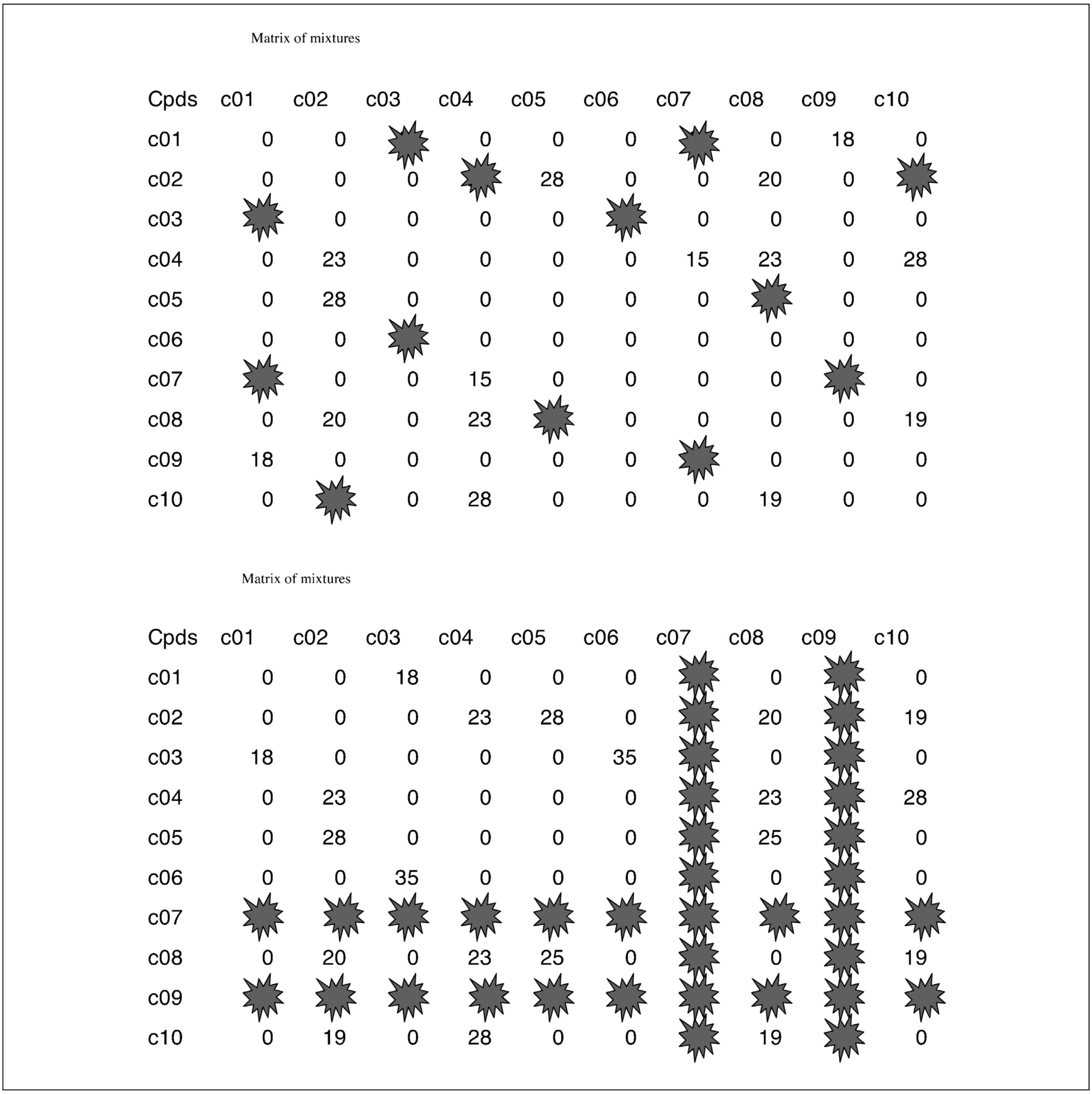
DOI: 10.1002/minf.201100129

# Existing and Developing Approaches for QSAR Analysis of Mixtures

Eugene N. Muratov,*[a, b] Ekaterina V. Varlamova,[a] Anatoly G. Artemenko,[a] Pavel G. Polishchuk,[a] and Victor E. Kuz'min[a]



Matrix of mixtures

| Cpds | c01 | c02 | c03 | c04 | c05 | c06 | c07 | c08 | c09 | c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| c01 | 0 | 0 | ✹ | 0 | 0 | 0 | ✹ | 0 | 18 | 0 |
| c02 | 0 | 0 | 0 | ✹ | 28 | 0 | 0 | 20 | 0 | ✹ |
| c03 | ✹ | 0 | 0 | 0 | 0 | ✹ | 0 | 0 | 0 | 0 |
| c04 | 0 | 23 | 0 | 0 | 0 | 0 | 15 | 23 | 0 | 28 |
| c05 | 0 | 28 | 0 | 0 | 0 | 0 | 0 | ✹ | 0 | 0 |
| c06 | 0 | 0 | ✹ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| c07 | ✹ | 0 | 0 | 15 | 0 | 0 | 0 | 0 | ✹ | 0 |
| c08 | 0 | 20 | 0 | 23 | ✹ | 0 | 0 | 0 | 0 | 19 |
| c09 | 18 | 0 | 0 | 0 | 0 | 0 | ✹ | 0 | 0 | 0 |
| c10 | 0 | ✹ | 0 | 28 | 0 | 0 | 0 | 19 | 0 | 0 |

Matrix of mixtures

| Cpds | c01 | c02 | c03 | c04 | c05 | c06 | c07 | c08 | c09 | c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| c01 | 0 | 0 | 18 | 0 | 0 | 0 | ✹ | 0 | ✹ | 0 |
| c02 | 0 | 0 | 0 | 23 | 28 | 0 | ✹ | 20 | ✹ | 19 |
| c03 | 18 | 0 | 0 | 0 | 0 | 35 | ✹ | 0 | ✹ | 0 |
| c04 | 0 | 23 | 0 | 0 | 0 | 0 | ✹ | 23 | ✹ | 28 |
| c05 | 0 | 28 | 0 | 0 | 0 | 0 | ✹ | 25 | ✹ | 0 |
| c06 | 0 | 0 | 35 | 0 | 0 | 0 | ✹ | 0 | ✹ | 0 |
| c07 | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ |
| c08 | 0 | 20 | 0 | 23 | 25 | 0 | ✹ | 0 | ✹ | 19 |
| c09 | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ | ✹ |
| c10 | 0 | 19 | 0 | 28 | 0 | 0 | ✹ | 19 | ✹ | 0 |

**Abstract**: This review is devoted to the critical analysis of advantages and disadvantages of existing mixture descriptors and their usage in various QSAR/QSPR tasks. We describe good practices for the QSAR modeling of mixtures, data sources for mixtures, a discussion of various mixture descriptors and their application, recommendations about proper external validation specific for mixture QSAR modeling, and future perspectives of this field. The biggest problem in QSAR of mixtures is the lack of reliable data about the mixtures' properties. Various mixture descriptors are used for the modeling of different endpoints. However, these descriptors have certain disadvantages, such as applicability only to 1:1 binary mixtures, and additive nature. The field of QSAR of mixtures is still under development, and existing efforts could be considered as a foundation for future approaches and studies. The usage of non-additive mixture descriptors, which are sensitive to interaction effects, in combination with best practices of QSAR model development (e.g., thorough data collection and curation, rigorous external validation, etc.) will significantly improve the quality of QSAR studies of mixtures.

**Keywords:** Mixture descriptors · QSAR/QSPR · Predictive modeling · Synergism

QSAR/QSPR (Quantitative Structure-Activity/Property Relationship): theoretical analysis devoted to establishing the dependence of a compound's activity/property on its chemical structure. In our review, we use the terms QSAR and QSPR interchangeably.

Mixture descriptors/descriptors of mixtures: specific structural descriptors whose values are unique to each mixture. In our review, we use these terms interchangeably.

Predictive modeling: QSAR modeling focused on the development of externally predictive QSAR models, i.e., models which are able to make correct predictions for new compounds.

Synergism: the combined effect of two or more chemicals is greater than the sum of their individual effects.

## 1 Introduction

Ever since the days of mammoth hunting, early man realized that a joint effort is far more productive than a solitary one. Subsequently, this insight has been reflected in numerous proverbs and phrases, e.g., "two heads are better than one", "all for one and one for all".[1] No endeavor demonstrates this clearer than scientific research. Mixtures found wide applications in many fields of science because their effect can be stronger than the action of single compounds. Modern materials, detergents, etc., usually represent an ensemble of several individual compounds. Mixtures are also widely used in medicinal chemistry and related fields. For instance, combining virus inhibitors might overcome the disadvantages of monotherapy.[2] Given the variety of viral strains, there are two ways of battling viruses.[3] One option is the creation of a new drug that would be active against all strains of the target virus. "Dirty drugs" would be more effective than "magic bullets"[4] in achieving this aim. Another alternative is the use of drug mixtures for antiviral treatment, which is expected to be the predominant method in the future.[3] Combining drugs, especially those with different modes of action, could combat the resistance phenomenon. The interest in mixtures is still growing. Figure 1 depicts the increase in the number of publications from medicinal chemistry and related fields with titles containing the word "mixture". Most of these papers are devoted to overcoming antiviral or antibiotic resistance or to increase effectiveness of existing drugs by combining them synergistically. For instance, Nikolaeva and Galabov[2] showed that by using mixtures of picornavirus inhibitors, the same antiviral effect, and thus, a higher selectivity ratio, could be achieved at lower concentrations than those required if these drugs were used individually. Coutinho et al.[5] describe another example of the successful application of mixtures when pure compounds failed to achieve the desired effect. It was shown that the addition of a *Turnera ulmifolia* ethanol extract to amikacin, neomycin, and tobramycin significantly increased their antimicrobial activity against two *E. coli* strains normally resistant to these antibiotics. A similar success of a mixture of chlorpromazine with amikacin, kanamycin, and tobramycin indicated the involvement of an efflux system in the resistance to these aminoglycosides. Results suggest that such mixture therapy could be used as a new weapon against antibiotic resistance.[5] There are many more studies that demonstrate the success of mixtures in cases where single compounds failed,[6–9] and there is growing interest in the application of mixture therapy for the treatment of different diseases.

Another argument in favor of mixture therapy could be derived from the analysis of structures of the 200 bestsell-

[a] *E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min*
*Laboratory of Theoretical Chemistry, Department of Molecular Structure, A. V. Bogatsky Physical Chemical Institute, National Academy of Sciences of Ukraine,*
*Lustdorfskaya Doroga 86, Odessa 65080, Ukraine*
*tel: +380487662394, fax: +380487662394*
**e-mail: 00dqsar@ukr.net*
*murik@email.unc.edu*

[b] *E. N. Muratov*
*Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, Eshelman School of Pharmacy, University of North Carolina,*
*Beard Hall 301, CB#7568, Chapel Hill, NC, 27599, USA*
*tel: +19199663459, fax: +19199660204*

ing drugs.[10] After excluding 53 monoclonal antibodies, enzymes, and proteins, 157 drugs remain, including 13 multicomponent (Figure 2) and 14 racemic mixtures (Figure 3). These drugs have broad applications ranging from antivirals to anxiolytics (Tables 1 and 2). Thus, as much as 14% of these bestselling drugs present examples of combinatorial therapy that can be illustrated by augmentin.

Augmentin is a prescription antibiotic used to treat a variety of infections in adults, children, and infants. Augmentin contains two different medications: amoxicillin and clavulanate potassium. It belongs to a group of medications known as aminopenicillins, which is part of a larger group of medications known as beta-lactam antibiotics. Amoxicillin inhibits bacterial cell wall synthesis, which eventually causes the bacteria to die. However, many bacteria have developed resistance to amoxicillin and similar antibiotics by producing enzymes called beta-lactamases. Beta-lactamases break the beta-lactam ring, making amoxicillin and similar antibiotics ineffective. The other component of Augmentin (clavulanate) is known as a beta-lactamase inhibitor. Clavulanate binds to bacterial beta-lactamase and stops these enzymes from breaking down the amoxicillin molecule. Clavulanate itself has no significant antibacterial activity; it merely helps to prevent amoxicillin from being broken down by bacteria that would otherwise be resistant to it. Essentially, clavulanate "augments" the activity of amoxicillin.

Drug-drug interactions, especially possible side effects and complications, are one of the biggest challenges in modern pharmacology and a serious problem in medical care units.[11] Mixtures began to play an important role in environmental protection studies because the environment

*Eugene N. Muratov* is a senior researcher in the Laboratory of Theoretical Chemistry in the A. V. Bogatsky Physical-Chemical Institute of National Academy of Sciences (NAS) of Ukraine, and a research assistant professor in the Division of Medicinal Chemistry and Natural Products in the Eshelman School of Pharmacy, UNC-Chapel Hill. He received his PhD in organic chemistry in 2004 from the A. V. Bogatsky Physical-Chemical Institute, Odessa, Ukraine. Then he continued to work in the Laboratory of Theoretical Chemistry in Odessa. In 2008, he joined the Laboratory for Molecular Modeling in UNC as a postdoc. His research interests are in the areas of cheminformatics (especially QSAR), computer-assisted drug design, antiviral research, and medicinal chemistry.

*Ekaterina Varlamova* is a PhD student of the Laboratory of Theoretical Chemistry in the A. V. Bogatsky Physical-Chemical Institute NAS of Ukraine. She received a Master Diploma cum laude from the Odessa National Polytechnic University in 2008. She will defend her PhD thesis on cheminformatics in 2012. Her research interests are in the areas of cheminformatics, especially QSAR of mixtures, and computer-assisted drug design.

*Anatoly Artemenko* is a senior researcher in the A. V. Bogatsky Physical-Chemical Institute of NAS of Ukraine. He received his PhD in organic chemistry in 2001 from the A. V. Bogatsky Physical-Chemical Institute of NAS of Ukraine. He is the President of Ukraine Premium for Young Scientists Laureate (2006). His research is supported by multiple STCU and INTAS grants. His current research interests are in areas of computer-assisted drug design, cheminformatics, medicinal chemistry, and data mining. He is the author of about 50 articles and book chapters.

*Pavel Polishchuk* is a research associate in the A. V. Bogatsky Physical-Chemical Institute of NAS of Ukraine. He received his PhD in bioorganic chemistry in 2009 from the A. V. Bogatsky Physical-Chemical Institute of NAS of Ukraine. His current research interests are in areas of computer-assisted drug design, cheminformatics, medicinal chemistry, and data mining. He is the author of about 20 articles and book chapters.

*Victor E. Kuz'min* is a professor and vice-director of the A. V. Bogatsky Physical-Chemical Institute of National Academy of Sciences of Ukraine. He received his PhD in organic chemistry in 1980 and defended his Doctor of Sciences (habilitation) thesis in 2004 at the A. V. Bogatsky Physical-Chemical Institute, Odessa, Ukraine. He has been the head of the Laboratory of Theoretical Chemistry since 1980. He was the scientific advisor for more than 10 PhD students. His research interests are in the areas of theoretical chemistry, cheminformatics, computer- assisted drug design, computational toxicology, and medicinal chemistry. His research is supported by multiple grants from the STCU, INTAS, and other funds.

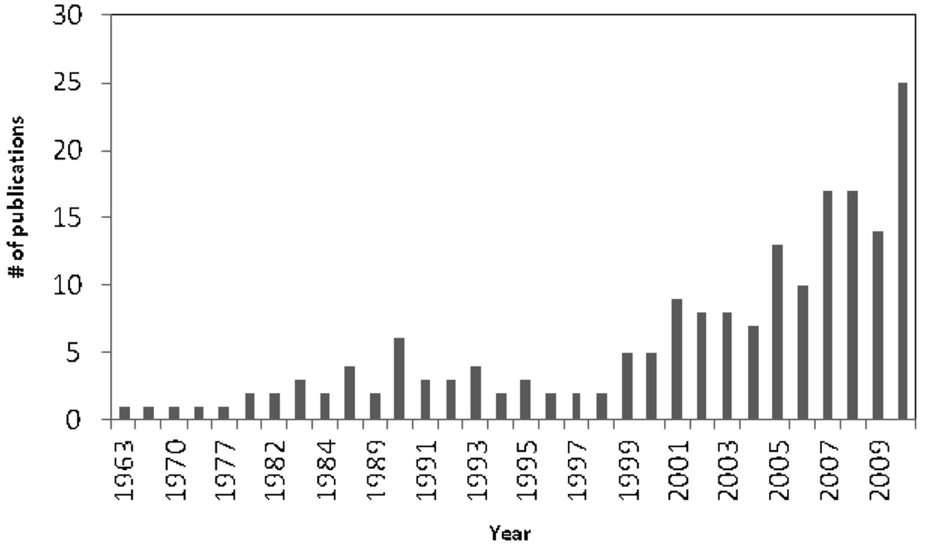


**Figure 1.** Number of annually published medicinal chemistry studies devoted to mixtures.

is exposed simultaneously or sequentially to chemical mixtures via multiple exposure routes, e.g., industrial, agricultural, and other activities. Many toxic effects are caused by chemical mixtures rather than by single chemicals. Compounds at much lower concentrations than the established Water Quality Standards can cause toxic effects when acting jointly with other chemicals.[12] However, until now, the vast majority of toxicity studies (both experimental and theoretical) dealt with single compounds and did not take potential effects of their combination into account. The importance of mixtures and forecasting their environmental impact has also been recognized by the US Environmental Protection Agency (EPA) in special guidelines devoted to health risk assessment of chemical mixtures.[13,14] Thus, the investigation and prediction of mixture toxicity represent one of most challenging problems in environmental chemical risk assessment.[15]

QSAR/QSPR modeling is a well-established cheminformatics approach that can aid in meeting these challenges (in this review, the terms QSAR and QSPR are used interchangeably). For many years, QSAR approaches have been intensively and successfully used for the analysis and prediction of different activities and properties (e.g., antiviral and anticancer activity, toxicity, etc.[16–22]) of single compounds. All QSAR approaches are based on the simple postulate that any property of a chemical compound is a function of its structure, and its corollary that compounds with similar structures are expected to have similar biological activities. These principles must be applicable not only to single compounds but also to their mixtures. This means that QSAR can be successfully used for predicting mixture properties from the mixture's composition and its constituents' structures. While modern QSAR methodology is quite successful when dealing with individual compounds, no mature QSAR workflows exist for the analysis of mixtures,[23] even though such tasks are becoming more widespread and important.[24] Thus, developing approaches for the QSAR analysis of mixtures is an emergent task. Mixture descriptors, i.e., specific structural descriptors whose values are unique to each mixture, are the key to the solution of this problem.

The purpose of this review is to summarize and critically analyze advantages and disadvantages of existing mixture descriptors and their application to various QSAR tasks. In addition, we will discuss best practices of QSAR modeling in relation to mixtures, possible sources of data, strategies for the external validation of QSAR models for mixtures, and our vision for the development of the field.

## 2 Best Practices of QSAR Modeling and Their Relation to QSAR of Mixtures

Although QSAR of mixtures does not deal with single compounds, it is still an integral part of standard QSAR analysis. The main difference is related to descriptor calculation and, to a lesser degree, to validation of obtained results, but the modeling techniques and all emerging trends remain fairly similar in both conventional and mixture QSAR. In this section, we shall remind the reader of the best practices of QSAR modeling that are equally applicable to analyzing both single compounds and mixtures. The summary of most common practices used in modern QSAR modeling can be found in an excellent recent review.[25]

High predictive power is the most important requirement for the acceptance of QSAR models.[26] Rigorous external validation, using compounds excluded from model building and selection (i.e., external set or n-fold external

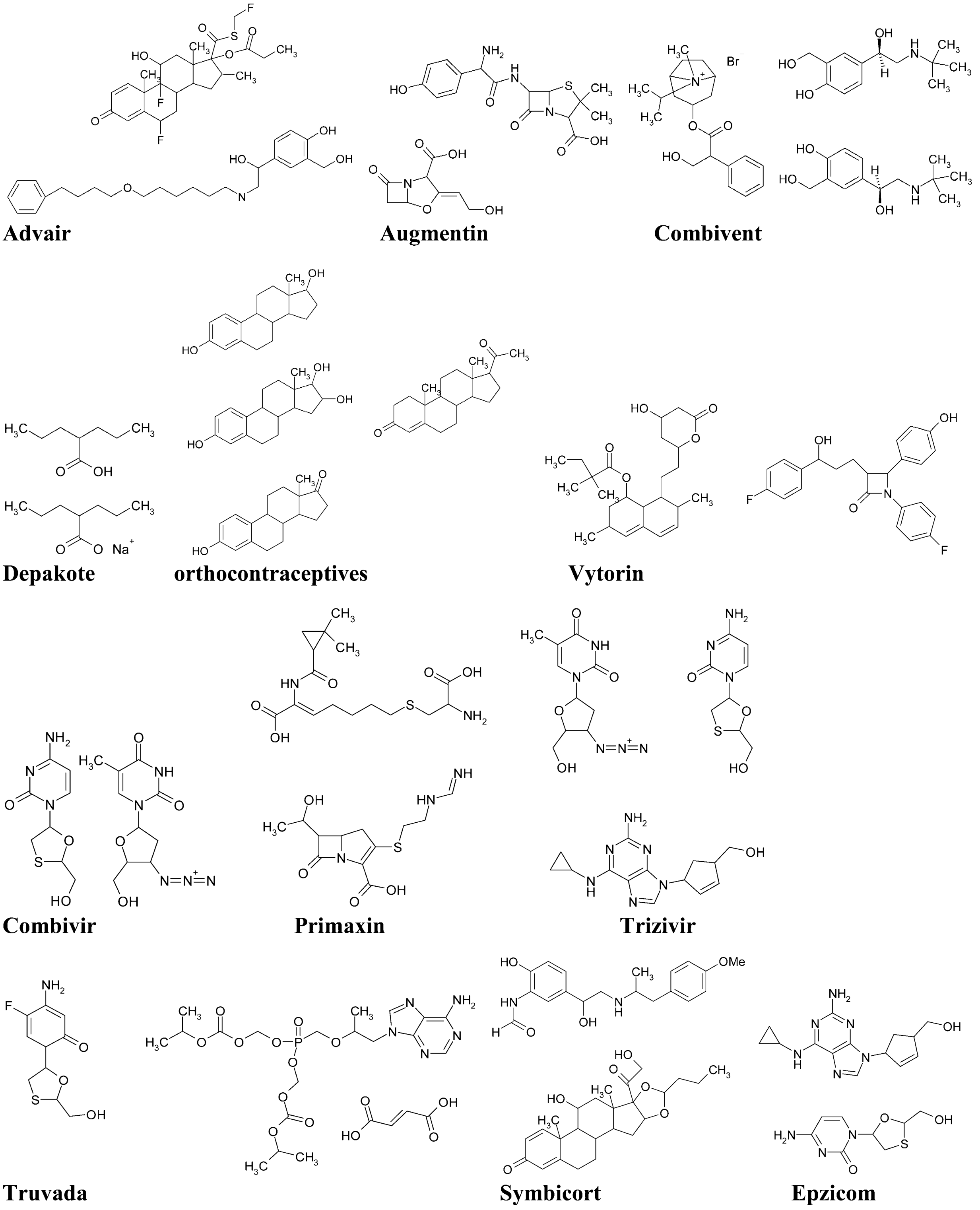


**Figure 2.** Topological structures of bestselling drug mixtures.

cross-validation), is currently the only way to correctly estimate a model's predictive power.[25] The importance of model validation and applicability domain (AD) could now be regarded as collective wisdom within the community of molecular modelers. The OECD (Organization for Economic Cooperation and Development) member countries adopted

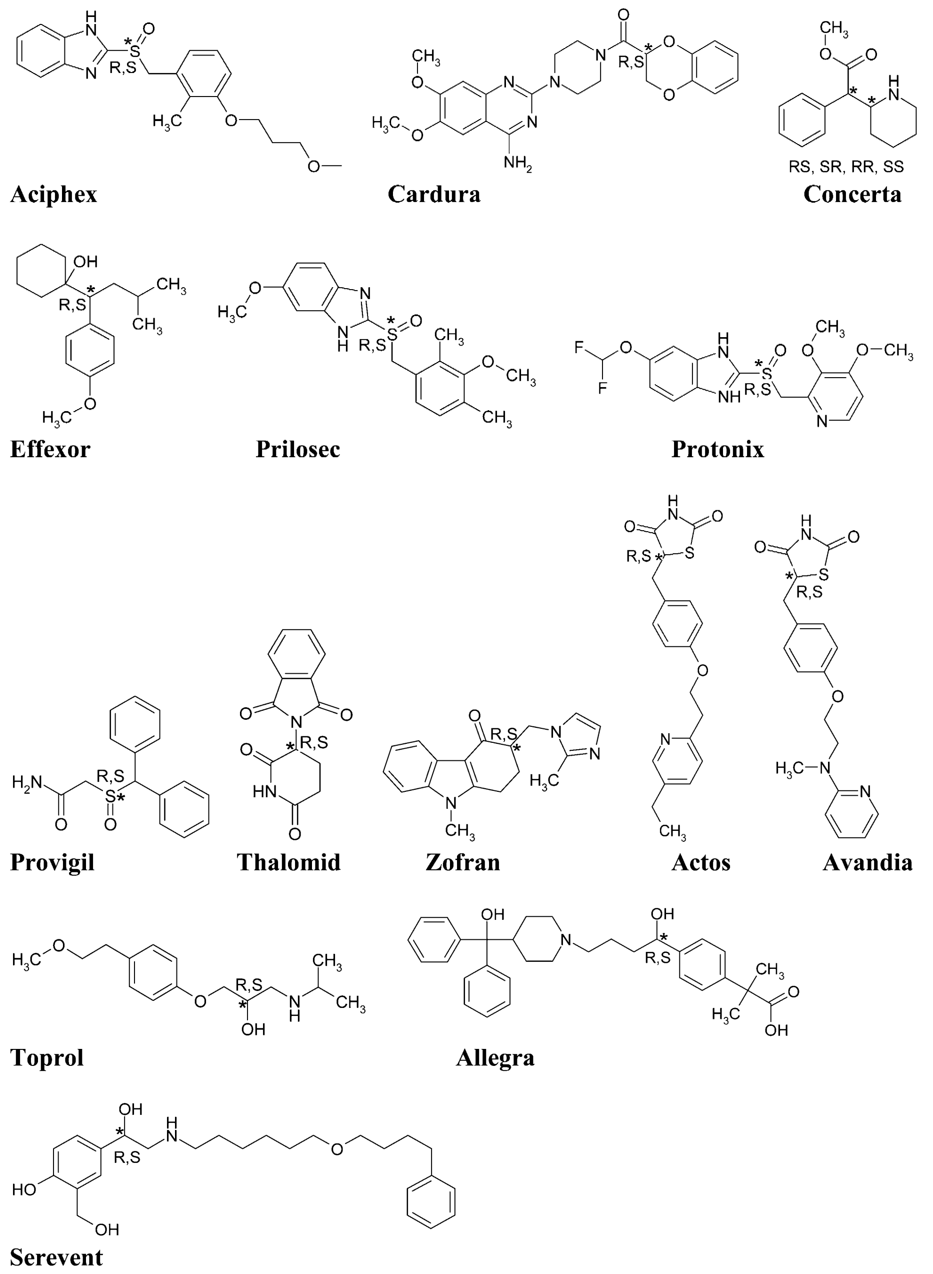


**Figure 3.** Topological structures of bestselling drugs that are racemic mixtures. Chiral centers responsible for stereoisomerism are marked by asterisks.

the following five principles that valid (Q)SAR models should follow to allow their use in regulatory assessment of chemical safety: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness, and predictivity; and (v) a mechanistic interpretation, if possible.[27] Also, in an effort to improve the quality of publications in the QSAR modeling field, the Journal of Chemical Information and Modeling published a special editorial highlighting the requirements for QSAR papers that authors should follow to publish their results in the journal.[28] It significantly decreased the number of low-quality QSAR publications in top cheminformatics journals.

Any QSAR investigation, including mixture-oriented ones, consists of four basic stages: (i) data collection and curation; (ii) model building (descriptor calculation, and establishing an empirical relationship between chemical structures represented by generated descriptors and investigated activity using appropriate statistical approach); (iii) rigorous external validation via new experiments; and (iv) model exploitation (model interpretation and prospective application to virtual screening and targeted molecular

**Table 1.** Bestselling drugs – mixtures of different compounds.

| Rank 2006 | Brand name(s) | Generic name | Disease/medical use | First approval date |
|---|---|---|---|---|
| 2 | Advair,Seretide | Fluticasone + Salmeterol | Asthma | 2000 |
| 47 | Vytorin | Ezetimibe + Simvastatin | Cholesterol | 2004 |
| 88 | Truvada | Tenofovir + Emtricitabine | HIV infection | 2004 |
| 89 | Symbicort | Budesonide + Formoterol | Asthma | 2000 |
| 100 | Augmentin | Co-amoxiclav | Bacterial infections | 1984 |
| 102 | Orthocontraceptives | Estrogen + Progesterone | Contraception | 1984 |
| 110 | Combivir | Lamivudine + Zidovudine | HIV infection | 1997 |
| 126 | Combivent | Ipratropium + Salbutamol | Chronic obstructive pulmonary disease | 1996 |
| 146 | Primaxin | Imipenem + Cilastatin | Bacterial infections | 1985 |
| 187 | Trizivir | Abacavir + Lamivudine + Zidovudine | HIV infection | 2000 |
| 194 | Epzicom,Kixeva | Abacavir + Lamivudine | HIV infection | 2004 |
| 77 | Depakote[a] | Valproate semisodium | Seizures | 1983 |

[a] Depakote is 1:1 mixture of valproic acid and its sodium salt

**Table 2.** Bestselling drugs – racemic mixtures.

| Rank 2006 | Brand name(s) | Generic name | Disease/medical use | First approval date |
|---|---|---|---|---|
| 13 | Effexor | Venlafaxine | Depression, Anxiety disorders | 1993 |
| 14 | Protonix,Pantozol,Pantoloc | Pantoprazole | Gastrointestinal disorders | 2000 |
| 23 | Avandia,Avandaryl Avandamet, | Rosiglitazone | Type 2 diabetes | 1999 |
| 24 | Actos | Pioglitazone | Type 2 diabetes | 1999 |
| 27 | AcipHex,Pariet | Rabeprazole | Gastrointestinal disorders | 1999 |
| 52 | Toprol,Seloken | Metoprolol | Hypertension | 1992 |
| 62 | Zofran | Ondansetron | Nausea and vomiting | 1992 |
| 72 | Prilosec,Losec | Omeprazole | Gastrointestinal disorders | 1989 |
| 124 | Allegra,Telfast | Fexofenadine | Allergic rhinitis | 1996 |
| 140 | Provigil | Modafinil | Sleepiness | 1998 |
| 176 | Serevent | Salmeterol | Asthma | 1994 |
| 177 | Cardura | Doxazosin | Hypertension | 1990 |
| 197 | Thalomid | Thalidomide | Erythema nodosum leprosum | 1998 |
| 118 | Concerta [a] | Methylphenidate | Attention-deficit hyperactivity disorder | 2000 |

[a] Concerta is a mixture of 2 racemates

design of novel compounds or mixtures with required properties) and, desirably, but very rarely in current practice, its usage in regulatory purposes.[24] Various errors related to these four stages, leading to unacceptable QSAR models, and recipes for avoiding some common mistakes in QSAR model development were described in several recent reviews.[25,26,29–31] Some steps that could be useful in developing more reliable QSAR studies of mixtures are summarized below. We also want to note that we do not explicitly detail external validation, AD estimation, or other procedures; rather, we emphasize their importance and refer the reader to articles with information regarding the point of interest.

(i) Data curation. Unfortunately, accuracy of chemical and experimental data is usually not discussed in QSAR manuscripts. Meanwhile, all databases (not to mention the original reported data) were created by people, and even the best public and commercial databases may contain errors.[32] Since it is senseless to launch complex modeling investigations if the underlying chemical structures or activity values are not correct, best practices for data preparation prior to initiating the modeling process have been described by Fourches et al.[23] Moreover, an additional OECD principle related to data curation before the model development was initiated.[23]: *"To ensure the consideration of (Q)SAR models for regulatory purposes, the models must be trained and validated on chemical datasets that have been thoroughly curated with respect to both chemical structure and associated target property values"*. We do believe that common acceptance of these practices in the future will positively affect the quality of QSAR research. Unfortunately, discussion of the data quality was absent in all QSAR studies of mixtures reviewed in our paper.

(ii) Model building: Almost every cheminformatics lab has its own reliable protocols of developing adequate QSAR models, e.g.,[33,34] but we will recommend the workflow described in[25,31] because of particular attention paid to rigorous internal and external cross-validation, estimation of AD, and *Y*-randomization as necessary steps of obtaining models. Consensus QSAR modeling is another required and important part of this workflow. It has recently become increasingly popular[35] and will be widely used in QSAR in future because the quality of predictions and AD of consensus model are usually higher than in conventional (non-consensus) models.

Correct AD estimation remains one of the most important and still unsolved problems in QSAR analysis of either single compounds or mixtures. It was well described in several publications[25,31,36–39] and will not be discussed in this review. However, the authors do hope that the solution of this problem will be one of the best achievements in cheminformatics in the not-so-distant future.

(iii) Rigorous external validation: Actually, rigorous external validation could be considered as a part of model development. We purposefully separated these stages to emphasize that every model must be externally validated using the molecules which had no involvement in either model development or selection. N-fold external cross-validation, when the whole modeling set alternately (by n turns) serves as the external validation set, currently is the easiest and most convenient way. If new experimental data are released after completing model development,[40] these data can be used for additional external validation.

(iv) Model exploitation and use for regulatory purposes: Before discussing this stage, we wish to remind the reader that QSAR modeling is an applied discipline and, therefore, except for benchmarking studies devoted to comparing newly developed methods with the best existing ones, modeling just for the sake of it is absolutely unacceptable, and experimental validation is still the most critical test of the actual utility of QSAR models. The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal. Moreover, the study[23] also illustrated that robust QSAR models could be used for initial experimental data curation, i.e., to question true positives as well as recover false negatives resulting from high throughput screening. In addition, we state that the model must be first predictive, while its interpretability is an important but secondary feature. In other words, a predictive and interpretable model has an advantage over a predictive and non-interpretable one, but both of them are acceptable, while a non-predictive and interpretable model is totally useless and unacceptable.

Although good practices of QSAR modeling are well-described in literature,[25,27–29,31,41] a large fraction of even recently published models, including the ones for mixtures, is still lacking reliability. It is closely related to the undesirably high population of "button pushers", i.e., researchers who conduct modeling without understanding and analyzing the data and modeling process itself, among the cheminformaticians. Prof. Bajorath attributes this to the ease of computational modeling on a technical level and the ability to carry out advanced calculations without critical assessment and understanding of their scientific foundations and limitations.[42] Numerous efforts, including OECD principles,[27] JCIM editorial,[28] and reviews published by other authors,[25,26, 29–31] promote the decrease in the number of low-quality QSAR papers and the size of the population of "button pushers", and we also hope that this critical review will also contribute to this process. We are aware that some recommendations listed below could be considered trivial by the experts in the field, but we hope that they will be useful to the beginners in cheminformatics or researchers from related fields, e.g., medicinal chemists or virologists whose data we are using for modeling.

The above discussion should explain why it should not come as a surprise that only a few QSAR studies of mixtures reported in Table 3 can be considered reliable. The vast majority of them will not be discussed in this review because of their obvious uselessness for any practical application. Their main drawbacks were: the predictivity (and sometimes robustness) of obtained models was not validated; they have no AD estimation and no proof of passing *Y*-randomization; and a very small set of congeneric compounds was frequently used for model development. An interested reader can find detailed descriptions of some of these studies in two well-written reviews[43,44] devoted to toxicity of mixtures and its modeling by QSAR. However, we wish to highlight studies[45,46] that exemplify unawareness of some researchers of the best practices of QSAR modeling. For instance, Wang et al.[46] analyzed 20 mixtures with one constant component and activity range of 0.84 logarithmic units (LU) only. In another study Yu et al.[45] developed a continuous QSAR model based on a dataset consisting of two clusters: 3 mixtures with p$EC_{50}$ values from 2.86 to 3.06 and 18 mixtures with p$EC_{50}$ values from 3.80 to 4.53. Certainly, very well-fitted ($R^2 = 0.93–0.95$) and absolutely unacceptable models were developed. These two examples serve as a perfect illustration that the wrong use of low-quality data will unavoidably result in worthless models. Unfortunately, QSAR studies of similar low quality are still published even in respectable journals casting a shadow on the entire field.

## 3 Data about Mixtures

The biggest problem in the field is a lack of data about mixtures and their properties available for modeling. We predict that ongoing rapid growth of large, publicly available databases such as PubChem[47] and ChEMBL[48] will eliminate this problem within next several years, as it has recently happened with the analogous problem in conventional QSAR.[49] As of June 2011, PubChem provides open access to over 30 million pure and characterized chemical compounds and close to 105 thousand complexes or mixtures.[50] However, these data contain a fair amount of duplicate, triplicate, etc., records spread among many assays reported in PubChem. Moreover, the PubChem crew focuses on the integration and repository of the data[50] without any curation efforts. Therefore, the quality of data in PubChem is very low.[51] Unfortunately, it is currently impossible to collect a reliable mixture dataset for QSAR analysis of

**Table 3.** Existing QSAR studies of mixtures.

| No. | Investigated activity | Number of compounds (modeling +external) | Descriptors | Statistical methods | Models | Validation | AD | Y-scrambling | Year /Ref. |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Toxicity to *Photobacterium phosphoreum* | 16 multicomponent mixtures | $K_{OW}$ | MLR | Single:<br>$R^2=0.93$<br>Mixtures:<br>$R^2=0.95$ | Internal | no | no | 2001[45] |
| 2 | Toxicity to *Photobacterium phosphoreum* | 8 single<br>84 binary mixtures | $K_{MD}$ | MLR | Single:<br>$R^2=0.934$<br>Mixtures:<br>$R^2_{test}=0.973$ | Internal (10 mixtures) | no | no | 2002[55] |
| 3 | Toxicity to *Photobacterium phosphoreum* | 21 mixtures | Hammett constant<br>The charge of the carbon atom in the carbon chain | MLR | $R^2=0.89$–0.92 | no | no | no | 2003[63] |
| 4 | Toxicity to *Photobacterium phosphoreum* | 94 binary mixtures | $K_{OW}$<br>Lewis acidity<br>Lewis basicity | MLR | $R^2_{adj}=0.95$ | Internal | no | no | 2003[56] |
| 5 | Toxicity to *Vibrio fischeri* | 34 single<br>33 multicomponent mixtures | $K_{OW}$ | MLR | Single:<br>$Q^2=0.88$–0.93<br>Mixtures:<br>$Q^2=0.94$<br>Single & Mixtures:<br>$Q^2=0.93$–0.95 | Internal | no | no | 2004[64] |
| 6 | Deviation of the experimental mixture density (MED) from the “ideal” calculated mixture density (MCD) (Delta) | 271 binary mixtures (4679 points) | 3D Dragon<br>3D Cerius | Neural network<br>k-NN | $R^2_{test}=0.754$–0.976 | External | no | yes | 2006[57] |
| 7 | Toxicity to *Photobacterium phosphoreum* | 12 single<br>50 binary mixtures | $K_{OW}$<br>Frontier orbital energy gap, Dipole moment, etc. | MLR | Single:<br>$R^2=0.961$<br>Mixtures:<br>$R^2=0.906$<br>$R^2=0.850$ | no | no | no | 2007[15] |
| 8 | Permeability coefficient ( $k_p$) | 24 multicomponent mixtures (288 points) | log*P*, *MW*, Indicator variable for the various vehicles (VEH), Molar refractivity (*MR*), other additive integral descriptors. | MLR | $Q^2_{25\%}=0.66$–0.80 | Internal | no | no | 2007[85] |
|  | Isolated perfused porcine skin flap (IPPSF) | 5 multicomponent mixtures (50 points) |  |  | $Q^2_{25\%}=0.06$–0.69 |  |  |  |  |
| 9 | Toxicity to *Scenedesmus obliquus* | 20 binary mixtures (one component is constant) | Log *P*, Frontier orbital energy gap | MLR | Mixtures:<br>$R^2_{test}=0.86$ | Internal | no | no | 2008[46] |
| 10 | Infinite dilution activity coefficients | 411 binary mixtures | Cerius | PLS, BPNN | $R^2_{test}=0.90$–0.94 | External | no | yes | 2008[58] |
| 12 | AChE inhibition | 17 compounds (2 achiral, 11 stereoisomers, and 14 racemic mixtures) | SiRMS<br>Lattice model<br>Dragon | PLS | $Q^2_{ext}=0.95$–0.96 | External | yes | yes | 2009[75] |
| 13 | Anti-Polio activity | 8 single compounds and 146 binary mixtures | SiRMS | PLS | $Q^2_{ext}=0.86$ | External | yes | yes | 2009[76] |
| 14 | Excess molar volume | 271 binary mixtures 4679 points) | Cerius 2 | BPNN, ENN | $R^2_{test}=0.95$ | External | No | Yes | 2010[59] |

mixtures from PubChem. Some information about the mixtures can be found within ChEMBL;[48] there are 356 mixtures of organic molecules not including salts, etc., but there is no single target that contains enough information to perform QSAR modeling of mixtures. Some data about the anticancer activity of mixtures could be found in the NCI database.[52] However, excluding more than 4000 salts, the remaining ca. 300 mixtures have an activity distribution that does not allow development of QSAR models. The same situation is observed for the DTP AIDS Antiviral Screen database,[53] where compounds were checked for evidence of anti-HIV activity. In total, slightly more than 200 mixtures were tested and only 6 of them were reported as confirmed actives. There are 1248 drug product mixtures reported in Thomson Reuters Integrity.[54] This includes mixtures of not only small-molecule drugs, but also of biologics and vaccine (e.g., DTaP/Hib vaccine). Experimental data are spread between many targets and assays and may not be available for all mixture records, and not all of the drug mixtures have associated generic names (i.e., some are only referred to by abbreviation or code name).

Some very limited and disparate data could be collected from the literature. For instance, toxicity to *Photobacterium phosphoreum* dataset of ca. 100 compounds could be extracted from two papers by Lin et al.[55,56] Three datasets of reasonable size (271–411 binary mixtures) related to various properties of liquids were published by Ajmani et al.[57–59] Vapor-liquid equilibrium data for 101 pure compounds and 261 binary mixtures were compiled by Oprisiu et al.[60] from the Korean Thermophysical Properties Databank.[61] Five hundred fifty mixtures created by 33 compounds used for discovery of novel small-molecule combinations with strong anti-inflammatory activity were investigated by Small et al.[62] This dataset is unique because, unlike other studies, the mixtures are not binary but contain two to nine components. Some data about the antiviral action of mixtures were published (for instance, studies[2,6–9]), but sometimes only one combination of drugs was investigated per publication, and it is difficult to collect even a single dataset about antiviral activity of mixtures.

## 4 Existing Mixture Descriptors and Their Application in QSAR Studies

A key feature in building QSAR models of mixtures is the use of appropriate mixture descriptors. Adequacy of the descriptors of mixtures will affect model quality much more than the choice of statistical approach. All published studies could be split into several groups depending on the descriptor type: (i) descriptors based on the partition coefficient for mixture; (ii) integral (whole-molecule) additive descriptors (weighted sum of descriptors of individual components); (iii) integral non-additive descriptors of mixtures (mixture components are taken into account in a different manner from the additive scheme); and (iv) fragment non-additive descriptors (structural parts of different mixture components are simultaneously taken into account in same descriptor).

The study of Lin et al.[63] does not belong to any of the aforementioned groups. The authors were trying to predict two different endpoints: (i) toxicity of 14 mixtures of aldehydes with malonitrile and (ii) toxicity of seven mixtures of cyanogenetic toxicants with acetaldehyde. The Hammet constant ($\sigma_P$) of individual aldehydes was used in the first case, and the charge of the carbon atom (C*) in the carbon chain of cyanogenetic toxicants was used in the second study. Thus, in both cases, standard chemical descriptors were used for describing the mixture. The use of single-compound descriptors was possible because one of the components of the binary mixtures was constant for all members of the modeling set; thus, this methodology has no general appeal because no special mixture descriptors were used.

### 4.1 Descriptors Based on Partition Coefficient for Mixtures

*N*-octanol/water partition coefficient ($K_{ow}$) for single compounds and partition coefficient ($K_{mix}$) for mixtures were used as descriptors in the studies.[45,64] To describe mixture partitioning, Equation 1 derived by Verhaar[65] was used to calculate $K_{mix}$. The intention of this surrogate parameter was to simulate the bioconcentration of a mixture of micropollutants, i.e., the process by which they penetrate from water into an organism and concentrate there, without the need to actually identify and quantify all compounds present in both aqueous and biomimetic samples. The aim is to provide one concentration-type number that can easily be related to the expected total body burden in aquatic animals. This number should be obtainable by measuring the total amount of xenobiotics in only the biomimetic sample.[65]

It should be realized that for mixtures of compounds of different partitioning behavior there is no simple distribution coefficient linking aqueous concentration to the concentration in a hydrophobic phase. This can be shown easily for a simple, two-constituent mixture. The partitioning behavior of a mixture can be described with the following Equation 1:

$$K_{\mathrm{mix}} = \frac{W}{V} \times \frac{\sum_{i=1}^{n} \frac{Q_i}{1+\frac{W}{V\cdot K_i}}}{\sum_{i=1}^{n} Q_i - \sum_{i=1}^{n} \frac{Q_i}{1+\frac{W}{V\cdot K_i}}} \qquad (1)$$

where $K_{\mathrm{mix}}$ is the partition coefficient of the mixture, $W$ is the volume of the aqueous phase, $V$ is the volume of the lipid phase, n is the number of compounds in the mixture, $Q_i$ is the total amount of compound $i$ in the system and $K_i$ is the partition coefficient of compound $i$.

If one applies this formula to a 1:1 mixture of two compounds with partition coefficients $K_i$ of 100 and 10000, respectively, for a *W/V* ratio of 105, $K_{mix}$ is approximately 4820, whereas for a *W/V* ratio of 1, $K_{mix}$ is approximately 199. It can equally be derived that, for an equimolar mixture, in order to arrive at a $K_{mix}$ that is effectively independent of the *W/V* ratio, *W/V* must at least be five time higher than $K_i$ of the most lipophilic compound of the mixture. This also means that for mixtures of unknown composition, the W/V ratio must be chosen such that it reflects the $K_i$ of the most lipophilic compound expected or the most lipophilic compound that is to be considered important. This *W/V* ratio restriction also ensures that there will be no significant decrease in aqueous concentration, as required for biomimetic extraction.[65]

Empore $C_{18}$ disk/water partition coefficient ($K_{SD}$) was used as descriptor by Lin et al.[55] Log$K_{SD}$ was calculated according to Equation 2 introduced by Verhaar et al.:[65]

$$\log K_{SD} = 0.995\ \log K_{OW} + 0.70 \qquad (2)$$

where $K_{OW}$ is *n*-octanol/water partition coefficient, which was calculated by the Fragment Constant Methodology of Hansch and Leo.[66] The model (Equation 2) was developed using only 17 compounds without any validation, even internal, and estimation of its applicability domain (AD). Thus, the ability of this model to predict log$K_{SD}$ is highly questionable.

Empore $C_{18}$disk/water partition coefficient for mixtures ($K_{MD}$) was calculated according to Equation 1 and has all the drawbacks inherent in this methodology and described above. In another study,[56] the autors used two following types of descriptors:

(i) *n*-octanol/water partition coefficient ($K_{OW}$) for single compounds and n-octanol/water partition coefficient for mixtures ($K_{MOW}$), which was calculated according to the Equation 1.

(ii) the Lewis acidity (*A*) and basicity (*B*) which were calculated using Equations 9 and 10.

$$A = \log K_{bw} - \log K_{cw} \qquad (3)$$

$$B = \log K_{chw} - \log K_{tw} \qquad (4)$$

where $K_{bw}$ is the di-*n*-butyl ether–water partition coefficient, $K_{cw}$ is the cyclohexane–water partition coefficient, $K_{chw}$ is the chloroform–water partition coefficient, and $K_{tw}$ is the carbon tetrachloride–water partition coefficient.

The partition coefficients log$K_{mow}$ (*n*-octanol/water partition coefficient for mixtures), log$K_{mbw}$ (di-*n*-butyl ether–water partition coefficient for mixtures), log$K_{mcw}$ (cyclohexane–water partition coefficient for mixtures), log$K_{mchw}$ (chloroform–water partition coefficient for mixtures), and log$K_{mtw}$ (carbon tetrachloride–water partition coefficient for mixtures) were calculated according to Equation 1. If $K_{mbw}$, $K_{mcw}$, $K_{mchw}$, and $K_{mtw}$ were used as partition coefficients instead of $K_{bw}$, $K_{cw}$, $K_{chw}$, and $K_{tw}$ in Equations 3 and 4; the difference among these partition coefficients is mainly caused by the joint effect of hydrogen bond in mixtures (Equations 5 and 6) that can be quantified as following:

$$A^{MH} = \log K_{mbw} - \log K_{mcw} \qquad (5)$$

$$B^{MH} = \log K_{mchw} - \log K_{mtw} \qquad (6)$$

where $A^{MH}$ and $B^{MH}$ is the joint effect of hydrogen-bond in mixtures which are analogous to Lewis acidity and Lewis basicity respectively.

All the studies, where partition coefficient for mixture based descriptors were used, were done with very small dataset and all the models were not properly validated. Utility of such descriptors is very low because it is impossible to perform QSAR analysis for large complex datasets using just $K_{mix}$ and related descriptors. Positive impact of such descriptors is that the property of mixture as a whole was intended to be considered.

### 4.2 Integral Additive Descriptors

This type of descriptors was not used very frequently, perhaps because intuitively the behavior of a mixture is much more complicated than is reflected by an additive scheme. Two following integral additive descriptors were used by Wang et al.:[46]

(i). *n*-octanol/water partition coefficient for single compounds (logP) and for the binary mixtures (log$P_{mix}$) which is estimated by Equation 7.

$$\log P_{mix} = (C_A \times \log P_A + C_B \times \log P_B + \cdots)/(C_A + C_B + \cdots) \qquad (7)$$

where log$P_A$ and log$P_B$ are *n*-octanol/water partition coefficients of components A and B respectively, and $C_A$ and $C_B$ are the concentrations of components A and B, respectively, in binary mixtures.

(ii). The frontier orbital energy gap for single compounds ($\Delta E$) and for mixtures ($\Delta E_{mix}$). According to Equation 8, $\Delta E$ for a single compound was defined as the difference between energy of its highest occupied molecular orbital ($E_{HOMO}$) and the lowest unoccupied molecular orbital ($E_{LUMO}$).

$$\Delta E = E_{LUMO} - \Delta E_{HOMO} \qquad (8)$$

The frontier orbital energy gap of binary mixture ($\Delta E_{mix}$) was calculated as a simple sum of $\Delta E$ of its individual components - compounds A and B (Equation 9).

$$\Delta E_{mix} = \Delta E_A + \Delta E_B \qquad (9)$$

where $\Delta E_A$ and $\Delta E_B$ are orbital energy gaps for compounds A and B, respectively.

Calculation of mixture descriptors used in the study[46] is clear and intuitevily understanable, however, they are suitable only for cases, when mixtures have an additive effect. But simple additive scheme is enough for the cases like this and QSAR analysis is not needed there.

The authors of study[57] reasoned that "the main problem here is to decide how to calculate descriptors for mixtures". The set of various 2D and 3D descriptors described by Todeschini[67] was calculated for the individual compounds using Dragon[68] and Cerius2[69] software. Descriptors of mixtures were calculated (Equation 10) as mole weighted average using the descriptor value and mole fraction of each component as follows:

$$\mathrm{MD} = \mathrm{R1D1} + \mathrm{R2D2} \tag{10}$$

where MD is the mixture descriptor, R1 and R2 are the mole fractions of the first and second components in the mixture, and D1 and D2 are the descriptors of the first and second components.

Although the mechanistic interpretation of Cerius 2 and Dragon descriptors is hard or even impossible, the authors of this review are pleased to note that all physical-chemical characteristics serving as a base for computing selected descriptors are intuitively related to the investigated property. Thus, according to selected descriptors, the following factors were important for the characterization of liquid mixture density: (i) inter- and/or intramolecular hydrogen bonding; (ii) electrophilic/nucleophilic and/or charge-transfer type interactions between the components in the mixture; and (iii) mutual interaction (accommodation and adjustment effects) of the components within the mixture. Unfortunately, the additive nature of mixture descriptors significantly decreased the value of this work. Thus, comparison of the results reported in study[57] with the results obtained on the same dataset using five non-additive mixture descriptors[59] showed that the latter were at least as good as several hundreds of Dragon and Cerius 2 additive descriptors for modeling but much better for mechanistic interpretation.

### 4.3 Integral Non-additive Descriptors

The most attractive feature of such descriptors is that they are not calculated from single compounds following the simple additive scheme, but instead attempt to consider the mixture as a whole. In our opinion, this more inclusive approach will increase the accuracy of models relating mixture composition and structure of its constituents to the investigated property.

Quantum chemical parameters obtained with Gaussian98[70] using ab initio MO theory at 6-31G* basis set level were used for calculating mixture descriptors by Zhang et al.[15] Both molecules of the binary mixture were pooled together for the calculation of descriptors, which included total energy (TE)/1500, $\triangle$TE (the difference of total energy, described by Equation 11), $E_{\mathrm{HOMO}}$, $E_{\mathrm{LUMO}}$, $\Delta E_{\mathrm{mix}}$, $\mu_{\mathrm{M}}$, $Q_{\mathrm{mM}}$ (mean absolute atomic charge of binary mixture, described by Equation 12), $q_{\mathrm{HM}}{}^{+}$, $q^{-}{}_{\mathrm{M}}$, $\mathrm{lgEnr}_{\mathrm{M}}$ and $\mathrm{GAPV}_{\mathrm{mM}}$ (the absolute value of molar volume difference, described by Equation 13; M represents binary mixture). Units of energy, charge, dipole, and molar volume were electron volt (eV), atomic charge unit (a.c.u.), and atomic unit (a.u.), bohr$^3$/mol, respectively.

$$\Delta \mathrm{TE} = \mathrm{TE}_{\mathrm{M}} - (\mathrm{TE}_i - \mathrm{Te}_j) \tag{11}$$

where $\mathrm{TE}_{\mathrm{M}}$ is the total energy of binary mixture of individual chemical $i$ and $j$, $\mathrm{TE}_i$ and $\mathrm{TE}_j$ are the total energy of individual chemical $i$ and $j$, respectively.

$$Q_{\mathrm{mM}} = \sum_{i}^{n} \frac{|q_i|}{n} \tag{12}$$

where $q_i$ is the charge of atom $i$, and $n$ is the summation of atom quantity of binary component.

$$\mathrm{GAPV}_{\mathrm{mM}} = |V_{\mathrm{m}i} - V_{\mathrm{m}j}|/1000 \tag{13}$$

where $V_{\mathrm{m}i}$, $V_{\mathrm{m}j}$ are the molar volumes of chemicals $i$ and $j$, respectively.

This study focused on the description of mixture as a whole, not derived from the descriptors of its components taken separately. To achieve this objective, the mixture was described by the ensemble of both components. This approach allowed to avoid limitations of additive schemes of mixture descriptors calculation as, for instance, those described by Wang et al.[46] However, the reported methodology was applied only to mixtures with 1:1 composition. In addition, the usage of three-dimensional (3D) representation of molecular structure and the use of quantum chemical descriptors is not the best choice for generating mixture descriptors because 3D descriptors are sensitive to the conformation of a molecule; very often this information is not available and mutual 3D orientation of compounds in a mixture is not known. Furthermore, it has been shown that 2D descriptors typically outperform 3D descriptors in QSAR modeling.[71] 3D descriptors are naturally preferred only for datasets consisting of racemates, enantiomers, and diastereomers with very different activity[24] when compound stereochemistry actually matters although studies have been reported in the literature on the development and use of chirality-sensitive 2D descriptors.[72] The advantages of using 2D descriptors were also discussed in, arguably, the most extensive comparative QSAR modeling studies by Gedeck.[73] The overall conclusion reached by the authors was that "none of the descriptors is best for all data sets; it is therefore necessary to test in each individual case, which descriptor produces the best model".

A unique set of mixture descriptors was reported in studies.[58,59] These investigations were devoted to QSPR analysis of infinite-dilution activity coefficients, excess molar volume and liquid density of different two-component mixtures. Mixture descriptors developed by the authors[58,59] reflect various intermolecular interactions between the components of mixture, taking into account the ratio of the constituents in the mixture. Parameters reflecting H-bonds (Equation 14), dipole-dipole interactions (Equation 15), and hydrophobic interactions (Equations 16, 17) were used.

H-bonds:

$$HBA-HBD = X_1X_2(2|A_1-A_2|+2|D_1-D_2|-A_1D_2-D_1A_2) \quad (14)$$

where $A_1$, $A_2$($D_1$, $D_2$) are the counts of hydrogen bond acceptors (donors) of first and second component respectively, and $X_1$, $X_2$ are the mole fractions of first and second component.

Dipole-dipole interactions:

$$Dip\_Mix = X_1X_2(Dip_1{}^2+Dip_2{}^2-2Dip_1Dip_2) \quad (15)$$

where, $Dip_1$ and $Dip_2$ are dipole moments of the first and second components respectively. It is well known that the energy of dipole-dipole interaction depends on the product of dipole moments. In Equation 15, $Dip_1{}^2$ and $Dip_2{}^2$ are responsible for dipole-dipole interactions of given constituent of mixture with itself and their doubled product $2Dip_1Dip_2$ is characterizing dipole-dipole interactions between different components of mixture.

Hydrophobic interactions were reflected by the descriptors based on the solvation/desolvation of mixture constituents in water and octanol:

$$FH_2O\_diff = X_1X_2|FH_1-FH_2| \quad (16)$$

$$FOct\_diff = X_1X_2|FO_1-FO_2| \quad (17)$$

where $FH_1$, $FH_2$ are desolvation free energies for water of first and second component, $FO_1$, $FO_2$ are desolvation free energies for octanol of first and second component; $X_1$, $X_2$ mole fractions of first and second component.

In addition, various descriptors accounting for effect of the shape (Equation 18) and size (Equation 19) of molecule in the mixtures were used:

$$Area\_ratio = X_1X_2(Ar_1/Ar_2) \quad \text{(ratio of area always > 1)} \quad (18)$$

$$Volume\_ratio = X_1X_2(V_1/V_2) \quad \text{(ratio of volume always > 1)} \quad (19)$$

$$\mathrm{Area\_Vol\_ratio} = \mathrm{X_1X_2(Area\_ratio/Vol\_ratio)} \quad (20)$$

where $Ar_1$, $Ar_2$ are molecular areas of first and second component; $V_1$, $V_2$ are molecular volumes of first and second component; and $X_1$, $X_2$ are mole fractions of first and second component.

Mixture descriptors developed by the authors[58,59] were capable to encode the most important non-covalent intermolecular interactions and may be sufficient to model a wide variety of properties of binary mixtures. However, we should note two drawbacks of this approach: (i) two different mixtures with $X_1=a$, $X_2=b$ and $X_1=b$, $X_2=a$, will have identical descriptors, but the properties of these mixtures could be rather different; and (ii) this approach is applicable only to binary mixture, where the components are dissolved in each other.

Ajmani et al.[58] developed and applied non-additive integral mixture descriptors to the dataset described earlier.[57] In addition to the deviation of the experimental mixture density,[57] excess molar volume was also modeled. The authors overcame the drawback of external test set rational selection inherent in their previous studies:[57,58] 25% of mixture points were included in the test set according to the mole fraction of the first component. However, the same compound could be the first component in one mixture and the second in another one, thus we can consider this approach as random. As in their previous two studies,[57,58] the authors succeeded to build robust and predictive QSPR models for both investigated endpoints and validate the utility of five developed mixture descriptors. However, the predictivity of the models, similarly to QMD-1 models from,[57] was limited by the missing points in the constitution of the given mixture because, according to the dataset splitting procedure, the points corresponding to different ratios of constituents of the mixture created by the same two compounds could be simultaneously present in both training and external test sets.

Concluding the analysis of the series of QSPR studies reported in,[57–59] we want to note that five non-additive integral mixture descriptors were as good as several hundreds of additive descriptors obtained using Dragon and Cerius 2 approaches for modeling but were significantly better than the latter for mechanistic interpretation. Moreover, seven descriptors reported in studies[58,59] encode the most important non-covalent intermolecular interactions and may be sufficient to model properties of binary mixtures, where the components are diluted in each other. However, their application is limited only to such systems and cannot be extended to more difficult and more interesting cases like drug-drug or ligand mixture-target interactions.

### 4.4 Fragment Non-additive Descriptors

Simplex representation of molecular structure (SiRMS)[33,74] has been used for the development of fragment non-additive descriptors of mixtures in two studies.[75,76] In the Simplex approach, any molecule can be represented as a system of different simplexes (tetratomic fragments of fixed composition, structure, chirality, and symmetry) on 1–4D levels.[77] Simplex approach proven itself as effective tool for solving various 1D-4D QSAR tasks for single chemicals[75,78–80] so Kuz'min et al.[75] decided to use 1–3D and develop special double 2.5D simplex descriptors for QSAR analysis of Acetylcholinesterase (AChE) inhibition by chiral organophosphates represented by single (R)- and (S)- isomers and their racemic mixture. In addition, 1–3D Dragon[67] and 3D Lattice Model (LM)[81] descriptors were used for comparison. All 3D structures were optimized using DFT B3LYP nonlocal correlation functional.[82,83] 3D Simplex, Dragon, and LM descriptors for single stereoisomers were calculated in a regular way. Mixture descriptors for racemic structure have been calculated as the half of the sum of corresponding descriptors for both (R)- and (S)-stereoisomers (Equation 21).

$$\mathrm{D}_{i(\mathrm{RS})} = 0.5 \times (\mathrm{D}_{i(\mathrm{R})} + \mathrm{D}_{i(\mathrm{S})}) \tag{21}$$

where $\mathrm{D}_{i(\mathrm{RS})}$ is i-th descriptor of racemic mixture; $\mathrm{D}_{i(\mathrm{R})}$ and $\mathrm{D}_{i(\mathrm{S})}$ are *i*-th descriptor of (R)- and (S)-isomers respectively. Certainly, all the drawbacks common for additive approach were suitable for these descriptors.

In comparison with additive 3D Simplex, Dragon, and LM descriptors, double 2.5D representation of molecular structure was expected to provide more adequate mixture descriptors. 2.5D appeared in the name because only topology of molecule (2D level) and chirality of phosphorus (3D property) have been taken into account. SiRMS allows ones to describe ensembles (mixtures) of molecules as a whole due to the application of unbounded simplexes.[33] In other words, a racemate can be considered as an ensemble of (R)- and (S)- isomers. Unbounded simplexes, where one part (of the simplex) represents (R)-, and the second part represents (S)-enantiomer, reflect structural species of the mixture, i.e., there are some simplexes that describe only racemate and they are absent for single enantiomers. It allows analyzing synergism, anti-synergism, or competition of racemic mixture components in AChE inhibition. Alongside with the racemate, two molecules have also been used for the description of achiral structure and individual stereoisomers (they were represented as a mixture of respectively two (R)- or (S)-isomers). Thus, every compound has been described by two molecules and such model of representation has been conventionally called as double 2.5D by Kuz'min et al.[75] Double 2.5D simplex descriptors are applicable only to mixtures with 1:1 ratio of components, which is a significant drawback of this approach.

The authors[75] developed robust and externally validated predictive models and even showed the supremacy of non-additive fragment mixture descriptors over 3D LM and 3D Dragon descriptors. However, chiral organophosphates did not show strong synergistic or antagonistic interactions and conventional 3D Simplex descriptors demonstrated the same performance as their mixture analogs. Moreover, very small size of the dataset (just 27 points) does not allow significant comparison of descriptors' predictivity. These models could be used only for the prediction of AChE inhibition for very limited class of highly similar chiral organophosphates.

The SiRMS approach was later modified[76] in order to make this method suitable for QSAR analysis of binary mixtures of any composition. The main difference with classical Simplex (or any other) approach is that the binary mixture is represented by two molecules simultaneously, when a pure compound is represented as a single molecule. Then simplex descriptors are calculated as usual. Bounded simplexes describe only single components of the mixture (compounds A or B), when unbounded simplexes can describe both the constituent parts and the mixture as a whole. It is necessary to indicate whether the parts of unbounded simplexes belong to the same molecule or to different ones. In the latter case, such unbounded simplexes will not reflect the structure of a single molecule, but will characterize a pair of different molecules. Actually, simplexes of this kind are, in fact, structural descriptors of mixtures of compounds (Figure 4). Special mark is used during descriptor generation to distinguish such "mixture" simplexes from ordinary ones. The mixture composition is taken into account, i.e., descriptors of constituent parts (compounds A and B) are weighted according to their molar fraction and mixture descriptors are multiplied on doubled molar fraction of deficient component. If in the same task both mixtures and pure compound have been considered, pure compound is considered as a mixture with composition $A_1B_0$. In this case only descriptors of pure compound A will be generated with the weight equal to 1. Thus, the structure of every mixture is characterized by both descriptors of the mixture as a whole and descriptors of its individual constituents.

Predictivity of developed models was validated using 8-fold external cross-validation, i.e., the authors wanted to model the addition of new drug to create a mixture with existing eight antivirals. Models obtained using Simplex descriptors of mixture significantly outperformed models based on additive descriptors. Feature net approach based on four additional inhibition concentrations ($IC_{30}$, $IC_{40}$, $IC_{60}$, $IC_{70}$,) of investigated compounds was not able to improve the quality of the consensus model developed for $IC_{50}$. Despite the aforementioned advantages, the study[76] also had two following drawbacks: in the majority of cases, both synergistic and antagonistic effects were constantly underestimated by the model and developed descriptors were only applicable to binary mixtures.

Simplex descriptors of mixtures as well as another type of mixture descriptors based on ISIDA substructural molecular fragments[84] were also used for the analysis of boiling point temperatures of biphasic mixtures.[60] The description of using ISIDA fragments for mixture description should appear in another paper that as of this writing is still under review by the Journal of Chemical Information and Modeling.[60]

Concluding this section, we should emphasize that two main problems of QSAR analysis of mixtures are: (i) lack of data concerning the action or property of mixtures, and (ii) adequate description of any mixture by the system of (structural) parameters. Concerning the methodologies mentioned above, additive descriptors of mixtures, where the latter were characterized by mole weighted average descriptors of the constituents, have the following disadvantages: (i) there is no rational basis for this approach, other than the expectation that conventional descriptors can be significant in the explanation of a property of mixture; and (ii) the consideration of (inter)action effects is impossible and, thus, only simple tasks with additive or very close to additive effects can be investigated by this approach. Advantages of the additive approach are: (i) the process of descriptor generation is simple and intuitively understandable; and (ii) this approach is not property-oriented, i.e., additive descriptors could be applied (bearing in mind the drawbacks of this methodology) to any investigated activity or property, and, sometimes, this method showed good results, like in the study of Ajmani et al.[57] More complicated and adequate mixture descriptors developed by the authors[58,59] are capable of encoding the most important non-covalent intermolecular interactions, which is their significant advantage. However, they are system- and property-specific, i.e., are applicable only to the systems (binary mixtures where the components are dissolved in each other) and properties that were described by Ajmani et. al.[58,59] (or highly similar to them). Another serious disadvantage of this approach is that two mixtures of different composition could be described by identical descriptors, i.e., this approach is not universal. Simplex and ISIDA mixture descriptors are free of the mentioned drawbacks; they can be applied to any property of interest; interaction or joint effect of components can be captured by these descriptors. However, in their current version, these approaches could be applied only to binary mixtures. From the methodological point of view these two approaches appear better than others, but serious improvement is still needed for both of them.

We should also state that no single QSAR study or methodology has been reported that can be recommended as a reliable uniform tool for the analysis of mixtures. In addition to poor dataset size and the drawbacks of the descriptors used (additive scheme, limited applicability, etc.), reported studies also contain methodological disadvantages shared with the conventional QSAR modeling, e.g., lack of external validation, absence of AD, etc. However, we did not only recognized and criticized the common drawbacks of existing studies, but also analyzed their origin and developed recommendations (described in Best Practices of QSAR Modeling, Proper External Validation of QSAR Models for Mixtures and Future Perspectives sections) for how to overcome some of them. We also want to state that the entire QSAR of mixtures field is still under development; existing approaches can be considered as a starting point, and perhaps the foundation, for the next generation of both descriptors and studies.

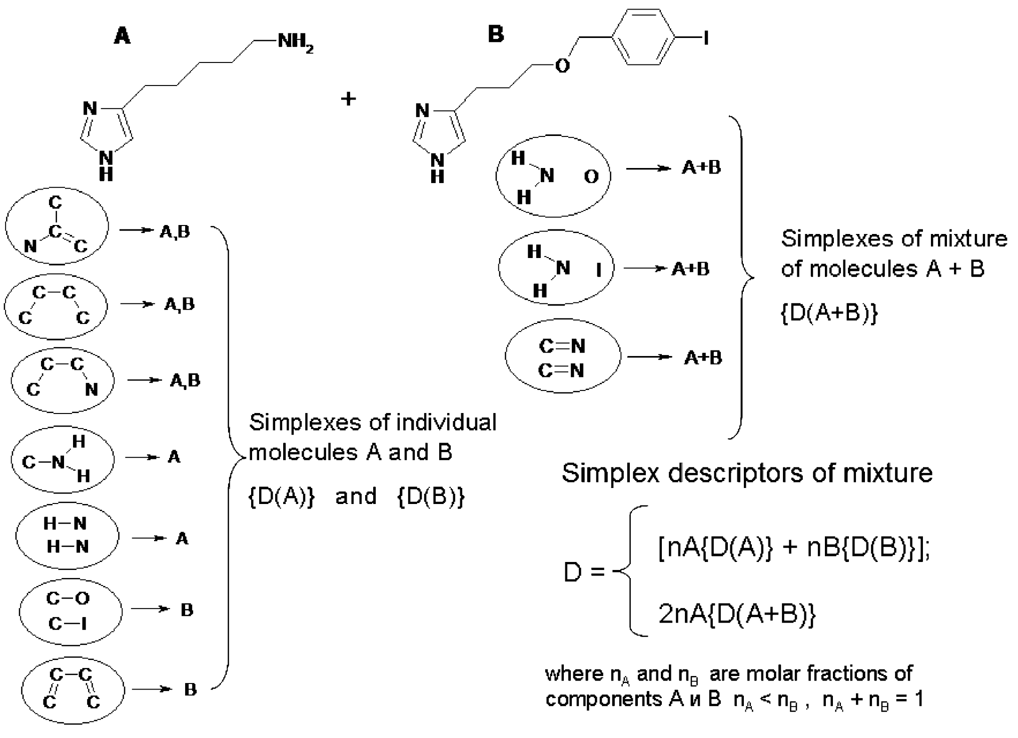

A
B
+
Simplexes of individual molecules A and B
{D(A)} and {D(B)}
Simplexes of mixture of molecules A + B
{D(A+B)}
Simplex descriptors of mixture
D = [nA{D(A)} + nB{D(B)}]; 2nA{D(A+B)}
where $n_A$ and $n_B$ are molar fractions of components A и B $n_A < n_B$ , $n_A + n_B = 1$

**Figure 4.** Simplex descriptors of mixtures.

## 5 Proper External Validation of QSAR Models for Mixtures

As in classical QSAR modeling, rigorous external validation is a required and it is an integral part of QSAR modeling of mixtures.[41] However, proper external validation of QSAR models for mixtures, especially when mixtures of the same compounds with different ratios are present several times in the dataset, is less straightforward than in traditional QSAR. The conventional external cross-validation procedure, when the points (compounds) are randomly placed in the external set (or fold) is unacceptable because it leads to an over-optimistic estimation of the predictive power of the developed models. Indeed, if both training and external sets include data points for the same mixture, the model's true predictive performance will not be estimated properly. Thus, new, more rigorous protocols for external validation must be developed specifically for QSAR modeling of mixtures.[60]

Depending on the initial data and potential application of developed models, three different strategies of external validation (Figure 5) could be used: (i) "points out" – prediction of the investigated property for any composition of mixtures from the modeling set, (ii) "mixtures out" – filling of missing cells in the initial data (mixtures) matrix, i.e., prediction of the investigated property for mixtures with unknown activity created by pure compounds from the modeling set, and (iii) "compounds out" – prediction of the investigated property for mixtures formed by novel pure compound(s) absent in the modeling set.

(I) "Points out". This strategy is applicable if and only if several mixtures with the same components with different ratios are present in the modeling set. Data points are randomly placed in each fold of the external cross-validation set (Figure 5A). Every mixture is present simultaneously in both training and external sets. This method is the simplest and the easiest one; it will adequately reflect the capability of models to predict only existing mixtures with novel composition and, therefore, its usefulness is fairly low. The "points out" strategy with small variations has been used in several studies.[57–60]

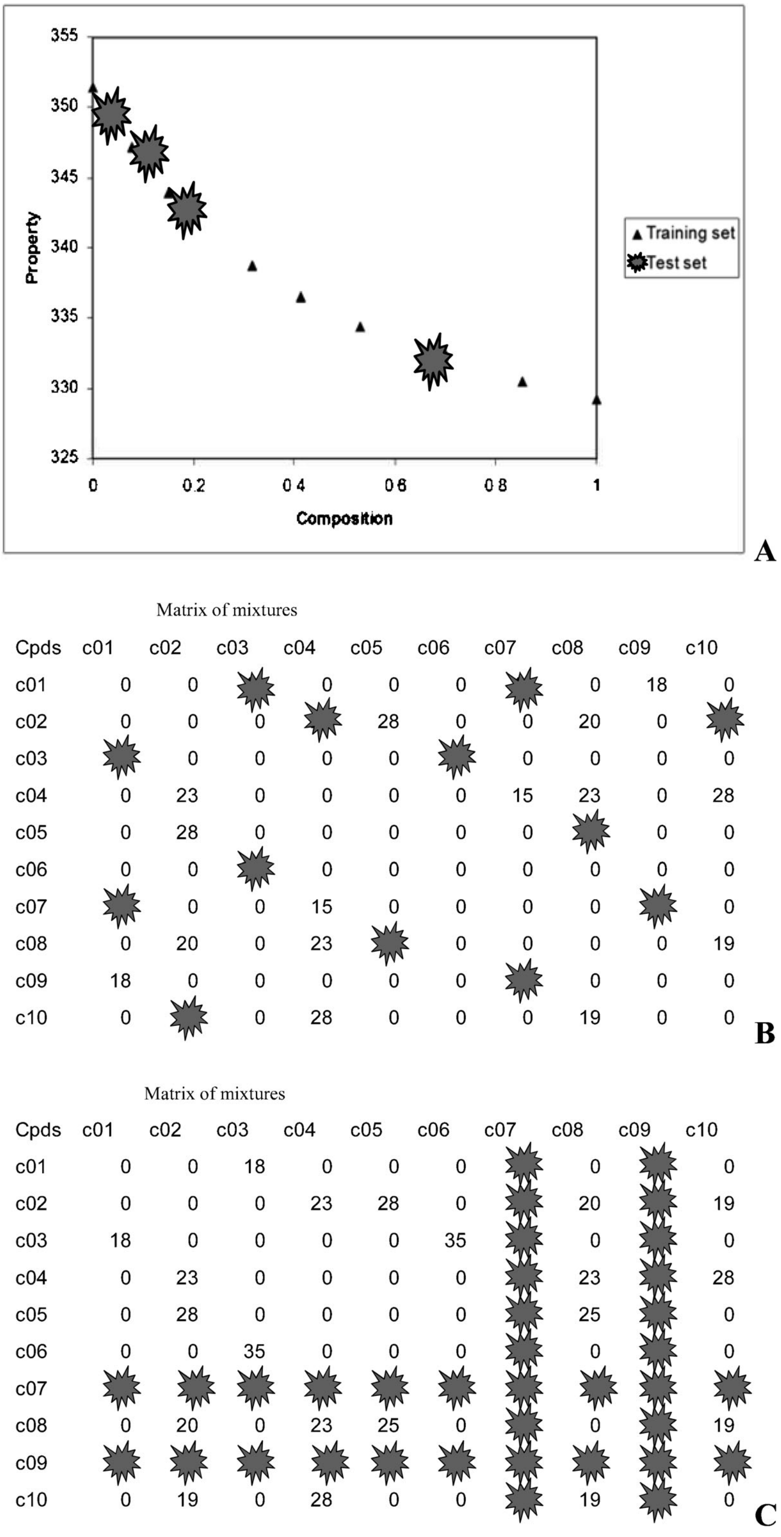
(II) "Mixtures out". All data points corresponding to mixtures composed of the same constituents but in different ratios are simultaneously removed and placed in the same external fold (Figure 5B). Thus, every mixture is present either in the training or external set, but never in both sets. The "mixtures out" strategy needs supervision if there are several compositions for the same mixture, but the process could be unsupervised if there is only one composition for the mixture of the given two compounds. Such a way of external cross-validation will gives a higher error of prediction than for the "points out" strategy. However, it will not be limited by already known mixtures and will be useful for estimating the accuracy of prediction of missing data in the mixture matrix formed by modeling set compounds. Such matrices are fairly sparse, and a gap-filling exercise could be useful. "Mixtures out" strategy and its variations have been used in studies.[57,60]

(III) "Compounds out". Pure compounds and their mixtures are simultaneously placed in the same external fold (Figure 5C). Thus, every mixture in the external set contains at least one compound which is absent in the training set. This differs from the classical external CV algorithm in that the folds are not created randomly, but supervised in order to keep the number of both pure compounds and their mixtures amongst the folds more or less constant. The supervision is especially needed in cases when one pure compound can participate in only one mixture and another compound can create plenty of mixtures; here, the classical random algorithm is unable to handle such a situation during external folds creation. Moreover, despite a supervised process of folds creation, some folds could be predicted badly because they are still anisotropic, and a considerable lack of information in the training set can be observed for some external folds. Every mixture will be placed in the external set n times, where n is a number of components in the mixture, except in the cases when several pure compounds - constituents of the given mixture - will belong to the same external fold. This procedure simulates the addition of a novel component to existing matrix of mixtures. This is the most rigorous method of external validation in QSAR modeling of mixtures. Although the error of prediction for this strategy is expected to be the largest, QSAR models passing "compounds out" strategy will be able to predict the investigated property for mixtures created by new pure compound(s) beyond the modeling set. The "compounds out" strategy has been used in studies.[60,76]

## 6 Perspectives for QSAR Modeling of Mixtures

We shall discuss now the trends relevant to future developments in the field of QSAR modeling of mixtures and note that some of these trends are also applicable to any QSAR study. Moreover, the latest trends of conventional QSAR analysis (i.e., careful collection and understanding of the data, thorough data curation, rigorous internal and external validation, and application of developed models for virtual screening of large databases for the discovery of new "hits" or "alerts", etc.), which are almost absent in current mixture studies, will significantly improve the quality of developed models.

We specially want to note that the purpose of this review was not to demonstrate that all existing studies and methodologies have certain drawbacks. Our chief goal was to explain the complexity of the problem to develop rigorous and predictive QSAR models of mixtures and stimulate the development of new and improved descriptors of mixtures to increase the reliability of the models. We assert that QSAR is and continues to be very useful in modeling and predicting the properties of mixtures. We are continu-



Matrix of mixtures

| Cpds | c01 | c02 | c03 | c04 | c05 | c06 | c07 | c08 | c09 | c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| c01 | 0 | 0 | ✱ | 0 | 0 | 0 | ✱ | 0 | 18 | 0 |
| c02 | 0 | 0 | 0 | ✱ | 28 | 0 | 0 | 20 | 0 | ✱ |
| c03 | ✱ | 0 | 0 | 0 | 0 | ✱ | 0 | 0 | 0 | 0 |
| c04 | 0 | 23 | 0 | 0 | 0 | 0 | 15 | 23 | 0 | 28 |
| c05 | 0 | 28 | 0 | 0 | 0 | 0 | 0 | ✱ | 0 | 0 |
| c06 | 0 | 0 | ✱ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| c07 | ✱ | 0 | 0 | 15 | 0 | 0 | 0 | 0 | ✱ | 0 |
| c08 | 0 | 20 | 0 | 23 | ✱ | 0 | 0 | 0 | 0 | 19 |
| c09 | 18 | 0 | 0 | 0 | 0 | 0 | ✱ | 0 | 0 | 0 |
| c10 | 0 | ✱ | 0 | 28 | 0 | 0 | 0 | 19 | 0 | 0 |

Matrix of mixtures

| Cpds | c01 | c02 | c03 | c04 | c05 | c06 | c07 | c08 | c09 | c10 |
|---|---|---|---|---|---|---|---|---|---|---|
| c01 | 0 | 0 | 18 | 0 | 0 | 0 | ✱ | 0 | ✱ | 0 |
| c02 | 0 | 0 | 0 | 23 | 28 | 0 | ✱ | 20 | ✱ | 19 |
| c03 | 18 | 0 | 0 | 0 | 0 | 35 | ✱ | 0 | ✱ | 0 |
| c04 | 0 | 23 | 0 | 0 | 0 | 0 | ✱ | 23 | ✱ | 28 |
| c05 | 0 | 28 | 0 | 0 | 0 | 0 | ✱ | 25 | ✱ | 0 |
| c06 | 0 | 0 | 35 | 0 | 0 | 0 | ✱ | 0 | ✱ | 0 |
| c07 | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ |
| c08 | 0 | 20 | 0 | 23 | 25 | 0 | ✱ | 0 | ✱ | 19 |
| c09 | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ |
| c10 | 0 | 19 | 0 | 28 | 0 | 0 | ✱ | 19 | ✱ | 0 |

**Figure 5.** Strategies of external validation for QSAR modeling of mixtures: "points out" (A); "mixtures out" (B); "compounds out" (C). External set selection is highlighted by asterisks.

ing to tune Simplex descriptors of mixtures for the analysis of multi-component mixtures of any composition and their application to various mixture datasets. We highly encourage and welcome the effort of Prof. Varnek's group in the University of Strasbourg in developing ISIDA-based mixture descriptors,[60] the intention of Prof. Livingston's group in the University of Portsmouth to account for intermolecular interactions which occur between ligands and receptors by mixture descriptors developed in his group,[59] as well as any other steps to enrich and improve QSAR of mixtures field.

The face of QSAR modeling has changed dramatically in recent years. QSAR has shifted away from the original simple and interpretable linear models developed using just a few descriptors towards complex multi-parametric and, quite often, non-linear approaches. However, in the case of current mixture modeling (i.e., rather small datasets), linear approaches and non-additive chemically transparent parameters will still be useful and even preferable to multi-parametric monsters because of the possibility of interpreting obtained results followed by targeted design of new mixtures with improved properties.

Future development of mixture therapy will push theoretical chemistry to create adequate tools for description and QSAR analysis of dual or multi-component mixtures with different composition. Indubitably, a key point of QSAR of mixtures is a correct description of the mixture. This will result in a great need for chemical descriptors that are capable of adequately representing the mixture consisting of two or more compounds, especially when the response is nonlinear. In addition, these mixture descriptors must be capable of dealing with strong synergistic and antagonistic effects. Unfortunately, nobody, including the authors of this review, is able to define what are the best descriptors of a mixture, but we can say that these descriptors should not represent the weighted sum of the descriptors of a mixture's constituents. One possible solution is conglomerate-based descriptors. This approach is based on the hypothesis that fluctuations in mixture properties from the additive scheme are driven by intermolecular interactions of mixture constituents. Such interactions could lead to the creation of various conglomerates which can cause synergism, antisynergism, etc. Thus, the key idea of the offered approach is to model mixtures as complex molecular systems and to describe them by possible constituents' conglomerates. A mixture will be defined by descriptors of individual compounds weighted according to their mole fractions and by conglomerates (descriptors of the mixture) weighted according to the probability of their realization.

For instance, for mixture $A_x+B_y+C_z$, one will have the construction $A_xB_yC_z$, where $x+y+z=100\%$. Then, the set of potential conglomerates will be: $A_2$, $B_2$, $C_2$, $AB$, $AC$, $BC$, $A_2B$, $A_2C$, $AB_2$, $BC_2$, $ABC$, etc. Conglomerates of the same constituents ($A_2$, $B_2$, $C_2$) are included. The maximal number of constituents in conglomerate is three. The probability of creation of every conglomerate depends on the composition of mixture and is determined by traditional combinatorial relationships. The number of realization variants for conglomerate $A_n$ for a mixture containing $x$% of compound A is $x!/(x-n)!n!$; for conglomerate $B_\mathrm{m}-y!/(y-m)!m!$; and for $C_k-z!/(z-k)!k!$. For heterogeneous conglomerates, for instance, $A_nB_m$, it is equal to $x!y!/[(x-n)!n!][(y\ -m)!m!]$. The probability of existence of a given conglomerate is determined by the ratio of variants of its realization to the total number of variants for all considered conglomerates. Thus, conglomerates are multiplied with corresponding probabilities and used as descriptors of mixtures in QSAR/QSPR. Obviously, such an approach will enlarge the number of generated descriptors. Even for binary mixtures, there are seven possible types of conglomerates: ($A_2$, $A_3$, $B_2$, $B_3$, $AB$, $A_2B$, $AB_2$). However, this is not a problem for modern modeling techniques. Moreover, because there is no way to sort compounds in a mixture, the real number of conglomerates will be smaller. For instance, one cannot distinguish conglomerates $\mathrm{A_2B}$ and $\mathrm{AB_2}$, and values of these descriptors will be averaged. Overall, four conglomerates could be considered for a binary mixture: $(A_2+B_2)/2$; $(A_3+B_3)/2$; $(A_2B+AB_2)/2$, and $AB$. We hope that the conglomerate approach for mixture descriptor generation will increase the quality of QSAR/QSPR models and will contribute to the understanding of the action of mixtures by estimating the influence of concrete conglomerates on the investigated property.

We also forecast that a lot of structural information about the biological targets and their interactions with drugs (and their mixtures) will be revealed and collected in near future. It will solve the problem of the lack of data that is crucial for QSAR of mixtures at the moment. The availability of structural data and the constantly growing computing power allow one to conclude that theoretical methods in the future medicinal chemistry will not be limited only to QSAR techniques, but structure based approaches such as docking and molecular dynamics computations will be increasingly required and useful.[3]

## 7 Summary

- The application of mixtures is expected to significantly increase in the nearest future. This trend creates an emerging need for new theoretical approaches employing QSAR analysis of mixtures, especially focused on the adequate description of chemical mixtures.
- The biggest problem in the field is the lack of data. However, we are expecting that ongoing rapid growth of large, publicly available databases will eliminate this problem within the next several years.
- The adequate description of any mixture by the system of (structural) parameters is the second main problem. The usage of non-additive mixture descriptors that are (i) based on structural features of both constituents and the mixture as a whole and (ii) sensitive to interaction effects will help to overcome this problem.
- The latest trends of conventional QSAR analysis, i.e., careful collection and understanding of the data, thorough data curation, rigorous internal and external validation, and application of developed models for virtual screening of large databases, which are almost absent in

current mixture studies, will significantly improve the quality of mixture QSAR models.

- External validation of QSAR models for mixtures is less straightforward than in traditional QSAR. The conventional external cross-validation procedure is unacceptable because it leads to over-optimistic estimation of the predictive power of developed models. The "compounds out" strategy, where a given single compound and all its mixtures are simultaneously placed in the same external fold, is the most rigorous way of external validation in QSAR modeling of mixtures.
- We specially want to note that our review should be considered not merely as a warning that all existing studies and methodologies have certain drawbacks, but also as a stimulus to develop new and improved descriptors of mixtures and to increase the reliability of the modeling part.
- We assert that QSAR is and continues to be very useful in modeling and especially predicting the properties of mixtures. The field of QSAR of mixtures is still under development, and existing efforts could be considered a starting point and perhaps the foundation for future approaches and studies. We encourage and welcome any efforts directed to the development of new, and the improvement of existing, QSAR approaches for mixtures.

## Acknowledgements (financial and competing interest disclosure)

E. Muratov gratefully acknowledges the financial support from *NIH* (grant GM6694) and *EPA* (RD 83382501 and R832720). Finally, the authors would like to thank the anonymous reviewers for highly valuable comments and suggestions that helped us significantly improve and enrich this paper. The authors are grateful to Ms. *Mary La* for her help in editing this manuscript.The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties, apart from those disclosed. There are no any potential conflicts of interest with the subject matter or materials discussed in the manuscript. No writing assistance was utilized in the production of this manuscript.

## References

[1] A. Dumas, *Les Trois Mousquetaires*, Baudry, Paris, **1844**.

[2] L. Nikolaeva, A. S. Galabov, *Acta virologica* **1999**, *43*, 303–311.

[3] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, V. E. Kuz'min, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *Future Med. Chem.* **2011**, *3*, 31–43.

[4] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran et al., *Nature* **2009**, *262*, 175–182.

[5] H. D. M. Coutinho, J. G. M. Costa, E. O. Lima, V. S. Falcao-Silva, J. P. Siqueira, *Biol. Res. Nurs.* **2010**, *11*, 332–335.

[6] N. A. Ilyushina, N. V. Bovin, R. G. Webster, E. A. Govorkova, *Antiviral Res.* **2006**, *70*, 121–131.

[7] N. A. Ilyushina, A. Hay, N. Yilmaz, A. C. M. Boon, R. G. Webster, E. A. Govorkova, *Antimicrobial Agents Chemother.* **2008**, *52*, 3889–3897.

[8] N. A. Ilyushina, E. Hoffmann, R. Salomon, R. G. Webster, E. A. Govorkova, *Antiviral Ther.* **2007**, *12*, 363–370.

[9] L. Nikolaeva, A. S. Galabov, *Acta virologica* **2000**, *44*, 73–78.

[10] http://en.wikipedia.org/wiki/List_of_bestselling_drugs

[11] P. L. Smithburger, S. L. Kane-Gill, N. J. Benedict, B. A. Falcione, A. L. Seybert, *Ann. Pharmacother.* **2010**, *44*, 1718–1724.

[12] S. Xu, N. Nirmalakhandan, *Water Res.* **1998**, *32*, 2391–2399.

[13] *Guidelines for the Health Risk Assessment of Chemical Mixtures*, US EPA, Washington, **1986**, p. 29.

[14] *Supplementary Guidance for Conducting Health Risk Assessment of Chemical Mixtures*, US EPA, Washington, **2000**, p. 209.

[15] L. Zhang, P. Zhou, F. Yang, Z. Wang, *Chemosphere* **2007**, *67*, 396–401.

[16] H. Jenssen, T. J. Gutteberg, T. Lejon, *J. Pept. Sci.* **2005**, *11*, 97–103.

[17] A. Kovatcheva, A. Golbraikh, S. Oloff, Y. Xiao, W. Zheng, P. Wolschann, G. Buchbauer, A. Tropsha, *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 582–595.

[18] H. Kubinyi, *J. Cancer Res. Clin. Oncol.* **1990**, *116*, 529–537.

[19] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, L. G. Gorb, M. Qasim, J. Leszczynski, *J. Comp.-Aided Mol. Des.* **2008**, *22*, 747–759.

[20] E. N. Muratov, A. G. Artemenko, V. E. Kuz'min, V. P. Lozitsky, A. S. Fedchuk, R. N. Lozitska, Y. A. Boschenko, T. L. Gridina, *Antiviral Res.* **2005**, *65*, A62–A63.

[21] J. R. Votano, M. Parham, L. M. Hall, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, *J. Med. Chem.* **2006**, *49*, 7169–7181.

[22] S. Zhang, A. Golbraikh, A. Tropsha, *J. Med. Chem.* **2006**, *49*, 2713–2724.

[23] D. Fourches, E. N. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

[24] E. N. Muratov, A. G. Artemenko, E. V. Varlamova, P. G. Polischuk, V. P. Lozitsky, A. S. Fedtchuk, R. N. Lozitska, T. L. Gridina, L. S. Koroleva, V. N. Sil'nikov et al., *Future Med. Chem.* **2010**, *2*, 1205–1226.

[25] A. Tropsha, *Mol. Inf.* **2010**, *29*, 476–488.

[26] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

[27] The report from the expert group on (Quantitative) Structure-Activity Relationships[(Q)SARs] on the principles for the validation of (Q)SARs. In *OECD Series on Testing and Assessment*, Organisation for Economic Co-operation and Development, Paris, **2004**, p. 206.

[28] W. L. Jorgensen, *J. Chem. Inf. Model.* **2006**, *46*, 937–937.

[29] J. Dearden, M. Cronin, K. Kaiser, *SAR QSAR Environ. Res.* **2009**, *20*, 241–266.

[30] A. Doweyko, *J. Comp.-Aided Mol. Des.* **2008**, *22*, 81–89.

[31] A. Tropsha, A. Golbraikh, *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.

[32] D. Young, T. Martin, R. Venkatapathy, P. Harten, *QSAR Comb. Sci.* **2008**, *27*, 1337–1345.

[33] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *J. Comp.-Aided Mol. Des.* **2008**, *22*, 403–421.

[34] A. Lagunin, A. Zakharov, D. Filimonov, V. Poroikov, *SAR QSAR Env. Res.* **2007**, *18*, 285 – 298.

[35] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha, *J. Chem. Inf. Model.* **2006**, *46*, 1984 – 1995.

[36] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *Altern. Lab. Anim.* **2005**, *33*, 445 – 459.

[37] I. V. Tetko, I. Sushko, A. K. Pandey, A. Tropsha, H. Zhu, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.

[38] A. Tropsha, in *Burger's Medicinal Chemistry and Drug Discovery,* Vol. 1, 7ed. (Ed: D. Abraham), Wiley, New York, **2010**, pp. 505 – 533

[39] A. Tropsha, A. Golbraikh, in *Handbook of Chemoinformatics Algorithms* (Eds: J.-L. Faulon, A. Bender), Chapman and Hall, London, **2010**, pp. 175 – 212

[40] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, I. V. Tetko, *J. Chem. Inf. Model.* **2008**, *48*, 766 – 784.

[41] A. Tropsha, P. Gramatica, V. Gombar, *QSAR Comb. Sci.* **2003**, 22,69 – 77.

[42] J. Bajorath, *J. Comp.-Aided Mol. Des.* **2012**, *26*, 11–12

[43] R. Altenburger, M. Nendza, G. Schuurmann, *Environ. Toxicol. Chem.* **2003**, *22*, 1900 – 1915.

[44] M. Tichy, M. Cikrt, Z. Roth, M. Rucki, *SAR QSAR Environ. Res.* **1998**, *9*, 155 – 169.

[45] H. Yu, Z. Lin, J. Feng, T. Xu, L. Wang, *Acta Pharmacol. Sin.* **2001**, *22*, 45 – 49.

[46] C. Wang, G. Lu, Z. Tang, X. Guo, *J. Environ. Sci.* **2008**, *20*, 115 – 119.

[47] http://pubchem.ncbi.nlm.nih.gov/.

[48] ChEMBLdb. https://www.ebi.ac.uk/chembl/

[49] A. Williams, V. Tkachenko, C. Lipinski, A. Tropsha, S. Ekins, *Drug Disc. World* **2010**, *10*, 33 – 39.

[50] E. Bolton, J. Chen, S. Kim, L. Han, S. He, W. Shi, V. Simonyan, Y. Sun, P. Thiessen, J. Wang et al., *J. Cheminform.* **2011**, *3*, 32.

[51] J. H. Hsieh, X. S. Wang, D. Teotico, A. Golbraikh, A. Tropsha, *J. Comp.-Aided Mol. Des.* **2008**, *22*, 593 – 609.

[52] http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html.

[53] *DTP AIDS Antiviral Screen*, http://dtp.nci.nih.gov/docs/aids/aids_data.html.

[54] Thomson Reuters Integrity, http://thomsonreuters.com/productsservices/science/science_products/a-z/integrity.

[55] Z. Lin, H. Yu, D. Wei, G. Wang, J. Feng, L. Wang, *Chemosphere* **2002**, *46*, 305 – 310.

[56] Z. Lin, P. Zhong, K. Yin, L. Wang, H. Yu, *Chemosphere* **2003**, *52*, 1199 – 1208.

[57] S. Ajmani, S. Rogers, M. Barley, D. Livingstone, *J. Chem. Inf. Model.* **2006**, *46*, 2043 – 2055.

[58] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *QSAR Comb. Sci.* **2008**, *27*, 1346 – 1361.

[59] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *Mol. Inf.* **2010**, *29*, 645 – 653.

[60] I. Oprisiu, E. Muratov, E. Varlamova, A. Artemenko, G. Marcou, P. Polischuk, V. Kuz'min, A. Varnek, *Mol. Inf.* **2012**, in press?

[61] J. W. Kang, K. P. Yoo, H. Y. Kim, H. Lee, D. R. Yang, C. S. Lee, *Int. J. Thermophys.* **2001**, *22*, 487 – 494.

[62] B. G. Small, B. W. McColl, R. Allmendinger, J. Pahle, G. Lopez-Castejon, N. J. Rothwell, J. Knowles, P. Mendes, D. Brough, D. B. Kell, *Nature Chem. Biol.* **2011**, *7*, 902 – 908.

[63] Z. Lin, K. Yin, P. Shi, L. Wang, H. Yu, *Chem. Res. Toxicol.* **2003**, *16*, 1365 – 1371.

[64] D. B. Wei, L. H. Zhai, H.-Y. Hu, *SAR QSAR Environ. Res.* **2004**, *15*, 207 – 216.

[65] H. J. M. Verhaar, F. J. M. Busser, J. L. M. Hermens, *Environ. Sci. Technol.* **95**, *29*, 726 – 734.

[66] C. Hansch, A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, **1979**.

[67] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors,* Wiley-VCH, Weinheim, **2000**.

[68] http://146.107.217.178/lab/edragon/index.html.

[69] http://accelrys.com/.

[70] http://www.gaussian.com/.

[71] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 1 – 9.

[72] A. Golbraikh, A. Tropsha, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 144 – 154.

[73] P. Gedeck, B. Rohde, C. Bartels, *J. Chem. Inf. Model.* **2006**, *46*, 1924 – 1936.

[74] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, I. L. Volineckaya, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *J. Med. Chem.* **2007**, *50*, 4205 – 4213.

[75] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, E. V. Varlamova, L. G. Gorb, J. Wang, J. Leszczynski, *QSAR Comb. Sci.* **2009**, *28*, 664 – 677.

[76] E. N. Muratov, E. V. Varlamova, V. E. Kuz'min, A. G. Artemenko, L. Nikolaeva-Glomb, A. S. Galabov, *Antivir. Res.* **2010**, *86*, A62.

[77] V. E. Kuz'min, A. G. Artemenko, V. P. Lozitsky, E. N. Muratov, A. S. Fedtchouk, N. S. Dyachenko, L. N. Nosach, T. L. Gridina, L. I. Shitikova, L. M. Mudrik et al., *Acta Biochim. Polon.* **2002**, *49*, 157 – 168.

[78] A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. A. Kovdienko, A. I. Hromov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *J. Antimicrob. Chemother.* **2007**, *60*, 68 – 77.

[79] V. E. Kuz'min, A. G. Artemenko, R. N. Lozitska, A. S. Fedtchouk, V. P. Lozitsky, E. N. Muratov, A. K. Mescheriakov, *SAR QSAR Environ. Res.* **2005**, *16*, 219 – 230.

[80] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, L. G. Gorb, M. Qasim, J. Leszczynski, *Chemosphere* **2008**, *72*, 1373 – 1380.

[81] V. E. Kuz'min, A. G. Artemenko, N. A. Kovdienko, I. V. Tetko, D. J. Livingstone, *J. Mol. Model.* **2000**, *6*, 517 – 526.

[82] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785 – 789.

[83] B. Miehlich, A. Savin, H. Stoll, H. Preuss, *Chem. Phys. Lett.* **1989** *157*, 200–206.

[84] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. Tetko, G. Marcou, *Curr. Comp.-Aided Drug Des.* **2008**, *4*, 191 – 198.

[85] J. E. Riviere, J. D. Brooks, *SAR QSAR Environ. Res.* **2007**, *18*, 31 – 44.

Received: September 13, 2011
Accepted: February 4, 2012
Published online: March 13, 2012