

Query Details

[Back to Main Page](#)

1. Please check whether the author names and affiliations are correct.

yes, they are correct

2. Please check the sentence "CNNs take in" for clarity.

can be changed in:

CNNs store an image in the input neurons, ...

3. Please check whether the edit made to the term "crate" is correct.

yes

4. Please check whether the edit made to the sentence "CNNs take in" is correct.

the edit is correct; however I proposed another formulation in AQ2

5. Reference "18" was not cited anywhere in the text. Please provide in text citation or delete the reference from the reference list.

citation should be in 4.1.2 at the end of this sentence:

"Often the Quantitative Structure–Activity Relationship (QSAR) literature presents models developed on the few good–quality available datasets, but in most of the real applications data should be found from the prime sources (laboratory studies) and checked to be coherent, of quality, and really comparable [1, 18]."

6. Please provide missing volume number and page range for reference [9].

it is a book

QSAR Methods

Giuseppina Gini

Email : giuseppina.gini@polimi.it

Affiliationids : Aff1, Correspondingaffiliationid : Aff1

Aff1 DEIB, Politecnico di Milano, Milan, Italy

Abstract

This chapter introduces the basis of computational chemistry and discusses how computational methods have been extended from physical to biological properties, and toxicology in particular, modeling. Since about three decades ago, chemical experimentation is more and more replaced by modeling and virtual experimentation, using a large core of mathematics, chemistry, physics, and algorithms. Animal and wet experiments, aimed at providing a standardized result about a biological property, can be mimicked by modeling methods, globally called in silico methods, all characterized by deducing properties starting from the chemical structures. Two main streams of such models are available: models that consider the whole molecular structure to predict a value, namely QSAR (quantitative structure–activity relationships), and models that check relevant substructures to predict a class, namely SAR. The term in silico discovery is applied to chemical design, to computational toxicology, and to drug discovery. Virtual experiments confirm hypotheses, provide data for regulation, and help in designing new chemicals.

Key words

Computer models

Chemometrics

Toxicity prediction

SAR

QSAR

Model interpretation

1. Starting from Chemistry

“All science [AQ1](#) is computer science.” When a New York Times article in 2007 used this title, the general public was aware that the introduction of computers has changed the way that experimental sciences have been carried out so far. Chemistry and physics are the best examples of such a new way of making science.

A new discipline, *chemoinformatics*, has been in existence for the past few decades [[1](#), [2](#)]. Many of the activities performed in chemoinformatics are information retrieval [[3](#)], aimed at searching for new molecules of interest when a single molecule has been identified as being relevant. However, chemoinformatics is more than “chemical information”; it requires strong algorithmic development.

It is useful to remember that models of atoms were defined by analogy with different systems; Thomson in 1897 modeled the atom as a sphere of positive electricity with negative particles; Rutherford in 1909 adapted the solar system model with a dense positively charged nucleus surrounded by negative electrons. Finally, in the 1920s, the electron cloud model was defined; in this model, an atom consists of a dense nucleus composed of protons and neutrons surrounded by electrons. A molecule is an electrically neutral group of two or more atoms held together by covalent bonds, sharing electrons. The valence model naturally transforms a molecule into a graph, where the nodes are atoms and the edges are bonds. This graph representation is usually called 2D chemical structure.

The graph theory, whose basic definition has been established back in the eighteenth century, initially evolved through chemistry. Two scientists in particular, Alexander C. Brown and James J. Sylvester, developed the molecular representation as nodes (atoms, indicated by their name) and bonds. The edges are assigned weights according to the bond: single, double, triple, or aromatic where electrons are delocalized. Today, hydrogen atoms are implicitly represented in the graph since they are assumed to fill the unused valences [[4](#)].

A common representation of the graph is the adjacency matrix, a square matrix with dimension N equal to the number of atoms. Each position (i, j) in the matrix specifies the absence (0 value) or the presence of a bond connecting the atoms i and j , filled with 1, 2, or 3 to indicate simple, double, or triple bond, 4 for amide bond, and 5 for aromatic bond. The diagonal elements are always zero. An example of a matrix representation is in Fig. [1](#).

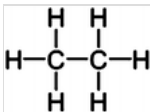
Fig. 1

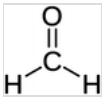
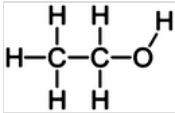
Adjacency matrix of 2-methylbutane pictured after hydrogen elimination; only the five Ccarbon atoms are considered and numbered

The Simplified Molecular Input Line Entry Specification (SMILES) [[5](#)] is a very popular string representation of a molecule. This string is formally defined in a context-free language and lists bonds and atoms encountered in the depth-first visit of the chemical graph, using parentheses to enclose branches. Hydrogens are left out. Table [1](#) shows examples of SMILES representations.

Table 1

Examples of molecules with their SMILES code

SMILES	Name	Formula	Graph
CC	Ethane	CH ₃ CH ₃	

SMILES	Name	Formula	Graph
<chem>C=O</chem> <chem>O=C</chem>	Formaldehyde	CH_2O	
<chem>CCO</chem> <chem>OCC</chem> <chem>C(C)O</chem> <chem>C(O)C</chem>	Ethanol	$\text{CH}_3\text{CH}_2\text{OH}$	

The SMILES notation suffers the lack of a unique representation, since a molecule can be encoded beginning anywhere; in Table 1, four SMILES strings, all correct, represent ethanol. Therefore, a method of encoding a molecule was quickly developed that provided an invariant SMILES representation called canonical SMILES [6].

Recent developments in chemical notations are the InChI (International Chemical Identifier) codes, supported by the International Union of Pure and Applied Chemistry (IUPAC), which can uniquely describe a molecule, at different levels of detail, but is not intended for human readability [7].

What about the real shape of molecules? They are 3D objects, and as such they should be represented. Considering the 2-methylbutane molecule, illustrated as a simple drawing in Fig. 1, its formula, SMILES, and 3D conformation are illustrated in Fig. 2a-c.

Fig. 2

The 2-methylbutane molecule: (a) chemical formula, (b) SMILES, (c) 3D conformer (from NIH PubChem)

Defining the 3D shape of a molecule is a basic topic of computational chemistry.

2. Computational Chemistry

Computational chemistry is a branch of chemistry that uses computers to assist in solving chemical problems, studying the electronic structure of molecules, and designing new materials and drugs. It uses the results of theoretical chemistry, incorporated into programs, to calculate the structures and properties of molecules. The methods cover both static and dynamic situations: accurate methods—ab-initio methods, and less-accurate methods—called *semi-empirical*.

It all happened in the second half of the last century.

1. In the early 1950s, the first semiempirical atomic orbital calculations.
2. The first ab initio calculations in 1956 at MIT.
3. Nobel prize for Chemistry, in 1998, assigned to John Pople and Walter Kohn, for Computational Chemistry.
4. Nobel prize for Chemistry assigned in 2013 to chemistry assigned to Martin Karplus, Michael Levitt, and Arieh Warshel for their development of multiscale models for complex chemical systems.

Computational chemistry is a way to move away from the traditional approach of solving scientific problems using only direct experimentation, but it does not remove experimentation. Experiments produce new data and facts. The

uses theories to produce new facts in a manner similar to the real experiments. It is now possible to simulate in the computer an experiment before running it.

In modeling chemical processes, two variables are important, namely time and temperature. It is necessary to make dynamic simulations and to model the force fields that exist between atoms and explain the bonds breaking. This task usually requires solving the quantum mechanics equations.

A hierarchy of simulation levels provides different levels of details. The study of the fundamental properties without the introduction of empirical parameters is the so-called *ab initio* methods. Those computations consider the electronic and structural properties of the molecule at the absolute zero temperature. They are computationally expensive, so the size of the molecules is limited to a few hundred atoms. When the *ab initio* methods cannot be used, it is possible to introduce empirical parameters and we have the so-called molecular dynamics methods [8].

Today, the applications of quantum mechanics to chemistry are widely used. The most notable is the Gaussian software (<https://gaussian.com/>), developed at the Carnegie Mellon University of Pittsburgh. This program gained a large popularity since the Nobel Prize for chemistry in 1998 was assigned to Pople, one of the inventors of Gaussian.

2.1. Molecular Simulation

Using computers to calculate the intermolecular forces, it is possible to compute a detailed “history” of the molecules. Analyzing this history, by the methods of statistical mechanics, affords a detailed description of the behavior of matter [8]. Three techniques are available:

1. Molecular Dynamics (MD) simulation. In this technique, the forces between molecules are calculated explicitly, and the motion of the molecules is computed using a numerical integration method. The starting conditions are the positions of the atoms (taken from crystal structure) and their velocities (randomly generated). Following Newton's equations, from the initial positions, velocities, and forces, it is possible to calculate the positions and velocities of the atoms at a small time interval later. From these new positions, the forces are recalculated and another step in time made. Following an equilibration period of many thousands of time steps, during which the system “settles down” to the desired temperature and pressure, a production period begins where the history of the molecules is stored for later analysis.
2. Monte Carlo (MC) simulation. Monte Carlo simulation resembles the Molecular Dynamics method in that it also generates a history of the molecules in a system, which is subsequently used to calculate the bulk properties of the system by means of statistical mechanics. However, the procedure for moving the atoms employs small random moves used in conjunction with a sampling algorithm to confine the random walk to thermodynamically meaningful configurations.
3. Molecular Mechanics (MM) modeling. MM is a method for predicting the structures of complex molecules, based on the energy minimization of ~~the~~^{its} potential energy function, obtained empirically, by experiment, or by the methods of quantum chemistry. The energy minimization method is an advanced algorithm to optimize the speed of convergence. The method's main advantage is its computational cheapness.
4. Any of those methods is necessary to optimize the 3D structure of the molecule before constructing models that use the 3D shape instead of the graph representation of the molecules.

2.2. Biological Effects of Chemicals and Toxicology

Since the nineteenth century, the practice of animal experimentation was established in physiology, microbiology, and surgery. The explosion in molecular biology in the second half of the twentieth century increased the importance of *in vivo* models [9].

All models have their limitations; their prediction can be poor, and their transferability to the real phenomena they model can be unsatisfactory. So extrapolating data from animal models to the environment or to human health depends on the degree to which the animal model is an appropriate reflection of the condition under investigation.

These limitations are, however, an intrinsic part of all modeling approaches. Most of the questions about animal models are ethical more than scientific; in public health, the use of animal models is imposed by regulations and is unlikely that any health authority will allow novel drugs without supporting animal data.

3. Bioassays for Toxicity

Toxicity is the degree to which a substance can damage an organism. Toxicity is a property of concern for every chemical substance. Theophrastus Phillipus von Hohenheim (1493–1541) Paracelsus wrote: “All things are poison and nothing is without poison; only the dose makes a thing not a poison.”

The relationship between dose and its effects on the exposed organism is of high significance in toxicology. The

process of using animal testing to assess toxicity of chemicals has been defined in the following ways:

- Toxicity can be measured by its effects on the target.
- Because individuals have different levels of response to the same dose of a toxin, a population-level measure of toxicity is often used which relates the probabilities of an outcome for a given individual in a population. Example is LD_{50} : the dose that causes the death of 50% of the population.
- When the dose is individuated, define “safety factors.” For example, if a dose is safe for a rat, one might assume that a small percentage of dose would be safe for a human, allowing a safety factor of 100, for instance.

This process is based on assumptions that usually are very crude and presents many open issues. For instance, it is more difficult to determine the toxicity of chemical mixtures (gasoline, cigarette smoke, waste) since the percentages of the chemicals can vary, and the combination of the effects is not exactly a summation.

Perhaps the most common continuous measure of biological activity is the $\log (IC_{50})$ (inhibitory concentration), which measures the concentration of a particular compound necessary to induce a 50% inhibition of the biological activity under investigation. Similarly, the median lethal dose, LD_{50} , is the dose required to kill half the members of a tested population after specified test duration. LD_{50} is not the lethal dose for all subjects, only for half of them [10].

The dose-response relationship describes the change in effect on an organism caused by differing levels of doses to a chemical after a certain exposure time. A dose-response curve is an x - y graph relating the dose to the response of the organism.

- The measured dose is plotted on the X axis, and the response is plotted on the Y axis.
- The response may be a physiological or biochemical response; LD_{50} is used in human toxicology; IC_{50} inhibition concentration and its dual EC_{50} effect concentration is used in pharmacology.

Usually, the logarithm of the dose is plotted on the X axis, and in such cases, the curve is typically sigmoidal, with the steepest portion in the middle. Figure 3 shows an example of the dose-response curve for LD_{50} .

Fig. 3

A curve for $\log LD_{50}$. (Source: http://www.dropdata.org/RPU/pesticide_activity.htm)

Today also, *in vitro* testing is available. It is the scientific analysis of the effects of a chemical on cultured bacteria or mammalian cells. Experiments using *in vitro* systems are useful in the early phases of medical studies where the screening of large number of potential therapeutic candidates may be necessary, or in making fast tests for possible pollutants. *In vitro* systems are nonphysiological and have important limitations, as their results poorly correlate with *in vivo*. However, there are substantial advantages in using *in vitro* systems to advance mechanistic understanding of toxicant activities and the use of human cells to define human-specific toxic effects.

Both *in vivo* and *in vitro* methods use the chemical substance; in case of initial steps in designing new drugs or materials, the chemical substance is only designed, not yet available. Other methods that need only the chemical structures are needed, and they are collectively named *in silico*.

4. In Silico Methods

Animal testing refers to the use of nonhuman animals in experiments. Worldwide, it is estimated that the number of vertebrate animals annually used for animal experiments is in the order of tens of millions. In toxicity, animal tests are called *in vivo* models; they give doses for some species and are used to extrapolate data to human health or to the environment. As we said above, the extrapolation of data from species to species is not obvious. For instance, the lethal doses for rats and mice are sometimes very different.

How to construct a model that relates a chemical structure to the effect was investigated even before computers were available. The term *in silico* today covers the methods devoted to this end, and *in silico* refers to the fact that computers are used, and computers have silicon in their hardware.

The most known *in silico* methods are the (Q)SARs—(quantitative) structure–activity relationships methods, based on the premise that the molecular structure is responsible for all the activities and aimed at finding the dose or the substructure responsible for the activity [11, 12, 13]. QSAR is based on the concept of similarity, as similar molecules are expected to have similar effects, and learns this relationship from many examples. Read across is another *in silico* method, strictly based on finding the most similar molecule and adapting its experimental value to the molecule under investigation. Figure 4 illustrates the *in vivo*, *in vitro*, and *in silico* methods as a whole.

Fig. 4

The methods available for testing molecules. After chemical synthesis, *in vivo* and *in vitro* produce data to compute the standardize effect. Without chemical synthesis, molecular dynamics studies the 3D structures, (Q)SAR predicts the properties, and read across translates the properties of a molecule to a similar one

4.1. QSAR

Quantitative structure–activity relationship (QSAR) models seek to correlate a particular response variable of interest with molecular descriptors that have been computed or even measured from the molecules themselves. QSAR methods were first pioneered by Corwin Hansch [14] in the 1940s, which analyzed congeneric series of compounds and formulated the QSAR Eq. 1:

$$\text{Log } 1/C = a p + b s + c E_s + \text{const} \quad 1$$

where C = effect concentration, p = octanol–water partition coefficient, s = Hammett substituent constant (electronic), and E_s = Taft's substituent constant.

$\text{Log } P$, octanol–water partition coefficient [15, 16], is the ratio of concentrations of a compound in the two phases of a mixture of two immiscible solvents at equilibrium. It is a measure of the difference in solubility of the compound in these two solvents. Normally one of the solvents is water while the second is hydrophobic such as octanol. With high octanol/water partition coefficient, the chemical substance is hydrophobic, and preferentially distributed to hydrophobic compartments such as cell membrane, while hydrophilic are found in hydrophilic compartments such as blood serum. $\text{Log } P$ values today are often predicted.

The definitions of the chemical structure and of the function remain a challenge today, but relating structure to property is widely adopted in drug discovery and in risk assessment.

Sometimes, the QSAR methods take a more specific name as QSPR (quantitative structure–property relationship) when used for physicochemical properties, as the boiling point, the solubility, logP. They all correlate a dependent variable (the effect or response) with a set of independent variables (usually calculated properties, or descriptors). They are statistical models and can be applied to predict the responses for unseen data points without the need of using the real chemical, even without the chemical has been synthesized.

4.1.1. Molecular Descriptors

The generation of informative data from molecular structures is of high importance in chemoinformatics since it is often used in statistical analyses of the molecules. There are many possible approaches to calculate molecular descriptors [17] that represent local or global salient characteristics of the molecule. Different classes of descriptors are:

- Constitutional descriptors, which depend on the number and type of atoms, bonds, and functional groups.
- Geometrical descriptors, which consider molecular surface area and volume, moments of inertia, shadow area projections, and gravitational indices.
- Topological indices, based on the topology of molecular graph [4]. Only the structural information is used in generating the description. Examples are the Wiener index (the sum of the number of bonds between all nodes in a molecular graph) and the Randic index (the branching of a molecule).
- Physicochemical descriptors, which estimate the physical properties of molecules. Examples are water solubility and partition coefficients, as logP.
- Electrostatic descriptors, such as partial atomic charges and others depending on the possibility to form hydrogen bonds.
- Quantum chemical descriptors, related to the molecular orbital and their properties.
- Fingerprints. Since subgraph isomorphism (substructure searching in chemoinformatics) in large molecular databases is quite often time consuming, substructure screening was developed as a rapid method of filtering out those molecules that definitely do not contain the substructure of interest. The method used is Structure–Key Fingerprints; the fingerprint is a binary string encoding a molecule, where the 1 or 0 in a position means that the substructure of this position in the dictionary is present or not. The dictionaries are designed according to knowledge of chemical entities that can be of interest for the task at hand.

Another useful distinction considers the descriptors in a dimension space. 1D descriptors are atom and bond counts and molecular weight. 2D descriptors are derived from molecular topology and fragment count. 3D descriptors are calculated from quantum chemistry programs on the optimized 3D structure. Usually, 3D descriptors are not used in tasks that require rapid screening, because their computation can be quite expensive.

4.1.2. Model Construction of QSAR

The first step in model construction is the preparation of a dataset containing the available experimental data about the selected endpoint. This step is very demanding, as few datasets found in the literature are complete and validated. Often the QSAR literature presents models developed on the few good–quality available datasets, but in most of the real applications data should be found from the prime sources (laboratory studies) and checked to be coherent, of quality, and really comparable [1, 18].

The second step is computing and selecting the chemical descriptors. In the early days of QSAR, only a few descriptors a priori considered as crucial were used. Today, hundreds or thousands of 2D descriptors are easily computed; often it is preferred to compute a large number of them and then select the few ones that are more relevant considering simple methods as the build–up method (adding one at a time), or the build–down method (removing one at a time), or optimization methods for variable selection.

Whatever algorithm is then chosen to develop predictive models, it is important to take heed of the model quality statistics and ensure a correct modeling methodology is used to avoid overfitting or underfitting the training set. Numerous model statistics are available that can indicate if new data points, from which responses are to be predicted, can be applied to the model [19].

Two types of supervised learning models are of interest: classification and regression. Classification methods assign the target molecule to two or more classes—most frequently either biologically active or inactive. Regression methods attempt to use continuous data, such as a measured biological response variable, to correlate molecules

with that data so as to predict a continuous numeric value for new and unseen molecules using the generated model.

Another discrimination is between methods that are similarity-based, and tend to cluster similar molecules, and methods feature-based that use a set of molecular features (as the descriptors) to build the classification function.

A plethora of algorithms and methods is available for modeling, including statistical regression (linear discriminant analysis, partial least squares, and multiple linear regression) methods and machine learning methods (Naive Bayesian classifier, decision trees, recursive partitioning, random forest, artificial neural networks, and support vector machines).

In the last two decades, many methods derived from artificial intelligence became the methods of election in QSAR building [20], as they are not parametric and allow finding modeling functions without a priori decisions. Some more details of the most recent methods for model construction will be given in the following Subheading 5.

4.1.3. Model Performance Measures

In many cases, published QSAR models implement the leave-one-out cross-validation procedure and compute the cross-validated determination coefficient R^2 , called q^2 . If y_{pi} and y_i are the predicted and observed property values, y_{pi}^m and y_i^m respectively are the average values of the predicted and observed property values, and the determination coefficient is defined as in Eq. 2.

$$R^2 = 1 - \left(\text{SUM} (y_{pi} - y_i)^2 / \text{SUM} (y_{pi}^m - y_i^m)^2 \right) \quad 2$$

A high value of q^2 (greater than >0.5) is considered as an indicator or even as the ultimate proof that the model is predictive. A high q^2 is the necessary condition for a model to have a high predictive power; however, it is not a sufficient condition. In general, the procedure used is to divide the data set available into two parts, one used for training and the rest (often it is 20% of data) used for testing only. If the q^2 on the training and test sets are similar and high, it is assumed that the model is predictive.

From a statistic point of view, other parameters are more informative, and k-fold cross-validation is more consistent than the statistics on the testing sets [21].

Similar considerations apply to classifiers. Usually, the basic measures are true positive (TP), true negative (TN), false positive (FP), and false negative (FN) molecules in the training and test sets. From those numbers, important measures, as sensitivity and specificity, are computed (Eq. 3):

$$\text{Sensitivity} = \frac{TP}{TP + FN}; \text{Specificity} = \frac{TN}{TN + FP} \quad 3$$

A graphical view of sensitivity and specificity is the receiver operating characteristic (ROC) curve, while a more informative parameter that does not depend on class imbalance is the Matthews correlation coefficient (MCC), (Eq. 4), which returns a value between -1 and +1. Only positive values indicate that the model is better than random guessing.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TN + FP)(TN + FN)(TP + FP)(TP + FN)}} \quad 4$$

Regardless to the absolute values, we have to remember that the statistical performance of any model is related to the uncertainty and variability of the original data used to build the model.

4.1.4. Model Interpretation

Model interpretation is important; people aim at understanding the models from known basic principles. A low number of descriptors used and their role in a simple linear equation have been often considered as necessary to accept a QSAR result.

However, if the main aim of QSAR is prediction, the attention should be focused on the quality of the model, and not on its interpretation. Moreover, it is dangerous to attempt to interpret statistical models, since correlation does not imply causality. It is common to differentiate predictive QSARs, focused on prediction accuracy, from descriptive QSARs, focused on interpretability [22].

There is a trade-off between prediction quality and interpretation quality. Interpretable models are generally desired in situations where the model has to provide information about the problem domain and how to navigate

through chemistry space, allowing the medicinal chemist to make informed decisions. However, these models tend to suffer in terms of prediction quality as they become more interpretable. The reverse is true with predictive models in that their interpretation suffers as they become more predictive. Models that are highly predictive may use molecular descriptors that are not readily interpretable, as they are selected by algorithms to optimize the accuracy, or are expressed as implicit equations, as in case of neural networks. Highly predictive models can therefore be used as high-throughput models. If interpretability is of concern, other methods are available, as expert systems or SAR.

This view of the “two QSARs” is somehow obsolete as indicated by Polishchuk [22], who discusses how the concept of interpretation of QSAR models evolved. Interpretation is intended as the retrieval of useful knowledge from a model. Usually QSARs adopted the paradigm “model → descriptors → (structure),” that is, strictly based on the interpretability of the descriptors. Instead, the “model → structure” paradigm can be more general and useful. Often fingerprints descriptors are used, and they are directly related to the presence of substructure. Even in case not interpretable descriptors are used, any model can be made interpretable estimating the contributions of the molecules substructures. As a consequence, it is not important to use information about the machine learning method used, or the descriptors.

Livingstone [13] so concludes about the interpretability of QSAR models: “The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all accurate estimates of the chemicals activity.”

Both SAR and QSAR are predictive statistical models, and as such they suffer the problems of the statistical learning theory, the mathematics behind any inductive method that tries to gain knowledge from a set of data. It is well known that it is always possible to find a function that fits the data. However, such function could be very bad in predicting new data, in particular if data are noisy. Among the many functions that can accomplish the task of inducing a model, performance and simplicity characterize the most appreciated models. While performance is measured by accounting for the errors in prediction, simplicity has no unique definition; usually it means using few free parameter or descriptors. The limitations of all the inductive methods are mathematically expressed by the “no free lunch theorem.” “No free lunch theorem” is a popular name for the theorems demonstrated by Wolpert and Macready [23] stating that any two models are equivalent when their performance is averaged across all possible problems. The practical indication from these theorems is that, without assumptions on the phenomenon to study, the best algorithm does not exist. In other words, data cannot replace knowledge; the priors used explain the success in getting the prediction. Priors can derive from available knowledge expressed by the designer and given to the model, as in case of the relevant chemical descriptors, or can be automatically extracted looking at the structure of data, as in case of deep learning, presented in Subheading 5.

4.2. SAR

SAR (structure–activity relationships) typically makes use of rules created by experts to produce models that relate subgroups of the molecule atoms to a biological property. So, the SAR approach consists in detecting particular structural fragments of molecule already known to be responsible for the toxic property under investigation.

In the mutagenicity/carcinogenicity domain, the key contribution in the definition of such toxicophores comes from Ashby [24], who compiled a list of 19 structural alerts (SA) for DNA reactivity. Practically, SAs are rules that state the condition of mutagenicity by the presence or the absence of peculiar chemical substructures. SAs are sound hypotheses that derive from chemical and biological properties and have a sort of mechanistic interpretation; however, their presence alone is not a definitive method to prove the property under investigation, since the substituents present in some cases are able to change the classification.

SAs search is very common for properties as genotoxicity and carcinogenicity, where studies and implementations are available [25]. The extraction of SAs for other properties is seldom done. A few examples exist of automatic construction of such SAR systems, which use graph-mining approaches to mine large datasets for frequent substructures.

An automatic and freely available method for SAs extraction is SARpy (SAR in python), a data miner that automatically generates SAR models by finding the relevant fragments; it extracts a set of rules directly from data expressed in the SMILES notation, without any a priori knowledge [26]. Given a training set of molecular structures with their experimental activity labels, SARpy generates every substructure in the set and mines correlations between the incidence of a particular molecular substructure and the activity/inactivity of the molecules that contain it. This is done in three steps. First a recursive algorithm considers every combination of bond breakages and computes every substructure of the molecular input set. Then each substructure is validated as potential SA on the training set to assess its predictive power. Finally, the best substructures are kept as rules of the kind: “IF chemical contains <SAi> THEN <apply activity label>.” Moreover, the user can ask SARpy to also generate rules related to nontoxic substances and use them to better assign molecules to the nontoxic class. The extracted rule set becomes a SAR

model, applied to the target molecule for prediction; the model tags it as toxic when one or more SAs are present, as nontoxic if no SAs are found and rules of nontoxicity are present, as dubious in other cases.

5. New Trends in Building QSAR and SAR Models

The initial QSAR paradigm considered only congeneric compounds in the hypotheses that:

- Compounds in the series are closely related.
- Same mode of action is supposed.
- Basics biological activity are investigated.
- Linear relations are constructed.

Recently there has been an important change. Today, QSAR models are generated using a wide variety of statistical and machine learning methods, datasets containing a large variety of chemicals, and a large choice of molecular descriptors. QSARs may be simple equations or express nonlinear relationships between descriptors values and the property. Those changes have been necessary to cope with:

- Heterogeneous compound sets.
- Mixed modes of action.
- Complex biological endpoints.
- Large number of properties.

Alongside with new techniques for model building, new problems also emerged.

1. As more models are available for the same dataset, they can be used collectively to improve their performance; this requires defining the methods and the property of ensemble systems.
2. With larger datasets available for many properties of pharmaceutical or environmental interest, more rapid methods to deal with big data, in particular deep learning methods, became of interest in the QSAR community.
3. As a consequence of large adoption of machine learning and deep learning methods, computing and extracting features changed, as methods able to automatically compute features show performance similar or even better than methods using precomputed molecular descriptors.

5.1. Integrated and Ensemble Models

The development of computer programs able to contain in explicit form the knowledge about some domain was the basis of the development of “Expert Systems” in the 1970s [11]. Soon, expert systems moved from the initial rule-based representation to the modern modeling and interpretation systems. The starting “Machine Learning” community developed a way to make use of data in absence of knowledge which led to the development of Inductive Trees, well exemplified by C4.5 [27] and after by the commercial system CART.

Using different representations to reach a common agreement or a problem solution led to the idea of using computationally different methods on different problem representations, so to make use of their relative strengths. Examples are the hybrid neural and symbolic learning systems, and the neuro-fuzzy system that combines connectionist and symbolic features in form of fuzzy rules. While the neural representation offers the advantage of homogeneity, distribution, and parallelization and is effective with incomplete and noisy data, the symbolic representation brings the advantages of human interpretation and knowledge abstraction [28]. Independently, a similar evolution in the pattern recognition community proposed to combine classifiers. In this area, most of the intuitions started with a seminal work, about bagging classifiers [29], which opened the way to ensemble systems.

Combining the predictions of a set of classifiers has shown to be an effective way to create composite classifiers that are more accurate than any of the component classifiers. There are many methods for combining the predictions given by component classifiers, as voting, combination, ensemble, and mixture of experts [30]. Finally, it is possible to use a sequential approach, so to train the final classifier using the outputs of the input classifiers as new features. In QSAR, often only simple combinations, called consensus models, are used. Exploiting ensembles will offer many more ways of improving and comparing QSAR predictions [31].

Why ensembles works and why they outperform single classifiers can be explained considering the error in classifiers. Usually the error is expressed [32] as in Eq. 5:

$$\text{Error} = \text{noise} + \text{bias}^2 + \text{variance}$$

where *bias* is the expected error of the classifier due to the fact that the classifier is not perfect; *variance* is the expected error due to the particular training set used, and *noise* is irreducible.

Models with too few parameters can perform poorly, but the same applies to models with too many parameters. A model which is too simple, or too inflexible, will have a large bias, a model which has too much flexibility will have high variance. Usually, the bias is a decreasing function of the complexity of the model, while variance is an increasing function of the complexity, as illustrated in Fig. 5. The concepts of bias and variance are of help in understanding and getting rid of the noise. The predictor should be insensitive to the noise on the training data, to reduce variance, but flexible enough to approximate the model function, and so minimize bias. There is a trade-off between the two components of the error, and balancing them is an important part of the error reduction strategy.

Fig. 5

The error function for different complexities of the model

Integrating different models and using a stepwise strategy is today a practical way of assessing chemicals. In this case, both global QSAR models and local read across strategies [33] are combined in a weight-of-evidence approach [34].

5.2. Big Data and Deep Learning

Applying artificial intelligence methods to difficult datasets, where complex and nonlinear relationships are expected, has been tested, at the beginning of this century, on carcinogenicity data in the international Predictive Toxicology Challenge. The review of about 100 submitted models highlighted some appreciable results in a few models that, in a difficult endpoint as carcinogenicity, performed significantly better than random guessing and were judged by experts to have empirically learned a small but significant amount of toxicological knowledge [35].

With the availability of large datasets, in 2014, the Tox21 challenge organizers invited participants to build computational models to predict the toxicity of compounds for 12 toxic effects using data from high-throughput screening assay. The training set consisted of 12 K chemicals; not all the effects were available for each chemical. The winning model [36] adopted fingerprints and deep neural networks (DNNs), combined with other ML algorithms.

DNNs are NNs, so they implicitly define a mapping between an input (the chemical) and an output (the toxicity). They differ from classical NNs in that they use a large number of neurons, and most of them are hidden neurons (not coding the input values neither the output) which are organized in layers through weighed connections [37].

Weights are learned on the connections to adjust the output of the network iteratively during training. The number of the hidden layer determines the complexity of the function that the network can approximate; one hidden layer is enough to approximate any nonlinear function. DNNs can have thousands of neurons and tens of hidden layers to approximate complex functions and to take any information from the input.

Training such large networks, where hundreds or thousands of parameters should be optimized, is now possible through free packages [38] that automatically try different combinations, and through special hardware as GPUs, easily available as a service or for installation in personal computers.

DNNs are seldom used in QSAR, but are widely applied in chemical computations and de novo design. While the full process for drug design cannot be automated, recently DNNs have been used to shorten to 48 days the process of de novo design of antimicrobials. The authors report using deep learning classifiers and molecular dynamics simulations to screen, identify, synthesize, and test 20 candidates, of which two with high potential tested in vitro and in mice [39].

5.3. Modeling Without Descriptors

DNNs can learn the regularities in the data exploiting any information; co-linearity and redundancy in the input data do not create problems (differently from what holds in statistic methods) but improve the learning. This characteristic of DNNs makes them apt to a new way of developing QSAR: the full molecule representation can be given to input neurons, and features are automatically extracted from input data without the need of computing and selecting chemical descriptors.

Using DNNs as representation extractors has been pioneered in image analysis, where results of a particular kind of DNN, the convolutional NN (CNN), surpass the results of applying classical algorithms to statistical features and even bypass the human ability to categorize images. CNNs take in [A02](#) the input neurons the image, and apply to small pieces of it the same process: convoluting with a simple window and pooling to reduce the input image. This process continues until the reduced number of neurons that remain code the important image features; subsequent layers use them to make the recognition. In (Q)SAR, this method applies to the images representing the chemical graph [40, 41].

DNNs have found applications in text processing, where the problem is to create [A03](#) and exploit the semantic connections among words. Recurrent NN and RNN have the ability to maintain for a given time the dependences among words. In (Q)SAR, the method can be applied to SMILES. Various cases of models developed on SMILES or on modified SMILES are available [42, 43, 44].

Finally, DNNs can take in input the standard representation of graphs, i.e., the adjacency matrix, plus basic information about the extra atoms and the kind of bonds. Graph convolutional NN (GCN) is the architecture of reference; it integrates convolutions and recurrent connections to extract subgraphs from the chemical graph [45].

An important consequence of the feature extraction property of DNNs is that the important features, which are sub-images, sub-strings, or sub-graphs, are readily available and can be compared with existing knowledge. The abovementioned DNN-based QSARs have provided an analysis about the results in terms of the paradigm “model → structure,” and this analysis shows a good agreement with the available knowledge and opens up fresh avenues to explore whether the newly found features characterize yet undiscovered properties.

6. Moving QSARs from Laboratory to Regulations

Toxicity testing typically involves studying adverse health outcomes in animals administered with doses of toxicants, with subsequent extrapolation to expected human responses. The system is expensive, time-consuming, low-throughput, and often provides results of limited predictive value for human health. The toxicity testing methods are largely the same for industrial chemicals, pesticides, and drugs and have led to a backlog of tens of thousands chemicals to which humans are potentially exposed but whose potential toxicity remains largely unknown. Today more than 170 million chemical substances are registered in the CAS registry (www.cas.org), and between 75,000 and 140,000 of them are on the market [46]. Experimental toxicity data are available for a small part of them. According to Johnson et al. [46], aquatic toxicity data are available for 11%, bioconcentration data for 1%, and persistence data for 0.2% of chemical substances marketed in the European Union (EU) and the United States (USA).

This potential risk has urged national and international organizations in making a plan for assessing the toxicity of those chemicals, both for human and environment safety. In USA, for instance, EPA (Environmental Protection Agency) applies the Toxic Substances Control Act (TSCA) and routinely uses predictive QSAR based on existent animal testing to authorize new chemicals. Recently in the US, a new toxicity testing plan, “Human Toxome Project,” has been launched which will make extensive experimentation using predictive, high-throughput cell-based assays (of human organs) to evaluate perturbations in key pathways of toxicity. There is no consensus about this concept of “toxicity pathway” that in the opinion of many should be instead “disruption of biological pathways.” The target of the project is to gain more information directly from human data, so to check in a future, with specific experiments, the most important pathways. In the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation for industrial chemical has been introduced together with specific regulations for cosmetics, pesticide, and food additives. REACH is accepting, still with restrictions, QSAR models as well as read across.

Different regulations apply for different use for the chemicals or for different compartments, as air pollutants, industrial products, food, drinking water, cosmetics and detergents, pesticides, and drugs. In all those cases, there is only limited international agreement on the regulations and doses. Europe is proceeding with the concept of “One

Substance One Assessment,” in order to harmonize different regulations.

Many open issues still make it uneasy to accept QSARs for assessing toxicology. One of the main problems is the difficulty in making a statistical model apt for indicating causal links. Even though the QSAR interpretation can individuate the most important phenomena involved, this interpretation is far from constructing cause–effect relationships. This difficulty is somehow solved in case a full path of foreseen transformations of the chemical in the interacting organism is constructed; recent work in modeling pathways and AOP tries to construct this explanation.

While the acceptance of (Q)SARs is debated on the claim that it cannot substitute experimental practices, ethical issues, and in particular the concern about making experiments on animals, are urging (Q)SAR adoption. Those two conflicting trends will be briefly described.

6.1. Statistical Models and Causal Reasoning

QSARs are both statistical models that exploit variable correlations, and knowledge-based models, that discover how patterns in input data explain the output pattern [47]. Those correlations are at the basis of the explanations provided by QSAR models, which give importance to functional subgroups and make warnings about possible holes in the applicability domain.

Learning causal relationships from raw data has been on philosophers' wish list since the eighteenth century. Hume argued that causality cannot be perceived and instead we can only perceive correlation. And indeed the basic biological experiments aim at finding a correlation (positive or negative) between some features and effect.

Humans understand causation to be a relation between events in which the presence of some events causes the presence of others. Human inference of causal relationships relies primarily on universal cues, such as spatiotemporal contingency, or reliable covariation between effects and their causes, as well as on domain-specific knowledge. Accordingly, most theories of causation invoke an explicit requirement that a cause precedes its effect in time. Yet temporal information alone cannot distinguish genuine causation from spurious associations.

Causation is a transitive, nonreflexive, and antisymmetric relation. That is:

1. If A is a cause of B and B is a cause of C , then A is also a cause of C .
2. An event A cannot cause itself.
3. If A is a cause of B then B is not a cause of A .

Computational models of causality, which combine together networks, Bayesian probability, and Markov assumptions, as in the Causal Bayes Networks [48] are still computationally hard and do not have yet applications in real toxicology.

A kind of simulation approach is more feasible, as adopted in recent studies about the Adverse Outcome Pathway (AOP), aimed at identifying the workflow from the molecular initiating event to the final observed outcome. It integrates other terms often used in explaining the toxicology, as “mode of action,” or “mechanistic interpretation.” Unfortunately, there is no unique AOP for a class of chemicals, as different biological pathways are usually observed or supposed.

6.2. Ethical Issues

Toxicity testing typically involves studying adverse health outcomes in animals subjected to high doses of toxicants with subsequent extrapolation to expected human responses at lower doses.

The system is expensive, time-consuming, with low-throughput, and often provides results of limited predictive value for human health.

Conversely each year a huge number of new substances are synthesized and possibly sent to the market. Is it really ~~AQ4~~ necessary to test all of them on animals? Even more, is it necessary to synthesize them or would it be better to in silico assess their properties before making them, using a proactive strategy?

The Declaration of Bologna, in 1999, called the 3 R (for reduce, refine, replace), proposed a manifesto to develop alternative methods that could save millions of animals. In this scenario, the ethical issues however are advocated also by authorities that have to protect humans and that see the animals as a more ethical use than that of humans.

The stakeholders in the toxicity assessment are:

- Scientists and producers: They want modeling of the process and discovery of properties. In other words, build knowledge, and translate it rapidly in products and drugs.
- Regulators and standardization organizations: They want to be convinced by some general rule (mechanism of action). In other words, reduce the risk of erroneous evaluations. Be fast and conservative in decisions taking.

- Public, media, and opinion makers: They want to be protected against risk at 100%. Part of the population is strongly against the use of animal models.

There are many models, many techniques to build (Q)SAR models, and also many reasons to build up a model. The intended use of a model can greatly affect its development. In new chemicals and new drug design, the target is to identify active compounds, so avoiding false positives. Final users and regulators, instead, have an opposite need, as they want to surely mark toxic compounds, so they want to avoid false negatives. Sensitivity and specificity can be in turn the most important parameter.

Good and validated in silico models can benefit multiple actors. In any case, the output of the model, despite its good predictive value, is not sufficient; documentation enabling the user to accept or not the prediction is necessary.

6.3. Towards a Larger Acceptance of QSAR Models

Experts today when asked to assess the properties of a chemical substance prefer to experimentally test it. If testing is not possible or allowed for all the properties of interest, they prefer finding the results of similar substances and adapt them to the target chemical. This read across method makes it central that the similarity measure adopted is well suited to the problem domain; the definition of a proper similarity measure is still challenging [49].

The common perception is that a predictive model cannot match the observation. This topic has been central in the philosophy of science; however, here a pragmatic approach is necessary. In practice, can QSARs be used to improve expert evaluation?

In the public sector, asking experts to provide advice in decision-making is common. For instance, EPA asks panels of experts to assess substances for which data are incomplete or contradictory. The experts have to express their opinions about the probability that a certain outcome will happen. Those judgments are subjective and are often expressed through words (as likely) whose meaning in terms of probability changes from expert to expert. Morgan [50] analyzed the causes of expert bias, while Benfenati et al. [51] provided an example of how far the expert answers differ when experts are asked to assess the toxicity of some chemicals.

The use of QSAR is valuable, as it provides an expert-independent prediction, which can be the common basis for expert judgment.

Weak points about QSAR are still present and should be considered. The first weakness is that most of the QSARs are developed on small data sets (hundreds of chemicals), and so it is hard to believe that they can correctly cover all the chemical space; this problem is tackled by checking whether the target molecule is in the applicability domain. Another weakness is that the answer of the QSAR is a label, or a small interval of values, but this is not the probability that the molecule will be really toxic at that level. The sensitivity and specificity values depend on the training and test set, not on the percentage of positive and negative molecules in the environment. For this reason, other analyses are needed, and often they are carried out considering the exposure.

7. Conclusions

Alongside classical methods as in vivo and in vitro experiments, the use of computational tools is gaining more and more interest in the scientific community. Computational tools provide simulations, offer ways to speed up the molecular design, offer predictions, and somehow explanation of the mechanism. QSARs derive their methods from chemometrics and physical simulations, and combining them together with statistics to construct systems can be predictive.

QSARs can span from very predictive methods, which usually employ complex and nonlinear correlations, to explanatory QSARs that are focused on simple interpretability.

The usage of QSAR models is growing, since they provide the only affordable method to screen large quantities of data.

References

1. Brown N (2009) Chemoinformatics—an introduction for computer scientists. ACM Comput Surv 41(2):8. <https://doi.org/10.1145/1459352.1459353>
2. Gasteiger J, Engel T (2003) Chemoinformatics: a textbook. Wiley-VCH, Weinheim, Germany. ISBN: 978-3-527-30681-7

3. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? *Med Chem* 45(19):4350–4358. <https://doi.org/10.1021/jm020155c>
4. Balaban AT (1985) Applications of graph theory in chemistry. *J Chem Inf Comput Sci* 25:334–343. <https://doi.org/10.1021/ci00047a033>
5. 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. <https://doi.org/10.1021/ci00057a005>
6. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput Sci* 29:97–101. <https://doi.org/10.1021/ci00062a008>
7. Adam D (2002) Chemists synthesize a single naming system. *Nature* 417:369. <https://doi.org/10.1038/417369a>
8. Schlick T (2002) Molecular modeling and simulation: an interdisciplinary guide. Springer-Verlag, New York. ISBN 978-1-4419-6351-2
9. Chow PHK, Ng RTH, Ogden BE (2008) Using animal model in biomedical Research. 1st edition. World Scientific. <https://doi.org/10.1142/6454> A06
10. Balazs T (1970) Measurement of acute toxicity, in methods in toxicology. Blackwell Scientific Publications, Oxford and Edinburgh
11. Benfenati E, Gini G (1997) Computational predictive programs (expert systems) in toxicology. *Toxicology* 119:213–225. [https://doi.org/10.1016/s0300-483x\(97\)03631-7](https://doi.org/10.1016/s0300-483x(97)03631-7)
12. Hartung T (2009) Toxicology for the twenty-first century. *Nature* 460(9):208–212. <https://doi.org/10.1038/460208a>
13. Livingstone DJ (2000) The characterization of chemical structures using molecular properties. A survey. *J Chem Inf Comput Sci* 40:195–209. <https://doi.org/10.1021/ci990162i>
14. Hansch C, Malony PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with hammett substituent constants with partition coefficients. *Nature* 194:178–180. <https://doi.org/10.1038/194178b0>
15. 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 Comp Chem* 7:565–577. <https://doi.org/10.1021/ci00053a005>
16. Kubinyi H (2002) From narcosis to hyperspace: the history of QSAR. *Quant Struct Act Relat* 21:348–356. [https://doi.org/10.1002/1521-3838\(200210\)21:4<348::AID-QSAR348>3.0.CO;2-D](https://doi.org/10.1002/1521-3838(200210)21:4<348::AID-QSAR348>3.0.CO;2-D)
17. Karelson M (2000) Molecular Descriptors in QSAR/QSPR. Wiley-VCH, Weinheim, Germany. ISBN: 978-0-471-35168-9
18. 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 Cheminformatics* 10(60):1–13. <https://doi.org/10.1186/s13321-018-0315-6>
19. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer-Verlag, New York, NY. ISBN 978-0-387-84858-7
20. Gini G, Katritzky A (1999) Predictive toxicology of chemicals: experiences and impact of artificial intelligence tools. In: Proc. AAAI spring symposium on predictive toxicology, report SS-99-01. AAAI Press, Menlo Park, CAL. ISBN 978-1-57735-073-6

21. Héberger K, Rácz A, Bajusz D (2017) Which performance parameters are best suited to assess the predictive ability of models? In Roy K (ed) advances in QSAR modeling, Springer International. . ISBN 978-3-319-56850-8
22. Polishchuk PG (2017) Interpretation of quantitative structure–activity relationships models: past, Present and future. *J Chem Inf Model* 57(11):2618–2639. <https://doi.org/10.1021/acs.jcim.7b00274>
23. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1:67. <https://doi.org/10.1109/4235.585893>
24. Ashby J (1985) Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ Mutagen* 7:919–921. <https://doi.org/10.1002/em.2860070613>
25. Benigni R, Bossa C (2008) Structure alerts for carcinogenicity, and the salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat Res* 659(3):248–261. <https://doi.org/10.1016/j.mrrev.2008.05.003>
26. Ferrari T, Cattaneo D, Gini G, 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>
27. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufman, San Francisco; CA. <https://doi.org/10.1007/BF00993309>
28. Neagu C–D, Gini G (2003). Neuro–fuzzy knowledge integration applied to toxicity prediction. In Jain R, Abraham A, Faucher C, Jan van der Zwaag B (Eds), *Innovations in knowledge engineering*, advanced knowledge International Pty Ltd, Magill, South Australia, 311–342. ISBN 0 9751004 0 8
29. Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140. <https://doi.org/10.1023/A:1018054314350>
30. Polikar R (2006) Ensemble based systems in decision making. *IEEE Circ Syst Mag* 2006(6):21–45. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
31. Gini G, Garg T, Stefanelli M (2009) Ensembling regression models to improve their predictivity: a case study in QSAR (quantitative structure activity relationships) with computational chemometrics. *Appl Artif Intell* 23:261–281. <https://doi.org/10.1080/08839510802700847>
32. Friedman J (1997) On bias, variance, 0/1 loss and the curse of dimensionality. *Data Mining Knowl Discov* 1:55–77. <https://doi.org/10.1023/A:1009778005914>
33. 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 QSAR Environ Res* 25(12):1–13. <https://doi.org/10.1080/1062936X.2014.976267>
34. Benfenati E, Chaudhry Q, Gini G, Dorne JL (2019) Integrating in silico models and read–across methods for predicting toxicity of chemicals: a step–wise strategy. *Environ Int* 131:105060. <https://doi.org/10.1016/j.envint.2019.105060>
35. Toivonen H, Srinivasan A, King RD, Kramer S, Helma C (2003) Statistical evaluation of the predictive toxicology challenge 2000–2001. *Bioinformatics* 19(10):1183–1193. <https://doi.org/10.1093/bioinformatics/btg130>
36. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 3:80. <https://doi.org/10.3389/fenvs.2015.00080>
37. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444. <https://doi.org/10.1038/nature14539>
38. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. arXiv:1412.6980v9 [cs.LG] 30 Jan 2017

39. Das P, Sercu T, Wadhawan K, Padhi I, Gehrman S, Cipcigan F, Chenthamarakshan V, Strobelt H, dos Santos C, Chen P-Y, Yang YY, Tan JPK, Hedrick J, Crain J, Mojsilovic A (2021) Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng* 5:613–623. <https://doi.org/10.1038/s41551-021-00689-x>
40. 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. arxiv.org/abs/1706.066892017
41. Gini G, Zanoli F (2020) Machine learning and deep learning methods in ecotoxicological QSAR modeling. In: Roy K (ed) *Ecotoxicological QSARs*. Humana Press, Springer, New York, pp 111–149. ISBN 978–1–0716–0150–1
42. Goh G, Hodas N, Siegel C, Vishnu A (2018) SMILES2vec: an interpretable general-purpose deep neural network for predicting chemical properties. [arXiv:1712.02034](https://arxiv.org/abs/1712.02034)v2 [stat.ML]
43. Gini G, Zanoli F, Gamba A, Raitano G, Benfenati E (2019) Could deep learning in neural networks improve the QSAR models? *SAR QSAR Environ Res* 30(9):617–642. <https://doi.org/10.1080/1062936X.2019.1650827>
44. Chakravarti SK, Radha Mani AS (2019) Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. *Front Artif Intell* 2:17. <https://doi.org/10.3389/frai.2019.00017>
45. Gini G, Hung C, Benfenati E (2021) Big data and deep learning: extracting and revising chemical knowledge from data. In: Basak S, Vracko M (eds) *Big data analytics in Chemoinformatics and bioinformatics (with applications to computer-aided drug design, cancer biology, emerging pathogens and computational toxicology)*. Elsevier, Amsterdam
46. Johnson AC, Jin X, Nakada N, Sumpter JP (2020) Learning from the past and considering the future of chemicals in the environment. *Science* 367:384–387. <https://doi.org/10.1126/science.aay6637>
47. Gini G (2018) QSAR, what else? In: Nicolotti O (ed) *Computational toxicology: methods and protocols*, vol 1800. Springer, Clifton, NJ, pp 79–105. ISBN 978–1–4939–7899–1.
48. Pearl J (2003) Statistics and causal inference: a review. *Test J* 12:281–345. <https://doi.org/10.1007/BF02595718>
49. G. Gini G (2020) The QSAR similarity principle in the deep learning era: confirmation or revision?, *Found Chem* 22: 383–402. DOI: <https://doi.org/10.1007/s10698-020-09380-6>
50. Morgan MG (2014) Use (and abuse) of expert elicitation in support of decision making for public policy. *PNAS* 111(20):7176–7184. <https://doi.org/10.1073/pnas.1319946111>
51. Benfenati E, Belli M, Borges T, Casimiro E, Cester J, Fernandez A, Gini G, Honma M, Kinzl M, Knauf R, Manganaro A, Mombelli E, Petoumenou MI, Paparella M, Paris P, Raitano G (2016) Results of a round-robin exercise on read-across. *SAR QSAR Environ Res* 27(5):371–384. <https://doi.org/10.1080/1062936X.2016.1178171>