constant or missing values can be removed in order to reduce the number of considered variables. Additionally, variable reduction can be performed defining a threshold of the correlation among the considered descriptors. Methods based on correlation, besides reducing the number of descriptors, can be used to decrease the redundancy and multicollinearity of the data. One method for unsupervised variable reduction, available in alvaDesc, is the so-called V-WSP [27] algorithm. This algorithm is based on the analysis of the correlation matrix and is a modification of the WSP algorithm for design of experiments (DOE) [59].

4 Structural Keys and Molecular Fingerprints

Molecular fingerprints can be of two types, structural keys and hashed molecular fingerprints.

Molecular fingerprints describe a molecule considering different local aspects of its structure; specifically, structural keys identify the presence or absence of a defined list of structural features, while

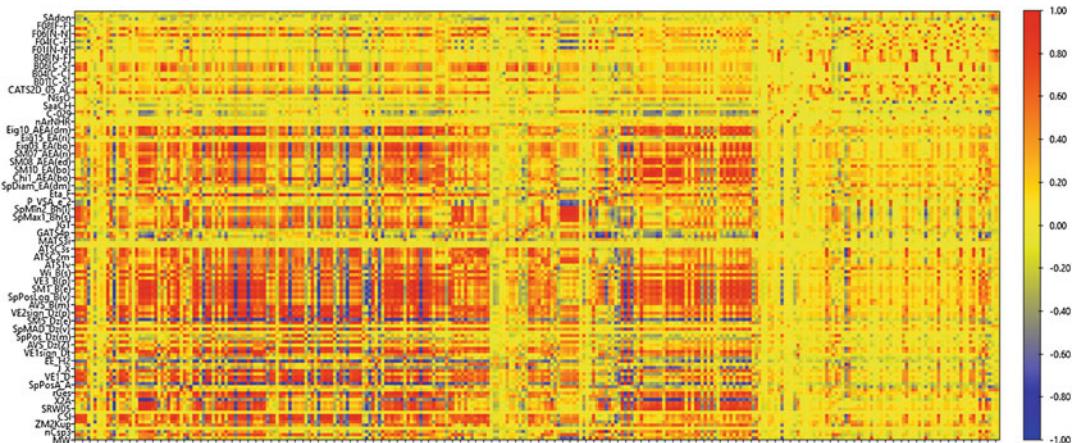


Fig. 2 alvaDesc correlation map performed on 2194 descriptors

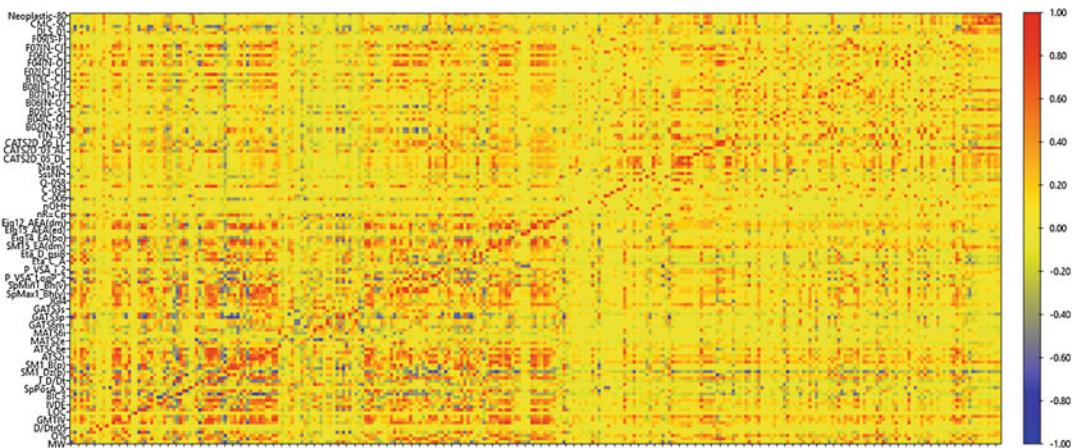


Fig. 3 alvaDesc correlation map performed on 574 descriptors remaining after the application of V-WSP algorithm for variable reduction. Correlation threshold has been set to 0.9

hashed fingerprints describe the molecule identifying all possible fragments.

Molecular fingerprints can be used to identify common substructures and were originally proposed for fast database searching, but this characteristic has been exploited even for building QSAR models, specifically those based on local similarity as in the kNN approach [8, 60].

4.1 Structural Keys

Structural keys are defined as a set of structural features (e.g., atom-centered fragments, functional groups, pharmacophoric atom-types) that have been prepared in order to discriminate among molecules and eventually to highlight chemical properties of a molecule.

The alvaDesc software includes different structural keys, most of them are included in the molecular descriptor blocks, like

functional group counts, atom-centered fragments, and the CATS 2D descriptors belonging to the pharmacophore descriptors block.

Additionally, alvaDesc provides the calculation of the MACCS 166 fingerprint [25]. The MACCS 166 fingerprint is a fixed-size boolean vector reflecting the presence/absence of a set of 166 well-defined molecular features.

4.2 Hashed Chemical Fingerprints

Hashed chemical fingerprints do not have a predefined list of structural features to be found but explore the molecular structure storing all possible identified substructures following a set of rules. Since the number of substructures identifiable in a molecule set is not predefined, a hashing function is used to reduce a variable-size boolean vector to a fixed-size one.

Hashing function is deterministic; this means that, under a predefined set of rules, a specific fragment will always be associated to a defined set of bits in the fingerprint.

Hashing function has the advantage to transform an indefinite set of structural features to a fixed-length vector but introduce the so-called bit collision; this means that two different fragments may share one or more bits among their bit sets.

A key feature of hashed fingerprint is the so-called darkness. Darkness of a fingerprint represents the percentage of bits set to one. The average darkness of a dataset is a relevant property since high values of darkness will lead to higher chances of false-positive matches, while low values of darkness is an indicator that the fingerprint size could be lowered.

Hashed fingerprints do not allow reversible-decoding, which means that it is not possible to recreate the original substructure starting from a fingerprint.

Nevertheless, fingerprints encode an almost exhaustive set of patterns with respect to structural keys, resulting in a more detailed description of a molecular structure in almost all situations.

Two hashed fingerprint types are included in alvaDesc:

- Extended-connectivity fingerprints (ECFP) [26]
- Path fingerprints (PFP)

Both extended-connectivity fingerprints (ECFP) and path fingerprints (PFP) calculation can be customized using a set of parameters:

- Fingerprint size
- Number of bits per pattern
- Minimum fragment length
- Maximum fragment length
- Fragment occurrences

Fingerprint size is the length of the boolean vector which affects the darkness. Increasing the size lowers the darkness and reduces the chance of false-positive matches, which leads to the need for more space to store the fingerprints.

The number of bits per pattern is the number of bits used to encode a substructure. (e.g., using 2 bits per pattern each substructure will be hashed to 2 bits in the fingerprints). Increasing the number of bits per pattern reduces the chance of different fragment collision of the same bits and increases the darkness.

Minimum fragment length is the smallest size of the detected substructures.

Maximum fragment length is the biggest size of the detected substructures, and it affects the darkness. Increasing the maximum fragment length leads to more substructures identified and encoded in the fingerprint.

If fragment occurrences parameter is taken into account, the fingerprint generation process stores multiple pattern occurrences; otherwise it encodes only the presence/absence of molecular substructures.

In addition to the general fingerprint parameters, atom-type identification can be customized considering the following atom parameters:

- Atom-type
- Aromaticity
- Attached hydrogens
- Connectivity (total)
- Total bond order
- Connectivity (no H)
- Charge
- Ring memberships in SSSR
- Smallest ring size in SSSR
- Bond order
- Atom-type

The selected atom parameters affect the fragments identified during hashed fingerprint calculation; e.g., if atom-type parameter is selected, the substructure is identified encoding atoms considering their atomic number. This means that if two substructures are composed of atoms with different atomic numbers, then the two substructures will be bound to different bits of the fingerprint. If atom-type parameter is not selected, atoms are not differentiated based on their atomic number, and all atoms are considered as belonging to the same atom-type.

The identified fragments can be exported from alvaDesc as SMARTS strings (SMARTS is a line notation developed by

Daylight Chemical Information Systems for representing molecular substructures).

5 Dealing with Disconnected Structures

QSAR has been widely used for the evaluation of physicochemical and ecotoxicological properties of full-connected organic molecules. Recently QSAR has been applied to evaluate molecular properties of non-full-connected molecular structures, such as salts, mixtures, ionic liquids, and metal complexes. QSAR has been used for the prediction of nonadditive physicochemical properties like density, bubble temperature, and azeotropic behavior [61] and for the evaluation of toxicological properties of ionic liquids [62–64].

Available data on non-full-connected structures is a valuable help for the development of QSAR models, and it is associated with the development of methods to correctly represent those types of molecules with molecular descriptors and fingerprints [65, 66].

With respect to non-full-connected structures, molecular descriptors can be grouped in two [17].

The first group includes all the descriptors with a mathematical definition which can be applied to non-full-connected structures preserving their chemical meaning. Molecular weight and cyclomatic number are examples of molecular descriptors that have algorithms and meaning that are preserved on disconnected structures. In the same way functional groups, atom-centered fragments, fingerprints, and structural keys are additive descriptors, and their interpretation is identical for full-connected and non-full-connected structures.

The second group collects those descriptors that have a definition and meaning that cannot be directly extended to non-full-connected structures. In this case, a deeper analysis of molecular descriptor algorithm should be carried out; alternatively a linear combination of the calculated values can be applied to every full-connected constituent of the compound [65].

Many software tools cannot calculate molecular descriptors on non-full-connected structures, while alvaDesc provides six different theoretical approaches for the calculation of molecular descriptors on such structures:

- Standard
- Maximum descriptor value
- Minimum descriptor value
- Average descriptor value
- Sum of descriptor values
- Retain the biggest fragment

Standard approach has been defined considering the mathematical definition of all the implemented algorithms for descriptor calculation. Every algorithm has been checked and eventually modified in order to be applicable not only to full-connected structures but even to structures composed of more than one disjoint substructures.

This approach has been implemented in order to provide as more information as possible when considering a disconnected structure since it considers all the disjoint substructures together as a unique entity. Conversely maximum descriptor value, minimum descriptor value, and retain the biggest fragment approaches consider only one of the disjoint substructures included in the whole molecule. Average descriptor value and sum of descriptor values approaches consider every disjoint substructure as an isolated molecule, then the results obtained on the isolated substructures are merged using the average or sum approach.

Maximum descriptor value approach considers every disjoint substructure as a single molecule. Molecular descriptors are calculated on all disjoint substructures separately; the maximum value obtained on all disjoint substructures is retained.

$$MD = \max(\mathbf{x})$$

where \mathbf{x} is the array including all the disjoint structures in the original molecule.

Minimum descriptor value approach considers every disjoint substructure as a single molecule. Molecular descriptors are calculated on all disjoint substructures separately; the minimum value obtained on all disjoint substructures is retained.

$$MD = \min(\mathbf{x})$$

where \mathbf{x} is the array including all the disjoint structures in the original molecule.

Average descriptor value approach considers every disjoint substructure as a single molecule. Molecular descriptors are calculated on all disjoint substructures separately; the average of the obtained values is retained.

$$MD = \frac{\sum_i x_i}{|\mathbf{x}|}$$

where $|\mathbf{x}|$ is the cardinality of the vector \mathbf{x} (i.e., the number of disjoint substructures in the original molecule) and x_i is the descriptor value of the i -th substructure.

Sum of descriptor values approach considers every disjoint substructure as a single molecule. Molecular descriptors are calculated on all disjoint substructures separately; the sum of the obtained values is retained.

$$MD = \sum_i x_i$$

where x_i is the descriptor value of the i -th substructure.

Finally, using retain the biggest fragment approach means that the molecular descriptors are calculated only for the biggest fully connected structure included in the original molecule. The biggest fragment is defined as the biggest fully connected structure with the highest number of non-hydrogen atoms. In case of equivalency, the structure with the highest molecular weight is retained.

6 Conclusions

The evaluation of ecotoxicological properties using in silico techniques has become increasingly important over the last few decades. The opportunity to evaluate the toxicological properties of chemicals avoiding animal testing, together with the possibility to significantly reduce the costs needed to evaluate even a single physicochemical property, has made QSAR an indispensable technique.

QSAR models can be defined using different representations of molecules; the most commonly used ones have been described in this chapter. Molecular descriptors, structural keys, and hashed fingerprints represent different approaches to codify chemical structures in a mathematical way.

While structural keys can easily be interpreted due to the fact that they identify a well-defined list of chemical fragments, their application can be limited if considered in isolation, since they provide a limited representation of a chemical structure. Conversely, hashed molecular fingerprints have been proposed to describe the whole chemical structure and can be successfully used for QSAR analysis.

Furthermore, as discussed in this chapter, molecular descriptors variability is as wide as the number of possible combinations of structure representations, atom/bond weighting schemes, and mathematical functions. Nevertheless, despite their complexity, molecular descriptors are playing a fundamental role in chemical representation and in QSAR, not only for the prediction of ecotoxicological properties but generally as a tool to analyze chemical data.

References

- Ihlenfeldt WD, Bolton EE, Bryant SH (2009) The PubChem chemical structure sketcher. *J Cheminform* 1(1):1–9
- Kim S, Thiessen PA, Bolton EE, Bryant SH (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res* 43(W1):W605–W611
- Davies M et al (2015) ChEMBL web services: streamlining access to drug discovery data and

- utilities. *Nucleic Acids Res* 43(W1):W612–W620
4. Gaulton A et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954
 5. Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45(1):177–182
 6. Worth AP (2009) The role of Qsar methodology in the regulatory assessment of chemicals. In: Recent advances in QSAR studies. Springer, Dordrecht; New York
 7. Cassotti M, Ballabio D, Consonni V, Mauri A, Tetko IV, Todeschini R (2014) Prediction of acute aquatic toxicity toward *Daphnia magna* by using the GA-kNN method. *Altern Lab Anim* 42(1):31–41
 8. Cassotti M, Consonni V, Mauri A, Ballabio D (2014) Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards *Daphnia magna*. *SAR QSAR Environ Res* 25(12):1013–1036
 9. Khan PM, Roy K, Benfenati E (2019) Chemometric modeling of *Daphnia magna* toxicity of agrochemicals. *Chemosphere* 224:470–479
 10. Tebby C, Mombelli E, Pandard P, Péry ARR (2011) Exploring an ecotoxicity database with the OECD (Q)SAR Toolbox and DRAGON descriptors in order to prioritise testing on algae, daphnids, and fish. *Sci Total Environ* 409(18):3334–3343
 11. Grisoni F, Consonni V, Vighi M (2018) Acceptable-by-design QSARs to predict the dietary biomagnification of organic chemicals in fish. *Integr Environ Assess Manag* 15(1):51–63
 12. Khan K, Roy K (2017) Ecotoxicological modelling of cosmetics for aquatic organisms: a QSTR approach. *SAR QSAR Environ Res* 28(7):567–594
 13. Holmquist H, Lexén J, Rahmberg M, Sahlin U, Palm JG, Rydberg T (2018) The potential to use QSAR to populate ecotoxicity characterisation factors for simplified LCIA and chemical prioritisation. *Int J Life Cycle Assess* 23(11):2208–2216
 14. Khan K, Roy K, Benfenati E (2019) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. *J Hazard Mater* 369:707–718
 15. Fourches D, Muratov E, Tropsha A (2010) Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research. *J Chem Inf Model* 50(7):1189–1204
 16. Todeschini R, Consonni V (2009) Molecular Descriptors for Chemoinformatics. Vol. 1. Alphabetical Listing; Vol. 2. Appendices, References. Wiley-VCH, Weinheim
 17. Mauri A, Consonni V, Todeschini R (2017) Molecular descriptors. In: Leszczyński J, Kaczmarek-Kedziera A, Puzyn T, Papadopoulos MG, Reis H, Shukla MK (eds) Handbook of computational chemistry. Springer International Publishing, Switzerland, pp 2065–2093
 18. Moriwaki H, Tian YS, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. *J Cheminform* 10(1):1–14
 19. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466
 20. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J Chem Inf Comput Sci* 43(2):493–500
 21. Willighagen EL et al (2017) The chemistry development kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9(1):1–19
 22. RDKit: Open-source cheminformatics; <http://www.rdkit.org>
 23. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: an easy approach to molecular descriptor calculations. *Match Commun Math Comput Chem* 56(2):237–248
 24. Alvascence srl (2019) alvaDesc (software for molecular descriptors calculation). Available at: <https://www.alvascence.com/>
 25. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 42(6):1273–1280
 26. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754
 27. Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergent M, Todeschini R (2014) A novel variable reduction method adapted from space-filling designs. *Chemom Intell Lab Syst* 136:147–154
 28. Berthold MR et al (2008) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications, vol 11(1). Springer, Berlin/Heidelberg, pp 319–326
 29. Sushko I et al (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25(6):533–554

30. Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? *QSAR Comb Sci* 27(11–12):1337–1345
31. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29(6–7):476–488
32. Randić M (1996) Molecular bonding profiles. *J Math Chem* 19(3):375–392
33. Guha R, Willighagen E (2012) A survey of quantitative descriptions of molecular structure. *Curr Top Med Chem* 12(18):1946–1956
34. Todeschini R, Gramatica P (1997) The Whim theory: new 3D molecular descriptors for Qsar in environmental modelling. *SAR QSAR Environ Res* 7(1–4):89–115
35. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J Chem Inf Comput Sci* 42(3):682–692
36. Wiener H (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69(1):17–20
37. Plavšić D, Nikolić S, Trinajstić N, Mihalić Z (1993) On the Harary index for the characterization of chemical graphs. *J Math Chem* 12(1):235–250
38. Randić M (1975) On characterization of molecular branching. *J Am Chem Soc* 97(23):6609–6615
39. Randić M (2001) The connectivity index 25 years after. *J Mol Graph Model* 20(1):19–35
40. Moreau JL, Broto P (1980) Autocorrelation of molecular structures: application to SAR studies. *Nouv J Chim* 4:757–764
41. Broto P (1984) Molecular structures: perception, autocorrelation descriptor and sar studies. *Eur J Med Chem* 19:66–70
42. Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* 37(1–2):17–23
43. Schneider G, Neidhart W, Giller T, Schmid G (1999) ‘Scaffold-Hopping’ by topological pharmacophore search: a contribution to virtual screening. *Angew Chemie Int Ed* 38(19):2894–2896
44. Renner S, Fechner U, Schneider G (2006) Alignment-free pharmacophore patterns – a correlation vector approach. In: Langer T, Hoffmann RD (eds) *Pharmacophores and pharmacophore searches*. Wiley-VCH, Weinheim, pp 49–79
45. Ertl P, Rohde B, Selzer P (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem* 43(20):3714–3717
46. Ertl P (2008) Polar Surface Area. In: Mannhold R (eds) *Molecular Drug Properties. Measurement and Prediction*. Wiley-VCH, Weinheim, pp 111–126
47. Moriguchi I, Hirono S, Nakagome I, Hirano H (1994) Comparison of reliability of log P values for drugs calculated by several methods. *Chem Pharm Bull* 42(4):976–978
48. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J Phys Chem A* 102(21):3762–3772
49. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* 1(4):337–341
50. Jolliffe IT (2002) *Principal component analysis*. Springer-Verlag, New York
51. Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1):1–14
52. Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605
53. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J Chem Inf Comput Sci* 25(2):64–73
54. Kier LB, Hall LH (1990) An electrotopological-state index for atoms in molecules. *Pharm Res* 7(8):801–807
55. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* 35(6):1039–1045
56. Kier LB, Hall LH (1981) Derivation and significance of valence molecular connectivity. *J Pharm Sci* 70(6):583–589
57. Gombar V, Kumar A, Murthy MS (1987) Quantitative structure activity relationships part ix. A modified connectivity index as structure quantifier. *Indian J Chem Sect B Org Chem Incl Med Chem* 26(12):1168–1170
58. Burden FR (1989) Molecular identification number for substructure searches. *J Chem Inf Comput Sci* 29(3):225–227
59. Santiago J, Claeys-Bruno M, Sergent M (2012) Construction of space-filling designs using WSP algorithm for high dimensional spaces. *Chemom Intell Lab Syst* 113:26–31

60. Rojas C et al (2017) A QSTR-based expert system to predict sweetness of molecules. *Front Chem* 5:53
61. Ajmani S, Rogers SC, Barley MH, Livingstone DJ (2006) Application of QSPR to mixtures. *J Chem Inf Model* 46(5):2043–2055
62. Varnek A, Kireeva N, Tetko IV, Baskin II, Solov'yev VP (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J Chem Inf Mod* 47(3):1111–1122
63. Roy K, Das RN, Popelier PLA (2014) Quantitative structure-activity relationship for toxicity of ionic liquids to *Daphnia magna*: aromaticity vs. lipophilicity. *Chemosphere* 112:120–127
64. Roy K, Das RN, Popelier PLA (2015) Predictive QSAR modelling of algal toxicity of ionic liquids and its interspecies correlation with *Daphnia* toxicity. *Environ Sci Pollut Res* 22 (9):6634–6641
65. Oprisiu I, Novotarskyi S, Tetko IV (2013) Modeling of non-additive mixture properties using the Online CHEmical database and Modeling Environment (OCHEM). *J Cheminform* 5(1):1
66. Mauri A, Ballabio D, Todeschini R, Consonni V (2016) Mixtures, metabolites, ionic liquids: a new measure to evaluate similarity between complex chemical systems. *J Cheminform* 8 (1):1–3

INDEX

A

Acetylcholinesterase, 79, 514, 534, 536, 569, 577
Acid rain, 617
Activated sludge, 13, 307, 389, 391, 392, 395, 401, 566
Active pharmaceutical ingredients (APIs), 236–238, 331–354
Activity cliff analysis, 30, 58, 104, 105, 107, 112
Acute *Daphnia magna* toxicities, 547, 595, 596, 600–603, 610
Acute toxicity tests, 13, 19, 133, 152, 153, 157, 424, 449, 457, 460, 556, 578, 712, 713, 731
Adverse effects, 15, 16, 35, 81, 90, 112, 119, 151, 153, 249, 332, 334, 358–360, 364, 366, 371, 399, 420, 482, 484, 516, 568, 572, 617, 619, 625, 721, 750, 772, 781, 785
Aggregated Computational Toxicology Online Resource (ACToR), 39, 715, 716, 738, 749
Agrochemical, viii, 28, 35, 39, 41, 359, 438, 460, 495, 515, 520, 523, 537, 622, 623, 709, 712, 734, 738, 746
Air pollution filter, 488
Algal toxicity, 592–595, 599–601, 603, 604, 606–610, 696, 697
Aliivibrio fischeri, 309, 622
Allergenicity, 412
Allium cepa, 309, 315, 322
Alternative, vii, viii, 8, 10, 13, 20, 28, 35, 36, 42, 44, 46, 47, 56, 112, 116, 157, 158, 218, 241, 261, 264, 265, 290, 292, 323, 333, 346, 349, 360, 365, 375, 376, 379, 388, 398, 401, 431, 449, 469, 481, 485, 492, 505, 514, 523, 537, 545–630, 693, 694, 710, 714, 779, 790, 802, 815
Alzheimer's disease, 88, 535, 619
Ames test, 118, 119, 129, 130, 132, 140, 144, 145, 244, 313, 413, 418, 419, 761, 762, 770
Ammonia, 489
Analog, 157, 299, 424
Analogue approach, 245, 290, 299, 302, 550–552
Androgenic, 311
Animal experiments, 18, 616, 630
Antagonistic, 82, 84, 306, 312, 439, 442, 446, 457, 462, 466, 574, 577, 579, 622
Applicability domain (AD), viii, 33, 35, 41, 60, 100, 144, 145, 243, 246, 259, 262, 264, 272, 284, 285, 367, 372, 381, 395, 402, 403, 410, 411, 428,

449, 454, 492, 493, 502, 532, 534, 552, 567, 609, 620, 623, 624, 626–628, 630, 649, 650, 696, 762, 764, 766, 768, 777

Aquatic, 2, 35, 120, 151, 200, 236, 291, 313, 333, 358, 387, 406, 438, 479, 519, 546, 563, 615, 645, 686, 712, 763

ecosystems, 21, 617, 753

invertebrates, 159, 291, 565, 617

organisms, 43, 45, 151, 152, 154, 159, 200, 358, 479, 563, 565, 575, 577, 617, 619, 625

Aromaticity, 426, 623, 803, 809, 814

Arsenic, 83, 479

Artificial neural network (ANN), 33, 118, 169, 170, 179, 186–188, 200, 201, 208, 210, 370, 371, 413–415, 419, 428, 449, 450, 453, 504, 532, 534, 537, 616, 620, 621, 645–647, 694, 698, 702

Atmosphere, 311, 317

Atomic property fields (APF), 222

Autocorrelation descriptors, 805–808

B

Bacteria, 42, 44, 45, 48, 129, 203, 310, 363, 376, 389, 398, 413, 416, 422, 425, 429, 461, 463, 480, 489, 568, 574, 627, 628, 692, 721

Baseline toxicity, ix, 309, 312, 334–336, 339–347, 349–354, 562, 565, 577, 578, 684

Behavioral effect, 86–88

Benzophenone, 373

BFGS Algorithm, 620

Bioassay, 262, 306–309, 320–322, 374, 377, 422, 531, 533, 719

Biocidal Product Regulation (BPR), vii, 14, 15, 359–362, 365, 387, 388, 398, 400, 401, 403

Biocides, ix, vii, viii, 14, 15, 20, 357–382, 387–401, 517, 565, 600, 622, 686, 761, 778, 782

Bioconcentration factor (BCF), 16, 120, 243, 246, 294, 296–299, 301, 302, 428, 519, 524, 525, 538, 563, 573, 593, 713, 714, 741, 761, 767–769

Biodegradation, vii, 13, 297, 376, 425, 428, 429, 562, 564, 566, 713, 715

Biodiversity, 6, 615, 617

Biological Assays and Diseases (BAD), 306, 321, 322

Biological data curation, 101

Bioluminescent bacteria, 43

Biomonitoring, ix, 80, 82, 84, 91

Biosphere, vii, 152, 617
 Birds, 49, 78, 80, 81, 83, 85, 87–91, 120, 121, 187, 520, 523, 567, 617, 619, 645, 712, 723
 Birth Defects Systems Manager (BDSM), 717
 Body armor, 488
 Buckypaper, 488

C

Cadmium, 83, 461–463, 479, 619
 Cancer, 81, 132, 183, 309, 359, 406, 490, 535, 619, 640, 717
 Carbon nanotubes (CNTs), ix, 13, 223, 227–230, 477–506, 791
 Carcinogenic, 10, 16, 28, 43, 44, 79, 85, 86, 118, 264, 308, 316, 318, 321, 406, 407, 415–417, 422, 478, 483, 490, 569, 717, 723, 728–730
 Carcinogenic effects, 28, 85, 86, 406, 483, 723
 Carcinogenicity, vii, 35, 115, 117, 118, 187, 238, 244, 293, 318, 369, 371, 413, 416–418, 439, 732–734, 740, 741, 743, 751, 761, 771
 Carcinogenicity, mutagenicity and reprotoxicity (CMR) assessment, 116, 117, 119, 145, 763, 771
 Carcinogenic potency database (CPDB), 255, 717, 733
 Cardiovascular diseases, 619
 Case studies, x, ix, viii, 57, 292–294, 296, 299, 302, 545–558, 591–611, 696, 739, 741, 782
 Categorization of chemicals, 157, 594, 596, 599, 600, 603, 607, 610, 712, 742
 Category, vii, 163, 165, 167, 200, 244, 245, 250–259, 290, 291, 293–295, 298, 299, 302, 321, 339, 342, 361, 372, 374, 420, 427, 548, 550–552, 592, 593, 597, 599, 600, 603, 607, 609, 610, 630, 686, 690, 700, 727
 Category approach, 245, 290, 291, 294, 299, 550, 551, 592, 593, 607, 727
 Cationic, x, 42, 424, 623, 624, 681–702
 Cationic structure, 623
 Cell transformation assay (CTA), 255, 318–320, 322
 Characterization factors (CFs), 321, 562, 563
 Chemical abstracts service (CAS), 249, 250, 290, 296, 335, 553–557, 716, 721–723, 726, 728, 731, 801
 Chemical carcinogenesis research information system (CCRIS), 132, 717
 Chemical categories, 244, 245, 251, 253, 254, 257–259, 290, 291, 294, 295, 299, 420, 548, 685
 Chemical data curation, 101, 102, 106
 Chemical mixture, 84, 91, 316, 438–440, 442–444, 447–449, 455, 460, 462, 463, 465–469, 573, 577, 578, 580
 Chemicals, 6, 27, 55, 77, 97, 151, 177, 195, 218, 235, 271, 290, 306, 333, 358, 388, 406, 437, 478, 513, 545, 562, 592, 615, 640, 662, 682, 709, 759, 790, 801
 Chemical safety, 8–10, 119, 294, 299, 438, 450, 546, 740

Cheminformatics, 29, 215, 412, 720, 732, 751, 789–798
 Chlorobenzenes, 478, 483
 Chlorophenols, 389, 478, 479, 483
 Chromosome aberrations (CA), 315, 316, 751
 Chronic kidney disease, 619
 Chronic toxicity tests, 152, 153
 C3H10T1/2 clone 8 mouse embryonic fibroblasts, 318, 791
 Class 1 (non-polar narcotic), 334, 593, 596, 603, 605, 610
 Class 2 (polar narcotic), 334, 593, 596, 603, 610
 Classification, Labelling and Packaging (CLP) Regulations, vii, 7, 15, 19, 21, 546, 739
 CNT, 228–230, 488, 489
 COMBASE, ix, 387–403, 565, 775, 778
 Comet test, 309, 314, 315, 322
 Comparative molecular field analysis (CoMFA), 216–219, 222, 225, 373, 414, 521, 522, 524–529, 533, 537, 570
 Comparative molecular moment analysis (CoMMA), 220
 Comparative molecular similarity indices analysis (CoMSIA), 218, 220, 222–225, 230, 414, 524–526, 533, 537
 Comparative residue interaction analysis (CoRIA), 220
 Component-based approach, 443
 Conceptual density functional theory (CDFT), 581, 663, 664
 Concrete, 10, 20, 81, 488
 Confidence predictors, ix, 272, 273
 Conformal prediction, ix, 60, 62, 271–285
 Conformity score, 274–276
 Congenital anomalies, 619
 Consensus toxicity estimates, 550, 552–555, 558
 Contaminants, vii, 20, 28, 41, 42, 45, 48, 49, 78, 80, 85, 86, 89, 153, 236, 240, 358, 377, 379, 381, 461, 482, 489, 490, 492, 497, 501, 505, 566, 569, 715, 783
 Contaminants of emerging concern (COEC), 45, 48
 Contamination, 20, 27, 152, 425, 439, 480, 617
 Continuous molecular fields (CMFs), 221
 Copper, 461, 462, 479, 480
 CORAL, 44, 103, 115, 121, 369, 498, 500, 501, 772
 Cosmetics, vii, viii, 28, 35, 41, 43, 45, 46, 49, 117, 159, 360–364, 366, 367, 371, 377, 381, 480, 482, 546
 Cross-cutting legislation, 21
 Cytotoxicity, vii, 42–45, 310, 319, 417, 465, 624, 662, 747, 791, 793

D

Danish (Q) SAR Database, 39, 298, 717, 718, 783
Daphnia magna, 43, 246, 248, 249, 273, 276, 277, 297, 309, 332, 373, 376, 422, 522, 593, 596, 600–602, 610, 621, 622, 626–628, 712, 715, 802

- Databases, 37, 62, 79, 97, 112, 154, 178, 222, 242, 296, 323, 335, 360, 389, 410, 440, 503, 550, 562, 611, 630, 641, 710, 783, 792, 810
- Data curation of large-size datasets, 98
- Data gaps, vii, 43, 47, 48, 50, 111, 245, 249, 251, 258, 259, 291, 292, 295, 296, 359, 370, 439, 449, 469, 481, 546, 547, 549, 550, 552, 553, 556, 623, 701, 710, 714, 727
- Data sets, viii, 18, 57, 59–63, 154, 157, 165, 166, 168, 169, 202, 206, 246, 262, 353, 370, 371, 407, 409, 413, 415–417, 420, 423, 582, 592, 593, 596, 599, 600, 602, 604, 606, 608–610, 667, 668, 676, 740
- Decision process, 307, 323
- Decision support system (DSS), 323
- Decision tree (DT), 33, 113, 115, 116, 134, 165, 166, 169, 170, 182, 183, 201, 202, 274, 298, 334, 450, 626, 697
- Deep learning (DL), ix, 111–146
- Density functional theory (DFT), 200, 225, 228, 373, 415, 417, 426, 495, 500, 571, 572, 662
- Detergents, 42, 358, 360, 406, 482, 726
- Developmental and reproductive toxicology database (DART), 255, 371, 718, 737, 750
- Developmental toxicity (DevTox), 35, 38, 119, 371, 713, 715, 717–719, 737, 761, 762
- 3D-HoVAIFA, 221
- Diabetes, 83, 84
- Dialkyl phthalate esters, 478
- Dichlorodiphenyltrichloroethane (DDT), 37, 78, 81, 83, 480, 782
- Dieldrin, 37, 480
- Difficult substances, 557
- Disability Adjusted Life Year (DALY), 323
- Distributed Structure-Searchable Toxicity database (DSSTox), 39, 256, 263, 719, 733
- Distribution coefficient ($\log D$), 333, 344–353, 495, 593, 601, 603, 607
- 3D-QSAR, ix, 215–230, 522, 526–528, 535
- Duplicate analysis, 104, 105, 107
- Dye, ix, vii, 309–311, 317, 361, 370, 405–431, 478–481, 497
- Dynamic lattice-oriented molecular modeling system (DYLOMMS), 216
- E**
- ECHA, vii, 20, 157, 158, 245, 248, 255, 256, 292, 360, 361, 366, 388, 683, 701, 736, 739, 764, 774, 790, 793
- Ecological Structure Activity Relationships Predictive Model (ECOSAR), 40, 46, 159, 253, 341, 360, 368, 369, 376, 423, 563, 592, 602, 626, 629, 684, 693
- Ecology, 40, 78, 79, 81, 151, 163, 422, 720
- ECOSAR software, 563, 626
- Ecosystem Pollution, 617, 619
- Ecotoxic, x, 179, 204, 205, 537, 563, 570, 615, 639–655
- Ecotoxicity, x, ix, 13, 18, 28, 35, 36, 42, 43, 45, 46, 48–50, 56, 161, 170, 195–211, 219, 291, 299, 333, 335–337, 339, 343, 348, 360, 362, 363, 367, 369, 375, 379, 381, 382, 387–403, 405–431, 442, 449, 454, 458, 460, 461, 463, 518, 520, 537, 562, 572, 591–611, 615, 616, 619, 623–628, 640, 641, 644, 647, 655, 692, 693, 695, 696, 699, 709–754
- Ecotoxicological properties, 47, 537, 765, 771, 772, 802, 815, 817
- Ecotoxicological risk, ix, viii, 3–23, 77–91, 241, 379, 518, 572, 615–630
- Ecotoxicological tests, 152–154
- Ecotoxicology, x, ix, viii, 3–23, 27–50, 77–91, 97–108, 111–146, 151–170, 205, 215–230, 235–265, 271–285, 289–302, 305–323, 331, 332, 357–382, 388, 389, 391, 437–470, 513–537, 561–582, 615–630, 641, 664, 666, 693, 700, 715, 759–785, 802, 815, 817
- ECOTOXicology knowledgebase (ECOTOX), 46, 161, 162, 255, 372, 375, 390, 391, 551
- EDC-MixRisk, 440
- 50% Effect concentration (EC_{50}), 30, 34, 43, 102, 152, 153, 204, 211, 310, 315, 320, 332, 335, 337, 338, 340, 343, 352, 373, 389, 391–393, 395, 401, 402, 439, 446, 447, 454, 469, 553–555, 562, 565, 573–575, 578, 594–596, 599–611, 622, 627, 649, 711
- Efficiency, 118, 216, 273, 276, 282–284, 403, 417, 425, 429, 485, 489, 493, 515, 523, 532, 546, 569, 666, 667
- Effluents, 16, 45, 153, 305–308, 321, 322, 360, 375, 387, 406, 422, 431, 478–482, 625
- Electromagnetic, 488
- Electronic Waste, 617
- Electrophilicity index, 663, 665, 666
- ELINCS database, 370
- Enalos Cloud platform, x, 789–798
- Endocrine disrupting chemicals, 80, 495
- Endocrine disruption, 16, 78, 81, 82, 90, 119, 308, 309, 311–313, 771, 780, 784
- Ensemble learning (EL), 165, 167
- Environment policy, 6
- Environment, 4, 35, 63, 78, 113, 151, 196, 236, 290, 305, 331, 359, 387, 406, 438, 478, 514, 516, 545, 562, 592, 640, 661, 682, 710, 761, 790, 804
- Environment Action Programme, 22
- Environmental fate, ix, 9, 17, 19, 56, 159, 248, 255–257, 291, 360, 377, 391, 425, 564, 565, 683, 700, 710, 717, 723, 735, 781
- Environmental footprint, 305, 323

- Environmental monitorization, 17, 152
 Environmental quality standards directive (EQSD), 20
 Environmental risk assessment (ERA), ix, 17, 18, 40, 197, 200, 235, 331, 424, 438, 483, 484, 563, 625, 687, 701, 726, 753
 Equipment, 78, 488
 ERM-PLS, 227
 Estrogenic, 308, 311, 568
 eTox, 723, 724, 726
 EU laws, ix, 4, 7, 8, 438
 EU legislation, 4, 18, 364, 441
 EuroMix, 440
 European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), 156, 255, 256, 292, 391
 European Chemical Agency (ECHA), vii, 20, 157, 158, 245, 248, 255, 256, 292, 360, 366, 388, 683, 701, 739, 764, 774, 776, 790, 793
 European Chemical Substances Information system (ESIS), 38, 721
 European Commission (EC), 4, 6, 8, 9, 13, 14, 16–21, 39, 81, 82, 152, 247, 296, 364, 365, 371, 391, 392, 438, 455, 482, 558, 624, 683, 721, 722, 729, 762, 763, 781, 785
 European Food Safety Authority (EFSA), 255, 256, 296, 370, 375, 388, 394, 402, 438, 480, 563, 739, 762–764, 770, 776, 783, 784
 European Union (EU), x, ix, vii, viii, 3–23, 157, 290, 292, 295, 312, 320, 331, 359, 360, 364, 388, 389, 391, 410, 412, 438–441, 469, 545, 690, 722, 724, 739, 740, 744
 EUToxRisk, 440
 Evaluations, ix, vii, 7, 8, 13, 17, 18, 44, 47, 63, 143, 151–159, 166, 170, 181, 182, 202, 227, 240, 241, 248, 252, 259, 260, 263, 264, 290, 294, 297, 307, 308, 310, 316, 318, 320, 321, 323, 361, 364, 365, 368, 374, 410, 414, 423, 438, 448, 454, 467, 484, 514, 518, 531, 546, 547, 596, 610, 616, 622, 627, 629, 661, 665, 693–695, 699, 710, 714–715, 725–728, 744, 750, 751, 753, 762, 768, 778, 779, 784, 790, 801, 802, 815, 817
 Excess toxicity, 335, 548, 549, 684, 685
 Exploratory data analysis (EDA), 547, 551–555
 Exposure, vii, 9, 42, 78, 153, 207, 229, 237, 290, 311, 332, 360, 394, 411, 438, 478, 515, 545, 565, 594, 616, 641, 694, 711, 760
 Extension TOxicology NETwork (EXTOXNET), 722, 723
 External Validation, viii, 33, 62, 202, 210, 374, 427, 454, 497–499, 601, 602, 616, 620, 622, 624, 626, 696, 719
- F**
- Feature selection, ix, viii, 32, 62, 113, 123, 177–189, 202, 227, 402, 453, 579, 626, 695, 699
- Filter methods, 181, 182, 189, 403
 Fish, 13, 35, 56, 78, 120, 152, 183, 200, 255, 291, 331, 358, 389, 424, 463, 514, 545, 563, 592, 617, 642, 685, 711, 763, 802
 Fish acute toxicity, 344, 346, 352, 554, 712, 715, 773, 774, 776, 777
 Foci, 309, 318, 320
 Fractional factorial design (FFD), 227
 Fragments, 37, 58, 112, 115, 121, 225, 227, 243, 253, 293, 298, 315, 414, 415, 428, 453, 517, 525, 526, 532, 533, 548, 571, 579, 622, 623, 640, 643, 650, 653–655, 683, 699, 700, 702, 714, 752, 767–769, 772, 776–780, 802–806, 808, 812–817
 Fraunhofer RepDose, 724, 725
- G**
- GA-MLRA, 168, 169, 225, 446, 530, 578
 Gap junctional intercellular communication (GJIC), 309, 316, 317
 Gastrointestinal disorders, 619
 Generalized concentration addition (GCA) models, 443, 446
 Genetic activity profile (GAP), 499, 726
 Genetic alterations in cancer (GAC), 725
 Genetic toxicity, 308, 309, 313, 733
 GENE-Tox, 38, 726, 750
 GERM, 219
 GETAWAY descriptors, 803, 808, 809
 Good laboratory practice regulation (GLP), 19, 21, 735
 Graphene oxide (GO), 185, 503
 Green algae, x, 42, 200, 310, 376, 454, 462, 591–611, 697
 Greenhouse, 485
 GRID, 216, 218, 220, 227
- H**
- 48-h algal EC₅₀, 600, 604, 611
 72-h algal EC₅₀, 599
 Harmful effects, 362, 461, 462, 479, 516, 616, 619, 629, 652, 711
 HASL, 219
 Hazard, 8, 35, 79, 161, 240, 290, 359, 388, 410, 438, 481, 523, 545, 565, 622, 687, 710, 785
 Hazard Evaluation Support System attached database (HESS DB), 727–728
 Hazardous Substances Data Bank (HSDB), 728, 739
 HBM4EU, 440
 Health hazard, 79, 573, 730
 Heat of formation, 461, 596
 Heavy metals, 87, 88, 238, 447, 461–462, 479, 480, 482, 490, 504, 617, 619
 Herbicides, 168, 460, 480, 492, 495, 497, 505, 571, 572, 577, 620, 622, 641

- Hexachlorobenzene (HCB), 37, 667, 672
 High production volume (HPV), 290, 299, 722, 724, 740
 HIV-1 PR, 223–225, 533
 Human and environmental risk assessment (HERA), 726–727
 Human androgen receptor (hAR), 312
 Human estrogen receptor (hER α), 312
 Human health, vii, 6, 8, 10, 16, 17, 19–23, 36, 42, 43, 45, 49, 79, 80, 82, 83, 86, 157, 159, 167, 237, 239, 240, 248, 255, 261, 290–293, 310, 320, 322, 364, 365, 391, 420, 478–480, 482–484, 514, 516, 535, 576, 615, 682, 683, 710, 729, 731, 739, 742, 743, 750, 790
 Human risk assessment (HRA), 316, 438, 745
 Hybrid in silico models, 196, 197, 200, 202, 203, 210, 211
 Hydrogen bonds, 202, 218, 222, 225, 227, 390, 411, 419, 430, 490, 491, 493, 495, 497, 499, 500, 503–505, 522, 524, 528, 533, 535, 571, 626, 643, 654, 655, 808
 Hydrogen sulphide (H₂S), 489
 Hydrophobicity, x, 206, 336, 338, 373, 419, 428, 519–521, 527, 529, 548, 549, 596, 605–607, 610, 626, 643, 652, 661–676, 683, 701
 Hydrophobicity parameter, 596, 663
 Hydrosphere, 617
 Hydroxybenzophenone, 597
 Hypertension, 619
- I**
- IAR203 cell line, 310, 317
 Immobilization, 152, 337, 343, 375, 596, 602, 611
 Immunotoxicity, 88–90, 228
 Impact category, 321
 Impurities, 15, 235–265, 365, 388, 547, 555, 556, 682, 760, ix
 Inception network, 127–129, 133
 Independent action (IA), 439, 443, 445–448, 459, 469, 573–576, 579, 721
 Industrial activities, 77, 617
 In silico, x, ix, xi, viii, 28, 44, 46, 116, 156, 161, 164, 187, 224, 241, 242, 245, 259–264, 294–296, 359, 368–370, 372, 374, 379, 380, 388, 418, 420, 422, 429, 448, 449, 454, 456, 465, 481, 484, 505, 506, 514, 523, 526, 535, 538, 556, 561–583, 616, 629, 649, 693, 696, 710, 714, 752, 759, 766, 782, 790, 798, 801, 817
 In silico modeling, 197, 582, 715
 Integrated approaches to testing and assessment (IATA), 292, 592
 Integrated risk information system (IRIS), 729–731, 740
 Integrated testing and assessment strategies (ITS), 35, 47, 264, 375, 546, 547, 553, 556, 558, 740, 741
 Interactions, 35, 84, 115, 168, 200, 216, 306, 334, 420, 439, 480, 522, 548, 567, 662, 697, 743, 781, 796, 808
 Interface, 119, 120, 155, 196, 203, 204, 230, 259, 720, 795, 797, 798, 804
 Internal validation, 33, 208, 367, 497
 International Agency for Research on Cancer (IARC) monograph, 117, 416, 428–429, 726, 728, 729, 732
 International Programme on Chemical Safety (IPCS), 81, 438, 718, 736
 International Toxicity Estimates for Risk (ITER), 731, 732, 743
 Interspecies QSAR, 593, 595, 596, 601–603, 610, 715
 Ionic liquids, x, 28, 41–43, 48, 103, 200, 201, 481, 615, 618, 623, 624, 803, 815
 Iopromide, 479
 ISO 8692, 594, 600
- J**
- Japan Existing Chemical Database (JECDB), 371, 731–733, 740
 Jupyter, 63–89
- K**
- Klimisch code, 547
 K-Nearest Neighbor (k-NN), 114, 165, 166, 168, 169, 221, 248, 296, 374, 409, 450, 520, 564, 621, 793
 k-Nearest Neighbor Molecular-Field Analysis (kNN-MFA), 221
 KNIME workflows, 100, 102, 106–108, 412
- L**
- Lazar, 119, 248, 249, 416
 Lead (Pb), 83, 216, 240, 479, 525, 531, 554, 619, 714, 721, 813
 Leadslope, 373, 732–735
 Learning, 32, 63, 86, 114–116, 122, 124, 133–140, 145, 163, 166, 167, 180, 414, 415, 450, 504, viii
 Legislation, 4, 6, 7, 18, 21, 22, 120, 157, 239, 364, 438, 492, 514, 518, 572, 616, 682, 727, 771, 783
 Legislative act, 6–8, 20, 21
 Life cycle assessment (LCA), 306, 321, 322
 Lincomycin, 479
 Lindane, 480
 Linear model, 32, 43, 178–180, 184, 465, 504, 523, 623–625, 695
 Linear solvation energy relationship (LSER), 495, 503, 549
 Lithosphere, 617
 Loudspeaker, 488
 Lowest observed effect concentration (LOEC), 335, 346, 351, 353

M

- MACCS, 802, 813
 MACCS fingerprints, 802, 813
 Machine learning, 32, 34, 56, 61, 63, 111–146, 151–170, 178, 180–182, 195–211, 222, 245, 248, 275, 276, 450, 578, 621, 629, 650, 698, 715, 754
 Machine learning and ecotoxicology, 167–169
 Machine learning and QSAR, 163–165
 Magnets, 488
 Mammals, 80, 83, 85, 86, 89, 230, 299, 358, 418, 515, 617, 662
 Management of adverse effect, 336
 Matched molecular pairs (MMP), 105, 107
 MCF-7 cell line, 310, 312
 MDL, 117, 734–735
 Mean electrostatic potentials (MEP), 220
 Mechanisms/models of action, 84, 85, 91, 228, 439, 537, 592, 594, 595, 599, 607, 609, 684, 697
 Medicinal products, 18–19, 236, 237
 Mercury (Hg), 83, 479, 619
 Metals, 20, 87–89, 157, 238, 405, 438, 439, 447–462, 469, 479, 480, 482, 490, 504, 615, 617, 619, 626
 Micronuclei, 315, 316
 Mixture toxicity assessment, ix, 439, 442–448, 456, 457, 467, 468, 470
 ML algorithms applied to QSAR, 165
 Modeling, 28, 55, 97, 113, 154, 197, 216, 292, 367, 389, 407, 461, 481, 518, 567, 602, 616, 640, 686, 710, 762, 791, 804
 Mode of action (MoA), 157, 185, 244, 245, 249, 291, 293, 310, 318, 338, 372, 424, 443–445, 454, 458, 466, 516, 548–551, 556, 557, 562, 563, 572, 574, 577, 592, 689, 714, 740
 Molecular descriptors, x, 29–31, 58, 59, 103, 104, 117, 118, 120, 145, 155, 164, 165, 168, 177, 197, 205, 222, 242, 243, 373, 374, 408, 452–453, 463, 466, 493, 501, 502, 514, 520, 534, 571, 579, 625, 626, 629, 641, 643, 646, 647, 649–656, 663, 664, 666, 667, 695, 698, 791, 794, 801–817
 Molecular fingerprints, 295, 621, 802, 811–815, 817
 Molecular interaction fields (MIF), 216, 219, 220, 227, 522
 Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA), 225
 Molecular shape analysis (MSA), 219
 Monitor, 196, 258, 485
 MTT, 319–311
 Multi-Attribute Decision Making (MADM), 323
 Multi criteria decision making approach, 368
 Multiple linear regression (MLR), 32, 34, 44, 61, 168, 169, 187, 225, 227, 393, 396, 398, 402, 409, 413–415, 417, 419, 424, 426–431, 449, 450, 453, 492, 497, 501, 502, 504, 519, 526, 527, 532–534, 537, 578, 579, 582, 603, 604, 623, 624, 627, 663, 666, 667, 672–675, 695, 714

Multiple linear regression (MLR) analysis, 520

Multi-scale, 639–656

Multi-walled CNT (MWCNT), 486–488, 490–492, 496–505, 704, 795, 798

Mutagenicity, 117–119, 129–131, 133–143, 145, 168, 293, 296, 313, 371, 407, 413–416, 418, 419, 481, 568, 628, 717, 726, 731, 740, 741, 751, 761–763, 770, 771, 782

N

Naïve Bayes (NB), 114, 165, 166, 183, 621

Nanoforms regulation, 13, 14

Nanoinformatics, 789–798

Nanomaterials, vii, viii, 13, 15, 28, 30, 41–45, 48, 49, 178, 215–230, 291, 293–295, 410, 438, 480, 490, 728, 790

Nanoparticles, 28, 42–44, 48, 195–207, 217, 218, 481, 500, 615, 628, 693, 799–798

Nanostructures, ix, 197, 198, 215–230, 499, 790

Narcosis, 333, 334, 336, 339, 341, 342, 409, 458, 663, 701

National Toxicology Program (NTP), 467, 717, 726, 735–738, 745, 747

Neonicotinoid, x, 513–537

Neural networks (NN), 33, 113, 114, 121–124, 165, 167, 169, 179, 183, 187, 188, 200, 201, 208, 210, 273, 370, 371, 409, 413, 415, 419, 423, 427, 431, 449, 453, 502, 504, 524, 532, 537, 567, 616, 620, 621, 627, 645, 694, 698, 774

Neurotoxicity, 22, 86–88, 90, 535, 627, 738

Neutral red (NO), 309–311, 322, 747

Non-linear QSAR, 177–189

Non-polar narcosis, 341, 342, 409, 663

Non-testing, 235–265, 290, 293, 296, 388, 492, 563, 740, 790, 793

Non-toxic environment, 22, 23

No obeserved effect concentration (NOEC), 332, 335, 346–354, 389, 424, 439, 442, 446, 575, 711

Normalization of chemical structures, 99–100

O

Obesity, 83, 84, 619

Octanol/water partition coefficient (log KOW), 14, 56, 291, 333, 376, 428, 453, 458, 465, 504, 519, 527, 534, 547–549, 551, 552, 556, 573, 577, 605, 662, 663, 675, 683, 685, 723, 769, 770, 780, 809

Oil spills, 36, 617

Online Chemical Database (OCHEM), 105, 154–161, 169, 568, 804

Optimised Strategies for Risk Assessment of Industrial Chemicals through Integration of Non-test and Test Information (OSIRIS), 40, 740, 741

Organic pollutants, 36, 48, 83, 86, 369, 426, 480, 491–493, 496, 497, 502, 505, 580, 626

- Organization Environmental Footprint (OEF), 321, 322
 Organization for Economic Cooperation and Development (OECD), 19, 30, 60, 154, 245, 272, 292, 335, 361, 389, 409, 450, 547, 563, 592, 621, 684, 709, 790
 criteria for scientific validity of QSARs, 409, 550
 principles, 19, 34, 60, 61, 368, 409, 450, 621, 626, 630, 696
 TG 201, 594, 595, 599, 712
 TG 202, 596, 602, 712
O
 Organophosphorus, x, 79, 88, 513–538, 569, 572, 574, 575
P
 Paints, 360, 478, 615, 617
 Parabens, 358, 362, 374, 598, 599, 607, 608, 625
 Parkinson’s disease, 535, 619, 738
 PatchDock, 225
 PBT assessment, 119–121
 PCP ingredients, 46, 368
 Perfluoroctanesulfonates (PFOS), 37, 478, 486
 Perfumes, 358, 478, 615
 Persistent, bioaccumulative, and toxic (PBT), 8, 9, 27, 41, 116, 117, 119–121, 145, 299, 358, 365–369, 375, 376, 546, 721, 763, 764, 771
 Persistent, bioaccumulative, and toxic/very persistent and very bioaccumulative (PBT/vPvB), 119, 375, 546
 Personal care products (PCP), xi, vii, viii, 42, 45, 46, 48, 357–381, 479, 482, 625, 693, 711, 782
 Pesticides, x, vii, 7, 36, 78, 79, 81, 83, 87, 88, 120, 121, 132, 157, 159, 255, 375, 390, 394, 402, 439, 442, 447, 458, 460, 469, 478–482, 492, 495, 497, 505, 513–537, 561–582, 592, 599, 600, 607, 610, 615, 617–622, 629, 639–656, 715, 722, 723, 738, 739, 745, 764, 785
 Pharmaceuticals, 7, 28, 87, 113, 161, 216, 235, 331, 359, 407, 438, 479, 615, 682, 709
 PHASE, 222, 223
 Photodegradation, 374, 377, 425, 564
Pimephales promelas, 46–48, 336, 372, 626, 628, 649, 666, 667, 672, 715
 Plant protection products regulation, 15, 16
 Plasticizers, vii, 36, 478, 615
 Plastic waste, 483, 617
 Pollinators, 619, 713
 Pollutants, x, ix, vii, 20, 28, 35, 36, 41, 48, 49, 78–80, 83, 85–90, 151–153, 305, 307, 311, 322, 334, 369, 377, 381, 388, 426, 462, 477–506, 573, 578, 580, 617, 619, 620, 623, 626, 628, 629, 665, 693, 709, 748
 Pollution, vii, 6, 20, 42, 77, 167, 236, 411, 441, 461, 478, 480–482, 485, 488, 490, 492, 494, 497, 506, 615, 617–619, 712, 714, 721, 728, 782
 Polychlorinated biphenyls (PCBs), 37, 83, 85, 89, 91, 358, 480
 Polychlorinated dibenzofurans, 37, 480
 Polychlorinated dibenzo-p-dioxins, 480, 483
 Polycyclic aromatic hydrocarbons (PAH), 20, 83, 85, 86, 118, 478, 483, 498, 499, 628, 629
 Polyethylene, 488
 Polymer, 197, 209, 362, 481, 546, 682
 Potential risks, 15, 17, 18, 77, 82, 363, 377, 381, 482, 572, 616, 683
 Predicted no effect concentration (PNEC), 8, 19, 294, 332, 546–548, 553, 554
 Prediction, 20, 33, 55, 100, 111, 156, 178, 196, 216, 236, 271, 291, 318, 359, 387, 408, 439, 481, 514, 546, 562, 592, 616, 640, 663, 687, 710, 762, 790, 815
 Predictive Effect Concentration (PEC), 18, 19, 546, 740
 Pre-screening, 592, 593, 596, 597, 599–602, 606–608
 Primary aliphatic amines, 603
 Principal component analysis (PCA), 32, 216, 368, 394, 498, 695, 803, 809, 810
 Prioritization, 46, 47, 113, 130, 241, 302, 368, 369, 410, 411, 565, 621, 682, 690, 740, 762–764, 770, 802
 Product Environmental Footprint (PEF), 321, 322
 Project COMBASE, 387–403
 Project DEMETRA, 120, 760–762
 Project JANUS, 763, 770–772
 Project OptiTox, 783–784
 Project toDIVINE, 763, 780–781, 784
 Project VERMEER, 781, 782, 784
 Protein corona, 796–798
Pseudokirchneriella subcapitata, 45, 276–283, 592, 605, 608, 626
 Public health, 23, 365, 731, 735
 Python, 59, 64, 66–70, 73, 276, 277
Q
 QSARINS, 367–369, 625
 Quantitative Ion Character–Activity Relationship (QICAR), 504
 Quantitative nanostructure–property relationship (QNPR), 499
 Quantitative structure–activity/property relationship (QSAR/QSPR), ix, 155, 202, 217, 218, 407, 408, 410, 431, 481, 616, 617
 Quantitative structure–activity relationship (QSAR), 13, 28, 55, 97, 112, 157, 177, 195, 215, 242, 275, 332, 359, 388, 407, 449, 481, 514, 546, 562, 592, 617, 640, 661, 710, 759, 790, 801
 of heavy metals and their mixtures, 461–462
 of mixture, 450
 of mixture of agrochemicals, viii, 28, 41, 359, 460, 537

- Quantitative structure-activity relationship (QSAR)
 (cont.)
 models, 20, 29, 55, 98, 115, 161, 177, 195, 222, 243, 298, 354, 360, 388, 407, 450, 522, 550, 567, 621, 640, 662, 714, 750, 759, 802
 of organic chemical mixtures, 462–466, 469
 of pharmaceutical mixtures, 454–459
 toolbox, 248–260, 262, 296, 375, 391, 563, 596, 684, 728
- Quantitative structure–activity/toxicity (QSAR/QSTR), 359
- Quantitative structure–property relationship (QSPR), 29, 97, 98, 154, 168, 169, 178, 218, 409, 426, 431, 477–506, 579, 624, 662, 698, 699, ix
- Quantitative structure–toxicity relationship (QSTR), 29, 31, 34, 42–47, 97, 98, 100–106, 178, 196, 202–211, 367, 411, 424, 466, 521, 522, 524, 620, 621, 627, 628, 630, 666, 667, 671, 675, 714, 791
- Quantum chemical descriptors, 43, 225, 230, 413, 419, 426, 429, 519, 570–572, 596, 624, 699
- Quaternary ammonium compounds, 362, 372
- R**
- Rainbow trout, 368, 394–398, 400, 402, 519, 562, 567, 622, 760, 765, 775, 776
- Random forest (RF), viii, 33, 113–119, 165–167, 169, 170, 199, 201, 202, 273–276, 284, 413, 450, 621, 719
- Raphidocelis subcapitata, 422, 463, 592
- REACH regulation, 7–15, 19, 20, 22, 119, 157, 248, 290, 292, 627, 629, 782
- Reactivity, 156, 228, 245, 366, 372, 410, 417, 427, 429, 491, 519, 547–549, 551, 555, 572, 580–582, 610, 663, 664, 682, 810
- Read-across, 47, 242, 290, 370, 546, 562, 592, 683, 727, 762, 790
- Read-across assessment framework (RAAF), 229, 547
- Registration, Evaluation, Authorization, Evaluation and Restriction of Chemical (REACH), xi, vii, 7–17, 19–22, 46, 119, 156–158, 163, 238, 248, 290–293, 388, 410, 492, 545–558, 616, 627, 629, 683, 685, 690, 701, 702, 739, 740, 753, 763, 782, 783, 798, 802
- Registry of Industrial Toxicology Animal-data (RITA), 743–745
- Regression, viii, 32, 42–44, 46, 47, 61, 113, 116, 120, 142, 145, 155, 163, 164, 168, 177, 179, 183, 200, 222, 225, 227, 259, 273, 297, 409, 413–415, 419, 422–424, 429–431, 449, 450, 453, 461, 462, 465, 495, 496, 503, 520–522, 526, 527, 530, 533, 537, 567, 582, 606, 624, 666–675, 694, 695, 697–698, 701, 714, 715
- RepDose database, 370, 724–725
- Repeated-dose, 294, 295, 370, 371, 724, 727, 741
- Reproducibility, 56, 57, 60, 61, 63, 73, 129, 291, 309
- Reproducible, ix, 55–73, 178, 257, 297, 323, 411, 765, 770
- Reproductive effects, 45, 81, 83
- Research reproducibility, 56, 73
- Risk assessment, 13, 28, 78, 113, 156, 197, 290, 316, 359, 388, 411, 438, 481, 515, 562, 592, 616, 685, 712, 781, 789
- Risk Assessment Information System (RAIS), 741–742
- Risk information exchange (RiskIE), 473, 742, 743
- Risk management, 13, 23, 47, 365–366, 483–485, 731, 753
- Robustness, 33, 63, 73, 182, 227, 245, 248, 284, 403, 410, 450, 492, 499, 532, 550, 605–607, 609, 610, 616, 624, 626, 627, 743, 795
- 3Rs principles, viii, 19, 21, 23, 309, 710, 753
- Rubber, 313, 359, 362, 483, 617
- S**
- Saccharomyces cerevisiae*, 312
- Salmonella typhimurium, 44, 129, 249, 309, 313, 419, 628, 737
- Sampling, 78, 79, 307, 698
- SARpy, 115, 118, 120, 121, 131, 772–774, 777, 780, 784
- Scientific Committee on Consumer Safety (SCCS), 364
- Scrape loading and dye transfer, 316
- Screening, viii, 36, 37, 43, 47, 106, 123, 130, 164, 222, 224, 225, 291, 293, 302, 368, 376, 389, 410, 411, 419, 431, 537, 556, 564, 569, 580, 592, 593, 596, 621, 625, 628, 649, 711, 740, 742, 746, 753, 760, 762–764, 770, 789, 790, 792, 796, 797
- Self-Consistent Atomic Property Fields by Optimization (SCAPFOLD), 222
- Self-Organizing Molecular-Field Analysis (SOMFA), 221
- Sensors, 489, 490, 534
- Sentinels, ix, 77–91
- Similar action (dose/concentration addition), 443, 444, 446
- Similarity, viii, 32, 47, 88, 104, 105, 107, 112, 114, 116, 183, 218, 220, 230, 242, 244, 262, 275, 291, 293, 295, 296, 298, 301, 370, 412, 414, 420, 422, 548, 550, 551, 557, 567, 623, 624, 717, 734, 751, 766–770, 774, 775, 780, 781, 808, 812
- Simplified molecular-input line-entry system (SMILES), 102, 106, 119, 131, 132, 144, 154, 250, 296, 335, 412, 418, 498, 500, 524, 536, 549, 564, 578, 595, 596, 611, 642, 717, 765, 772, 773, 776, 777, 792, 794, 795
- Single Cell Gel Electrophoresis (SCGE), 314
- Single-walled carbon nanotubes (SWNTs), 227, 493, 503
- Single-walled CNTs (SWCNT), 228, 229, 485–487, 490, 492–496, 503, 505

- SMARTS notation, 596, 603, 609, 611, 814
 Soil organisms, 153
 Solar cells, 488
 Solubilities, 14, 41, 42, 169, 202, 227, 243, 291, 299, 340, 341, 407, 420, 485, 547, 555, 556, 572, 595, 596, 603, 607, 662, 666, 687, 701, 723, 771, 774, 780
 Solutions, 17, 22, 23, 106, 137, 185, 201, 262, 426
 Solvents, vii, viii, 36, 42, 202, 227, 238, 239, 264, 313, 418, 427, 462, 478, 485, 495, 615, 626, 627, 697, 723, 782
 Standardization, 41, 80, 103, 107, 154, 276, 292, 310, 361, 592, 621, 743, 804
 Standardization of ecotoxicological tests, 154–161
 Statistical parameters, 396, 520, 536, 603, 620, 630
 Steric and electrostatic alignment (SEAL), 218
 Structural alerts (SAs), 112, 244, 254, 257, 296–298, 420, 443, 568, 592, 597, 599, 600, 607, 714, 727, 768, 769, 772, 778, 780
 Structural keys, 805, 811–813, 815, 817
 Structure activity-relationships (SARs), 112, 115, 118–120, 372, 412, 531, 533, 564, ix, viii, 10, 27–50, 154, 159, 177, 195, 215, 243, 248, 263, 293, 332, 333, 369, 407, 417, 430, 449–459, 519, 524, 525, 533, 546, 564, 568–570, 577–579, 592, 623, 625, 640, 710, 714, 759, 790
 Structure of the database, 154
 Structure of the modeling process, 155
 Sulfamethoxazole, 459, 479
 Support vector machine (SVM), viii, 113, 114, 165, 170, 179, 183, 187, 201, 227, 273, 450, 492, 504, 532, 537, 621, 623, 714
 Surfactants, 47, 90, 367, 372, 478, 615, 624, 627, 687, 697
 Synergistic, 15, 16, 84, 306, 439, 442, 454, 456, 466, 574–576, 579, 622
- T**
- Terrestrial ecosystems, 615, 617
 - Test guidelines (TGs), 18, 159, 161, 335, 546, 592, 594, 610, 611, 627, 694, 712, 713
 - Test protocols, 602
 - Tetracyclines, 457, 479, 490, 491, 595, 600, 602, 607
 - Tetrahymena pyriformis*, 424, 666, 668, 674, 675
 - Textiles, 22, 43, 406, 412, 416, 418, 431, 479, 480, 483, 487
 - Thiochemicals, x, 545–558
 - The three-step strategy
 - Toll-like receptors (TLR), 228–230
 - Tools, 28, 63, 89, 97, 113, 151, 178, 197, 241, 285, 294, 306, 359, 389, 407, 439, 493, 514, 564, 616, 650, 685, 714, 759, 791, 802
- ToxCast, 256, 568, 720, 746–749, 783
 Toxicants, 36, 48, 49, 87–89, 339, 342, 381, 445, 466, 557, 562, 577
 Toxicity, 10, 28, 79, 97, 111, 152, 177, 197, 228, 241, 273, 290, 306, 333, 358, 388, 406, 438, 481, 514, 562, 592, 616, 640, 661, 683, 710, 761, 790, 802
 Toxicity estimation software tool (T.E.S.T), 246, 248, 374, 564, 693
 Toxicity Reference Database (ToxRefDB), 38, 256, 749
 The Toxic mode of action (MOA), 157, 548, 550–552, 557, 566, 714, 740
 Toxicodynamics, 534, 536, 548, 582, 785
 Toxicokinetics, 56, 369, 445, 448, 548, 551, 575, 582, 737, 784
 Toxicology Data Network (TOXNET), 298, 718, 726, 748–750
 Toxicology testing in the 21st Century (Tox21), 720, 745, 746, 783
 Toxic ratios, 334, 457, 562
 Toxic Substances Control Act Test Submissions (TSCATS), 750
 TOXMAP, 748–750
 Transformation products, vii, 306, 372, 377, 399, 454, 460, 562–563, 565–570, 572, 582, 583
 Trend analysis, 245, 259
 Tributyltins (TBT), 362, 376
 Triclosan, 358, 377, 380, 458, 625
 Tumors, 81, 85, 86, 309, 310, 316–318, 322, 717, 725, 726, 735, 751
- U**
- United States Environmental Protection Agency (US EPA), 28, 116, 161, 246, 360, 374, 438, 479, 482, 563, 592, 616, 715, 740, xi
- USETox, 321
- US FDA chemical evaluation and risk estimation system (CERES), 740, 750, 751
- UV filters, 47, 358, 362, 368, 372
- V**
- Validity, 181, 273, 274, 281, 282, 409, 421, 550, 558, 629
 - Variable reduction, 803, 811, 812
 - Varnishes, 615
 - Verhaar scheme, 334, 338, 339, 341, 593, 596
 - Veterinary medicinal products, 18
 - Vibrio fischeri*, *see Aliivibrio fischeri*
 - Virtual models for property evaluation of chemicals within a global architecture (VEGA), 119, 120, 131, 247, 248, 262, 295, 297, 301, 388, 401, 403, 759–770, 772–781, 783–785

830 | ECOTOXICOLOGICAL QSARs Index

Virtual screening, 22, 164, 431, 537, 628, 790, 792, 796, 797
Viruses, 89, 133, 225, 489, 529, 728
VITIC, 723, 751, 752

W

Waste, 6, 16, 17, 22, 45, 236, 238, 240, 241, 438, 478, 479, 482–484, 617, 625
Waste disposal, 617
Waste oil, 617
Wastewater, 17, 236, 238, 305–323, 333, 358, 369, 388, 425, 431, 479, 563, 625, 753
Wastewater treatment plant (WWTP), 305, 306, 321, 322, 358, 369, 625
Water filter, 489
Water flea, 309, 310, 322, 376
Water Framework Directive (WFD), 20, 441
Water organisms, 153, 447
Weighted descriptors approach, 452, 466
Weighted Holistic Invariant Molecular (WHIMs) descriptors, 803, 805, 808, 809

Weight of evidence (WoE), 12, 157, 263, 294, 296, 416, 546, 552, 553, 556, 558, 593, 726, 764, 765, 770, 776, 780

Whole-mixture approach, 442, 443

Wiener index, 806, 808

WikiPharma, 752

Wildlife, 16, 35, 77–91, 161, 332, 438, 720

World Health Organization (WHO), 17, 81, 82, 323, 438, 728

Worst-case prediction, 550, 551

Wrapper methods, 181, 183, 184, 189

Y

Yeast, 256, 312, 721

Z

Zinc, 99, 380, 461, 479, 489, 801

Zinc oxide, 489