

Published in final edited form as:

*J Med Chem.* 2014 June 26; 57(12): 4977–5010. doi:10.1021/jm4004285.

## QSAR Modeling: Where have you been? Where are you going to?

**Artem Cherkasov<sup>1</sup>, Eugene N. Muratov<sup>2,3</sup>, Denis Fourches<sup>2</sup>, Alexandre Varnek<sup>4</sup>, Igor I. Baskin<sup>5</sup>, Mark Cronin<sup>6</sup>, John Dearden<sup>6</sup>, Paola Gramatica<sup>7</sup>, Yvonne C. Martin<sup>8</sup>, Roberto Todeschini<sup>9</sup>, Viviana Consonni<sup>9</sup>, Victor E. Kuz'min<sup>3</sup>, Richard Cramer<sup>10</sup>, Romualdo Benigni<sup>11</sup>, Chihae Yang<sup>12</sup>, James Rathman<sup>12,13</sup>, Lothar Terfloth<sup>14</sup>, Johann Gasteiger<sup>14</sup>, Ann Richard<sup>15</sup>, and Alexander Tropsha<sup>2,\*</sup>**

<sup>1</sup>Vancouver Prostate Centre, University of British Columbia, Vancouver, BC, V6H3Z6, Canada

<sup>2</sup>Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA

<sup>3</sup>Department of Molecular Structure and Cheminformatics, A.V. Bogatsky Physical-Chemical Institute National Academy of Sciences of Ukraine, Odessa, 65080, Ukraine

<sup>4</sup>Department of Chemistry, L. Pasteur University of Strasbourg, Strasbourg, 67000, France

<sup>5</sup>Department of Physics, Lomonosov Moscow State University, Moscow, 119991, Russia

<sup>6</sup>School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool L33AF, UK

<sup>7</sup>Department of Structural and Functional Biology, University of Insubria, Varese, 21100, Italy

<sup>8</sup>Martin Consulting, Waukegan, IL, 60079, USA

<sup>9</sup>Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, 20126, Italy

<sup>10</sup>Tripos, Inc., St. Louis, MO, 63144, USA

<sup>11</sup>Environment and Health Department, Istituto Superiore di Sanita', Rome, 00161, Italy

<sup>12</sup>Altamira LLC, Columbus OH 43235, USA

<sup>13</sup>Department of Chemical and Biomolecular Engineering, the Ohio State University, Columbus, OH 43215, USA

<sup>14</sup>Molecular Networks GmbH, 91052 Erlangen, Germany

<sup>15</sup>National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27519, USA

### Abstract

Quantitative Structure-Activity Relationship modeling is one of the major computational tools employed in medicinal chemistry. However, throughout its entire history it has drawn both praise

\*Corresponding Author: Alexander Tropsha, Ph.D., alex\_tropsha@unc.edu; phone: +19199662955.

and criticism concerning its reliability, limitations, successes, and failures. In this paper, we discuss: (i) the development and evolution of QSAR; (ii) the current trends, unsolved problems, and pressing challenges; and (iii) several novel and emerging applications of QSAR modeling. Throughout this discussion, we provide guidelines for QSAR development, validation, and application, which are summarized in best practices for building rigorously validated and externally predictive QSAR models. We hope that this Perspective will help communications between computational and experimental chemists towards collaborative development and use of QSAR models. We also believe that the guidelines presented here will help journal editors and reviewers apply more stringent scientific standards to manuscripts reporting new QSAR studies, as well as encourage the use of high quality, validated QSARs for regulatory decision making.

---

**Where have you been?**

**Where are you going to?**

**I wanna know what's new**

**I wanna go with you.**

*Chris Rea, The Blue Café, 1998*

## Introduction

More than fifty years have passed since the field of Quantitative Structure-Activity Relationships (QSAR) modeling was founded by Corwin Hansch.<sup>1</sup> Initially conceptualized as the logical extension of physical organic chemistry, QSAR modeling has grown, diversified, and evolved from application to small series of congeneric compounds using relatively simple regression methods to the analysis of very large datasets comprising thousands of diverse molecular structures using a wide variety of statistical and machine learning techniques. More than fifty years of continuous improvements, interdisciplinary breakthroughs, and community-driven developments were needed to make QSAR one of the commonly employed approaches to modeling the physical and biological properties of chemicals in use today. In fact, the analysis of published literature indicates that the continuing growth of chemical data and databases especially in the public domain has stimulated the concurrent growth in QSAR publications (Figure 1).

QSAR modeling is widely practiced in academy, industry, and government institutions around the world. Recent observations suggest that following years of strong dominance by the structure-based methods, the value of statistically-based QSAR approaches in helping to guide lead optimization is starting to be appreciatively reconsidered by leaders of several larger CADD groups.<sup>2</sup> QSAR models find broad application for assessing potential impacts of chemicals, materials, and nanomaterials on human health and ecological systems. An area of active QSAR expansion is in the use of predictive models for regulatory purposes by government agencies, where a still growing number of specialized regulatory tools and databases are being developed and validated.

Obviously, QSAR modeling is a computational field but its major beneficiaries and the ultimate judges are medicinal chemists. Whenever computational scientists begin to address

problems within a primarily experimental scientific domain there is always a challenge of finding the proper interface and balance between computational and the respective experimental domain expertise. Of course, it is computational scientists who are expected to create computational tools that would be valuable for experimental researchers. One may then pose a question as to once the tools are created who should use them: still the "developers" who have learned enough of the domain expertise to make the application of the tools meaningful or the "users" who have learned enough of computational science to use the tools properly. We believe that generally speaking there is no definitive answer to this question as the answer depends on the level of expertise and experience of the researcher. However, we would like to point out that both fields are highly sophisticated and it will require a significant effort to accumulate high level of knowledge in both areas. In fact, very few scientists have achieved prominence both as computational and experimental medicinal chemists. We tend to think that the highest level of success will be enabled through the collaboration between computational and experimental scientists who have deep knowledge of their respective fields but have also made efforts to understand and develop working knowledge of the complimentary field.

This paper is written by a group of cheminformatics experts with a deep knowledge of the theory of QSAR modeling as well as extended experience in various types of QSAR applications, especially in designing compounds with the desired biological activities. Our objective is to provide an overview of the state-of-the-art QSAR methods and applications to a diverse community of readers, including by default mostly experimental medicinal chemists but also scientists practicing QSAR modeling. Therefore, we spend fair amount of effort to describe major concepts and methodologies in the field (as they have emerged throughout the entire history of QSAR) but then emphasize the applications of QSAR in medicinal chemistry. We describe several studies (published in *J. Med. Chem.*, *Nature Chem. Biol.*, and other high-profile journals) where cheminformaticians and medicinal chemists worked together to discover novel molecules with unique biological activities; this was achieved by developing QSAR models and employing them for virtual screening followed by experimental validation. In our opinion, these case studies were successful because of the distribution of labor where computational experts ensured the highest quality of models and experimental chemists guided by the predictions were in a position to skillfully synthesize and test compounds predicted as hits. We posit that such examples are of high value for all readers of this Journal because they prove the capabilities of computational techniques to guide chemical synthesis and biological testing when the number of possible experiments significantly exceeds technical capabilities of researchers (i.e., it is impossible to synthesize and/or test all compounds). We further suggest that both developers and users of QSAR modeling should employ or be familiar with the best practices for data curation, analysis, and modeling; the former group should follow them and the latter group should be aware of them to evaluate if they can rely on a particular model in planning their experiments. Therefore, we have endeavored to reach out to the broadest readership of this Journal and help readers to understand and appreciate the state-of-the-art practices in the QSAR field, as well as its successes, and, most importantly, its limitations.

The amount of chemogenomics data generated by experimentalists is exploding. Databases incorporating millions of compounds with associated bioactivities are available in the public

domain, whereas HTS platforms are becoming more and more common in academic structures. We believe that computational approaches are critical for accessing, querying, mining, modeling, screening such enormous ensemble of chemical data, and not only cheminformaticians can use those approaches but medicinal chemists as well. In such context, our review is an overall, recapitulative summary of what QSAR modeling was at its origin, what QSAR modeling is today, and what QSAR modeling is likely to be in the next coming years. We believe it is of high interest not only for cheminformaticians and medicinal chemists but for both computational and experimentalist communities in general.

As in any evolving field, QSAR has experienced successes, suffered failures, and responded to emerging trends. This paper aims to discuss: (i) the historical development of QSAR (Part 1) including the founding pioneers, initial concepts, and important milestones in the evolution of the field; (ii) current trends, unsolved problems, and pressing challenges (Part 2); and (iii) several novel and emerging applications of QSAR science (Part 3). Throughout this discussion, in an effort to build on past lessons learned, we provide some guidelines for use and application, and recommend best practices for developing validated and externally predictive QSAR models.

Obviously, it is not possible to address all aspects of this rich and expanding field and acknowledge all contributions made over the years by many of our outstanding colleagues. Thus, this paper should not be perceived as a comprehensive monograph covering the entire discipline of QSAR modeling. Rather, our international group of co-authors working in industry, government agencies, and academia has made an attempt to share our expertise and collective wisdom concerning some most important, in our opinion, general aspects and best practices of building, validating, and employing QSAR models using examples mostly drawn from our own research. We hope that the readers will both gain an appreciation for the challenges of developing truly rigorous and useful QSAR models (*i.e.*, "will wanna know what's new"), as well as share the excitement of the authors concerning new opportunities offered by this evergrowing research area (*i.e.*, "will wanna go" into the field!).

## 1. History and evolution of QSAR

### 1.1. In principio erat verbum, et verbum erat QSAR

**Origin of QSAR**—Many mark the founding of modern QSAR practice to the 1962 publication of Hansch *et al.*<sup>1</sup> This paper represented the culmination of a fifteen year struggle to understand the basis of the structure-activity relationships (SARs) of plant growth regulators – most of that time spent in the pursuit of a suitable Hammett relationship, the reigning methodology for explaining substituent effects on chemical reactivity. Unable to obtain a robust relationship, Hansch followed up on the arguments of Veldstra<sup>3</sup> and investigated the effect of lipophilicity on biological potency.<sup>4</sup> He wisely ignored Veldstra's complex methodology and turned to octanol-water partition coefficients<sup>5</sup> to serve as a surrogate measure of lipophilicity.

In the meantime, Fujita was employing quantum-chemical calculations to account for activity variations in plant growth regulators.<sup>6</sup> On the recommendation of Fukui, Fujita

accepted a post-doctoral fellowship at Pomona in mid-1961 and started to experimentally measure octanol-water partition coefficients (logP). Hansch and Fujita soon realized that logP was an additive property, *i.e.*, that the partial contribution of a substituent to the logP of one molecule is often the same as the contribution of that substituent to the logP of another molecule. They used the term  $\pi$  for this substituent effect on hydrophobicity.<sup>7</sup>

Of course, Hansch and Fujita did not work in a scientific vacuum: in the 1950s, the power of the Hammett equation to account for reactivity differences dominated the explanation of substituent effects. A 1953 article by Jaffe included several hundred of such equations.<sup>8</sup> In the late 1950s, Taft extended the concept of linear free energy relationships to propose and fit an equation that included not only electronic effects of substituents but also steric effects.<sup>9</sup>

In contrast, biochemical pharmacologists at the time were focusing on the effect of partition coefficients of molecules on drug absorption.<sup>10</sup> This approach can be traced back to Overton<sup>11</sup> and Meyer<sup>12</sup> some fifty years earlier and by Collander in the 1930s.<sup>13</sup> Notably, Fieser, an eminent organic chemist of the mid-1900s, showed graphically the relationship between the antimalarial potency of naphthoquinones and their ether-water distribution coefficients.<sup>14</sup> He also observed a constant optimum lipophilicity for different series of molecules.

Although Kauzmann's 1959 review article<sup>15</sup> prompted biochemists to endorse the central role of hydrophobicity in determining protein structure, with early work further emphasizing the role of partitioning to the biological target, hydrophobicity as a governing factor in the biological potency of small molecules only gradually entered the vocabulary of QSAR. Having found that the relationship between logP and biological potency was no clearer than was that between Hammett's sigma ( $\sigma$ ) and potency, Hansch and Fujita included both terms in a new equation.<sup>4</sup> The publications that followed successfully demonstrated a computational approach to modeling quantitative effects of substituents on potency.<sup>16,17</sup> Part of the attraction of the work is that the substituent effects were based on model equilibria, partition coefficients, and pKa that are easy to understand. In addition, values for these substituent effects were found to be largely transferable from one series of molecules to another. Some of the attention to the publication was that the fit involved the use of a computer, the power of which was only then becoming appreciated. Less attention in early work was paid to the power of statistics to distinguish between possible explanations.

**Evolution of QSAR**—QSAR concepts have long been used in the design of medicinal chemistry series. Craig suggested that the properties of possible substituents be plotted versus each other and that substituents be chosen to sample the full range of plotted values.<sup>18</sup> Topliss invented schemes for the stepwise exploration of a series according to the physical chemical property that governs the increase or decrease in potency of each new compound.<sup>19</sup> Hansch and Unger used cluster analysis to group possible aromatic substituents, suggesting that a good series contain one molecule from each cluster.<sup>20</sup>

In 1956, Fujita and Hansch published tables of  $\pi$  values generated from careful measurements of logP. These were useful for calculating the relative logP of members of a

series. However, comparing the optimum logP value within a series would still require the logP of the corresponding parent molecules. To address this need, Hansch and Leo used an extensive database of experimental logP values for hundreds of diverse parent molecules to parameterize a fragment-additive approach for predicting logP that was automated in the CLOGP program.<sup>21</sup> Over the years many other methods for computing logP have been devised but the fragment-additive approach remains one of the most commonly employed.

The parabolic relationship between potency and logP did not fit all datasets. Hence, Kubinyi suggested a bi-linear equation that allows for different slopes at low and high logP values.<sup>22</sup> At approximately the same time, Martin and coworkers observed the same bi-linear property of model-based equations of ionizable compounds, where both the logP and pKa vary within the series.<sup>23</sup> A key feature of these equations is that the relationship may be independent of logP at high logP values.

As the Hansch-type approach to QSAR was being established, more fundamental quantum-mechanical calculations were becoming increasingly feasible, providing an alternative means for exploring electronic and steric determinants of activity among closely related chemicals. In one of the earliest examples, the Pullmans showed how the carcinogenic potential of aromatic hydrocarbons is related to their electronic structure in predicting the potential for bay-region metabolic activation to DNA-reactive diol-epoxide intermediates. Further work in this area was limited by the computational demands of quantum chemical calculations, the greater expertise required, and the targeted nature of these methods of modeling local stereo-electronic features. Hence, the methods did not easily lend themselves to more globally applicable, higher throughput QSAR methods for evaluating diverse chemical structures.

In the early 1980's, Klopman developed an approach to break a molecule into constituent 2D fragments, to auto-generate such fragments for large numbers (hundreds to thousands) of molecules in a training set, and to attempt to correlate the frequency of each of these fragments with biological activity.<sup>24</sup> The approach was a breakthrough at the time in that it created an efficient computational method for representing and correlating easily interpretable structural features for a large number of chemicals; hence, it began to tackle the challenge of creating so-called global QSAR models for prediction of biological activity.

The treatment of steric effects of substituents had been another long-standing problem in traditional QSAR. Early investigations used Taft  $E_s$  values. Hansch and Kutter<sup>25</sup> as well as Charton<sup>26</sup> showed that  $E_s$  parameters of symmetric substituents were related to the radius of the substituent. Verloop and colleagues<sup>27</sup> took this further and calculated five shape descriptors for substituents. Hansch also frequently used molar refractivity of the substituent as a measure of its bulk. It was not until the development of CoMFA<sup>28</sup> (Comparative Molecular Field Analysis) and other 3D approaches, however, that electrostatic potential interaction energies across a series of related, superimposed structures were effectively taken into consideration. CoMFA was the first successful demonstration of a 3D QSAR technique for correlating molecular field approaches with biological activities; it was also the first commercial product that employed the partial least square (PLS) method.<sup>29</sup>

As datasets became larger and more structurally diverse, descriptors that were designed to be applied within a common reaction mechanism framework were no longer sufficient. The simplest solution was to use indicator variables to distinguish one series (*i.e.*, assumed mechanism) from another.<sup>30</sup> However, a more general solution was to generate molecular descriptors that could serve the same purpose as indicator variables by delineating series of molecules whose activity was more likely to be governed by common mechanisms: for instance, the CASE fragments<sup>31</sup> and the widely used molecular connectivity indices<sup>32</sup>. They do not discard the notion of using descriptors to directly model activity within a series, but rather attempt to capture more general determinants of activity variations both across and within series based on the principle that the entire chemical structure of a molecule dictates its properties. The work of Rosenkranz and Klopman in extending the application of CASE (now CASETOX), and later MultiCASE, to a wide diversity of toxicity modeling challenges exemplifies the power of the substructure-based approach to model previously unexamined activities and to elucidate the structural basis of activity, both across and within congeneric series of chemicals. More recently, in drawing on the more general aspects of this approach, new use has been made of descriptors that originate from the field of substructure searching; for example MDL MACCS keys,<sup>33</sup> and circular fingerprints.<sup>34</sup>

The traditional application of QSAR to a series of congeners requires that each molecule in the dataset has measurable biological activity, *i.e.*, a quantitative non-zero potency value. Discriminant analysis or a logistic regression method can be applied more generally to modeling binary or categorical responses (*e.g.*, active/inactive, mutagenic/non-mutagenic).<sup>35</sup> Recursive partitioning enables the consideration of many more descriptors and larger sets of molecules and, as a result, has become the basis for newer classification methods, such as random forest.<sup>36</sup> The support vector machines<sup>37</sup> and Bayesian classifiers<sup>38</sup> have also been successfully applied.

The field has also progressed in the development of objective methods for assessing models' reliability and prediction confidence. Statistical procedures were adopted early on to avoid chance correlations, *e.g.*, using random variable simulations, as explained in the seminal paper by Topliss and Edwards.<sup>39</sup> Although much of this early work concentrated on the fit of data to an equation, researchers now evaluate models by their ability to accurately forecast activity for test set molecules that were not used in model development.<sup>40</sup>

Quantum chemistry remains a powerful tool for exploring fundamental reactivity determinants in QSAR, as well as for calculation of *ab initio* properties (*e.g.*, dipole moment) or whole-molecule reactivity indicators, such as  $E_{\text{HOMO}}$  and  $E_{\text{LUMO}}$ . Its application at higher levels of *ab initio* theory however is typically limited to isolated systems (gas phase) and relatively few molecules. Quantum chemistry and its semi-empirical and molecular mechanics implementations are currently used in 3D QSAR approaches, such as *in silico* virtual ligand screening and profiling in drug discovery. In addition, such methods can be used to examine the relative stability of conformers, which, in turn, can influence 3D-dependent properties employed in QSAR. Mekenyan and coworkers have incorporated conformation-dependent property distributions into QSAR approaches to model toxicological endpoints, demonstrating their importance and relevance to QSAR modeling.<sup>41</sup>

In more than 50 years of active development, the field of QSAR modeling has grown tremendously with respect to the diversity of both methodologies and applications. The subsequent sections of this review provide more details concerning some recent evolutions and current trends in the QSAR field, future directions, and some innovative applications of models in multiple domains as diverse as the prediction of bioprofiles of nanomaterials or the calculation of chemical and physical properties of molecular mixtures.

## 1.2. Molecular descriptors

Chemical descriptors are at the core of QSAR modeling and so many different types of chemical descriptors reflecting various levels of chemical structure representation have been proposed so far. These levels range from molecular formula (so-called 1D) to the most popular among chemists two-dimensional structural formula (2D) to three-dimensional, conformation-dependent (3D) and even higher levels taking into account mutual orientation and time-dependent dynamics of molecules (4D and higher; for discussion see study of Polanski).<sup>42</sup> A comprehensive collection of molecular descriptors, along with their definitions, mathematical formulas, examples, and references, was published in 2009.<sup>43</sup> We briefly discuss here the most popular 2D descriptors to provide the necessary context for QSAR as a tool to predict bioactivity of compounds from their structure followed by the discussion of most popular and evolving 3D QSAR approaches.

**2D descriptors for QSAR modeling (2D QSAR)**—The two-dimensional representation of a molecule, commonly referred to as topological representation, defines the connectivity of atoms in the molecule in terms of the presence and nature of chemical bonds. Such two-dimensional representation enables the definition of so-called molecular 2D-descriptors. The main advantages of these QSAR parameters are that they (i) contain simple and useful information about the molecular structure, (ii) are invariant to molecule roto-translation, and (iii) can be calculated avoiding structure optimization. In general, 2D descriptors do not uniquely characterize molecular topology, and, as a consequence, they do not always allow the reconstruction of the molecule. Therefore, suitably defined ordered sequences of 2D descriptors can be used to characterize molecules with higher discrimination.

A molecular graph is a topological representation of a chemical compound; it is usually denoted as  $G = (V, E)$ , where  $V$  is a set of vertices which correspond to the molecule atoms and  $E$  is a set of elements representing the binary relationship between pairs of vertices; unordered vertex pairs are called edges, which correspond to bonds between atoms. When a molecular graph is obtained excluding all hydrogen atoms, it is called an H-depleted molecular graph, whereas a molecular graph where also hydrogens are included is called H-filled molecular graph (or, simply, molecular graph).

Graph-theoretical matrices are the most common mathematical tool to encode structural information provided by molecular graphs, a huge number of which were proposed in the last decades. Graph-theoretical matrices can be both vertex matrices, if both rows and columns refer to graph vertices (atoms) and matrix elements encode some property of pairs of vertices, or edge matrices, if both rows and columns refer to graph edges (bonds) and

matrix elements encode some property of pairs of edges. Off-diagonal entries of the matrix encode different information about the pairs of vertices such as their connectivity (adjacency matrix), topological distances (distance matrix), and sums of the weights of the atoms along the connecting paths (weighted matrices). The matrix diagonal entries can be equal to zero or encode chemical information about the vertices (augmented matrices); besides the most common atomic properties, also Local Vertex Invariants, which are numerical quantities derived from the molecular topology and used to characterize properties of molecule atoms (*e.g.*, vertex degree, vertex distance sum, atom eccentricity), are frequently encountered as the atomic weightings.

Graph invariants are mathematical quantities derived from a graph representation of the molecule and representing graph-theoretical properties that are preserved by isomorphism, *i.e.*, properties with identical values for isomorphic graphs. A graph invariant may be a characteristic polynomial, a sequence of numbers, or a single numerical index obtained by the application of algebraic operators to graph-theoretical matrices and whose values are independent of vertex numbering or labeling. In the last few years, several formulas and algorithms dealing with molecular graph information have been proposed and applied to different molecular matrices and various weighting schemes, leading to several new classes of related graph invariants.<sup>44</sup>

Single indices derived from a molecular graph are usually called topological indices. These are numerical quantifiers of molecular topology that are mathematically derived in a direct and unambiguous manner from the structural graph of a molecule. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching, and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. In fact, topological indices are usually divided into two categories: topostructural and topochemical ones. Topostructural indices encode only information about the adjacency and distances between atoms in the molecular structure; topochemical indices quantify information about topology but also specific chemical properties of atoms such as their chemical identity and hybridization state. Topological indices are mainly based on distances between atoms calculated by the number of intervening bonds and are thus considered through-bond indices; they differ from geometrical descriptors which are, instead, considered through-space indices because they are based on inter-atomic geometric distances.

Another important class of graph invariants is represented by the so-called autocorrelation indices. They were first introduced by Moreau-Broto<sup>45</sup> in 1980 to define a relationship between atoms as a function of their spatial separation; a review of the autocorrelation descriptors is given by Consonni and Todeschini.<sup>46</sup> The most common autocorrelation descriptors can be obtained by taking the molecule atoms as the set of discrete points in space and an atomic property as the function evaluated at those points. The autocorrelation descriptor is then the integration of the products of the function calculated at atom  $x$  and atom  $x+k$ , where  $k$  is the lag, *i.e.*, the topological distance. This descriptor expresses how numerical values of the function at intervals equal to the lag are correlated.

Randić suggested<sup>47</sup> a list of attributes for topological indices that should: 1) have structural interpretation, 2) have good correlation with at least one property, 3) preferably discriminate among isomers, 4) can be applied to local structure, 5) be preferably independent, 6) be simple, 7) not be based on experimental properties, 8) not be trivially related to other descriptors, 9) be possible to construct efficiently, 10) use familiar structural concepts, 11) have the correct size dependence, 12) change gradually with gradual change in structures. Most of the topological descriptors have the above characteristics, which is why they have been prolifically applied in characterizing the structural similarity/dissimilarity of molecules and in QSAR/QSPR modeling.

**3D descriptors for QSAR modeling (3D QSAR)**—Taken literally, many QSAR expressions suggest that biological selectivity results from each target forming highly specific interactions such as hydrogen bonds with a ligand. However it also seems that ligands binding preferences emerge primarily from non-covalent field effects exerted in the spatial vicinity of those ligands. Thus, systematic sampling of those field differences should yield molecular descriptors particularly well-suited for QSAR. This was the vision that motivated the creation of 3D QSAR, in its original and still predominant CoMFA formulation.<sup>28</sup>

The most important challenge related for CoMFA implementation was the conflict between the thousands of molecular descriptors needed to sample a ligand field and the few biological responses, almost a heretical situation when contrasted with the familiar precepts of good QSAR practice, as detailed elsewhere in this review. Thus the critical event in CoMFA's development was a private discussion at the 1982 QSAR Gordon Conference, where Wold first expounded the PLS approach to Cramer.<sup>29</sup> More than ten thousand references to "3D QSAR" and/or "CoMFA", which a Google Scholar search today produces, including consistent reports of successful potency predictions suggests that PLS has indeed been useful in describing biological differences as effects of ligand field differences.<sup>48,49</sup> However, the contrasts between 3D QSAR methodology and those QSAR precepts do remain: in addition to that sacrilegious descriptor-to-structure count ratio, a CoMFA model requires high collinearity of ligand field variations for its robustness, whereas auto-scaling 3D QSAR's molecular descriptors can prevent successful model generation.

The biggest challenge in performing CoMFA is related to the alignment protocol of training- and test-set ligands, as well as the selection of both conformation and orientation of each ligand to be included in the QSAR model. In practice, this task can often become a slow, tedious, and somewhat *ad hoc* pursuit of higher statistical criteria (*e.g.*,  $q^2$  value). However, considering many uncertainties (including the biological potencies being fit), the coarseness of the ligand field sampling, the structural representativeness of any training set selection, as well as unavoidable subjectivity of individual ligand alignments, such emphasis on  $q^2$  as a metric of comparative model quality (as opposed to a threshold of model acceptability) is inappropriate. It should instead be appreciated that every statistically acceptable QSAR model, resulting from the systematic variation of the training set composition (and alignment), represents a valid alternative interpretation, and that exploring any differences among such models may improve the understanding of the causative effects of field differences. Of course, the potential benefit of such exploration also depends upon the cost

of generating each individual model, favoring automatic and robust alignment methodologies.

The requirements of 3D QSAR modeling to operate on 3D descriptors and PLS methodology contributed to the availability of its methodological extensions dependent on specialized software vendors. Thus, CoMSIA,<sup>50</sup> which extends ligand field varieties from CoMFA's steric and electrostatic to hydrogen-bonding and hydrophobic effects, is available from Tripos. Use of pharmacophoric constraints to facilitate 3D QSAR considering multiple conformations, based on the original molecular field generating software GRID,<sup>51</sup> is provided by a family of programs from Molecular Discovery.<sup>52,53</sup> Cresset software employs extrema in ligand fields as guides in ligand alignment.<sup>54</sup> Schrodinger package offers PHASE as its 3D QSAR capability.<sup>55</sup> Two approaches that use a complementary biological target field to refine ligand-based 3D QSAR models are COMBINE<sup>56</sup> and AFMoC<sup>57</sup> programs; two other pioneering 3D QSAR methodologies, HASSLE<sup>58</sup> and MSTD<sup>59</sup> also deserve mentioning. Finally, though more "pseudo-receptor" –like rather than 3D QSAR approach (the latest incarnation of the pioneering COMPASS 3D QSAR methodology) QMOD,<sup>60</sup> shows noteworthy promise.

Notably, none of these innovations, however useful, have greatly alleviated 3D QSAR's ligand-alignment bottleneck. Indeed, as our understanding of the physics of ligand-target interaction improves, the criteria for physicochemically meritorious alignments become less certain. Even a target-bound ligand structure, which 3D QSAR cannot compete with as a source for structural inspiration, is undermined as the "gold standard" for 3D QSAR by an increasing awareness that the dynamic ligand interactions that may be critical for selective *in vivo* activity can be invisible in the non-physiologically static crystalline environment.

In considering and developing any methodology we should not lose sight of fundamental basis of any QSAR investigation: the only possible cause of any difference in biological activity of two structures is their structural difference – regardless of how complex may be the physicochemical and biological interactions that subsequently connect that structural difference to an observed biological effect. And the basis for 3D QSAR is that this causative difference in ligand structure is best expressed, primarily and initially, as a causative difference in ligand fields. These considerations suggest a different goal in ligand alignment for the purpose of 3D QSAR: to minimize the relative importance of grid locations distant from any ligand structural difference and therefore non-causative, by aligning training and test set ligands in a way that maximizes the steric overlap of their structurally identical moieties.

Accordingly, two new methods have emerged for the 3D QSAR alignments, the topomer protocol<sup>61</sup> and still evolving "template" protocol.<sup>62</sup> Both approaches are extremely facile, in effect converting 3D QSAR from one of the most tedious and therefore costly CADD approaches into one of the easiest, on the verge of becoming almost completely automatable. More importantly, the reported predictive performances of the topomer protocol in discovery projects were uniformly encouraging. The topomer similarity has consistently forecasted biological similarity, for example in published "lead-hopping"<sup>63</sup> and off-target<sup>64</sup> applications. And the standard error of pIC<sub>50</sub> taken from 144 predictions followed by

synthesis and testing in four different discovery organizations, was reported with extraordinarily low 0.6 value (or, expressed as an error in predicted potency ratios, 4x).<sup>65–67</sup>

The practice of 3D QSAR is inherently limited to local models (as herein defined elsewhere). However, it can be expected, that with the latest explosive expansion of public databases such as ChEMBL and PubChem and with further evolution of alignment protocols, that limitation will slowly recede.<sup>68</sup> One encouraging indicator of that is the use of 3D QSAR (topomer CoMFA) models derived entirely from PubChem SAR data to successfully overcome the cytochrome P450 (CYP) liability in drug design.<sup>66</sup>

### 1.3. Challenges in QSAR modeling

Since the seminal paper by Hansch and co-workers,<sup>1</sup> great strides have been made in successful application of QSAR modeling as well as in the development of QSAR methodology itself, reflected in numerous articles published in the last decade.<sup>69–73</sup> Such efforts in developing new methods have provided guidance on recommended procedures to be followed for optimal results. Unfortunately, despite these efforts, prediction errors due to poor application of statistical methods and recommended guidelines perpetuate in the development and use of QSARs. In that regard, Dearden et al.<sup>71</sup> in 2009 reported and discussed a total of 21 types of such problems; these are briefly reiterated here (for supporting references to the examples below, see).

**Failure to take account of data heterogeneity**—Sometimes *in vivo* data from two or more different species (or obtained with different protocols) are used in the same QSAR model. It may be that it is necessary to obtain an adequately-sized dataset, but it potentially compromises the integrity of the model and should be avoided if at all possible. It may seem that such considerations do not apply to the QSPR (Quantitative Structure-Property Relationship) modeling of physical-chemical properties, but this is not necessarily the case. For example, aqueous solubility can be determined in pure water, or as un-dissociated species (intrinsic solubility), or at specified pH. Another example - the flash point values, which are dependent on the size of the sample, the heating rate, the use of open- or closed sample cup, the presence or absence of stirring, and the energy and type of ignition source (*e.g.*, spark or flame).

**Use of inappropriate endpoint units**—It is often not recognized that for most QSAR purposes related to molecular activities, the endpoint format should be expressed in molar units (*e.g.*, mol.kg<sup>-1</sup>), not weight units (*e.g.*, mg.kg<sup>-1</sup>); otherwise strict molecule-to-molecule activity comparisons are not possible. If, however, data are reported as, say, percentage response at a fixed dose of 10 mg.kg<sup>-1</sup>, then such a conversion is not possible, and the data may not be appropriate for QSAR interpretation.

**Use of confounded descriptors**—If two descriptors in a given QSAR solution are highly collinear, they contribute the same information twice, thus confounding the statistical association and making it more difficult to deduce a mechanistic interpretation of the QSAR. Thus, Dearden et al. showed that, although two highly correlated topological descriptors correlated well separately, and with positive coefficients, the toxicity of alkyl ethers to

mouse, when the two descriptors were used together, one had a negative coefficient. There is no definitive value above which collinearity is unacceptable, but Dearden et al.<sup>71</sup> recommend that  $r^2$  values of  $>0.8$  be carefully scrutinized.

**Use of non-interpretable descriptors**—One of main values of a QSAR model is possible insight into mechanism of action under study, although it must always be recognized that a correlation does not provide any guarantee of causality. Currently, there are thousands of molecular descriptors available for QSAR purposes, and it is difficult to discern a clear physical-chemical interpretation for many of them. Sometimes descriptors involved in reported QSAR models are not clearly defined or identified, and often no mechanistic interpretation is given. Modelers should employ descriptors that are reasonably interpretable, particularly when the aim of a study is to yield insight into mechanisms of action or where such association can improve model's acceptance and use.

**Errors in descriptor values**—With few exceptions (*e.g.*, atom counts and molecular weight, provided that they are correctly calculated), descriptor values, whether measured or calculated, can contain errors. For example, for a simple organic chemical, 4-nitrophenol, seven published measured logP values range from 0.76 to 2.08, while seven logP values calculated from available software (VCCLAB, [www.vcclab.org](http://www.vcclab.org)) range from 1.35 to 1.93, with the generally accepted "best" measured value being 1.91. In cases where experimental values are debatable, scientific practice requires reporting error bars or uncertainties in modeled values where these are known or can be estimated. In other cases, blatant mistakes in structure representations and resultant computed properties provide yet additional sources of model errors.

**Poor transferability of QSARs**—It is necessary that any published QSAR model can be independently validated and used by others for predictive purposes. Transferability of QSAR is the equivalent of "lateral validation", a concept pioneered by Hansch et al.<sup>74</sup> Later, Hartung et al.<sup>75</sup> suggested five criteria for transferability, the prime one being that descriptor values can be reproduced. Often, however, those criteria are not satisfied either due to inadequate documentation or lack of readily available descriptor-generating software.

**Inadequate or undefined applicability domain**—The applicability domain (AD) of a QSAR model has been defined as "the response and chemical structure space in which the model makes predictions with a given reliability".<sup>76</sup> Essentially this means that a test chemical must be reasonably similar to some training-set chemicals, or valid prediction cannot be expected. However, very few QSAR publications actually provide sufficient information (*e.g.*, training set structures and descriptor values) for the assessment of model applicability domain.

**Unacknowledged omission of data points**—Many QSAR data sources originate from the literature, and often only a selected number of data points are used in the model. However, it is sometimes the case that data are pruned without explanation and this can lead to the suspicion that pruning (or removal of outliers) was done to improve the model. Of course, outliers can occur in any data set, and it is acceptable to remove them, provided that a good explanation, independent of the modeled result, can be given for their removal.<sup>70</sup>

**Use of inadequate data**—Data inadequacy includes matters already mentioned, such as heterogeneity, use of inappropriate units, lack of information on the AD, and unacknowledged omission of data points. It can also include the use of incorrect, misspelled, or insufficiently defined chemical names, incorrect CAS numbers, and incorrect structures. For instance, a QSAR study of skin absorption<sup>77</sup> listed chloroxylenol and 4-chlorocresol among the chemicals examined; chloroxylenol has 18 structural isomers and 4-chlorocresol has two. Young et al.<sup>78</sup> found that incorrect structures in six public and private databases ranged from 0.1% to 3.4%.

**Replication of chemicals in a data set**—It is unfortunately quite common for chemicals to be replicated in the training and the test sets.<sup>79</sup> Such co-occurrence can falsely improve the apparent predictive power of the QSAR solution. Replication can occur because of different names, CAS numbers, or structure codes for the same chemical, or because of different activity or property values for the same chemical. Oftentimes, however, replication occurs upon indiscriminate "desalting" of a structure file prior to descriptor generation, after which the parent and salt (with the same or different activities) map to the same structure. It is therefore essential that datasets are carefully checked and curated (including merging or removal of duplicates) before use. More details on chemical data curation are provided in Section 2.1.

**Narrow range of endpoint values**—The range of endpoint values of a QSAR training set should be significantly greater than the experimental error in the values. The experimental error among *in vivo* data can often exceed half a log unit; as a rule of thumb, Geddeck et al.<sup>80</sup> recommend an endpoint value range of at least 1.0 log unit to obtain a good QSAR model. Of course, it is not always possible to achieve such a wide range of endpoint values, either through paucity of data or because of the nature of the endpoint (*e.g.*, melting points of a given class of chemicals). In such cases, closer consideration of the data and model performance statistics, including external validation, are required.

**Over-fitting of data**—The well-known "Topliss and Costello rule"<sup>19</sup> states that, to minimize the risk of chance correlations, the ratio of training set chemicals to descriptors should be at least 5:1 when using simple linear regression methods. This rule, as well as the standard requirements for basic statistical practices, is still widely broken. A glaring example is provided by the modeling of the aquatic toxicities of 12 aliphatic alcohols with 9 descriptors.<sup>81</sup> With the use of non-linear multivariate techniques, such as artificial neural networks, over-training and over-fitting can be a problem if care is not taken, although rigorous statistical tests are available allowing for the proper processing of cases where the number of descriptors far exceeds the number of chemicals.

**Use of excessive numbers of descriptors in a QSAR**—Even if the "Topliss and Costello rule" is not broken, use of a large numbers of descriptors in a QSAR can make interpretation and explanation the model more difficult. Generally speaking, the principle of "Occam's Razor" is completely applicable to QSAR, *i.e.*, one seeks a QSAR with the smallest number of descriptors that yield a reasonable model. Often, a small number of simplest molecular descriptors affords a model that outperforms significantly more complex

ones. For instance, Oprea<sup>82</sup> reported that the length of the molecule gave the best correlation with fiber affinity; this simple model even outperformed CoMFA.

**Inadequate or missing statistic measures**—The statistical measures are used to indicate the goodness-of-fit and predictivity of a QSAR, and so are vital for assessing its validity. However, even now QSARs are reported without any statistics (*e.g.*, see the study of Ghasemi and Saaidpour),<sup>83</sup> and many still appear with inadequate statistics, which makes the assessment of model's validity difficult. As well as the correlation ( $R$ ) and determination ( $R^2$ ) coefficients, and standard error of the estimate(s), it is useful to have the adjusted determination coefficient ( $R^2_{adj}$ ), which allows for comparison between QSARs with different numbers of descriptors and can indicate when a given QSAR model incorporates too many descriptors. In addition, the internally cross-validated  $R^2$  ( $Q^2$ ) and the Fisher statistic or variance ratio provide an indication of a chance correlation. In addition, the regression coefficients of each of the descriptors, although rarely reported, are valuable for indicating whether a particular descriptor contributes significantly to a linear regression. The probability that the descriptor is there by chance, should generally be less than 0.05 (*i.e.*, 5%) to be considered statistically significant; otherwise this descriptor should be discarded.

**Incorrect calculation**—Editors and reviewers usually assume that all calculations reported in a manuscript have been made correctly, and it is often impossible to check otherwise. But nonetheless incorrect published QSARs have been identified to date, and it is difficult to assess how widespread this problem actually is. Dearden et al.<sup>71</sup> reported two such instances, and there are probably many more.

**Lack of descriptor auto-scaling**—Descriptor values often cover widely different numerical ranges, which necessitates the use of auto-scaling methods. When no scaling is employed, it is difficult to determine the relative contribution of each descriptor to the QSAR and those descriptors with large numerical values can dominate the model compromising its statistical validity. Many published models do not employ auto-scaling, but its use is highly recommended.

**Misuse or misrepresentation of statistics**—Even if QSAR practitioners are not statisticians, the basic rules of good statistical practice should be used by all and should be enforced by reviewers, editors, and publishers. The study of aquatic toxicity of alcohols<sup>81</sup> has already been mentioned, where unjustified incorporation of additional descriptors did result in significant model improvement. Yaffe et al.<sup>84</sup> used fuzzy ARTMAP statistics to obtain a standard error of prediction of 0.08 log unit in their QSPR study of aqueous solubility. However, the experimental error in aqueous solubility measurements is estimated to be 0.6 log unit, and hence, the reported QSAR with a prediction error lower than the error of measurement is most likely a result of over-fitting.

**No consideration of distribution of residuals**—QSAR model predictions can contain two types of errors - random and systematic. Random error is an indication of the irreproducibility in the response data and/or the descriptor values, and can be reduced by careful selection of these properties. Systematic errors usually result from the bias in measurement or calculation, and can be identified by a simple plot of residuals against

measured response values. If systematic error is absent, the residuals should exhibit random distribution around zero line. If the plot shows a marked bias to one side of the zero line, or shows a regular variation of residuals with measured response values, the systematic error is present, and should be eliminated if possible. It would be useful to have residual plots reported or included in QSAR publications.

**Inadequate training and/or test set selection**—For best results, training set data should be well-distributed over the full range of endpoint values. This is often not possible, for various reasons, but very poor distribution, such as two clusters of chemicals, or one or two chemicals far removed from the others, will exert strong model leverage and must be avoided. Adequate distribution of properties and endpoint values within the test set is also crucial. Test set chemicals must be reasonably similar to some of the training set chemicals, and yet too great similarity can give an overly optimistic indication of a QSAR's predictive capability.<sup>85</sup>

**Inadequate QSAR model validation**—It is now widely accepted that to rigorously assess the predictivity of a QSAR model, some external validation is required, *i.e.*, to use the QSAR model to predict endpoint values for an external test set of chemicals that were not included in model training.<sup>86</sup> Tropsha and Golbraikh<sup>87</sup> recommended that the process of training and test set selection and external validation should be carried out a number of times so as to identify the ranges of external predictivity of a model. Perhaps even more stringent approach to external validation should be based on a "time-split" selection as advocated in a recent study by Sheridan.<sup>88</sup>

**Lack of mechanistic interpretation**—It is not always possible to provide a mechanistic interpretation of a QSAR model. Furthermore, it should be borne in mind that the existence of even a very good correlation does not imply causality. Nevertheless, mechanistic interpretations are often helpful, for example in guiding future synthesis of drug candidates. An report of the Organization for Economic Cooperation and Development (OECD)<sup>89</sup> recommended that the following questions to be asked about possible mechanistic basis of a QSAR model: (i) Do the descriptors have any physicochemical interpretation that is consistent with a known mechanism? (ii) Can any literature references be cited in support of the purported mechanistic basis of the developed QSAR? If the responses to both questions are positive, one may have some confidence in the proposed mechanism of action.

To summarize, a rich history of QSAR calls for the proper use of well-established statistical practices and "best practice" rules unifying the standards of data processing and model interpretation, and aiming to avoid the above-described common mistakes and missteps.

## 2. Current trends in QSAR methodology

### 2.1 Chemical data curation

One of fundamental assumptions of any QSAR or cheminformatics study is the correctness of input data generated by experimental scientists. As obvious as it seems, the presence of incorrect or imprecise data in modeling sets is increasingly considered a major concern for building computational models, particularly where the activity signal is sparse or potency

variation limited, and a QSAR pattern is not easily discerned. A recent study<sup>90</sup> demonstrated that on average, there are two chemical annotation errors per each medicinal chemistry publication, with an overall error rate for the compounds indexed in the popular commercial WOMBAT database<sup>91</sup> being as high as 8%. In another recent study,<sup>78</sup> authors investigated several public and commercial databases and reported error rates in chemical structure annotation ranging from 0.1 to 3.4%. Two main types of errors in input data can be considered: (i) directly related to chemical structures; and (ii) related to associated experimental measurements.

Recent publications<sup>78,92,93</sup> clearly pointed out the importance of chemical data curation in the context of QSAR modeling. They suggest that having erroneous structures represented by erroneous descriptors could have a detrimental effect on models' performance. Thus, the authors demonstrated that rigorous manual curation of structural data, and elimination of questionable data points, often leads to substantial increase in model predictivity. This conclusion becomes especially important in light of the studies of Olah et al.<sup>90</sup> and Tiikkainen and Franke,<sup>94</sup> emphasizing the significant error rate in medicinal chemistry articles and bioactivity databases respectively.

Surprisingly, there are very few published reports describing how the primary data quality influences the performances of QSAR models. Beyond the calls on the importance of data (mostly chemical) curation discussed by Williams and Ekins,<sup>95</sup> only the studies conducted by Young et al.,<sup>78</sup> Southan et al.,<sup>96</sup> and Fourches et al.<sup>92</sup> described the compendium of issues directly related to chemical data curation. Fourches et al.<sup>92</sup> developed a guideline of best practices for preparing chemical data prior to any other stage of the predictive QSAR modeling workflow (see Figure 2). Organized into a solid functional process, different curation steps (see Figure 2B) allow both the identification and correction of structural errors, sometimes at the expense of removing incomplete or confusing data records. They include the removal of inorganics, organometallics, counterions, and mixtures that most QSAR descriptor generation programs are ill-equipped to handle or that lead to confounding duplicates when simplified (*e.g.*, desalted). Additional curation elements include structural cleaning (*e.g.*, detection of valence violations), ring aromatization, normalization of specific chemotypes, and standardization of tautomeric forms. Post processing entails deletion of duplicates resulting from curation, standardization and normalization, and manual checking of complex cases.

**Removal of mixtures**—Treatment of mixtures is not a simple computational issue and various situations are encountered in the workflow: (i) the mixture contains a large organic compound and several smaller moieties, either organic or inorganic (*e.g.*, complexes or non-stoichiometric hydration) – if it is reasonable to assume that the experimentally determined biological activity associated with the record is directly and only caused by the largest molecule (and is not affected by the smaller components), the largest compound can be kept and the smaller moieties should be deleted; (ii) several similar organic compounds with close molecular weights, such as in the case of mixtures of stereo or geometric isomers: usually, the deletion of the entire record is recommended (unless the active ingredient or isomer is known and can be selected manually), because it is not possible to determine which component has to be retained for modeling using simple rules and automated

software; manual intervention is usually required and a representative isomer may be chosen; and (iii) a formulation or impure compound where the active or major ingredient is known, but the "inert" or impurity ingredients are unknown – these cases usually call for a judgment call.

**Normalization of specific chemotypes**—Very often the same functional group may be represented by different structural patterns in the same dataset. For example, nitro groups have multiple mesomers and, thus, can be represented using two double bonds between nitrogen and oxygens in their neutral forms (*i.e.*, a hypervalent N state), or a single bond linking the nitrogen and the protonated oxygen, or linking both nitrogen and oxygen atoms that are oppositely charged. For QSAR modelers, these situations may lead to inconsistent modeling outcomes depending on how descriptor-generation programs process these cases. Similarly, although ring aromatization and the normalization of carboxyl, nitro, and sulfonyl groups are relatively obvious, more complex cases like anionic heterocycles, quaternary ammonium ions, poly-zwitterions, tautomers, etc., require a deeper analysis and multiple normalization steps.

**Removal of duplicates**—Rigorous statistical analysis of any dataset assumes that each compound is unique. However, structural duplicates are often present, especially in large datasets. As a result, QSAR models built on such collections may have artificially skewed predictivity. Hence, duplicates must be pre-processed and removed prior to any modeling efforts. Once duplicates are identified, the analysis of their associated properties is mandatory, requiring some manual curation. For a given pair of duplicates, if their experimental properties are identical, one should be straightforwardly erased. However, if their experimental properties are different, there are several scenarios to consider: (i) the property value may be wrong for one compound due to, *e.g.*, a human error when the database was built; (ii) both experimental properties are correct but the previous curation tasks (for example, the removal of counterions in salts) have modified the substance records to create such duplicates; and (iii) there are experimental replicates in the dataset with different reported property values. All three cases will require some additional scrutiny and evaluation to determine the best course of action. In the case where one value is known of suspected of being in error, rejecting the entry is the obvious course; where desalting led to a duplicate, the property associated with the original salt form (as opposed to the unmodified parent) should be deleted; and in the case of experimental replicates, results must be appropriately averaged or aggregated to produce a single result.

**Final manual checking**—The last step of data curation entails the manual inspection of complex molecular structures. Common errors identified during the manual cleaning procedure may have different origins: (i) the structure is wrong – a rapid check of both the compound's IUPAC name (if available) and its structure is essential to identify possible errors concerning the scaffold and positions of substituents (*e.g.*, due to manual errors or program bugs<sup>78</sup> in the conversion of SMILES into 2D structures); (ii) the normalization of bonds is incomplete – common mistakes are related to the presence of different representations of the same functional groups, *i.e.*, despite the normalization procedure, some very specific cases can still be present and thus, the corresponding chemotypes must

be corrected manually; (iii) some duplicates may still be present despite the use of automated software to remove them, for instance, some tautomers can still be found; and (iv) other errors – wrong charges, presence of explicit hydrogens in a hydrogen depleted structure, incorrect bonds, etc.

Some general rules following the curation workflow were also formulated: (i) it is risky to calculate chemical descriptors directly from SMILES, whereas it is preferable to compute descriptors (integral, fragments, etc.) from curated 2D (or 3D if necessary) chemical structures where all chemotypes are strictly normalized; (ii) structural comparison across available databases may facilitate the detection of incorrect structures; (iii) even small differences in functional group representations can lead to significant errors in models; (iv) locating and processing of structural duplicates is one of the mandatory steps in QSAR analysis, although such searches based on chemical name or CAS number only are incomplete and inefficient; and (v) nothing can replace hands-on participation in the process since some errors obvious for chemists are still not obvious for computers.

Given the clear importance of data curation prior to QSAR modeling, we recommend an additional principle for QSAR model validation, stating that "to ensure the consideration of QSAR models for regulatory purposes, the chemical datasets used to train and validate these models must be thoroughly curated with respect to both chemical structure and associated target property values."

## 2.2. QSAR in toxicity prediction

**Special challenges**—Applications of structure-activity relationships (SAR) to modeling and predicting toxicity endpoints are not fundamentally different from those used in other fields and employ almost every existing SAR approach, ranging from Structural Alerts, to SAR heuristics (expert judgment), to QSAR for congeneric and non-congeneric sets, to combinations of models (consensus).<sup>97–100</sup>

Toxicity prediction, however, also poses special challenges. Applications of QSAR modeling to fields such as drug discovery are usually focused on sifting through large numbers of potential drug candidates for compounds that are active at a well characterized enzyme target, where some knowledge of the target interaction and the chemical space of known ligands constrains the search. In the field of toxicology, QSAR methods are typically applied towards the more elusive goal of predicting potential toxicity outcomes for *in vitro* cell cultures or *in vivo* animal test systems, where the toxicity endpoint (*e.g.*, mutagenicity, developmental toxicity, cancer) tends to be less well understood, and is likely to encompass multiple mechanisms and pathways to adverse outcome, *i.e.*, where no single interaction mechanism is known or anticipated. Other significant challenges pertain to the chemical knowledge base used for model building (*i.e.*, the training set) and the chemical space to which models will be applied (*i.e.*, the prediction space) and for what aim (mechanism elucidation, screening, prioritization, safety assessment, etc.).

In contrast to the design of drugs and pesticides, where a chemical-activity space of interest is usually populated to serve as a training set for model building, a researcher in toxicity modeling is most often constrained to work with whatever limited data are publicly

available, with the goal of predicting whether a chemical is potentially harmful. In particular, within a regulatory or safety assessment workflow, where exposure of humans or ecosystems to each of hundreds to thousands of diverse chemical compounds is a distinct possibility, there is not only greater weight placed on individual QSAR predictions, but regulatory action most often requires a greater body of supporting evidence accompanying each prediction.<sup>101,102</sup>

It is generally accepted that QSAR success in modeling toxicity is more likely when one or more of the following conditions are met: (i) compounds within the training set are structurally similar (*i.e.*, congeneric), implying that a single target-mediated mechanism, even if unknown, is more likely; (ii) the toxicity endpoint being modeled is either non-target specific (*e.g.*, narcosis in aquatic toxicity due to membrane concentration effects), or subject to relatively well-understood chemical reactivity principles (*e.g.*, electrophilic theory of carcinogenicity); (iii) the toxicity endpoint is linked to a well-defined molecular target (*e.g.*, estrogen receptor) or phenotype (*e.g.*, cleft palate malformation, or liver tumors in rats); or (iv) toxicity data are available for a sufficiently large number of diverse chemicals to capture all or most of the possible structure-activity associations, representing multiple possible adverse outcome pathways within the same dataset (*e.g.*, genotoxicity). Hence, a defining challenge for QSAR applied to toxicology is that of balancing the highest possible endpoint resolution with the need for sufficient statistical representation, with the latter closely tied to the number of chemical-activity pairs in the training set. To meet this challenge, toxicity endpoints have sometimes been aggregated to what toxicologists may consider biologically meaningless extremes (*e.g.*, lumping all possible developmental malformations and outcomes, such as cleft palate and fetal death, to one endpoint, "developmental toxicity") to yield the largest possible chemical training set. On the other hand, data for organ or species-specific toxicity phenotypes (*e.g.*, mouse liver tumors, rat cleft palate, etc.) tend to be available for far fewer compounds, resulting in insufficiently robust QSAR models.

In practice, the use of prior knowledge of biological or chemical mechanisms in guiding and constraining a QSAR modeling study, or the use of *in vitro* test data in conjunction with structural features and properties have proven to be critical for overcoming the perennial challenge of "not enough data". Examples include skin sensitization QSAR models built on clear mechanistic and chemical reactivity principles,<sup>103</sup> and a recent proposal of Benigni<sup>104</sup> to strategically combine the use of well-established structural alerts with the results of *in vitro* mutagenesis and cell transformation assays for the prediction of genotoxic and non-genotoxic chemical carcinogens.

**QSAR and computational toxicology**—A number of advances and new initiatives in the growing field of "computational toxicology" have the potential to move QSAR approaches beyond current limitations, as well as to extend models into areas of toxicology previously considered as intractable (*e.g.*, reproductive toxicity). Notable progress in computer technologies, computational chemistry and cheminformatics, as well as increasingly sophisticated statistical and machine-learning approaches, have fueled much of the methodological advancements in this field. Equally, if not more important, however, have been major initiatives on the toxicity data side of the equation, both in the better capture, representation, and utilization of existing toxicity data, and in the generation of new

data.<sup>102–104</sup> There have been great strides in the chemical structure annotation, curation, and ontological organization of *in vitro* and *in vivo* public toxicity datasets to serve as data-mining resources, as well as for use in training and validating QSAR models.<sup>105,106</sup>

An example of a highly curated public toxicity reference database, capturing many levels of resolution of *in vivo* toxicology, is the U.S. Environmental Protection Agency's (EPA) ToxRefDB.<sup>105</sup> ToxRefDB is publicly available within EPA's Aggregated Computational Toxicology Resource (ACToR),<sup>106</sup> the latter focused on the larger goal of aggregating publicly available chemical toxicity and bioactivity data (or linkages to such data) into a central hub for supporting computational toxicology research.

Examples of public resources that focus more specifically on accurate chemical structure annotation and the construction of summary toxicity endpoint data for use in QSAR development are EPA's Distributed Structure-Searchable Toxicity (DSSTox) project,<sup>107</sup> the Istituto Superiore di Sanita's ISSTOX project,<sup>108</sup> eTOX project,<sup>109</sup> and DrugMatrix molecular toxicology reference database<sup>110</sup> from NTP. In addition, with increasing regulatory pressures to reduce reliance on animal testing, particularly in Europe, there is a proliferation of publicly available, on-line or downloadable QSAR resources. Examples include ToxTree,<sup>111</sup> OpenTox,<sup>112</sup> eTOX Project ([www.etoxproject.eu/](http://www.etoxproject.eu/)), DrugMatrix from EPA (<https://ntp.niehs.nih.gov/drugmatrix/index.html>), and the OECD QSAR Toolbox.<sup>113</sup>

On the toxicity data generation side of the equation, EPA's ToxCast program<sup>114</sup> and the multi-Agency Tox21 program (a collaboration between EPA, the National Toxicology Program, the National Institutes of Health's Chemical Genomics Center – NCGC, and the US Food and Drug Administration – FDA)<sup>115</sup> are employing quantitative high-throughput screening (qHTS) approaches to test thousands of environmental and commercial chemicals against tens to hundreds of biological assays potentially relevant to, and informative of, toxicity pathways. A primary objective of these programs is to use qHTS results in conjunction with toxicity databases and knowledge bases pertaining to *in vivo* toxicity to build pathway-based models relating *in vitro* results to *in vivo* biology.<sup>116</sup> Curated chemical structure inventories are publicly available for these testing programs through the DSSTox website,<sup>107</sup> and all qHTS results for ToxCast are being made publicly available through the EPA ACToR website.<sup>106</sup>

To date, the QSAR community has had limited engagement with these data. A few investigators have employed machine learning and a wide range of traditional QSAR statistical methods to analyze the ToxCast Phase I data (consisting of results for approximately 300 compounds, mostly pesticides, tested in >500 qHTS assays) either alone (as "biological descriptors") or in combination with chemical structure descriptors to model selected *in vivo* endpoints represented within ToxRefDB.<sup>117,118</sup> These purely statistical efforts applied to a complex toxicity data set have been, for the most part, unsuccessful. The latter is not surprising after taking into account the structural diversity and limited size of this Phase I data set, the noise within the qHTS data, the relatively low hit rate across qHTS assays (generally 10% or less), and the relatively low incidence of *in vivo* positives for the modeled phenotypic endpoints. However, when rational data constraints (*e.g.*, rejecting incomplete experimental protocols), prior biological knowledge (*e.g.*, literature-reported

results implicating particular gene targets), and pathway-based hypotheses were employed, several studies have successfully demonstrated significant *in vitro* to *in vivo* correlations, most remarkably for reproductive toxicity.<sup>119,120</sup> In other studies, Tropsha and co-workers have demonstrated that qHTS data, when used as biological descriptors (without regard to biological relevance to the modeled endpoint), added significant information content beyond what could be achieved with chemical descriptors alone and improved overall model performance for predicting *in vivo* toxicity.<sup>121,122</sup>

What seems clear from these results is that QSAR approaches can potentially benefit from the new information content contained within qHTS data, information that extends beyond purely chemical structure analogy and into the biological realm, but that some prior knowledge, biological (or chemical) mechanism considerations and hypotheses are needed to guide QSAR modeling efforts into productive areas.

In addition, although qHTS *in vitro* to *in vivo* models have met with some initial success, they have thus far failed to integrate QSAR approaches that could potentially guide the development and improve models' performance. Greater opportunities will present themselves with the expansion to nearly 1,000 chemicals in ToxCast Phase II<sup>119</sup> and with qHTS screening of the larger Tox21 library, consisting of more than 8,000 diverse structures across 50–100 selected assays being run at the NCGC. These new qHTS data and computational toxicology initiatives represent an area of open possibility and challenge for QSAR, to better integrate with biologically based models and to extend its reach in chemical space and in modeling toxicity at more refined levels of biological organization.

**A "QSAR21" approach using MoA QSAR**—Large *in vitro* and *in vivo* data sets, such as being generated within Tox21, probe diverse biological pathways to reveal assay-endpoint signatures. However, when focusing exclusively on the biological aspects, computational toxicology modelers are in danger of making the same mistakes that chemists made in the early days of QSAR: focusing too narrowly on the chemical side, while reducing complex biological phenomena into overly-simplistic numeric values. In the early phases of Tox21 and ToxCast data analysis, due in part to the small size of the chemical landscape considered, biological models mostly focused on linking biology to biology in relating *in vitro* to *in vivo* outcomes, with some limited success.<sup>116</sup> Whilst this approach brings a way to formulate mechanistic pathways for a large, diverse inventory of chemicals, it fails to utilize the value of the underlying chemical information in helping to discern mechanistically driven, meaningful patterns in the data. It also confines models to the experimental data realm only, where HTS data are required inputs. In the end, it is only by creating linkages across the full progression of "genotype ↔ chemotype ↔ phenotype" that mechanistic approaches will produce actionable knowledge in modeling chemically-induced toxicity on the basis of chemical structure inputs.

The MoA QSAR approach has been recently applied to build on the mode-of-action (MoA) concept of classifying chemicals to establish a collection of biologically similar compounds for a given phenotype training set based on *in vivo* toxicity data. In choosing predictors, the MoA QSAR also employs biological assay results as descriptors in addition to chemical structure-based features and properties. The biological descriptors can include qHTS results

where assay targets (genotypes) proposed to be relevant to toxicity mechanisms are grouped by co-occurrence in known or hypothesized molecular pathways. The presumption is that the new qHTS assay data are effectively populating the vast data space of toxicity pathways, and as this landscape becomes more mature, it becomes possible to infer more robust and biologically based connections from chemical structure to toxicity endpoints. With the chemical activity landscape bounded by these mechanistic principles, chemical elements may be more easily discernible and a modeler has greater freedom to employ unbiased statistical approaches to reveal chemical features and determinants of activity. A molecular initiating event can trigger numerous cellular responses, with key events leading up to the organ responses. The pathway information data-mined from qHTS assay results are also used as guiding principles during MoA formulation. Although there are subsequent events following the initial chemical action at the molecular level, some high level events can be used to group chemicals via related MoAs. The set of chemical classes that are highly enriched within this group of related MoAs are defined as chemical MoA categories; each class within a chemical MoA category, in turn, is represented by a "chemotype", a representation that incorporates chemical structure, physicochemical properties, and biological information all together. A chemotype thus serves to link a chemical structure to a toxicity pathway. A chemotype, at minimum, is a structural alert for a given toxicity endpoint, but augmented with chemical reactivity within an MoA context. Thus the chemotype inherits biological information and can be used to group chemical structures based on biological and chemical similarities. The chemotypes carrying the MoA information guide the process of constructing training sets by providing mechanistic interpretations. MoA QSAR uses this link between biological event and chemical group to identify more mechanistically biased training sets that ultimately relate to phenotypic effects.

Results from models are then combined to obtain one prediction outcome by employing quantitative decision methods, including naive Bayes or Dempster-Shaffer Theory approaches. The prediction outcome obtained by such combination approaches is designed to give more robust and improved predictivity while maintaining model interpretability. The MoA QSAR approach combined with a decision theory based on a Bayesian treatment has been successfully applied to modeling of bacterial mutagenicity, clastogenicity, tumorigenicity, developmental fetal toxicity and specific malformation endpoints, *in vivo* skin irritation, and skin sensitization in safety assessment in a regulatory setting.<sup>123,124</sup>

Within the regulatory workflow for assessing potential chemical hazards, an important requirement of information is that it can support the decisions that a regulator needs to make by a clear rationale within a reasonable time frame. Transparency does not just apply to the ability to access and scrutinize underlying information sources and model details, but also to clear communication of the basis for the rationale in both biological and chemical terms. The MoA QSAR/Decision Theory approach meets the needs of toxicologists and regulators due to its transparent reasoning anchored in relatively simple conveyances of molecular structure linked to biological mechanisms. The "QSAR21" paradigm enabled by the MoA QSAR/Decision Theory approach, offers a means to bridge chemistry and biology along a mechanistic framework, promising more accurate and usable models for regulatory applications. To fully capitalize on these advances, however, will require QSAR

practitioners to gain more intimate knowledge of, and engagement with the biological data, both at the *in vitro* and *in vivo* level.

**Present and future role of QSAR in toxicology**—Despite the great promise of computational toxicology approaches, there continue to be areas of chemistry and chemical risk assessment in which relevant test compounds are either unavailable (such as in early phases of chemical design or pre-manufacture review), or qHTS test results are unattainable with current technologies (*e.g.*, volatiles, reactives, insolubles, metabolites). Such problematic areas have been and will continue to be heavily reliant upon QSAR. With heightened pressures to regulate new and existing commercial and environmental chemicals, and decreasing resources for testing, QSAR methods are being increasingly used in screening, testing prioritization, pollution prevention initiatives, green chemistry, hazard identification, and risk assessment. To be fully accepted by end-users (toxicologists, regulators, industry), however, these QSARs must meet a range of needs, including relevance to regulatory schemes, transparency, biological plausibility, and understandability by non-developers.<sup>125</sup>

### 2.3. QSAR prediction of metabolism

**Historical prospective**—The physician Kahn illustrated in a popular scientific book series the view of a human as a powerful machine using metaphors from industrial society.<sup>126</sup> Kahn showed that during a typical 70 year life span, a human consumes up to 1,400-fold mass of bodyweight in foodstuffs, with both nutritional processing and clearance of toxic substances governed largely by endogenous metabolism. Hydrophilic substances undergo limited biotransformation and can be excreted unchanged. Lipophilic compounds are extensively metabolized but poorly excreted. In the course of evolution, enzymes developed that preferentially act on lipophilic xenobiotics and transform them to more hydrophilic, easily excretable metabolites. Unfortunately, very lipophilic compounds such as insecticides and other persistent organic pollutants (*e.g.*, DDT, chlordane, polychlorobiphenyls, etc.) are less easily metabolized and eliminated, thus leading to bioaccumulation.

Driving forces for the progress in metabolism research during the past five decades are largely due to the tremendous progress in analytical instrumentation and the increasing awareness of the impact of metabolism on unwanted drug effects. Pharmacokinetic consequences may be observed because of the following factors: (i) a drug might induce one or multiple enzymes in metabolism, resulting in a time-dependent therapeutic response over days or weeks; (ii) a drug or metabolite can inhibit a metabolic pathway, resulting in complex kinetics; (iii) the physicochemical properties of the drug metabolites might differ significantly from the parent drug, *e.g.*, higher polarity may result in faster urinary excretion, whereas high lipophilicity may lead to retention in tissue and bioaccumulation. A major issue in pharmacotherapy is that of severe adverse drug reactions (ADRs) and whether they are predictable, avoidable, and iatrogenic. Frequently, ADRs are related to metabolism and, therefore, special focus is placed on drug metabolism during the drug discovery and development process, as well as on pharmacovigilance. Drug-drug interactions due to metabolic inhibition or competition for storage binding sites may result in pharmacological

potentiation, whereas metabolic induction of drugdrug interactions may result in a decreased clinical response. Polymorphism of some drug-metabolizing enzymes may be responsible for a low metabolic capacity. Hence, phenotyping or genotyping of patients is increasingly considered an appropriate means to improve the patient's safety in pharmacotherapy.

In a recent publication, Testa et al.<sup>127</sup> reviewed the reactions and enzymes involved in the metabolism of drugs. Their analysis of the metabolic reactions of over 1,000 different substrates in three selected journals during the years 2004–2009 underlines the importance of cytochrome P450-catalyzed oxidations and UDP-glucuronosyl-catalyzed glucuronidations in drug metabolism. Nevertheless, the study demonstrates the role of other oxidoreductases, esterases, and transferases that significantly contribute to all drug metabolism reactions. Whereas almost 58% of metabolites in the first generation are produced by CYPs, their relative frequency decreases to 32% in the second generation, and 21% in the third generation of metabolites.

**Metabolism prediction models and resources**—The prediction of metabolites has to address at least three different kinds of selectivity questions. As the metabolic reaction can be catalyzed by different enzymes, the corresponding metabolism prediction models have to address enzyme selectivity first and foremost. Thus, in the case of cytochrome P450 the affinity toward different isoforms has to be modeled. Furthermore, CYPs mediate different reaction types and, therefore, the prediction of chemoselectivity is also mandatory. Finally, a particular reaction type might be applicable at multiple sites of a substrate. Therefore, the prediction of the regioselectivity of a reaction type is also required.

The prediction of the metabolism of a chemical on theoretical basis from first principles is not (yet) possible. The influence of solvation and flexibility of protein side chains are very complex phenomena to be directly calculated. Furthermore, individuals of a particular species might have differences in their metabolism due to enzyme polymorphism. Hence, a comprehensive knowledge base is required that can be used for inductive approximations.

A few metabolic reactions databases are currently available. The most widely used is the *Metabolite* database<sup>128</sup> that contains more than 70,000 reactions for different species (humans, rats, etc.) and is particularly reflective of the metabolism of drugs and drug-like compounds. Complementary to the *Metabolite* database (which originally was distributed by MDL), Accelrys offers a database named *Metabolism*.<sup>129</sup> *Metabolism* includes data from the primary literature and its initial scope was focused on metabolic pathways for agrochemicals. The third commonly used database of metabolic reactions is called ADME DB.<sup>130</sup> ADME DB is available as an online service only; its content is mainly focused on metabolism associated with cytochrome P450, with the corresponding data extracted from the primary literature.<sup>131,132</sup>

There are numerous studies focused on modeling cytochrome P450 substrates and inhibitors using QSAR methods, pharmacophore-based approaches, docking and molecular dynamics simulations (see a recent review by Kirchmair).<sup>133</sup> Chohan et al.<sup>134</sup> reviewed 61 QSAR studies that have been used to elucidate the molecular features that influence the binding and metabolism of a compound by the major phase 1 and phase 2 metabolizing enzymes;

Cytochrome P450 (CYP) and UDP-glucuronosyltransferase (UGT), respectively. Braga and Andrade<sup>135</sup> discussed a perspective of the utility of QSAR and QM/MM approaches on drug metabolism prediction including their present limitations and future perspectives in medicinal chemistry. Several software packages are also available for predicting metabolites, including the commercial products, METEOR,<sup>136</sup> META,<sup>137</sup> MetaSite<sup>138–140</sup>, and TIMES<sup>41</sup>, as well as the open-source Metaprint2D<sup>141,142</sup> and SMARTCyp<sup>143</sup> packages, among others.

**Current challenges**—Coverage and comprehensiveness of metabolic data represents a critical issue with respect to QSAR modeling; several reasons could be outlined:

1. Quantitative data on binding constants, kinetic parameters of metabolic reactions and product distribution metrics have to be measured under comparable experimental conditions to be suitable to derive Quantitative Structure-Metabolism Relationship models.
2. Experimental elucidation of chemical metabolites is a time- and cost-intensive process.
3. Experimental systems used to investigate metabolism *in vitro* are not necessarily comparable because the cells originate from different species.
4. Polymorphism of critical metabolic enzymes (such as cytochrome P450 2C19, etc.) might not be considered in older publications.
5. Clinical data (if available) must properly take into account differences in sex, age, diseases, medication, etc., of the study subjects.

As a consequence, the available metabolic datasets are typically heterogeneous. Thus, standardization of *in vitro* metabolism data would enable a paradigm shift from data-driven model building to model-driven data acquisition. Furthermore, publication of these data in a harmonized scheme that provides all necessary experimental information would be highly desirable. A structured format could also be automatically processed without the extensive need of manual curation. Open access of metabolism data to the scientific community would facilitate the validation and further improvement of prediction models. In order to support the input of potentially open access metabolism data, the software tool METIS<sup>144</sup> was developed under the contract from the European Commission, Joint Research Centre, Institute of Health and Consumer Protection (Ispra, Italy).

Moreover, the publication of metabolic QSAR models must be supplemented with the AD information; the validation of the published model should be transparent and convincing, and the choice of descriptors for reactivity modeling should be strongly justified as their repertoire is limited. A better understanding of metabolism requires expanding the view to a systems biological perspective including the processes of biological transport, regulation of CYP expression and consideration of their polymorphisms. Moving towards an integrated strategy combining standardization of experimental data with properly practiced *in silico* modeling appears as the most promising approach.

## 2.4. Interpretability of QSAR models

In the last decades, the focus of QSAR has shifted away from simple and interpretable linear models towards more complex multiparametric and nonlinear approaches.<sup>145</sup> This has resulted in what some perceive as a trade-off between predictive ability and interpretability of QSAR solutions,<sup>146</sup> as many highly predictive models are based on neural networks, support vector machines, and other "black box" approaches that do not lend themselves as easily to interpretability. However, the importance of interpretability to practical acceptance of QSAR solutions is well-established and is reflected in one of the OECD principles: "*To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with ... a mechanistic interpretation, if possible*".<sup>147</sup> For purposes of this discussion, the term "interpretability" refers to the ability of a user to understand and rationalize both the underlying model determinants of activity, i.e., what the model deems to be the primary predictors of activity, as well as the individual model predictions, in terms of chemical structure, reaction mechanisms, or known or plausible biological mechanisms of activity.

A predictive and interpretable model has a clear advantage over an equally predictive but non-interpretable one. The former allows for the targeted design of compounds with desired properties, lends itself to mechanistic interpretation and hypothesis generation, and can contribute to further understanding of mechanism(s) of action. When predictions from such a QSAR model can be independently supported, and are chemically and biologically plausible, they also can carry significantly greater weight in, for instance, a safety assessment workflow. The importance and benefits of building interpretable models have been demonstrated in several studies,<sup>148–154</sup> and some examples will be characterized herein. Below, we outline different constituent blocks of a general QSAR model workflow with respect to model interpretability.

**Role of molecular descriptors**—QSAR model interpretability strongly depends on the nature of chemical descriptors used. The use of well understood physical-chemical characteristics of a molecule as descriptors (van der Waals volumes and surfaces, lipophilicity, H-bond related parameters, etc.) can aid in model interpretation. Metrics related to the electronic configuration of a compound (partial charges, dipole moment, orbital energies, etc.) are also suitable for structural interpretation when placed in the context of reactivity hypotheses. In addition, various molecular fingerprints<sup>33,155,156</sup> and fragment descriptors such as MNA,<sup>151</sup> G-QSAR,<sup>157</sup> ISIDA,<sup>158</sup> and Simplex (SiRMS)<sup>159</sup> have a direct connection to molecular structure and, thus, have the potential to lend interpretability to a model.

QSAR models built with topological indexes reviewed above are often more difficult to relate to easily understood chemical concepts.<sup>160</sup> Autocorrelation descriptors pose a similar problem in that they encode an indirect relationship between molecular structure and descriptor values, such as BCUT, WHIM, GETAWAY, RDF, etc., although with some post processing, they can be used to convey insight into the role of branching, degree of structural nonhomogeneity and cyclicity of a molecule on activity.<sup>161</sup>

**Modeling techniques**—Some QSAR modeling techniques, such as linear regression and decision tree approaches are very straightforward to interpret given their intuitively understandable architecture; when combined with interpretable QSAR descriptors, these approaches allow for the development of ready-to-use structural rules and alerts. Other types of models, however, require additional operations to provide chemical insights.<sup>29</sup> Generally, all approaches enabling interpretation of QSAR solutions can be divided into model-specific and model-independent.

**Model-specific interpretation methods**—The estimation of weights of individual descriptors in a linear model is routinely made by considering the corresponding regression coefficients.<sup>162,163</sup> More sophisticated methods are also applied, *e.g.*, to interpreting PLS models, in this case relying on the analysis of a descriptor's contribution to the variation of the investigated property.<sup>29,164,165</sup> Similarly, linear support vector machine (SVM) models are amenable to this approach for providing interpretability of results.<sup>166</sup>

Some efforts have been made to interpret artificial neural network (ANN) models using their weights and biases.<sup>167</sup> Along a similar line, Kuz'min et al.<sup>168</sup> proposed an approach for interpretation of Random Forest models using the differences between mean activity values of compounds in the corresponding parent and child tree nodes.

**Model-independent interpretation approaches**—Similar to the concept of assessing leverage, the relative importance (or significance) of a given descriptor, even in a complex model, can be assessed by comparing the reduction in the overall predictive power of a model when one descriptor versus another is removed or altered. Originally suggested by Breiman<sup>169</sup> for interpreting Random Forest (RF) models, this approach was also adopted by Guha *et al.*<sup>170</sup> to evaluate neural networks-based solutions. Another approach analyzes local gradients or partial derivatives of descriptors as reflecting their contributions to the variation of the modeled property. These methods have been used to aid interpretation of PLS, RF, ANN, and SVM models, among others.<sup>171–173</sup>

**A universal approach to model interpretation**—To avoid many problems outlined in the previous sections, Polishchuk et al. proposed an approach to the interpretation of QSAR models that does not depend on the nature of the descriptors, utilized mathematical approach, and a type of endpoints (continuous or binary).<sup>174</sup> The gist of this methodology is very simple (Figure 3): activities of a target molecule ( $P_{\text{QSAR}}(\text{AB})$  in Figure 3a or  $P_{\text{QSAR}}(\text{ABC})$  in Figure 3b) and its virtual analog ( $P_{\text{QSAR}}(\text{A})$  in Figure 2a or  $P_{\text{QSAR}}(\text{A} \dots \text{C})$ , derived from this initial molecule by eliminating a molecular fragment with a pre-defined structure, are estimated using the QSAR model. The difference in the calculated activities of a target molecule and its virtual analog is considered as an influence (contribution) of eliminated fragment ( $P'(\text{B})$  in Figure 3). This simple approach is equally applicable for estimating the contributions of both terminal and central parts of a molecule. Notably, the contributions to activity are more sensitive to chemical descriptors, which are used to represent the molecule with and without the fragment, than to the modeling technique.

Concluding this section, we shall stress that, using the universal approach,<sup>174</sup> any QSAR model, despite of the complexity of the modeling technique or nature of the descriptors

used, can be formally interpreted in terms of significant chemical features that can be easily understood by medicinal chemists. Hence, all QSAR models should be primarily evaluated on the basis of well-established methods for assessing external predictive ability. Stated otherwise, a QSAR model with easily interpretable descriptors and poorer performance statistics should not be preferred over a more predictive model that is less easily interpreted.

## 2.5. Multi-task modeling

Often, chemicals have multiple biological activities (*cf.* polypharmacology) that may be interrelated (not to mention multiple physical properties that are frequently the subject of prediction by QSAR approaches). Typically, however, QSAR models are developed for each target property individually, without utilizing knowledge that can be extracted from QSAR models for other activities of the same compounds. Individual QSAR models of this sort should not be viewed as separate entities, but rather as nodes in a network of interrelated models. This concept is accounted for in an inductive knowledge transfer approach<sup>175</sup> realized in multi-task learning (MTL) and Feature Net (FN) methods. MTL<sup>176</sup> treats several tasks in parallel and uses a shared representation of data. This can be carried out using machine learning methods yielding models with several outputs, such as neural networks, PLS, or SVM with special kernels.<sup>177</sup> FN treats different tasks sequentially when predictions made by previously developed models are used as extra descriptors for the main task.<sup>175</sup>

An inductive knowledge transfer approach could significantly improve predictive performance of individual QSAR models (or, Single Task Learning (STL) models) built on small and structurally diverse datasets. Since the cost of obtaining new data is rather high, especially for *in vivo* experiments, the integration of available experimental data on related activities could serve as an alternative to costly and time-consuming experiments. The FN technique is widely used in QSAR studies whenever a successfully employed descriptor is reused. Thus, models employing parameters used by other QSAR models solutions (*e.g.*, logP, pKa, E<sub>homo</sub>, E<sub>lumo</sub>) as descriptors could formally be considered as FN. More generally, FN could be realized in the form of a multilayer network of models, in which the outputs of the models of previous layers are the inputs for the models of the next layers.<sup>178</sup>

Unlike FN, only a few applications of MTL in QSAR studies have been reported in the literature. For instance, higher performance of MTL and FN approaches over the conventional STL models was demonstrated by Varnek et al.<sup>175</sup> in QSAR modeling of tissue-air partition coefficients (logK) using backpropagation neural network with several output neurons. The initial datasets encompassed 11 different types of logK data points for humans (H) and rats (R). Only four datasets were of reasonable size (about 100 compounds), whereas the others included from 27 to 38 compounds. The output layer of the network contained either one (for STL and FN) or 11 (for MTL) neurons. In STL and MTL studies, only fragment descriptors were used as inputs, whereas in all FN calculations, all input neurons were fed by 10 predicted properties. The results of that work demonstrated that conventional STL allowed reasonable prediction of only 4 activities for which larger (about 100 compounds) data sets were available, whereas MTL and FN calculations generated models with acceptable quality for 9 activities.

Of note, MTL should not be confused with multi-target learning, in which the performance of models predicting binding of ligands to different biological targets is boosted by incorporating certain information concerning the proteins' structure.<sup>179,180</sup> Such information can be represented either directly as a set of target-specific descriptors or indirectly through special kernels for protein targets. In the latter case, descriptors for ligands and targets are combined in the feature spaces induced by kernels. In some publications the multi-target learning is erroneously associated with MTL,<sup>179–181</sup> although in reality this is an STL model for complex chemical objects, i.e., protein-ligand pairs. In our view, the term MTL should designate only the tasks based on common internal representation, which cannot be reduced to STL.

The growth of available experimental data and predictive structure-activity models will stimulate further development of the inductive learning transfer approach. We believe that in the future, isolated and unrelated QSAR models will be connected to the network of interrelated models solutions organized in the form of a "chemical brain" accommodating considerable volumes of experimental data and knowledge, which will significantly improve the quality of prediction of various chemical and biological properties. This expected development will further advance the integration of QSAR as part of Systems Chemical Biology.<sup>182</sup>

## 2.6. Experimental validation 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, experimental validation is the only indicator of actual utility of QSAR modeling. The general scheme of QSAR-guided research project inclusive of experimental validation is shown in Figure 4. Virtual screening (VS) approaches<sup>183</sup> are the inherent part of this workflow. They are used to identify chemical hits predicted to be active against selected target(s) from large chemical libraries (see Figure 5). With the continuous emergence of novel biological targets of therapeutic potential from drug discovery teams in industry and academia,<sup>184</sup> effective and accurate VS technologies continue to be in high demand. Meanwhile, this demand, especially in terms of both computational efficiency and accuracy, is amplified by the very rapid growth of chemical compound collections that are available for virtual screening even in the public domain.<sup>185</sup> For instance, both PubChem and ChemSpider, the two major collections of chemical structures on the web, currently include over 30 million compounds each, and ZINC, a database frequently used for virtual screening applications,<sup>186</sup> incorporates a total of approximately 21 million compounds. Fara et al.<sup>187</sup> discussed the integration of virtual (QSAR-based) and actual (physical) screening. However, as shown in Figure 5, modern VS workflow incorporates several critical filtering steps to eliminate compounds that are unlikely to be active: (1) sets of empirical rules (*e.g.*, Lipinski's,<sup>188</sup> QED drug-likeness),<sup>189</sup> (2) chemical similarity cutoff commonly computed using molecular fingerprints, (3) QSAR-based filter(s) (*e.g.*, only retain compounds predicted to be active or having predicted  $pK_i \geq 8$ ), and (4) chemical feasibility and/or purchasability.

Several important studies have been published recently in which QSAR-based predictions have been experimentally confirmed. These studies illustrate how useful and reliable computer-assisted approaches can be in assisting medicinal chemists to design novel compounds with controlled bioprofiles.

Keiser et al.<sup>190</sup> developed the Similarity Ensemble Approach (SEA) to compare targets utilizing the overall similarity between their known ligands. The authors applied this approach to prognosticate unknown drug-target interactions for demonstrating the potential use of computational approaches to study and predict drugs' polypharmacology. More than 3,600 FDA approved and investigational drugs were analyzed by SEA and thousands of unknown associations were discovered. Out of 30 experimentally tested associations, 23 precedently unknown drug-target interactions were confirmed (five of them characterized by a potency less than 100 nM).

Lounkine et al.<sup>191</sup> conducted a large-scale prediction and testing of drug potency on different side-effect targets. The authors extracted and curated experimental data for 285,000 ligands and 1,500 biological targets from the ChEMBL database. Then, they used the SEA similarity search approach<sup>190</sup> to predict the activity of 656 marketed drugs on 73 unintended 'side-effect' targets. Out of 1,644 significant drug-target associations predicted by SEA, there were 893 that were unknown and never reported before. The authors then conducted experimental tests for confirming these predictions. Out of 694 experimental tests, 478 drug-target associations (68.9%) were disproved and 65 found to be ambiguous. However, significant potencies (less than 30  $\mu$ M) were confirmed for the 151 remaining drug-target associations, especially for 48 compounds with sub-micromolar activities. Interestingly, the authors linked those targets with known side effects and successfully established previously unknown links between drugs and several side effects. This study demonstrated that QSAR-like predictions can successfully prognosticate ligand-target interactions, which can then be confirmed experimentally. One can argue that the use of additional filters such as applicability domain and prediction confidence scores could potentially have avoided the large number of unconfirmed ligand-target associations.

Besnard et al.<sup>192</sup> utilized cheminformatics methods to explore and design compounds with unique polypharmacology (see Figure 6). As both clinical efficacy and overall safety of a drug is determined by its activity profile towards many biological targets, there is a huge need for approaches capable of predicting and designing drugs with a specific multi-target behavior. The authors developed new methods that (i) generates one (or several) generation(s) of chemical analogues of a given parent drug with known properties, and (ii) predicts their polypharmacology using an ensemble of ligand-based QSAR models. Then, most interesting compounds with the preferred polypharmacology profiles are synthesized and confirmed experimentally. The authors explored the case study of an approved acetylcholinesterase inhibitor drug and its *in silico* generated analogues, all tested for their specific or promiscuous polypharmacology towards G-protein-coupled receptors. More than 800 ligand-target predictions of prospectively designed ligands were tested experimentally and 75% were confirmed.

Other examples of experimentally validated QSAR-based predictions have been published by the Tropsha group at the UNC. Recently, Hajjo et al.<sup>193</sup> have developed and validated binary classification QSAR models capable of predicting potential serotonin 5-HT<sub>2B</sub> actives that are known to cause valvular heart disease. The models were employed to screen the World Drug Index library and 122 compounds were prognosticated to be 5-HT<sub>2B</sub> actives. Ten of them were tested experimentally and nine were confirmed to be active. These QSAR models can thus be employed for predicting 5HT<sub>2B</sub>-related valvulopathy. A similar strategy<sup>194</sup> has been followed to design novel antimalarial compounds by modeling a dataset of 3,133 compounds defined as either active or inactive towards *P. falciparum*. The virtual screening of the ChemBridge library using the QSAR models led to the identification of 176 putative antimalarial compounds that were submitted for experimental validation along with 42 putative inactives as negative controls. The authors reported that 25 computational hits (14.2%) were confirmed to have antimalarial activities, whereas all 41 putative inactives were also confirmed as inactives. Importantly, confirmed hits featured novel chemical scaffolds that could be promising for developing novel antimalarial agents. In another case, the Tropsha group<sup>195</sup> generated QSAR models for 5-HT<sub>6</sub> receptors and utilized them for identifying novel actives in combination with predictions from the Connectivity Map (<http://www.broad.mit.edu/cmap/>). Thirteen common hits were tested experimentally and ten were confirmed as actives.

QSAR studies<sup>148–150,196,197</sup> of various antiviral activities represent another good example of targeted design of new compounds with desired properties using cheminformatics tools. Fifteen novel antiviral agents against influenza H3N2, herpes HSV-1, rhinovirus HRV-2, and coxsackievirus B3 were computationally designed as a result of interpretation of QSAR models. Then, their high activity and selectivity was confirmed by subsequent synthesis and experimental testing. These studies were summarized in review<sup>198</sup>.

In summary, a growing number of published QSAR studies include the experimental validation of predicted hits and this critical step should become a standard component of any QSAR investigation. Scientific journals should raise the bar for accepting and publishing papers that employ QSAR techniques. Importantly, Journal of Medicinal Chemistry already supports this trend by not accepting for publication any papers describing experimentally untested QSAR models and predictions.

### 3. Novel applications of QSAR and future trends

#### 3.1. QSAR modeling of peptides

Antimicrobial peptides (AMPs) represent a diverse class of natural peptides that form part of the innate immune system of mammals, insects, amphibians, and plants.<sup>199–202</sup> In the face of increasing antibiotic resistance by pathogens, AMPs have drawn significant attention as a prospective class of antimicrobial therapeutics<sup>203–206</sup> as they hold several notable advantages, including broad range of activity, low toxicity, and minimal development of resistance in target organisms.<sup>207,208</sup> More than fourteen peptides are currently in development or clinical trials, with two having demonstrated efficacy in Phase III clinical trials.<sup>209</sup>

Despite the fact that a broad spectrum of antimicrobial peptides have been reported and discussed, their structure-activity relationships are not well understood, largely because of substantial diversity in AMP structures and their non-specific mechanism of action.<sup>210</sup> The general view on characteristic features of the AMPs is typically focused on their cationic character, relatively high hydrophobicity/amphipathicity and short length.<sup>208,211</sup> Thus, the previous attempts of "*in silico*" modeling of peptide-based antibiotics were largely based on sequence-based approximation.

**Sequence-based AMP modeling**—The majority of the previous sequence-based modeling efforts was of qualitative nature and relied on available AMP sequences to (a) discover previously unknown natural peptides, and (b) to modify sequences of known AMPs to improve their therapeutic properties.

The AMP optimization methods relying on various sequence templates were recently reviewed by Fjell et al.<sup>209</sup> In short, such approaches imply systematic change of one or a few amino acids in the sequences of prominent AMPs such as cecropin, magainin, protegrin, lactoferricin, and bactenecin (and, recently, indolicidin<sup>212</sup> and brevinin)<sup>213</sup> to enhance their antimicrobial activity or reduce toxicity.<sup>214</sup> Rather commonly, template-based studies attempted to introduce unnatural amino acids into AMPs to increase stability.<sup>215,216</sup>

The sequence-based efforts did not result in drastic improvement of AMP properties.<sup>217,218</sup> However, such template-based studies (when one changes a part of the molecule, while keeping the rest intact, and records the overall response of the system) has brought the spirit of early-day QSAR into the field of AMP research.

**Residue-based QSAR modeling of AMPs**—The previous QSAR work on antimicrobial peptides was mainly focused on derivatives of three natural substances: bactenecin, protegrin and lactoferricin, and utilized a residue-based level of modeling. Thus, back in 1987 a set of based on the physical-chemical properties of the peptide z-descriptors, was proposed and used to investigate peptide variants of lactoferricin.<sup>219</sup> Several more recent studies have examined activities of lactoferricin derivatives against bacteria<sup>220–222</sup> and herpes simplex virus<sup>223</sup> in response to specific amino acids changes. Thus, Strom et al.<sup>222</sup> modeled a set of 20 peptides with such descriptors as alpha-helicity, HPLC retention time, net charge, molecular surface, and symmetry of charge and hydrophobicity distribution. Later, using an expanded set of peptides, good predictive accuracy was achieved using z-values on a larger set of peptide analogues where only a few amino acid substitutions were made.<sup>219,221</sup> However, predictions were much less accurate for the cases when more than one or two substitutions were made. In a recent study by Sánchez-Gómez<sup>224</sup>, residue-based QSAR descriptors were used to model membrane permeability of 8-to-12 amino acid long lactoferrin derivatives.

It should be noted that earlier QSAR attempts at AMP modeling suffered from similar shortcomings. Thus, modest numbers and sequence variation in AMP training sets did not typically allow large sets of QSAR descriptors to be employed (with a notable exception of an early work by Mee et al.)<sup>225</sup> and, hence, limited the use of more rigorous, nonlinear modeling techniques. Combined with traditional residue-based level of consideration of the

AMPs, these limitations typically did not enable the development of improved therapeutics that outperformed peptide variants in the training set. This situation has recently changed with the advances in high-throughput technologies of peptide synthesis, screening and analysis.<sup>202,226–229</sup>

### Atom-based QSAR modeling of AMPs with the Use of Machine-Learning

**Approaches**—In an important original study,<sup>230</sup> Cherkasov showed that AMP activity can be effectively quantified using machine learning QSAR and atomic levels of structural considerations. When large-scale training data were generated for the AMPs,<sup>227</sup> this assumption led to a series of studies that resulted in the development of synthetic peptides with significantly improved antibacterial activity and lowered toxicity.<sup>209,231–234</sup> This was achieved by employing artificial neural networks to build computational models of peptide activity based on the data from >1400 sequences, only biased in the content of certain amino acid types believed to be important for antibacterial activity. As a basis for enumeration of peptide structures, a set of 44 descriptors were employed, including 3D QSAR parameters that utilize atomic-scale molecular information, the so-named "inductive" QSAR descriptors.<sup>235,236</sup> Briefly, the "inductive" descriptors describe whole molecules based on the calculated effects of the atomic constituents of a molecule. Previously, these parameters allowed quantification of diverse sets of organic and organo-element molecules and free radicals.<sup>235–238</sup>

A total of 26 "inductive" descriptors of electronegativity, hardness, charge, substituent, and steric effects were used. An additional eighteen conventional molecular descriptors were also added, including numbers of hydrogen acceptors and donors, surface area, total and partial charges, and molecular weight. The developed QSAR models proved remarkably accurate in prediction of antimicrobial activity of nine amino-acid long training set AMPs<sup>231,234</sup> and allowed the creation of novel and improved peptide variants. In particular, 100,000 virtual peptides were created and QSAR models (pre-trained on experimental results for >1,400 AMPs) were used to predict their hypothetical antimicrobial effects. Based on the QSAR predictions for 100,000 sequence variants, 25 peptides were identified from each of the predicted activity quartiles (roughly corresponding to high-, medium-, low-, and inactive predictions by the QSAR). The selected 100 peptide candidates were then synthesized and assayed against 12 of the most dangerous "superbugs" including multidrug resistant strains of *Pseudomonas aeruginosa*, *Pseudomonas maltophilia*, *Staphylococcus aureus*, *Escherichia coli*, and *Enterobacter cloacae* among others.<sup>231</sup>

Remarkably, of 50 peptides that were predicted as most active, 49 peptides (or 98%) were actually found to be more active than the control antimicrobial peptide Bac2a, while the 2nd, 3rd and 4th quartiles were respectively 88%, 4% and 0% similar to or better than the control peptide Bac2a. Moreover, the best predicted peptides, when tested experimentally, not only demonstrated sub-micromolar *in vitro* activity against major, life-threatening human pathogens, but also showed significant activity in animal models.<sup>231</sup>

These results have unambiguously demonstrated that QSAR methodology is applicable to AMP data, and that the atomic level of consideration of AMPs combined with machine learning techniques results in practical models that deliver the most active peptides

identified to date. Not surprisingly, these findings initiated a broader interest of the QSAR community in the field of antimicrobial peptides. A number of recent studies have successfully utilized QSAR for analysis and discovery of AMPs using both atom- and sequence-based descriptors.<sup>202,239,240</sup> Such an approach is allowing rapid advancement of the field of AMPs and has resulted in the development of peptide candidates with improved therapeutic properties.<sup>228</sup>

### 3.2. QSAR modeling of chemical mixtures

It is a common knowledge that chemical mixtures have a very broad application in all fields of experimental science, as well as broad use in commercial, industrial, and pharmaceutical products. It is also feasible to forecast that the use of mixed formulations, reagents, and industrial releases into the environment will substantially increase in the future, especially in the medical field. At the same time, although modern QSAR is successful in dealing with individual compounds, there are no mature QSAR methodologies that could be directly used to model properties of mixtures, reflecting the lack of robust, well-benchmarked data pertaining to such properties.<sup>241</sup> To date, only a few published QSAR studies of mixtures could be considered reliable.<sup>242,243</sup> An interested reader can find detailed descriptions of studies devoted to mixtures and QSAR modeling of their properties elsewhere.<sup>241,244,245</sup> Herein, we will highlight those studies<sup>246,247</sup> that exemplify lack of awareness of some researchers of the best practices of QSAR modeling, which should apply equally to modeling of both individual compounds and chemical mixtures.

**Current challenges in QSAR modeling of chemical mixtures**—The lack of reliable data poses one of the biggest challenges for the development of QSARs for chemical mixtures. Some information can be found in the ChEMBL database,<sup>248</sup> which contains fragmental data on 356 organic mixtures. The NCI database<sup>249</sup> stores endpoints of anticancer activity for some mixtures, as does the DTP AIDS Antiviral Screen database<sup>250</sup> for anti-HIV activity.

Some limited and sparse data can also be found in the literature. For instance, toxicity of about 100 mixtures toward *Photobacterium phosphoreum* species can be extracted from papers by Lin et al.<sup>251,252</sup> Three datasets of reasonable size (271–411 binary mixtures) related to various properties of liquids were published by Ajmani et al.<sup>243,253,254</sup> Vapor-liquid equilibrium data for 101 pure compounds and 261 binary mixtures were compiled by Oprisiu et al.<sup>242</sup> using the Korean Thermophysical Properties Databank as a source.<sup>255</sup> 550 mixtures composed of 33 individual compounds were investigated by Small *et al.*<sup>256</sup> to discover strong anti-inflammatory combinations. It is expected, however, that the ongoing rapid growth of publicly available databases will begin to address this problem, and QSAR modeling of mixtures will advance, as it did for individual compounds some years ago.<sup>257</sup>

**Current approaches to QSAR modeling of mixtures**—QSAR modeling of organic mixtures requires the use of appropriate descriptors. All studies published to date on the subject can be divided into several groups depending on the descriptor type used: (i) descriptors based on the mixture partition coefficient<sup>258</sup> or biological descriptors,<sup>259</sup> (ii) additive molecular descriptors (weighted sum of descriptors of individual components), *e.g.*,

see work;<sup>260</sup> (iii) integral non-additive descriptors of mixtures (mixture components are taken into account in a different manner from the additive scheme), *e.g.*, see study;<sup>261</sup> and (iv) fragment non-additive descriptors (structural parts of different mixture components simultaneously taken into account in the same descriptor).<sup>262</sup>

We consider non-additive fragment descriptors to be the most promising for QSAR modeling of mixtures. For instance, the SiRMS approach<sup>159</sup> is suitable for QSAR analysis of binary mixtures of any composition. The method represents a mixture by two molecules considered simultaneously, where bounded simplexes (tetraatomic fragments) describe single components of the mixture individually, while unbounded simplexes describe the constituent parts of the mixture as a whole (see Figure 7). In this approximation, it is necessary to indicate whether the parts of unbounded simplexes belong to the same or different molecules. In the latter case, such unbounded simplexes will not reflect the structure of a single molecule, but will characterize a pair of different molecules.

A 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 (*e.g.*, compounds A and B) are weighted according to their molar fraction and mixture descriptors are multiplied on molar fraction of deficient component (see Figure 7). If in the same task both mixtures and pure compound have been considered, pure compound is regarded 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.

Analysis of the existing descriptors of mixtures demonstrates that additive descriptors of mixtures, where the latter are characterized by mole-weighted average descriptors of the constituents, have the following disadvantages: (i) they wholly rely on 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. 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 has shown good results, as in the study of Ajmani et al.<sup>254</sup>

More complicated mixture descriptors developed by Ajmani et al.<sup>243,253</sup> are capable of encoding important non-covalent intermolecular interactions, which is their significant advantage. However, they are system- and property-specific, *i.e.*, applicable only to particular systems (such as binary mixtures where the components are dissolved in each other) that were described by Ajmani et al.<sup>243,253</sup> Another serious disadvantage of this approach is that two mixtures of different composition could be described by identical descriptors, thus the descriptions are not sufficiently unique.

Simplex and ISIDA mixture descriptors<sup>242</sup> are free of the aforementioned drawbacks; they can be applied to any property of interest, and they are capable of capturing interaction or joint effect of components. However, in the current version, these approaches can be applied only to binary mixtures. From a methodological point of view, these two methods appear better than others, but major improvements are still required.

**QSAR modeling of mixtures**—In our view, no single QSAR study or methodology published to date can yet be recommended as a reliable tool or a "golden standard" for QSAR analysis of mixtures. In addition to the mentioned limitations in datasets sizes and drawbacks of mixture descriptors, many published reports contain significant methodological errors.

As in traditional QSAR, rigorous external validation is required for modeling mixtures. However, proper external validation is less straightforward for QSAR models of mixtures, especially when the same compounds with different ratios are present throughout the training set. The conventional external cross-validation procedure is also not always acceptable for those cases – if both training and external sets include data points for the same mixture, the model's true predictive performance will not be estimated properly. There is a consensus opinion that novel, more rigorous external validation protocols are required in the field.<sup>242</sup> Depending on the initial data and potential application of the developed models, three different strategies of external validation could be used: (i) "points out" – prediction of the investigated property for any composition of mixture 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 combining 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 (the most rigorous method of external validation in QSAR modeling of mixtures). Furthermore, careful collection and understanding of initial data, thorough its curation, rigorous internal and external validation, and application of developed models for virtual screening of large databases (which are mostly absent in current mixture studies) will significantly improve the quality of QSAR models of mixtures.

The field of QSAR modeling of mixtures is very new and is under active development. Given the importance and the increasing need for such models, efforts directed to the development of new methods, and the improvement of existing QSAR approaches for mixtures are welcomed and encouraged.

### 3.3. QNAR – quantitative nanostructure-activity relationships

Combinatorial chemistry and HTS technologies have been recently extended towards designing novel manufactured nanoparticles (MNPs).<sup>263,264</sup> With more than 1,000 manufacturer-identified, nanotechnology-based consumer products available on the market, nanotechnology is drawing worldwide attention for its numerous applications in various industrial areas, such as material science, medical research, cosmetics, or even clothing. Importantly, a significant portion of these efforts is directed towards the development of "green" products intended to achieve efficient and less polluting energy sources. In this

context, QSAR science has a role to play by: (i) facilitating the access, storage, search, and integration of all experimental results currently distributed in literature, databases, and other sources;<sup>265</sup> (ii) achieving externally predictive QSAR models to compute MNPs' properties based on their structural characteristics; and (iii) boosting the development and testing processes by identifying the most promising nanoparticles that require focused experimental investigations. The latter point is especially true due to the concerns about the safety of certain MNPs and the development of nanomedicine.<sup>266,267</sup> A growing compendium of experimental results shows that certain MNPs intended for industrial applications could cause toxic effects in humans.<sup>268–272</sup> Thus, computational tools capable of evaluating MNPs for their potential health risk are needed.

From a chemical perspective, MNPs are very different from small molecules in ways that make their modeling more challenging. Firstly, MNPs are characterized by high physical/structural complexity and diversity as they represent assemblies of inorganic and/or organic elements, sometimes mixed or coated. Moreover, the exact molecular stoichiometry may vary from one particle to another even in the same cohort, and there is a vast variety of particle categories with numerous potential applications, having ranges of desired and undesired physical, chemical, and biological properties.<sup>273</sup>

These factors help to explain why there are no systematic quantitative studies of MNPs in the literature. Thomas et al.<sup>265</sup> recently published a useful ontology for MNPs applicable in the field of cancer research, which allows for the integration of experimental results for nanoparticles. Similarly, the NanoTAB initiative<sup>274</sup> deals with the development of a common database and file exchange format for nanotechnology information.

There are also relatively few literature reports on computational modeling of MNPs,<sup>275</sup> especially in regard to nanotoxicology.<sup>276</sup> Liu et al.<sup>277</sup> used molecular dynamics simulations to reveal the overall changes in the structure of cellular membranes caused by the insertion of carbon nanotubes and to estimate affinity of drug-like molecules to the nanotubes. Puzyn et al.<sup>278</sup> recently introduced a term "nano-QSAR" and published a study on a small set of MNPs with metal-oxide cores.<sup>279</sup> Published studies and developing trends in QSAR modeling of nanomaterials have been summarized in few reviews.<sup>274,280</sup>

In a recent proof-of-concept study, Fourches et al.<sup>280</sup> introduced the terminology of Quantitative Nanostructure-Activity Relationships (QNARs) that employs classical machine-learning methods for establishing links between chemical descriptors and various measured activities of MNPs. Using an ensemble of 51 diverse NPs tested *in vitro* against four cell lines in four different assays at four different concentrations (resulting in 51x64 biological data points), they were able to identify three clusters of NPs for which QNAR models were obtained using four experimental descriptors: size, two measures of relaxivity, and zeta potential. The developed models resulted in an external accuracy of prediction as high as 73%. In another study, Fourches et al.<sup>280</sup> modeled 109 cross-linked iron oxide NPs decorated with small organic ligands to predict their uptakes by PaCa2 pancreatic cancer cells. Models afforded a reasonable predictive power of  $R^2 = 0.72$  using a 5-fold external cross-validation procedure. More recently, the same group succeeded in modeling a set of 84 decorated carbon nanotubes tested in six different *in vitro* assays. Predictive QNAR models

were obtained to accurately predict protein binding profiles and toxicological properties of these nanotubes. QNAR models were then utilized to screen a database of 200,000 ligands potentially attachable to carbon nanotubes in order to design new nanotubes with less protein binding affinities and less acute toxicity.

As nanomaterials continue to proliferate in many areas, computational methodologies such as QNAR modeling are expected to provide critical support to experimental studies to identify safe nanoparticles with desired properties. However, it is important to emphasize that such procedures require relatively large amounts of reliable and consistent experimental data where MNPs can be characterized by a set of physical chemical properties and tested in well-defined assays.

## 4. QSAR and regulatory decision support

### Historical context

QSAR has had a long-standing history of use within the areas of environmental research and regulation, particularly for food use, cosmetics, and industrial chemicals where regulations are often limited, while property and toxicity data are sparse or unavailable. For more than 30 years, EPA has made an extensive use of QSAR models for hazard identification among new industrial chemicals (chemicals subject to pre-manufacturing / premarketing notification) under the Toxic Substances Control Act (TSCA), particularly in the area of ecotoxicology. These approaches include the use of expert systems, QSAR modeling, nearest analogue and chemical class analyses, and prediction of mechanisms of toxicity among others. To enable these efforts, EPA has supported the development and public release of various QSAR and decision support tools, including the Estimation Program Interface (EPI) Suite,<sup>281</sup> which computes a variety of physical-chemical properties (based primarily on fragment-based QSAR approaches), and the Ecological Structure Activity Relationships (ECOSAR) software for predicting aquatic toxicity of compounds.<sup>282</sup> The latter tool has also been demonstrated to have some useful applicability to pharmaceuticals.<sup>283</sup>

Worldwide, QSAR methods have been used for identification of potential health hazards, screening and prioritization by various government agencies including Health Canada,<sup>284</sup> the FDA, and the European Union (EU) authorities.<sup>285</sup> In Canada, under the New Substances Provisions of the Canadian Environmental Protection Act (1999), regulators use QSAR predictions for assessing and prioritizing the Canadian inventory of existing substances (Domestic Substances List, DSL).<sup>284</sup> Within the EU, the Danish EPA utilizes QSAR-generated endpoints for ecological and health hazard assessments, developing an advisory list for self-classification of dangerous substances where experimental test results are incomplete or unavailable.

Within the EU, the New Chemicals Policy of the European Commission (REACH: Registration, Evaluation and Authorisation and Restriction of Chemicals)<sup>286</sup> has proposed a new system for managing chemical information in a single regulatory framework. According to that initiative, nearly 30,000 substances will be processed on a phased basis over a period of 11 years (2007–2018). An important part of this policy is the fostering of

research on the development and validation of alternative (non-animal) methods, including QSAR models and *in vitro* tests. Importantly, within the REACH initiative, it was considered important to develop an internationally recognized set of principles for QSAR validation, to increase the confidence in QSAR predictions, and to provide regulatory bodies with a scientific basis for making decisions on the acceptability of QSAR estimates.

### OECD principles

Some principles for assessing the validity of QSARs were proposed in 2002, as the "Setúbal Principles", at the international workshop in Setubal, Portugal. Two years later in 2004 those were modified by the OECD Work Programme on QSARs, as the "OECD principles for the Validation, for Regulatory Purposes, of QSAR Models".<sup>89</sup> The corresponding OECD principles are as follows:

"To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; (5) a mechanistic interpretation, if possible."

There are long-standing and ongoing debates in the scientific community regarding the so-called "mechanistic versus statistic" approaches to QSAR modeling. The first approach considers Principle 5 as the most important and focuses mainly providing a mechanistic basis for a model and its predictions, whereas the second approach follows the order of OECD Principles in assessing different model characteristics, with the complete model validation process (internal and external), requested by Principle 4, carried out before the interpretation of descriptors for their mechanistic meaning, if possible (Principle 5).

Also it was recommended that the results of (Q)SARs may be used instead of testing when the following conditions are met: (1) results are derived from a (Q)SAR model whose scientific validity has been established; (2) the substance falls within the applicability domain of the (Q)SAR model; (3) results are adequate for the purpose of classification and labeling and/or risk assessment, and, (4) adequate and reliable documentation of the applied method is provided.

### Available resources

The need for "adequate and reliable" documentation requires standardized QSAR reporting formats, such as proposed in "A (Q)SAR Model Reporting Format" (QMRF). The QMRF is a communication tool for reporting and assessing QSAR predictions that is meant to provide industry and regulators with reliable information. Access to this form is provided through the JRC (Joint Research Commission) QSAR Model Database, available through a web-interface.<sup>287</sup> A similar project led by the OECD has developed the "QSAR Toolbox", a software tool to facilitate grouping and category building and to assist in the development of read-across justifications and their transparent documentation, supporting the use of QSAR models in different regulatory frameworks.<sup>113</sup> This approach is based on chemical categorical grouping, emerging from "read-across" or trend analyses. It should be noted, however, that the "read across" technique implies that endpoint information for an untested

chemical can be predicted from the existing results for a compound deemed "similar" in some way (*e.g.*, activity, property, or structure). This clearly is a very simplistic approach that is not always able to accurately predict real trends in the data. This method is also known to be prone to failure when the notorious "activity cliff" occurs.<sup>288</sup> Clearly, further improvements in regulatory QSAR models are urgently needed.

To improve and promote the use of non-test methods for REACH, the European Commission and the European Chemicals Agency (ECHA), in cooperation with the stakeholders, has developed guidance and practical guides on "QSARs and grouping of chemicals", "How to report QSAR", "How to report read-across and categories", and recently on "How to avoid unnecessary testing on animals". ECHA also coordinates, together with OECD, the continuous development of the "QSAR Toolbox".<sup>113</sup>

It is important to underscore that QSAR models supporting hazard or risk assessment must be relevant, reliable and sufficient for their intended purpose. It is one of the most important goals for a scientific community to develop a common agenda and approaches in the area of regulatory QSAR.

## 5. Conclusions: Best practices and the future of QSAR modeling

The intention of bringing together the present collection of perspectives and reports on QSAR modeling, each largely and separately authored by an active practitioner in the field, is to provide both a historical perspective, to convey where we have been and how we have arrived at the present, as well as to give the reader a flavor for the broadly diverse nature and applicability of QSAR practice, spanning a wide range of scientific disciplines and practical applications. A reader who has ventured this far will have detected a few seemingly disparate viewpoints, as well as some general, common themes that run throughout the various sections. We conclude by briefly elaborating further on a few of these areas of active (and healthy) disagreement within the QSAR community, as well as revisiting some larger common themes, touched on separately by many of the coauthors, for which there is broad consensus moving forward.

### Best practices in QSAR modeling

General guidelines pertaining to QSAR modeling practice, governing elements of model construction, reporting, validation, and use have emerged from years of scientific practice and experience. Several sub-sections of this report have directly addressed these challenges, from a traditional practitioner (Section 1.3) and methodological modeling workflow standpoint (Sections 2.1 and 2.4), as well as from the standpoint of establishing guidelines for QSAR model acceptance for use in safety assessment in a regulatory setting (Section 3.4). Clearly, best practices are essential to ensuring the overall integrity and validity of any QSAR modeling study and, if not adhered to, can negatively impact the entire field. As with any maturing scientific discipline, the development of harmonized rules, standards and common practices are exceedingly useful. Given the multidisciplinary and diverse nature of the QSAR enterprise, however, these rules and practices must also be sufficiently encompassing and flexible to accommodate the wide range of problems to which QSAR is applied. Discussion of best practices in this review has primarily centered on four elements

of a data workflow: (i) data collection and curation; (ii) model building; (iii) rigorous external validation using compounds that were not part of the modeling set; and (iv) model use. The latter can include prospective application to virtual screening and targeted molecular design of novel compounds or mixtures with desired properties, as well as application to toxicity screening and safety assessment within a regulatory setting. The approaches employed in a data workflow, possible pitfalls of modeling, and some strategies for avoiding common errors in QSAR model development were described in several recent reviews<sup>40,71,145,289</sup> and are highlighted in Sections 1.3 and 2.1 of this review.

Almost every cheminformatics lab has its own protocols for developing reliable QSAR models (*e.g.*, discussed in studies).<sup>151,159</sup> The workflow described by Tropsha<sup>145</sup> is recommended because of particular attention paid to rigorous internal and external cross-validation, estimation of AD, and the Y-scrambling test as necessary steps of model building. Consensus QSAR modeling is another highly recommended part of this workflow given increasing evidence that the quality of predictions and AD of consensus models are usually higher than for individual QSAR models.<sup>290</sup> However, Thomas et al.<sup>118</sup> have shown that, it is often impossible to build predictive models even when the most sophisticated algorithms and rigorous modeling workflows are employed. In response to this finding, Golbraikh et al.<sup>291</sup> recently introduced the concept of "*dataset modelability*", *i.e.*, an *a priori* estimate of the feasibility to obtain externally predictive QSAR models for a dataset of bioactive compounds. This concept has emerged from analyzing the effect of "activity cliffs", *i.e.*, very similar compounds with very different activities, on the overall performance of QSAR models. In his seminal publication, Maggiora<sup>288</sup> suggested that the presence of "activity cliffs" in a dataset is significantly challenging for QSAR modeling. Later, SALI<sup>292</sup> and ISAC<sup>293</sup> scores were developed for identifying activity cliffs based on ligand- and structure-based approaches, respectively. Excellent perspectives recently published in the Journal of Medicinal Chemistry<sup>294,295</sup> are pointing out many issues posed by activity cliffs for cheminformatics investigations<sup>294</sup> and cover related topic of molecular similarity.<sup>295</sup> "MODELability Index" (MODI)<sup>291</sup> was proposed not only as a quantitative tool to quickly estimate whether predictive QSAR model(s) can be obtained for a given binary dataset but also as an attempt to answer the following questions: (i) how the number of activity cliffs in a given dataset correlates with the overall prediction performance of QSAR models for this dataset; (ii) whether such correlation is conserved across different datasets; (iii) whether one could use the fraction of activity cliffs in a datasets to assess the overall possibility of success or failure for QSAR modeling; (iv) why some datasets are modelable whereas others are not; and (v) how (and whether it is possible at all) to find the subset of compounds in overall non-modelable dataset, for which local QSAR models can be obtained.

### Model validation and use

Rigorous external validation must be considered an integral part of model development. We separate these stages only to emphasize that a model should be externally validated using molecules which had no involvement in either model development or selection. The simplest approach is *n*-fold external cross-validation, where the entire dataset is randomly divided into *n* parts (folds) and each part is used as an external set for the model developed, and

internally validated with the remaining compounds. The situation when new experimental data for new compounds become available after the model was built (exemplified by study)<sup>290</sup> is even more preferable; it additionally strengthens external validation.

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 achieving this goal. Moreover, the study<sup>92</sup> illustrated that robust QSAR models could be used for initial experimental data curation, *i.e.*, to question true positives as well as to recover false negatives resulting from the high throughput screening campaigns.

Once a QSAR model is statistically validated, applicability domain (AD) estimation remains one of the most difficult and pressing challenges for QSAR modelers. Assessing the reliability of a QSAR for prospective predictions of properties or activities of new chemicals is a crucial adjunct to any model. The challenges and current approaches to defining the AD within a modeling workflow have been well described in several publications.<sup>87,145,296,297</sup> We wish to add that since the AD derives from the model and descriptors to which it is applied, it carries the inherent limitations of that model. However, the same principles of validation applied to assessing a QSAR model can be iteratively applied to any proposed definition of the AD to determine its utility, demonstrating for external predictions that the model actually performs better within the AD than outside the AD. As with QSAR model validation, the ultimate proof of utility is judged by external performance.

### Model Interpretability

The issue of QSAR model interpretability is explicitly considered in Section 2.5, from a methodological vantage point, but is discussed in other contexts throughout this review, most particularly in Sections 1.3, 2.2, 2.3, and 3.4. This issue is intimately tied to fundamental aspects of QSAR modeling, including selection of model descriptors and methods, and in some settings largely determines whether a model is ultimately deemed valid and useful. In Section 2.5, the goal of model interpretability is considered alongside the goal of achieving the best model performance and external predictivity by objective statistical measures, with a valid argument made that the latter should be the first and primary objective of any modeling enterprise. In the discussion of the OECD Principles, the point is also made that model (or mechanistic) interpretability is desirable, but is a secondary goal to demonstrated model validity.

Elsewhere in the text, particularly in the context of the regulatory applications of QSAR in toxicity modeling and safety assessment, the point is made that the ability to provide mechanistic support for model predictions is highly desirable and can add biological and chemical plausibility and weight-of-evidence to individual predictions. One could also argue that introducing prior knowledge into the initial selection of model descriptors or classifiers, which by definition lends interpretability, is particularly important in overcoming the perennial problem of insufficient data for robust modeling of complex toxicity *in vivo* endpoints, weak model statistics, and large uncertainties in applying models to new chemicals (see Section 2.2). Hence, interpretability and validated predictive power of QSAR models used for regulatory decision support can decrease the uncertainties of the toxicity endpoints predictions (see Section 2.2 and 3.4).

## Promoting best practices

Although good practices of QSAR modeling have been well described in the literature<sup>40,71,87,92,145,298</sup>, many models published even in the recent literature fail to adhere to these practices. Bajorath attributes this phenomenon to the relative ease of computational modeling and the possibility to carry out advanced calculations without critical assessment and understanding of their scientific foundations and limitations.<sup>299</sup> An additional factor is the highly multidisciplinary nature of the QSAR modeling enterprise: biologically oriented journals tend to place greater emphasis on the endpoint aspects of modeling studies and journal editors and reviewers tend to be less able to judge the quality of the QSAR modeling study from a methodological standpoint; by the same token, editors and reviewers of QSAR or cheminformatics oriented journals are less equipped to judge the biological appropriateness of the QSAR modeling study, in terms of the endpoint and data chosen for study.

In an effort to improve the quality of publications in the QSAR modeling field, the Journal of Chemical Information and Modeling published an editorial highlighting the requirements for QSAR papers that authors should follow to publish their results in the journal.<sup>298</sup> Accompanied by more rigorous editorial and peer review standards, such guidelines have significantly decreased the number of low-quality QSAR publications in top cheminformatics journals. Additional efforts, including publication of the OECD principles, and reviews published by various authors<sup>40,71,87,145,289</sup> have all contributed to reductions in the publication of low-quality QSAR papers. There is a need to develop a minimal set of standards that any QSAR study should follow to be accepted by leading peer-reviewed journals. The proliferation of high-quality QSAR models could be helped by placing online such models developed by following best practices; recent appearance of portals such as OCHEM (<http://ochem.eu>), ChemBench (<http://chembench.mml.unc.edu>), and emerging NIH BARD project (<https://bard.nih.gov/>) are steps in this direction.

## The continuing importance of QSAR

As with any scientific discipline, there have been some voices in the community questioning the viability and practical utility of QSAR modeling due to instances of poor model external performance, lax scientific practices, and the advent of newer biologically-based models built using HTS data. Paraphrasing the famous Mark Twain's quote, we are confident that the "reports of [QSAR] death are an exaggeration". In this paper we have openly discussed challenges faced by QSAR modeling and offered guidelines for developing rigorous and properly validated QSAR models that, if followed, afford multiple and diverse successful applications of QSAR, as discussed herein. The enormous, continuing growth of data in various molecular sciences, from medicinal chemistry to nearly any "omics" discipline, growing application of QSARs in regulatory decision making, emerging applications in materials (including nanomaterials) informatics suggest a growing importance of the QSAR approach to molecular data modeling. The developing trends on minimizing animal use in biomedical research place additional focus on QSAR as a source of alternative predictors of *in vivo* effects in both animals and humans. We hope that this paper will help both computational and experimental chemists to develop reliable QSAR models and to use these models to optimally exploit the experimental data to guide future studies. In addition, we

hope that the guidelines presented here will help journal editors and reviewers apply more stringent scientific standards to manuscripts reporting new QSAR studies, as well as encourage the use of high quality, validated QSARs to provide reliable support for experimental study design and regulatory decision making.

We conclude with a final nod to Hansch, who founded the modern practice of QSAR over 50 years ago. In the latter half of his long and productive career, he championed the idea of "comparative QSAR" and cofounded a company Biobyte (with A. Leo) to implement the approach of comparing related QSAR models to glean new insights into common mechanistic drivers for activity. Over a 30 year period, Hansch and coworkers compiled over 17,000 QSARs from the literature, with approximately half pertaining to biological systems and the other half to mechanistic organic chemistry, and published several studies illustrating the power of this approach to generate new insights (see review<sup>300</sup> and references therein). Decades before the term "cheminformatics" entered the lexicon of QSAR modelers, and likely without the knowledge or use of sophisticated machine-learning approaches, such as described here, Hansch foresaw the power of large, searchable databases and put "inductive knowledge transfer" into practice. The field of QSAR is infinitely indebted to Hansch for extending careful investigations grounded in physical organic chemistry principles to the applied discipline of QSAR widely practiced today, and to the many investigators who have followed in his footsteps. We believe that he would be gratified to see that QSAR continues as a vibrant scientific enterprise and is advancing and contributing to many scientific disciplines along the paths he originally laid forth.

## Acknowledgments

This review combined a series of separately written, invited contributions from the various coauthors (some sections with multiple coauthors). Primary attributions for the various contributed sections are as follows: 1.1 – Y. Martin; 1.2 – V. Consonni, R. Todeschini, and R. Cramer; 1.3 – J. Dearden and M. Cronin; 2.1 – D. Fourches, E. Muratov, and A. Tropsha; 2.2 – A. Benigni, C. Yang, J. Rathman, and A. Richard; 2.3 – L. Terfloth, J. Gasteiger; 2.4 – V. Kuz'min and E. Muratov; 2.5 – I. Baskin and A. Varnek; 2.6 – D. Fourches and A. Tropsha; 3.1 – A. Cherkasov; 3.2 – E. Muratov and V. Kuz'min; 3.3 – D. Fourches; 4. – P. Gramatica; and 5 – A. Tropsha. Figures were done by D. Fourches and E. Muratov. Substantial assistance in editing of all contributions was provided by A. Cherkasov and A. Richard, and final editing was accomplished by A. Tropsha, who also takes primary responsibility for the final content. The senior author also appreciates the support provided to his lab by NIH (grants GM096967 and GM066940) and EPA (grant RD 83499901). The first author acknowledges the support from the Natural Sciences and Engineering Research Council of Canada (Discovery Grants Program).

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency and other research institutions. Mentioning of trade names or commercial products does not constitute endorsement or recommendation for use.

## Biographies

Artem Cherkasov is an Associate Professor of the Faculty of Medicine, University of British Columbia, Vancouver, Canada. His research interests include computer-aided drug design, structure-activity modeling, drug reprofiling, and development of novel cheminformatics and bioinformatics tools. Dr. Cherkasov has authored and co-authored more than 200 articles in peer reviewed journals and conference proceedings and 5 book chapters. During his tenure at UBC, Dr. Cherkasov has been an applicant or co-applicant on a number of successful grants totaling over 40 million dollars, and participated in starting several spinoff companies. His research has been featured in various media sources including "Financial

Times", "New Scientist", "Toronto Star", "Georgia Straight" newspapers, "Montecristo" magazine and other media sources.

Eugene N. Muratov is a research assistant professor in the Eshelman School of Pharmacy, University of North Carolina, Chapel Hill and a senior researcher in the Department of Cheminformatics and Molecular Structure in the A. V. Bogatsky Physical-Chemical Institute of National Academy of Sciences (NAS) of Ukraine. He received his MS in technology of organic substances from Odessa National Polytechnic University in 2000 and PhD in organic chemistry in 2004 from the A. V. Bogatsky Physical-Chemical Institute. Before obtaining research assistant professor position in UNC in 2012, he spent several years as a researcher in Odessa, Jackson, Strasbourg, and Chapel Hill. His research interests are in the areas of cheminformatics (especially QSAR), computer-assisted drug design, antiviral research, and medicinal chemistry.

Denis Fourches, PhD is a Research Assistant Professor at the UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, USA. Prof. Fourches received his PhD in Chemistry in 2007 from the University of Strasbourg, France. He came to UNC in 2008 as a postdoctoral fellow in Alexander Tropsha's group. In 2010, he became a research assistant professor at UNC, and in 2011, he was promoted as the Associate Director of the Laboratory for Molecular Modeling. His research interests are in the areas of cheminformatics, computer-assisted molecular design, computational (nano)toxicology, material informatics, and structural bioinformatics.

Alexandre Varnek got his PhD in physical chemistry from the Institute of Inorganic and General Chemistry of the Russian Academy of Sciences, Moscow. In 1988–1995, he was Associate Professor in theoretical chemistry at the Moscow Mendeleyev University of Chemical Technology. In 1995, Alexandre joined the University of Strasbourg, France, where he holds the position of a Professor in theoretical chemistry, Head of the Laboratory on Cheminformatics and the Director of the master program in cheminformatics. His research interests focus on the development of new approaches and tools for virtual screening and on "in silico" design of new compounds and chemical reactions.

Igor Baskin received his PhD (kandidat nauk) (1990) and habilitation (doctor nauk) (2010) from Lomonosov Moscow State University, Russia. After holding several positions at the Semenov Institute of Chemical Physics and Zelinsky Institute of Organic Chemistry of the Russian Academy of Sciences, Moscow, he joined in 2001 Lomonosov Moscow State University, where since 2005 he holds the position of a Leading Research Associate. He is regularly engaged as Visiting Scientist and Invited Professor at the University of Strasbourg, France. He has published more than 140 articles related to SAR/QSAR/QSPR methodology, medicinal chemistry, and molecular modeling. Igor Baskin is a member of the International Academy of Mathematical Chemistry since 2009. His current work focuses on the application of advances machine learning approaches in cheminformatics.

Mark Cronin is Professor of Predictive Toxicology at Liverpool John Moores University, England. He has over 25 years expertise in the development of quantitative structure-activity relationships and the application of cheminformatics to provide solutions in product

development and risk assessment. Specifically, he has developed models to predict the human health effects and environmental toxicity of industrial chemicals, pharmaceuticals and cosmetics ingredients etc. Current work aims to develop strategies to predict toxicity from a firm mechanistic basis. This will allow for the grouping of chemicals based on so-called Adverse Outcome Pathways and robust read-across to fill data gaps. He has published over 190 papers and 4 books and worked on numerous collaborative projects in the area of in silico ADMET.

John Dearden is Emeritus Professor of Medicinal Chemistry in the School of Pharmacy and Biomolecular Sciences at Liverpool John Moores University, UK. His research includes the prediction of drug activity and toxicity, the prediction of environmental toxicity and fate, and the prediction of ADME properties, using the QSAR (quantitative structure-activity relationship) approach. He has published about 260 scientific papers and book chapters, mostly in the field of QSAR. He was elected to honorary membership of the Royal Pharmaceutical Society of Great Britain in 1993 in recognition of his QSAR research in pharmaceutical sciences. He was the recipient of the 2004 International QSAR Award in recognition of significant contributions to the advancement of QSARs in environmental sciences.

Paola Gramatica is full professor of Environmental Chemistry at the Department of Theoretical and Applied Sciences in Insubria University (Varese). She is leader of the QSAR Research Unit in Environmental Chemistry and Ecotoxicology. Since 1994, PG researches are on QSAR modeling and screening/ranking of organic compounds of environmental concern, mainly for PBT properties. The main interest is on validation of QSAR models; she was member of the OECD Task Force of QSAR experts for QSAR model validation for application in the REACH regulation and was actively involved in the definition of the OECD Principles on this topic. More recently, she was leader of the WP on QSAR model development for emerging pollutants in the EU-FP7 Project CADASTER. Web site: <http://www.qsar.it>

Yvonne Connolly Martin obtained a BA from Carleton College and a PhD from Northwestern University. Her whole career (retired 2006) was at Abbott Laboratories. She is the author of "Quantitative Drug Design" book (1978 and 2010) and 100+ articles including highly cited ones on pharmacophore detection, 3D database searching, and molecular similarity. She is editor of several books, Perspectives Editor of the Journal of Computer-Assisted Drug Design and a titular member of the IUPAC Division VII. She served on several NIH study sections, various editorial boards, and as chair of 1977 Gordon Conference on QSAR, and as chair of the QSAR Society. Honors include AAAS fellow; life-time member of the Molecular Graphics and Modelling Society; and the Herman Skolnik award of the ACS CINF Division.

Roberto Todeschini is full professor (since 2000) of chemometrics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy), where he constituted the Milano Chemometrics and QSAR Research Group. His main research activities concern chemometrics, Quantitative Structure-Activity Relationships (QSAR), molecular descriptors, multicriteria decision making and software development. President of

the International Academy of Mathematical Chemistry and president of the Italian Chemometric Society, he is author of more than 180 publications on international journals and of the books "The Data Analysis Handbook", by I.E. Frank and R. Todeschini; Elsevier, 1994, "Handbook of Molecular Descriptors", by R. Todeschini and V. Consonni; Wiley-VCH, 2000, and "Molecular Descriptors for Cheminformatics", by R. Todeschini and V. Consonni; Wiley-VCH, 2009.

Viviana Consonni received her PhD in chemical sciences from the University of Milano in 2000 and is now full researcher of chemometrics and cheminformatics at the Department of Earth and Environmental Sciences of the University of Milano-Bicocca (Milano, Italy). She is a member of the Milano Chemometrics and QSAR Research Group and has 15 years' experience in multivariate analysis, QSAR, molecular descriptors, and software development. She is co-author of more than 40 publications in peer-reviewed journals and the books "Handbook of Molecular Descriptors" (2000) and "Molecular Descriptors for Cheminformatics" (2009) by R. Todeschini and V. Consonni, Wiley-VCH. In 2006, she obtained the International Academy of Mathematical Chemistry Award for distinguished young investigators and, in June 2009, was elected as the youngest Member of the Academy.

Victor E. Kuz'min is a full 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.

Richard D. Cramer has been Sr. VP for Science at Tripos, and now Certara, since 1983. From 1971 to 1983, he founded and led SmithKline's CADD activity. He received an AB in Chemistry & Physics from Harvard in 1963 and a PhD in Physical Organic Chemistry from MIT in 1967. The other four years were as bench chemist at Polaroid and as post-doc with E.J. Corey at Harvard. Dick is the inventor of 3D-QSAR, as CoMFA; yet he believes his more recent topomer innovations will ultimately have greater significance for drug discovery. Dick is also a pioneer SABRmetrician (baseball "QSAR") and was a founder and the chief technical officer of STATS, Inc, as detailed within the books "Moneyball" and "Numbers Game".

Romualdo Benigni received his education in chemistry at the University of Rome "La Sapienza". He then joined the Istituto Superiore di Sanita' (Italian National Institute of Health) in 1977, and where he remained except for two sabbaticals (New York University, 1988; Jawaharlal Nehru University in New Delhi, 2000). He worked experimentally in the field of molecular biology and environmental chemical mutagenesis. In the 1980's, he turned his attention to the statistical analysis and modeling of toxicological data, and to the study of the relationships between the structure of organic compounds and their toxicological

properties (mainly mutagenesis and carcinogenesis). He has published over 180 journal articles and book chapters, applying a wide variety of quantitative analysis techniques, including QSAR, to the examination of chemical toxicity information.

Chihai Yang is the Managing Director of Molecular Networks, GmbH and the Chief Scientific Officer of Altamira LLC. She is also a visiting scientist at US FDA CFSAN where she has been a fellow in the computational toxicology program (2008 – 2011). She previously was Chief Scientific Officer of Leadscape, Inc. She joined Leadscape in 2000 from her position as a tenured chemistry professor at Otterbein University and an adjunct professor at the Ohio State University. At Leadscape she developed Leadscape's predictive toxicology platform and QSAR methodology. She also developed ToxML, a public toxicity database standard, and developed a ToxML database entry tool in collaboration with the US FDA. Her interests include methods to assist decisions in early discovery as well as in safety assessment.

Jim Rathman is a Professor of Chemical and Biomolecular Engineering at The Ohio State University. He is also a co-founder and Managing Director of Altamira, LLC, a company based in Columbus Ohio that provides consulting, database development, and software tools for molecular informatics applications, with a focus on computational risk assessment of complex chemical systems. His research areas of interest include chemical informatics, statistical data analysis and experimental design, and computational modeling and simulation. Jim previously spent seven years in industrial research and development with Conoco Inc. and The Clorox Company.

Lothar Terfloth, Ph.D., Senior Research Scientist, joined Molecular Networks GmbH in 2007 after his position as a research scientist in Prof. Dr. Gasteiger's group at the Computer-Chemistry-Center University Erlangen-Nuremberg. He has more than ten years of experience in the field of cheminformatics, the application of artificial neural networks and machine learning methods in drug discovery. His current research interest is directed towards the field of computational toxicology and metabolism prediction. Dr. Terfloth received his diploma in chemistry and his Ph.D. from the University of Munster working on the investigation of chiral organolithium compounds, their characterization by NMR spectroscopy as well as ab initio and DFT calculations.

Prof. Dr. Johann Gasteiger studied chemistry and has had positions at the University of Munich, Germany, University of California, Berkeley, USA, and Technical University of Munich. In 1994 he moved to the University of Erlangen-Nuremberg, Germany where he cofounded the "Computer-Chemie-Centrum". He is one of the initiators of Cheminformatics in Germany and has produced more than 250 scientific publications in this field. His work has been recognized by several awards, including the Gmelin–Beilstein Medal of the German Chemical Society, awards from the ACS Division of Chemical Information and of the ACS Division of Computers in Chemistry as well as of the Chemical Structure Association. In 1997 he founded Molecular Networks GmbH, a company developing and distributing software for chemical applications.

Ann Richard, PhD, is a Research Chemist within the Environmental Protection Agency's Office of Research & Development. She was a charter member and has worked for the past 7 years in EPA's National Center for Computational Toxicology (NCCT). Her research activities have ranged from the application of computational chemistry and SAR methods to problems in environmental toxicology, with a focus on predictive toxicology to, more recently, development of cheminformatics capabilities and knowledge-informed, feature-based approaches in support of predictive toxicology. Within NCCT, she is lead for the DSSTox project and chemical library information management for the ToxCast and Tox21 projects, working to provide a foundation for improved toxico-chemoinformatics and SAR capabilities in predictive toxicology.

Alexander Tropsha, PhD, is K.H. Lee Distinguished Professor and Associate Dean for Research at the UNC Eshelman School of Pharmacy, UNC-Chapel Hill. Prof. Tropsha received his PhD in Chemical Enzymology in 1986 from Moscow State University, Russia. He came to UNC-Chapel Hill in 1989 as a postdoctoral fellow, and eventually became a full professor in 2003. His research interests are in the areas of Computer-Assisted Drug Design, Computational Toxicology, Cheminformatics, and Structural Bioinformatics. He has authored or co-authored more than 170 peer-reviewed research papers, reviews and book chapters and co-edited two monographs. His research is supported by multiple grants from the NIH, NSF, EPA, ONR, and industry. He is a member of editorial boards of several scientific journals and scientific advisory panels.

## Abbreviations used

<b>AD</b>	applicability domain
<b>ADR</b>	adverse drug reaction
<b>AMP</b>	Antimicrobial Peptide
<b>ANN</b>	Artificial Neural Network statistical approach
<b>CoMFA</b>	Comparative Molecular Field Analysis
<b>CYP</b>	cytochrome P450
<b>FN</b>	Feature Net
<b>(M)NP</b>	(Manufactured) Nanoparticles
<b>MODI</b>	MODeability Index
<b>MTL</b>	Multi-Task Learning
<b>OECD</b>	Organization for Economic Cooperation and Development
<b>PLS</b>	Partial Least Square statistical approach
<b>Q<sup>2</sup></b>	cross-validated determination coefficient
<b>qHTS</b>	quantitative High-Throughput Screening
<b>QNAR</b>	Quantitative Nanostructure-Activity Relationship

<b>(Q)SAR</b>	(Quantitative) Structure-Activity Relationship
<b>(Q)SPR</b>	(Quantitative) Structure-Property Relationship
<b>R</b>	correlation coefficient
<b>R<sup>2</sup></b>	determination coefficient
<b>REACH</b>	Registration, Evaluation and Authorisation and Restriction of Chemicals
<b>RF</b>	Random Forest statistical approach
<b>SEA</b>	Similarity Ensemble Approach
<b>STL</b>	Single Task Learning
<b>SVM</b>	Support Vector Machine statistical approach
<b>UGT</b>	UDP-glucuronosyltransferase
<b>VS</b>	Virtual screening

## References

1. Hansch C, Maloney P, Fujita T, Muir R. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*. 1962; 194:178–180.
2. Cramer RD. The Inevitable QSAR Renaissance. *J. Comput. Aided. Mol. Des.* 2012; 26:35–38. [PubMed: 22127732]
3. Veldstra H. The Relation of Chemical Structure to Bio-Logical Activity in Growth Substances. *Annu. Rev. Plant Physiol.* 1953; 4:151–198.
4. Hansch C. Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* 1969; 2:232–239.
5. Collander R. The Partition of Organic Compounds Between Higher Alcohols and Water. *Acta Chem. Scand.* 1951; 5:774–780.
6. Fujita T, Imai S, Koshimizu K, Mitsui T, Kato I. Plant Growth Activities of 5- and 8-Halogeno-Dihydro- and -Tetrahydro-1-Naphthoic Acids. *Nature*. 1959; 184:1415–1416. [PubMed: 14440844]
7. Fujita T, Iwasa J, Hansch C. A New Substituent Constant, II, Derived from Partition Coefficients. *J. Am. Chem. Soc.* 1964; 86:5175–5180.
8. Jaffe HH. A Reexamination of the Hammett Equation. *Chem. Rev.* 1953; 53:191–261.
9. Taft, R. Separation of Polar, Steric, and Resonance Effects in Reactivity. In: Newman, M., editor. *Steric Effects in Organic Chemistry*. Wiley; New York: 1956. p. 556
10. Hogben CAM, Tocco DJ, Brodie BB, Schanker LS. On the Mechanism of Intestinal Absorption of Drugs. *J. Pharmacol. Exp. Ther.* 1959; 125:275–282. [PubMed: 13642268]
11. Overton, E. *Studien Über Die Narkose*. Jena: Gustav Fischer; 1901.
12. Meyer H. Zur Theorie Der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft Der Anästhetica Bedingt Ihre Narkotische Wirkung? *Arch. für Exp. Pathol. und pharmacologie*. 1899; 42:109–118.
13. Collander R. The Permeability of Plant Protoplasts to Non-Electrolytes. *Trans. faraday Soc.* 1937; 33:985.
14. Fieser L, Ettlinger M, Fawaz G. Naphthoquinone Antimalarials; Distribution Between Organic Solvents and Aqueous Buffers. *J. Am. Chem. Soc.* 1948; 70:3228–3232. [PubMed: 18891827]
15. Kauzmann W. Some Factors in the Interpretation of Protein Denaturation. *Adv. Protein Chem.* 1959; 14:1–63. [PubMed: 14404936]

16. Hansch C, Muir R, Fujita T, Maloney P, Geiger F, Streich M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* 1963; 85:2817–2824.
17. Hansch C, Fujita T. r-s-p Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* 1964; 86:1616–1626.
18. Craig PN. Interdependence Between Physical Parameters and Selection of Substituent Groups for Correlation Studies. *J. Med. Chem.* 1971; 14:680–684. [PubMed: 5114063]
19. Topliss JG. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* 1972; 15:1006–1011. [PubMed: 5069767]
20. Hansch C, Unger SH, Forsythe AB. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* 1973; 16:1217–1222. [PubMed: 4747964]
21. Leo A, Jow PYC, Silipo C, Hansch C. Calculation of Hydrophobic Constant (log P) from  $\pi$  and  $f$  Constants. *J. Med. Chem.* 1975; 18:865–868. [PubMed: 1159707]
22. Kubinyi H. Quantitative Structure-Activity Relations. 7. The Bilinear Model, a New Model for Nonlinear Dependence of Biological Activity on Hydrophobic Character. *J. Med. Chem.* 1977; 20:625–629. [PubMed: 857018]
23. Martin YC, Hackbarth JJ. Theoretical Model-Based Equations for the Linear Free Energy Relations of the Biological Activity of Ionizable Substances. 1. Equilibrium- Controlled Potency. *J. Med. Chem.* 1976; 19:1033–1039. [PubMed: 966250]
24. Klopman G. Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* 1984; 106:7315–7321.
25. Kutter E, Hansch C. Steric Parameters in Drug Design. Monoamine Oxidase Inhibitors and Antihistamines. *J. Med. Chem.* 1969; 12:647–652. [PubMed: 5793157]
26. Charton M. Nature of the Ortho Effect. II. Composition of the Taft Steric Parameters. *J. Am. Chem. Soc.* 1969; 91:615–618.
27. Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In: Ariens, EJ., editor. *Drug Design*. New York: Academic Press; 1976. p. 165-207.
28. Cramer RD III, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* 1988; 110:5959–5967. [PubMed: 22148765]
29. Wold S, Ruhe A, Wold H, Dunn I. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM. J. Sci. Stat. Comput.* 1984; 5:735–743.
30. Martin YC, Lynn KR. Quantitative Structure-Activity Relations in Leucomycin and Lincomycin Antibiotics. *J. Med. Chem.* 1971; 14:1162–1166. [PubMed: 5000561]
31. Klopman G, Wang S. A Computer Automated Structure Evaluation (CASE) Approach to Calculation of Partition Coefficient. *J. Comput. Chem.* 1991; 12:1025–1032.
32. Hall LH, Kier LB. Issues in Representation of Molecular Structure. *J. Mol. Graph. Model.* 2001; 20:4–18. [PubMed: 11760002]
33. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* 2002; 42:1273–1280. [PubMed: 12444722]
34. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* 2010; 50:742–754. [PubMed: 20426451]
35. Martin YC, Holland JB, Jarboe CH, Plotnikoff N. Discriminant Analysis of the Relation Between Physical Properties and the Inhibition of Monoamine Oxidase by Aminotetralins and Aminoindans. *J. Med. Chem.* 1974; 17:409–413. [PubMed: 4830537]
36. Hawkins DM, Young SS, Rusinko A. Analysis of a Large Structure-Activity Data Set Using Recursive Partitioning. *Quant. Struct. Relationships.* 1997; 16:296–302.
37. Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/nondrug Classification. *J. Chem. Inf. Comput. Sci.* 2003; 43:1882–1889. [PubMed: 14632437]

38. Ajay A, Walters WP, Murcko MA. Can We Learn to Distinguish Between “Drug-Like ” and “Nondrug-Like” Molecules? *J. Med. Chem.* 1998; 41:3314–3324. [PubMed: 9719583]
39. Topliss JG, Edwards RP. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* 1979; 22:1238–1244. [PubMed: 513071]
40. Golbraikh A, Tropsha A. Beware of Q2! *J. Mol. Graph. Model.* 2002; 20:269–276. [PubMed: 11858635]
41. Mekenyan OG, Dimitrov SD, Pavlov TS, Veith GD. A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework. *Curr. Pharm. Des.* 2004; 10:1273–1293. [PubMed: 15078141]
42. Polanski J. Receptor Dependent Multidimensional QSAR for Modeling Drug–Receptor Interactions. *Curr. Med. Chem.* 2009; 16:3243–3257. [PubMed: 19548875]
43. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*. Mannhold, R.; Kubinyi, H.; Folkers, G., editors. Weinheim: Wiley-VCH; 2009. p. 1257
44. Ivanciuc O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Model.* 2000; 40:1412–1422.
45. Moreau G, Broto P. Autocorrelation of Molecular Structures. Application to SAR Studies. *Nouv. J. Chim.* 1980; 4:757–764.
46. Consonni, V.; Todeschini, R. Structure - Activity Relationships by Autocorrelation Descriptors and Genetic Algorithms. In: Lohdi, H.; Yamanishi, Y., editors. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. Hershey: IGI Global Publishers; 2010. p. 60-93.
47. Randić M. Generalized Molecular Descriptors. *J. Math. Chem.* 1991; 7:155–168.
48. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi F. Virtual Screening for R-Groups, Including Predicted pIC50 Contributions, Within Large Structural Databases, Using Topomer CoMFA. *J. Chem. Inf. Model.* 2008; 48:2180–2195. [PubMed: 18956863]
49. Doweiko AM. 3D-QSAR Illusions. *J. Comput. Aided. Mol. Des.* 2004; 18:587–596. [PubMed: 15729857]
50. Klebe G, Abraham U, Mietzner T. Molecular Similarity Indexes in a Comparative- Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological- Activity. *J. Med. Chem.* 1994; 37:4130–4146. [PubMed: 7990113]
51. Goodford PJ. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* 1985; 28:849–857. [PubMed: 3892003]
52. Pastor M, Cruciani G, Mclay I, Pickett S, Clementi S. GRIND-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* 2000; 43:3233–3243. [PubMed: 10966742]
53. Duran A, Zamora I, Pastor M. Suitability of GRIND-Based Principal Properties for the Description of Molecular Similarity and Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* 2009; 49:2129–2138. [PubMed: 19728739]
54. Low CMR, Vinter JG. Rationalizing the Activities of Diverse Cholecystokinin 2 Receptor Antagonists Using Molecular Field Points. *J. Med. Chem.* 2008; 51:565–573. [PubMed: 18201065]
55. Dixon SL, Smondyrev AM, Rao SN. PHASE: a Novel Approach to Pharmacophore Modeling and 3D Database Searching. *Chem. Biol. Drug Des.* 2006; 67:370–372. [PubMed: 16784462]
56. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* 1995; 38:2681–2691. [PubMed: 7629807]
57. Gohlke H, Klebe G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* 2002; 45:4153–4170. [PubMed: 12213058]
58. Doweiko AM. The Hypothetical Active Site Lattice. An Approach to Modelling Active Sites from Data on Inhibitor Molecules. *J. Med. Chem.* 1988; 31:1396–1406. [PubMed: 3290487]

59. Simon Z, Badilescu I, Racovitan T. Mapping of Dihydrofolate-Reductase Receptor Site by Correlation with Minimal Topological (steric) Differences. *J. Theor. Biol.* 1977; 66:485–495. [PubMed: 886879]
60. Varela R, Walters WP, Goldman BB, Jain AN. Iterative Refinement of a Binding Pocket Model: Active Computational Steering of Lead Optimization. *J. Med. Chem.* 2012; 55:8926–8942. [PubMed: 23046104]
61. Jilek RJ, Cramer RD. Topomers: a Validated Protocol for Their Self-Consistent Generation. *J. Chem. Inf. Comput. Sci.* 2004; 44:1221–1227. [PubMed: 15272829]
62. Cramer RD. R-Group Template CoMFA Combines Benefits of “Ad Hoc” and Topomer Alignments Using 3D-QSAR for Lead Optimization. *J. Comput. Aided. Mol. Des.* 2012; 26:805–819. [PubMed: 22661224]
63. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RD. “Lead Hopping”. Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* 2004; 47:6777–6791. [PubMed: 15615527]
64. Nisius B, Goller AH. Similarity-Based Classifier Using Topomers to Provide a Knowledge Base for hERG Channel Inhibition. *J. Chem. Inf. Model.* 2009; 49:247–256. [PubMed: 19434826]
65. Cramer RD. Rethinking 3D-QSAR. *J. Comput. Aided. Mol. Des.* 2011; 25:197–201. [PubMed: 21110063]
66. Wendt B, Mulbaier M, Wawro S, Schultes C, Alonso J, Janssen B, Lewis J. Toluidinesulfonamide Hypoxia-Induced Factor 1 Inhibitors: Alleviating Drug-Drug Interactions through Use of PubChem Data and Comparative Molecular Field Analysis Guided Synthesis. *J. Med. Chem.* 2011; 54:3982–3986. [PubMed: 21574568]
67. Tresadern G, Bemporad D, Howe T. A Comparison of Ligand Based Virtual Screening Methods and Application to Corticotropin Releasing Factor 1 Receptor. *J. Mol. Graph. Model.* 2009; 27:860–870. [PubMed: 19230731]
68. Wendt B, Uhrig U, Bos F. Capturing Structure-Activity Relationships from Chemogenomic Spaces. *J. Chem. Inf. Model.* 2011; 51:843–851. [PubMed: 21410249]
69. Walker JD, Jaworska J, Comber MH, Schultz TW, Dearden JC. Guidelines for Developing and Using Quantitative Structure-Activity Relationships. *Environ. Toxicol. Chem.* 2003; 22:1653–1665. [PubMed: 12924568]
70. Cronin MTD, Schultz TW. Pitfalls in QSAR. *J. Mol. Struct. THEOCHEM.* 2003; 622:39–51.
71. Dearden JC, Cronin MT, Kaiser KL. How Not to Develop a Quantitative Structure- Activity or Structure-Property Relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* 2009; 20:241–266. [PubMed: 19544191]
72. Chirico N, Gramatica P. Real External Predictivity of QSAR Models: How to Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.* 2011; 51:2320–2335. [PubMed: 21800825]
73. Gramatica, P. On the Development and Validation of QSAR Models. In: Reisfeld, B.; Mayeno, A., editors. *Computational Toxicology*. New York: Springer; 2013. p. 499-526.
74. Hansch, C.; Leo, A.; Hoekman, D. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. Heller, SR., editor. Washington, DC: ACS; 1995. p. 455
75. Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, Gribaldo L, Halder M, Hoffmann S, Roi AJ, Prieto P, Sabbioni E, Scott L, Worth A, Zuang V. A Modular Approach to the ECVAM Principles on Test Validity. *Altern. to Lab. Anim.* 2004; 32:467–472.
76. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJ, Tong W, Veith G, Yang C. Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern. to Lab. Anim.* 2005; 33:155–173.
77. Flynn, GL. Physicochemical Determinants of Skin Absorption. In: Gerrity, TR.; Henry, CJ., editors. *Principles of Route-to- Route Extrapolation for Risk assessment*. Elsevier; New York, NY: 1990. p. 93-117.
78. Young D, Martin D, Venkatapathy R, Harten P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* 2008; 27:1337–1345.

79. Coats, E. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. In: Kubinyi, H.; Folkers, G.; Martin, YC., editors. 3D QSAR in Drug Design. Vol. Vol. 3. Dordrecht: Kluwer Academic Publishers; 2002. p. 199-213.
80. Gedeck P, Rohde B, Bartels C. QSAR--How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* 2006; 46:1924–1936. [PubMed: 16995723]
81. Romanelli GP, Cafferata LFR, Castro EA. An Improved QSAR Study of Toxicity of Saturated Alcohols. *J. Mol. Struct. THEOCHEM.* 2000; 504:5.
82. Oprea, T. 3D QSAR Modeling in Drug Design. In: Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere, J., editors. Computational Medicinal Chemistry for Drug Discovery. New York: Marcel Dekker; 2004. p. 571-616.
83. Ghasemi J, Saaidpour S. QSPR Prediction of Aqueous Solubility of Drug-Like Organic Compounds. *Chem. Pharm. Bull. (Tokyo).* 2007; 55:669–674. [PubMed: 17409570]
84. Yaffe D, Cohen Y, Espinosa G, Arenas A, Giralt F. A Fuzzy ARTMAP Based on Quantitative Structure-Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* 2001; 41:1177–1207. [PubMed: 11604019]
85. Kuz'min VE, Muratov EN, Artemenko AG, Varlamova EV, Gorb L, Wang J, Leszczynski J. Consensus QSAR Modeling of Phosphor-Containing Chiral AChE Inhibitors. *QSAR Comb. Sci.* 2009; 28:664–677.
86. Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* 2012; 52:2570–2578. [PubMed: 23030316]
87. Tropsha A, Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* 2007; 13:3494–3504. [PubMed: 18220786]
88. Sheridan RP. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* 2013; 53:783–790. [PubMed: 23521722]
89. Organisation for Economic Co-operation and Development OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models. <http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf> [(accessed Sep 12, 2013)]
90. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Oprea, TI. WOMBAT: World of Molecular Bioactivity. In: Oprea, TI., editor. Chemoinformatics in Drug Discovery. New York: Wiley-VCH; 2005. p. 223-239.
91. Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, TI. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In: Schreiber, SL.; Kapoor, TM.; Weiss, G., editors. Chemical Biology: From Small Molecules to Systems Biology and Drug Design. Wiley-VCH; Weinheim: 2007. p. 760-786.
92. Fourches D, Muratov E, Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* 2010; 50:1189–1204. [PubMed: 20572635]
93. Olah MM, Bologa CG, Oprea TI. Strategies for Compound Selection. *Curr. Drug Discov. Technol.* 2004; 1:211–220. [PubMed: 16472248]
94. Tiikkainen P, Franke L. Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* 2012; 52:319–326. [PubMed: 22145975]
95. Williams AJ, Ekins S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discov. Today.* 2011; 16:747–750. [PubMed: 21871970]
96. Southan C, Varkonyi P, Muresan S. Quantitative Assessment of the Expanding Complementarity Between Public and Commercial Databases of Bioactive Compounds. *J. Cheminform.* 2009; 1:10–11. [PubMed: 20298516]
97. Ariens EJ. Domestication of Chemistry by Design of Safer Chemicals: Structure- Activity Relationships. *Drug Metab. Rev.* 1984; 15:425–504. [PubMed: 6386409]
98. Hansch C, Hoekman D, Leo A, Zhang L, Li P. The Expanding Role of Quantitative Structure-Activity Relationships (QSAR) in Toxicology. *Toxicol. Lett.* 1995; 79:45–53. [PubMed: 7570673]
99. Hansch C, Kim D, Leo AJ, Novellino E, Silipo C, Vittoria A. Toward a Quantitative Comparative Toxicology of Organic Compounds. *Crit. Rev. Toxicol.* 1989; 19:185–226. [PubMed: 2653732]

100. Votano J, Parham M, Hall L, Kier L, Oloff S, Tropsha A, Xie Q, Tong W. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis*. 2005; 19:365–377. [PubMed: 15388809]
101. Milan C, Schifanella O, Roncaglioni A, Benfenati E. Comparison and Possible Use of in Silico Tools for Carcinogenicity Within REACH Legislation. *J. Environ. Sci. Health. C. Environ. Carcinog. Ecotoxicol. Rev.* 2011; 29:300–323. [PubMed: 22107165]
102. Van Leeuwen K, Schultz TW, Henry T, Diderich B, Veith GD. Using Chemical Categories to Fill Data Gaps in Hazard Assessment. *SAR QSAR Environ. Res.* 2009; 20:207–220. [PubMed: 19544189]
103. Patlewicz G, Roberts DW, Uriarte E. A Comparison of Reactivity Schemes for the Prediction Skin Sensitization Potential. *Chem. Res. Toxicol.* 2008; 21:521–541. [PubMed: 18189364]
104. Benigni R. Alternatives to the Carcinogenicity Bioassay for Toxicity Prediction: Are We There Yet? *Expert Opin. Drug Metab. Toxicol.* 2012; 8:407–417. [PubMed: 22360376]
105. [(accessed Mar 13, 2013)] >ToxRefDB. <http://actor.epa.gov/toxrefdb/>
106. [(accessed Mar 13, 2013)] EPA Aggregated Computational Toxicology Resource (ACTOR). <http://actor.epa.gov/>
107. EPA, U. [(accessed Mar 13, 2013)] Distributed Structure-Searchable Toxicity (DSSTox) Database. <http://www.epa.gov/ncct/dsstox/>
108. Istituto Superiore di Sanita, ISSTOX Chemical Toxicity Databases. <http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7> [(accessed Mar 13, 2013)]
109. eTOX Project. <http://www.etoxproject.eu/> [(accessed Sep 2, 2013)]
110. [(accessed Sep 2, 2013)] NTP DrugMatrix. <https://ntp.niehs.nih.gov/drugmatrix/index.html>
111. ToxTree. [http://ihcp.jrc.ec.europa.eu/our\\_labs/predictive\\_toxicology/qsar\\_tools/toxtree](http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/toxtree) [(accessed Mar 13, 2013)]
112. OpenTox. <http://www.opentox.org/> [(accessed Mar 13, 2013)]
113. OECD QSAR Toolbox. [http://www.oecd.org/document/54/0,3746,en\\_2649\\_34379\\_42923638\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/54/0,3746,en_2649_34379_42923638_1_1_1_1,00.html) [(accessed Mar 13, 2013)]
114. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* 2007; 95:5–12. [PubMed: 16963515]
115. Kavlock RJ, Austin CP, Tice RR. Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment. *Risk Anal.* 2009; 29:485–497. [PubMed: 19076321]
116. Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, Reif DM, Rotroff DM, Shah I, Richard AM, Dix DJ. In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environ. Health Perspect.* 2010; 118:485–492. [PubMed: 20368123]
117. Benigni R, Bossa C, Giuliani A, Tcheremenskaia O. Exploring in Vitro/in Vivo Correlation: Lessons Learned from Analyzing Phase I Results of the US EPA's ToxCast Project. *J. Environ. Sci. Health. C. Environ. Carcinog. Ecotoxicol. Rev.* 2010; 28:272–286. [PubMed: 21069615]
118. Thomas RS, Black MB, Li L, Healy E, Chu T-M, Bao W, Andersen ME, Wolfinger RD. A Comprehensive Statistical Analysis of Predicting in Vivo Hazard Using High-Throughput in Vitro Screening. *Toxicol. Sci.* 2012; 128:398–417. [PubMed: 22543276]
119. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, Richard A, Rotroff D, Sipes N, Dix D. Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chem. Res. Toxicol.* 2012; 25:1287–1302. [PubMed: 22519603]
120. Martin MT, Knudsen TB, Reif DM, Houck KA, Judson RS, Kavlock RJ, Dix DJ. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biol. Reprod.* 2011; 85:327–339. [PubMed: 21565999]
121. Zhu H, Rusyn I, Richard A, Tropsha A. Use of Cell Viability Assay Data Improves the Prediction Accuracy of Conventional Quantitative Structure-Activity Relationship Models of Animal Carcinogenicity. *Environ. Health Perspect.* 2006; 116:506–513. [PubMed: 18414635]

122. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* 2011; 119:364–370. [PubMed: 20980217]
123. Chandler KJ, Barrier M, Jeffay S, Nichols HP, Kleinstreuer NC, Singh AV, Reif DM, Sipes NS, Judson RS, Dix DJ, Kavlock R, Hunter ES, Knudsen TB. Evaluation of 309 Environmental Chemicals Using a Mouse Embryonic Stem Cell Adherent Cell Differentiation and Cytotoxicity Assay. *PLoS One.* 2011; 6:e18540. [PubMed: 21666745]
124. Yang, C.; Richard, AM.; Arvidson, KB.; Worth, AP. The 51st Society of Toxicology Annual Meeting. San Francisco, CA: 2012. A Mode-of-Action-Based QSAR Approach to Improve Understanding of Developmental Toxicity. Abstract 1792
125. Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, Hulzebos E, Marchant C, Richard A, Woo Y-T, Yang C. The Expanding Role of Predictive Toxicology: An Update on the (Q)SAR Models for Mutagens and Carcinogens. *J. Environ. Sci. Health. C. Environ. Carcinog. Ecotoxicol. Rev.* 2007; 25:53–97. [PubMed: 17365342]
126. [(accessed Mar 13, 2013)] NLM Dream Anatomy. [http://www.nlm.nih.gov/dreamanatomy/da\\_g\\_IV-A-01.html](http://www.nlm.nih.gov/dreamanatomy/da_g_IV-A-01.html)
127. Testa B, Pedretti A, Vistoli G. Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discov. Today.* 2012; 17:549–560. [PubMed: 22305937]
128. Accelrys Metabolite. <http://accelrys.com/products/databases/bioactivity/metabolite.html> [(accessed Mar 13, 2013)]
129. Accelrys Metabolism. <http://accelrys.com/products/datasheets/metabolism.pdf> [(accessed Mar 13, 2013)]
130. Fujitsu ADME DB. [http://www.fqs.pl/chemistry\\_materials\\_life\\_science/products/adme\\_db](http://www.fqs.pl/chemistry_materials_life_science/products/adme_db) [(accessed Mar 13, 2013)]
131. Rendic S, Di Carlo FJ. Human Cytochrome P450 Enzymes: a Status Report Summarizing Their Reactions, Substrates, Inducers, and Inhibitors. *Drug Metab. Rev.* 1997; 29:413–580. [PubMed: 9187528]
132. Rendic S. Summary of Information on Human CYP Enzymes: Human P450 Metabolism Data. *Drug Metab. Rev.* 2002; 34:83–448. [PubMed: 11996015]
133. Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A, Glen RC. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* 2012; 52:617–648. [PubMed: 22339582]
134. Chohan KK, Paine SW, Waters NJ. Quantitative Structure Activity Relationships in Drug Metabolism. *Curr. Top. Med. Chem.* 2006; 6:1569–1578. [PubMed: 16918469]
135. Braga RC, Andrade CH. QSAR and QM/MM Approaches Applied to Drug Metabolism Prediction. *Mini Rev. Med. Chem.* 2012; 12:573–582. [PubMed: 22587770]
136. Marchant CA, Briggs KA, Long A. In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mech. Methods.* 2008; 18:177–187. [PubMed: 20020913]
137. Klopman G, Dimayuga M, Talafous J. META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *J. Chem. Inf. Comput. Sci.* 1994; 34:1320–1325. [PubMed: 7989397]
138. Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T, Vianello R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* 2005; 48:6970–6979. [PubMed: 16250655]
139. Berellini G, Cruciani G, Mannhold R. Pharmacophore, Drug Metabolism, and Pharmacokinetics Models on Non-Peptide AT1, AT2, and AT1/AT2 Angiotensin II Receptor Antagonists. *J. Med. Chem.* 2005; 48:4389–4399. [PubMed: 15974591]
140. Zamora I, Afzelius L, Cruciani G. Predicting Drug Metabolism: a Site of Metabolism Prediction Tool Applied to the Cytochrome P450 2C9. *J. Med. Chem.* 2003; 46:2313–2324. [PubMed: 12773036]
141. Boyer S, Arnby CH, Carlsson L, Smith J, Stein V, Glen RC. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.* 2007; 47:583–590. [PubMed: 17302400]

142. Carlsson L, Spjuth O, Adams S, Glen RC, Boyer S. Use of Historic Metabolic Biotransformation Data as a Means of Anticipating Metabolic Sites Using MetaPrint2D and Bioclipse. *BMC Bioinformatics*. 2010; 11:362. [PubMed: 20594327]
143. Rydberg P, Gloriam DE, Olsen L. The SMARTCyp Cytochrome P450 Metabolism Prediction Server. *Bioinformatics*. 2010; 26:2988–2989. [PubMed: 20947523]
144. METIS - Metabolic Information Input System. <http://www.molecularnetworks.com/products/metis> [(accessed Mar 13, 2013)]
145. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* 2010; 29:476–488.
146. Guha R. On the Interpretation and Interpretability of Quantitative Structure-Activity Relationship Models. *J. Comput. Aided. Mol. Des.* 2008; 22:857–871. [PubMed: 18784976]
147. Group QE. The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs. *Organ. Econ. Cooperation Dev. Paris*. 2004; 49:1–206.
148. Artemenko A, Muratov E, Kuz'min V, Kovdienko N, Hromov A, Makarov V, Riabova O, Wutzler P, Schmidtke M. Identification of Individual Structural Fragments of N,N'-(bis-5-Nitropyrimidyl)dispirotriperazine Derivatives for Cytotoxicity and Antiherpetic Activity Allows the Prediction of New Highly Active Compounds. *J. Antimicrob. Chemother.* 2007; 60:68–77. [PubMed: 17550890]
149. Kuz'min VE, Artemenko AG, Muratov EN, Volineckaya IL, Makarov Va, Riabova OB, Wutzler P, Schmidtke M. Quantitative Structure-Activity Relationship Studies of [(biphenyloxy)propyl]isoxazole Derivatives. Inhibitors of Human Rhinovirus 2 Replication. *J. Med. Chem.* 2007; 50:4205–4213. [PubMed: 17665898]
150. Kuz'min VE, Artemenko AG, Lozytska RN, Fedtchouk AS, Lozitsky VP, Muratov EN, Mescheriakov AK. Investigation of Anticancer Activity of Macrocyclic Schiff Bases by Means of 4D-QSAR Based on Simplex Representation of Molecular Structure. *SAR QSAR Environ. Res.* 2005; 16:219–230. [PubMed: 15804810]
151. Lagunin AA, Zakharov AV, Filimonov DA, Poroikov VV. A New Approach to QSAR Modelling of Acute Toxicity. *SAR QSAR Environ. Res.* 2007; 18:285–298. [PubMed: 17514571]
152. Kokurkina GV, Dutov MD, Shevelev SA, Popkov SV, Zakharov AV, Poroikov V. V Synthesis, Antifungal Activity and QSAR Study of 2- Arylhydroxynitroindoles. *Eur. J. Med. Chem.* 2011; 46:4374–4382. [PubMed: 21802177]
153. Lagunin A, Zakharov A, Filimonov D, Poroikov V. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction. *Mol. Inform.* 2011; 30:241–250.
154. Franke R. On the Interpretability of Quantitative Structure-Activity Relationships (QSAR). *Farmaco. Sci.* 1979; 34:545–570. [PubMed: 467631]
155. Bender A, Mussa HY, Glen RC, Reiling S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* 2004; 44:1708–1718. [PubMed: 15446830]
156. Bremser W. Hose — a Novel Substructure Code. *Anal. Chim. Acta.* 1978; 103:355–365.
157. Ajmani S, Jadhav K, Kulkarni SA. Group-Based QSAR (G-QSAR): Mitigating Interpretation Challenges in QSAR. *QSAR Comb. Sci.* 2009; 28:36–51.
158. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko I, Marcou G. ISIDA-Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided Drug Des.* 2008; 4:191–198.
159. Kuz'min VE, Artemenko AG, Muratov EN. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. *J. Comput. Aided. Mol. Des.* 2008; 22:403–421. [PubMed: 18253701]
160. Garcia-Domenech R, Galvez J, de Julian-Ortiz JV, Pogliani L. Some New Trends in Chemical Graph Theory. *Chem. Rev.* 2008; 108:1127–1169. [PubMed: 18302420]
161. Kuz'min VE, Konovorotskii YP. Relationship Between the Structural and Topological Characteristics of Molecules. *J. Struct. Chem.* 1986; 26:498–506.

162. Guha R, Jurs PC. Development of QSAR Models to Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* 2004; 44:1440–1449. [PubMed: 15272852]
163. Larsen SB, Jorgensen FS, Olsen L. QSAR Models for the Human H(+)/peptide Symporter, hPEPT1: Affinity Prediction Using Alignment-Independent Descriptors. *J. Chem. Inf. Model.* 2008; 48:233–241. [PubMed: 18092768]
164. Kuz'min VE, Muratov EN, Artemenko AG, Gorb L, Qasim M, Leszczynski J. The Effects of Characteristics of Substituents on Toxicity of the Nitroaromatics: HiT QSAR Study. *J. Comput. Aided. Mol. Des.* 2008; 22:747–759. [PubMed: 18385948]
165. Wold S, Sjostrom M, Eriksson L. PLS-Regression: a Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* 2001; 58:109–130.
166. Rosenbaum L, Hinselmann G, Jahn A, Zell A. Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring. *J. Cheminform.* 2011; 3:11. [PubMed: 21439031]
167. Guha R, Stanton DT, Jurs PC. Interpreting Computational Neural Network Quantitative Structure-Activity Relationship Models: a Detailed Interpretation of the Weights and Biases. *J. Chem. Inf. Model.* 2005; 45:1109–1121. [PubMed: 16045306]
168. Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA. Interpretation of QSAR Models Based on Random Forest Methods. *Mol. Inform.* 2011; 30:593–603.
169. Breiman LEO. Random Forests. *Mach. Learn.* 2001; 45:5–32.
170. Guha R, Jurs PC. Interpreting Computational Neural Network QSAR Models: a Measure of Descriptor Importance. *J. Chem. Inf. Model.* 2005; 45:800–806. [PubMed: 15921469]
171. Carlsson L, Helgee EA, Boyer S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* 2009; 49:2551–2558. [PubMed: 19824682]
172. Baskin II, Ait AO, Halberstam NM, Palyulin VA, Zefirov NS. An Approach to the Interpretation of Backpropagation Neural Network Models in QSAR Studies. *SAR QSAR Environ. Res.* 2002; 13:35–41. [PubMed: 12074390]
173. Marcou G, Horvath D, Solov'ev V, Arrault A, Vayer P, Varnek A. Interpretability of SAR/QSAR Models of Any Complexity by Atomic Contributions. *Mol. Inform.* 2012; 31:639–642.
174. Polishchuk P, Kuz'min V, Artemenko A, Muratov E. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inform.* 2013; 32:843–853.
175. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko I. V Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J.Chem.Inf.Model.* 2009; 49:133–144. [PubMed: 19125628]
176. Caruana R. Multitask Learning. *Mach. Learn.* 1997; 28:41–75.
177. Evgeniou T, Micchelli C, Pontil M. Learning Multiple Tasks with Kernel Methods. *J. Mach. Learn. Res.* 2005; 6:615–636.
178. Baskin II, Zhokhova NI, Palyulin VA, Zefirov AN, Zefirov NS. Multilevel Approach to the Prediction of Properties of Organic Compounds in the Framework of the QSAR/QSPR Methodology. *Dokl. Chem.* 2009; 427:172–175.
179. Erhan D, L'heureux P-J, Yue SY, Bengio Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* 2006; 46:626–635. [PubMed: 16562992]
180. Ning X, Rangwala H, Karypis G. Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets. *J. Chem. Inf. Model.* 2009; 49:2444–2456. [PubMed: 19842624]
181. Ning X, Karypis G. In Silico Structure-Activity-Relationship (SAR) Models From Machine Learning: A Review. *Drug Dev. Res.* 2011; 72:138–146.
182. Oprea TI, Tropsha A, Faulon J-LL, Rintoul MD. Systems Chemical Biology. *Nat. Chem. Biol.* 2007; 3:447–450. [PubMed: 17637771]
183. Reddy S, Pati P, Kumar P, Pradeep N, Sastry N. Virtual Screening in Drug Discovery -- a Computational Perspective. *Curr. Protein Pept. Sci.* 2007; 8:329–351. [PubMed: 17696867]
184. Frye S, Crosby M, Edwards T, Juliano R. US Academic Drug Discovery. *Nat. Rev. Drug Discov.* 2011; 10:409–410. [PubMed: 21629285]

185. Xie X-Q. Exploiting PubChem for Virtual Screening. *Expert Opin. Drug Discov.* 2010; 5:1205–1220. [PubMed: 21691435]
186. Irwin JJ, Shoichet BK. ZINC--a Free Database of Commercially Available Compounds for Virtual Screening. *J.Chem.Inf.Model.* 2005; 45:177–182. [PubMed: 15667143]
187. Fara DC, Oprea TI, Prossnitz ER, Bologna CG, Edwards BS, Sklar LA. Integration of Virtual and Physical Screening. *Drug Discov. Today Technol.* 2006; 3:377–385.
188. Leeson P. Drug Discovery: Chemical Beauty Contest. *Nature.* 2012; 481:455–456. [PubMed: 22281594]
189. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the Chemical Beauty of Drugs. *Nat. Chem. Biol.* 2012; 4:90–98.
190. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL. Predicting New Molecular Targets for Known Drugs. *Nature.* 2009; 462:175–181. [PubMed: 19881490]
191. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK, Urban L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature.* 2012; 486:361–367. [PubMed: 22722194]
192. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P, Norval S, Sassano MF, Shin AI, Webster LA, Simeons FRC, Stojanovski L, Prat A, Seidah NG, Constam DB, Bickerton GR, Read KD, Wetsel WC, Gilbert IH, Roth BL, Hopkins AL. Automated Design of Ligands to Polypharmacological Profiles. *Nature.* 2012; 492:215–220. [PubMed: 23235874]
193. Hajjo R, Grulke CM, Golbraikh A, Setola V, Huang X-P, Roth BL, Tropsha A. Development, Validation, and Use of Quantitative Structure-Activity Relationship Models of 5-Hydroxytryptamine (2B) Receptor Ligands to Identify Novel Receptor Binders and Putative Valvulopathic Compounds Among Common Drugs. *J. Med. Chem.* 2010; 53:7573–7586. [PubMed: 20958049]
194. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, Clark J, Connelly MC, Sigal M, Hodges D, Guiguemde A, Guy RK, Tropsha A. Discovery of Novel Antimalarial Compounds Enabled by QSAR-Based Virtual Screening. *J. Chem. Inf. Model.* 2013; 53:475–492. [PubMed: 23252936]
195. Hajjo R, Setola V, Roth BL, Tropsha A. Chemocentric Informatics Approach to Drug Discovery: Identification and Experimental Validation of Selective Estrogen Receptor Modulators as Ligands of 5-Hydroxytryptamine-6 Receptors and as Potential Cognition Enhancers. *J. Med. Chem.* 2012; 55:5704–5719. [PubMed: 22537153]
196. Muratov EN, Varlamova EV, Artemenko AG, Khristova T, Kuz'min VE, Makarov VA, Riabova OB, Wutzler P, Schmidtke M. QSAR Analysis of [(biphenyloxy)propyl]isoxazoles: Agents Against Coxsackievirus B3. *Future Med. Chem.* 2011; 3:15–27. [PubMed: 21428823]
197. Kuz'min VE, Artemenko AG, Lozitsky VP, Muratov EN, Fedtchouk AS, Dyachenko NS, Nosach LN, Gridina TL, Shitikova LI, Mudrik LM, Mescheriakov AK, Chelombitko VA, Zheltvay AI, Vanden Eynde J-J. The Analysis of Structure-Anticancer and Antiviral Activity Relationships for Macrocyclic Pyridinophanes and Their Analogues on the Basis of 4D QSAR Models (simplex Representation of Molecular Structure). *Acta Biochim. Pol.* 2002; 49:157–168. [PubMed: 12136936]
198. Muratov EN, Artemenko AG, Varlamova EV, Polischuk PG, Lozitsky VP, Fedchuk AS, Lozitska RL, Gridina TL, Koroleva LS, Sil'nikov VN, Galabov AS, Makarov Va, Riabova OB, Wutzler P, Schmidtke M, Kuz'min VE. Per Aspera Ad Astra: Application of Simplex QSAR Approach in Antiviral Research. *Future Med. Chem.* 2010; 2:1205–1226. [PubMed: 21426164]
199. Sima P, Trebichavsky I, Sigler K. Nonmammalian Vertebrate Antibiotic Peptides. *Folia Microbiol. (Praha).* 2003; 48:709–724. [PubMed: 15058182]
200. Sima P, Trebichavsky I, Sigler K. Mammalian Antibiotic Peptides. *Folia Microbiol. (Praha).* 2003; 48:123–137. [PubMed: 12800493]
201. Brogden KA. Antimicrobial Peptides: Pore Formers or Metabolic Inhibitors in Bacteria? *Nat. Rev. Microbiol.* 2005; 3:238–250. [PubMed: 15703760]

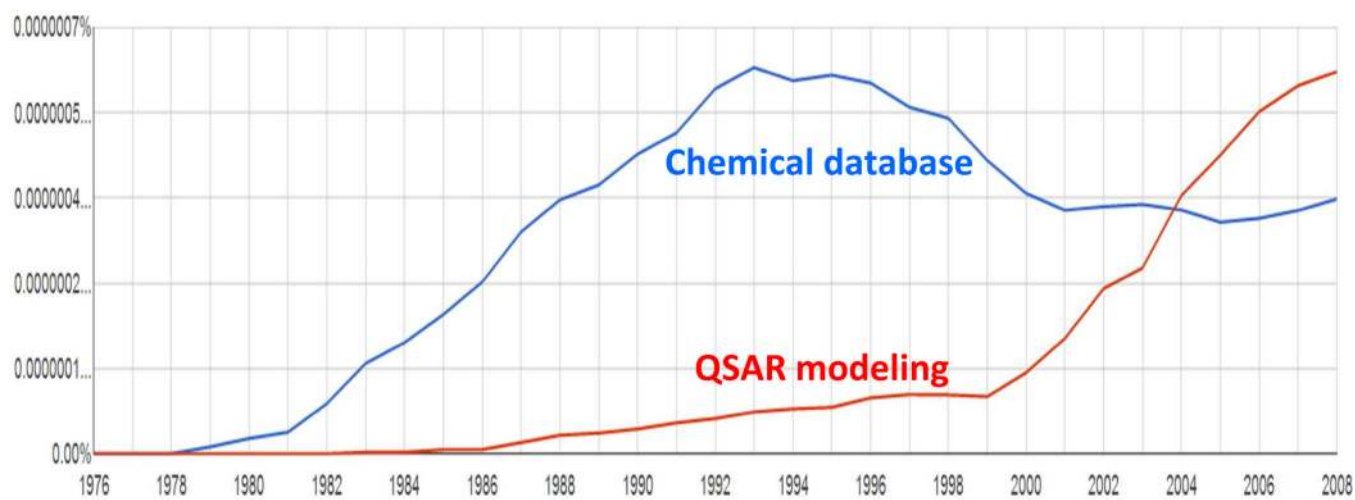
202. Taboureau O. Methods for Building Quantitative Structure-Activity Relationship (QSAR) Descriptors and Predictive Models for Computer-Aided Design of Antimicrobial Peptides. *Methods Mol. Biol.* 2010; 618:77–86. [PubMed: 20094859]
203. Hamilton-Miller JMT. Antibiotic Resistance from Two Perspectives: Man and Microbe. *Int. J. Antimicrob. Agents.* 2004; 23:209–212. [PubMed: 15164959]
204. Levy SB, Marshall B. Antibacterial Resistance Worldwide: Causes, Challenges and Responses. *Nat. Med.* 2004; 10:S122–S129. [PubMed: 15577930]
205. Koczulla AR, Bals R. Antimicrobial Peptides: Current Status and Therapeutic Potential. *Drugs.* 2003; 63:389–406. [PubMed: 12558461]
206. Finlay BB, Hancock REW. Can Innate Immunity Be Enhanced to Treat Microbial Infections? *Nat. Rev. Microbiol.* 2004; 2:497–504. [PubMed: 15152205]
207. Hancock RE. Cationic Peptides: Effectors in Innate Immunity and Novel Antimicrobials. *Lancet Infect. Dis.* 2001; 1:156–164. [PubMed: 11871492]
208. Yeaman MR, Yount NY. Mechanisms of Antimicrobial Peptide Action and Resistance. *Pharmacol. Rev.* 2003; 55:27–55. [PubMed: 12615953]
209. Fjell CD, Jenssen H, Cheung WA, Hancock REW, Cherkasov A. Optimization of Antibacterial Peptides by Genetic Algorithms and Cheminformatics. *Chem. Biol. Drug Des.* 2011; 77:48–56. [PubMed: 20942839]
210. Chapple DS, Hussain R, Joannou CL, Hancock REW, Odell E, Evans RW, Siligardi G. Structure and Association of Human Lactoferrin Peptides with Escherichia Coli Lipopolysaccharide. *Antimicrob. Agents Chemother.* 2004; 48:2190–2198. [PubMed: 15155221]
211. Powers J-PS, Hancock REW. The Relationship Between Peptide Structure and Antibacterial Activity. *Peptides.* 2003; 24:1681–1691. [PubMed: 15019199]
212. Nan YH, Bang J-K, Shin SY. Design of Novel Indolicidin-Derived Antimicrobial Peptides with Enhanced Cell Specificity and Potent Anti-Inflammatory Activity. *Peptides.* 2009; 30:832–838. [PubMed: 19428758]
213. Conlon JM, Ahmed E, Condamine E. Antimicrobial Properties of Brevinin-2-Related Peptide and Its Analogs: Efficacy Against Multidrug-Resistant *Acinetobacter Baumannii*. *Chem. Biol. Drug Des.* 2009; 74:488–493. [PubMed: 19793185]
214. Robinson JA. Protein Epitope Mimetics as Anti-Infectives. *Curr. Opin. Chem. Biol.* 2011; 15:379–386. [PubMed: 21419690]
215. Pathak S, Chauhan VS. Rationale-Based, de Novo Design of Dehydrophenylalanine-Containing Antibiotic Peptides and Systematic Modification in Sequence for Enhanced Potency. *Antimicrob. Agents Chemother.* 2011; 55:2178–2188. [PubMed: 21321136]
216. Taira J, Kida Y, Yamaguchi H, Kuwano K, Higashimoto Y, Kodama H. Modifications on Amphiphilicity and Cationicity of Unnatural Amino Acid Containing Peptides for the Improvement of Antimicrobial Activity Against Pathogenic Bacteria. *J. Pept. Sci.* 2010; 16:607–612. [PubMed: 20648478]
217. Tossi A, Sandri L, Giangaspero A. Amphipathic, Alpha-Helical Antimicrobial Peptides. *Biopolymers.* 2000; 55:4–30. [PubMed: 10931439]
218. Zelezetsky I, Tossi A. Alpha-Helical Antimicrobial Peptides--Using a Sequence Template to Guide Structure-Activity Relationship Studies. *Biochim. Biophys. Acta.* 2006; 1758:1436–1449. [PubMed: 16678118]
219. Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* 1987; 30:1126–1135. [PubMed: 3599020]
220. Lejon T, Strom MB, Svendsen JS. Antibiotic Activity of Pentadecapeptides Modelled from Amino Acid Descriptors. *J. Pept. Sci.* 2001; 7:74–81. [PubMed: 11277499]
221. Lejon T, Stiberg T, Strom MB, Svendsen JS. Prediction of Antibiotic Activity and Synthesis of New Pentadecapeptides Based on Lactoferricins. *J. Pept. Sci.* 2004; 10:329–335. [PubMed: 15214437]
222. Strom MB, Rekdal O, Stensen W, Svendsen JS. Increased Antibacterial Activity of 15-Residue Murine Lactoferricin Derivatives. *J. Pept. Res.* 2001; 57:127–139. [PubMed: 11168896]

223. Jenssen H, Gutteberg TJ, Lejon T. Modelling of Anti-HSV Activity of Lactoferricin Analogues Using Amino Acid Descriptors. *J. Pept. Sci.* 2005; 11:97–103. [PubMed: 15635641]
224. Sanchez-Gomez S, Japelj B, Jerala R, Moriyon I, Fernandez Alonso M, Leiva J, Blondelle SE, Andra J, Brandenburg K, Lohner K, Martinez de Tejada G. Structural Features Governing the Activity of Lactoferricin-Derived Peptides That Act in Synergy with Antibiotics Against *Pseudomonas Aeruginosa* in Vitro and in Vivo. *Antimicrob. Agents Chemother.* 2011; 55:218–228. [PubMed: 20956602]
225. Mee RP, Auton TR, Morgan PJ. Design of Active Analogues of a 15-Residue Peptide Using D-Optimal Design, QSAR and a Combinatorial Search Algorithm. *J. Pept. Res.* 1997; 49:89–102. [PubMed: 9128105]
226. Tapia VE, Ay B, Volkmer R. Exploring and Profiling Protein Function with Peptide Arrays. *Methods Mol. Biol.* 2009; 570:3–17. [PubMed: 19649587]
227. Hilpert K, Volkmer-Engert R, Walter T, Hancock REW. High-Throughput Generation of Small Antibacterial Peptides with Improved Activity. *Nat. Biotechnol.* 2005; 23:1008–1012. [PubMed: 16041366]
228. Taboureau O, Olsen OH, Nielsen JD, Raventos D, Mygind PH, Kristensen H-H. Design of Novispirin Antimicrobial Peptides by Quantitative Structure-Activity Relationship. *Chem. Biol. Drug Des.* 2006; 68:48–57. [PubMed: 16923026]
229. Mikut R. Computer-Based Analysis, Visualization, and Interpretation of Antimicrobial Peptide Activities. *Methods Mol. Biol.* 2010; 618:287–299. [PubMed: 20094871]
230. Cherkasov A, Jankovic B. Application of “Inductive” QSAR Descriptors for Quantification of Antibacterial Activity of Cationic Polypeptides. *Molecules.* 2004; 9:1034–1052. [PubMed: 18007503]
231. Cherkasov A, Hilpert K, Jenssen H, Fjell CD, Waldbrook M, Mullaly SC, Volkmer R, Hancock REW. Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective Against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chem. Biol.* 2009; 4:65–74. [PubMed: 19055425]
232. Jenssen H, Fjell CD, Cherkasov A, Hancock REW. QSAR Modeling and Computer-Aided Design of Antimicrobial Peptides. *J. Pept. Sci.* 2008; 14:110–114. [PubMed: 17847019]
233. Jenssen H, Lejon T, Hilpert K, Fjell CD, Cherkasov A, Hancock REW. Evaluating Different Descriptors for Model Design of Antimicrobial Peptides with Enhanced Activity Toward *P. Aeruginosa*. *Chem. Biol. Drug Des.* 2007; 70:134–142. [PubMed: 17683374]
234. Fjell CD, Jenssen H, Hilpert K, Cheung WA, Pante N, Hancock REW, Cherkasov A. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* 2009; 52:2006–2015. [PubMed: 19296598]
235. Cherkasov A. “Inductive” Descriptors. 10 Successful Years in QSAR. *Curr. Comput. Aided. Drug Des.* 2005; 21–42.
236. Cherkasov A. Can “Bacterial-Metabolite-Likeness” Model Improve Odds of “in Silico” Antibiotic Discovery? *J. Chem. Inf. Model.* 2006; 46:1214–1222. [PubMed: 16711741]
237. Cherkasov A, Jonsson M. Substituent Effects on Thermochemical Properties of Free Radicals. *J. Chem. Inf. Comput. Sci.* 1998; 38:1151–1156.
238. Cherkasov A, Jonsson M. A New Method for Estimation of Homolytic C-H Bond Dissociation Enthalpies. *J. Chem. Inf. Comput. Sci.* 2000; 40:1222–1226. [PubMed: 11045817]
239. Bhonsle JB, Venugopal D, Huddler DP, Magill AJ, Hicks RP. Application of 3D-QSAR for Identification of Descriptors Defining Bioactivity of Antimicrobial Peptides. *J. Med. Chem.* 2007; 50:6545–6553. [PubMed: 18062663]
240. Torrent M, Andreu D, Nogues VM, Boix E. Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS One.* 2011; 6:e16968. [PubMed: 21347392]
241. Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Kuz'min VE. Existing and Developing Approaches for QSAR Analysis of Mixtures. *Mol. Inform.* 2012; 31:202–221.
242. Oprisiu I, Varlamova E, Muratov E, Artemenko A, Marcou G, Polishchuk P, Kuz'min V, Varnek A. QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids. *Mol. Inform.* 2012; 31:491–502.

243. Ajmani S, Rogers SC, Barley MH, Burgess AN, Livingstone DJ. Characterization of Mixtures. Part 2: QSPR Models for Prediction of Excess Molar Volume and Liquid Density Using Neural Networks. *Mol. Inform.* 2010; 29:645–653.
244. Altenburger R, Nendza M, Schuurmann G. Mixture Toxicity and Its Modeling by Quantitative Structure-Activity Relationships. *Environ. Toxicol. Chem.* 2003; 22:1900–1915. [PubMed: 12924589]
245. Tichy M, Cikrt M, Roth Z, Rucki M. QSAR Analysis in Mixture Toxicity Assessment. *SAR QSAR Environ. Res.* 1998; 9:155–169.
246. Yu HX, Lin ZF, Feng JF, Xu TL, Wang LS. Development of Quantitative Structure Activity Relationships in Toxicity Prediction of Complex Mixtures. *Acta Pharmacol. Sin.* 2001; 22:45–49. [PubMed: 11730561]
247. Wang C, Lu G, Tang Z, Guo X. Quantitative Structure-Activity Relationships for Joint Toxicity of Substituted Phenols and Anilines to *Scenedesmus Obliquus*. *J. Environ. Sci. (China)*. 2008; 20:115–119. [PubMed: 18572533]
248. ChEMBL Database. <https://www.ebi.ac.uk/chembl/> [(accessed Mar 13, 2013)]
249. [(accessed Mar 13, 2013)] NCI Database. [http://dtp.nci.nih.gov/docs/3d\\_database/structural\\_information/structural\\_data.html](http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html)
250. [(accessed Mar 13, 2013)] DTP AIDS Antiviral Screen database. [http://dtp.nci.nih.gov/docs/aids/aids\\_data.html](http://dtp.nci.nih.gov/docs/aids/aids_data.html)
251. Lin Z, Yu H, Wei D, Wang G, Feng J, Wang L. Prediction of Mixture Toxicity with Its Total Hydrophobicity. *Chemosphere*. 2002; 46:305–310. [PubMed: 11827289]
252. Lin Z, Zhong P, Yin K, Wang L, Yu H. Quantification of Joint Effect for Hydrogen Bond and Development of QSARs for Predicting Mixture Toxicity. *Chemosphere*. 2003; 52:1199–1208. [PubMed: 12821001]
253. Ajmani S, Rogers SC, Barley MH, Burgess AN, Livingstone DJ. Characterization of Mixtures Part 1: Prediction of Infinite-Dilution Activity Coefficients Using Neural Network-Based QSPR Models. *QSAR Comb. Sci.* 2008; 27:1346–1361.
254. Ajmani S, Rogers SC, Barley MH, Livingstone DJ. Application of QSPR to Mixtures. *J. Chem. Inf. Model.* 2006; 46:2043–2055. [PubMed: 16995735]
255. Kang J, Kim H, Lee H, Yang D, Lee C. Development and Current Status of the Korea Thermophysical Properties Databank (KDB). *Int. J. Thermophys.* 2001; 22:487–494.
256. Small BG, McColl BW, Allmendinger R, Pahle J, Lopez-Castejon G, Rothwell NJ, Knowles J, Mendes P, Brough D, Kell DB. Efficient Discovery of Anti-Inflammatory Small-Molecule Combinations Using Evolutionary Computing. *Nat. Chem. Biol.* 2011; 7:902–908. [PubMed: 22020553]
257. J Williams A, Tkachenko V, Lipinski C, Tropsha A, Ekins S. Free Online Resources Enabling Crowd-Sourced Drug Discovery. *Drug Discov. World.* 2010; 10:33–39.
258. Wei DB, Zhai LH, Hu HY. QSAR-Based Toxicity Classification and Prediction for Single and Mixed Aromatic Compounds. *SAR QSAR Environ. Res.* 2004; 15:207–216. [PubMed: 15293547]
259. Patel HC, Duca JS, Hopfinger AJ, Glendening CD, Thompson ED. Quantitative Component Analysis of Mixtures for Risk Assessment: Application to Eye Irritation. *Chem. Res. Toxicol.* 1999; 12:1050–1056. [PubMed: 10563830]
260. Ajmani S, Jadhav K, Kulkarni SA. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* 2006; 46:24–31. [PubMed: 16426036]
261. Zhang L, Zhou P-J, Yang F, Wang Z-D. Computer-Based QSARs for Predicting Mixture Toxicity of Benzene and Its Derivatives. *Chemosphere*. 2007; 67:396–401. [PubMed: 17184822]
262. Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Nikolaeva- Glomb L, Galabov AS, Kuz'min VE. QSAR Analysis of Poliovirus Inhibition by Dual Combinations of Antivirals. *Struct. Chem.* 2013; 53:1665–1679.
263. Su G, Yan B. Nano-Combinatorial Chemistry Strategy for Nanotechnology Research. *J. Comb. Chem.* 2010; 12:215–221. [PubMed: 20131816]
264. Yan B. Nano-Combinatorial and Catalyst Screening Technologies. *Comb. Chem. high throughput Screen.* 2011; 14:146. [PubMed: 21271980]

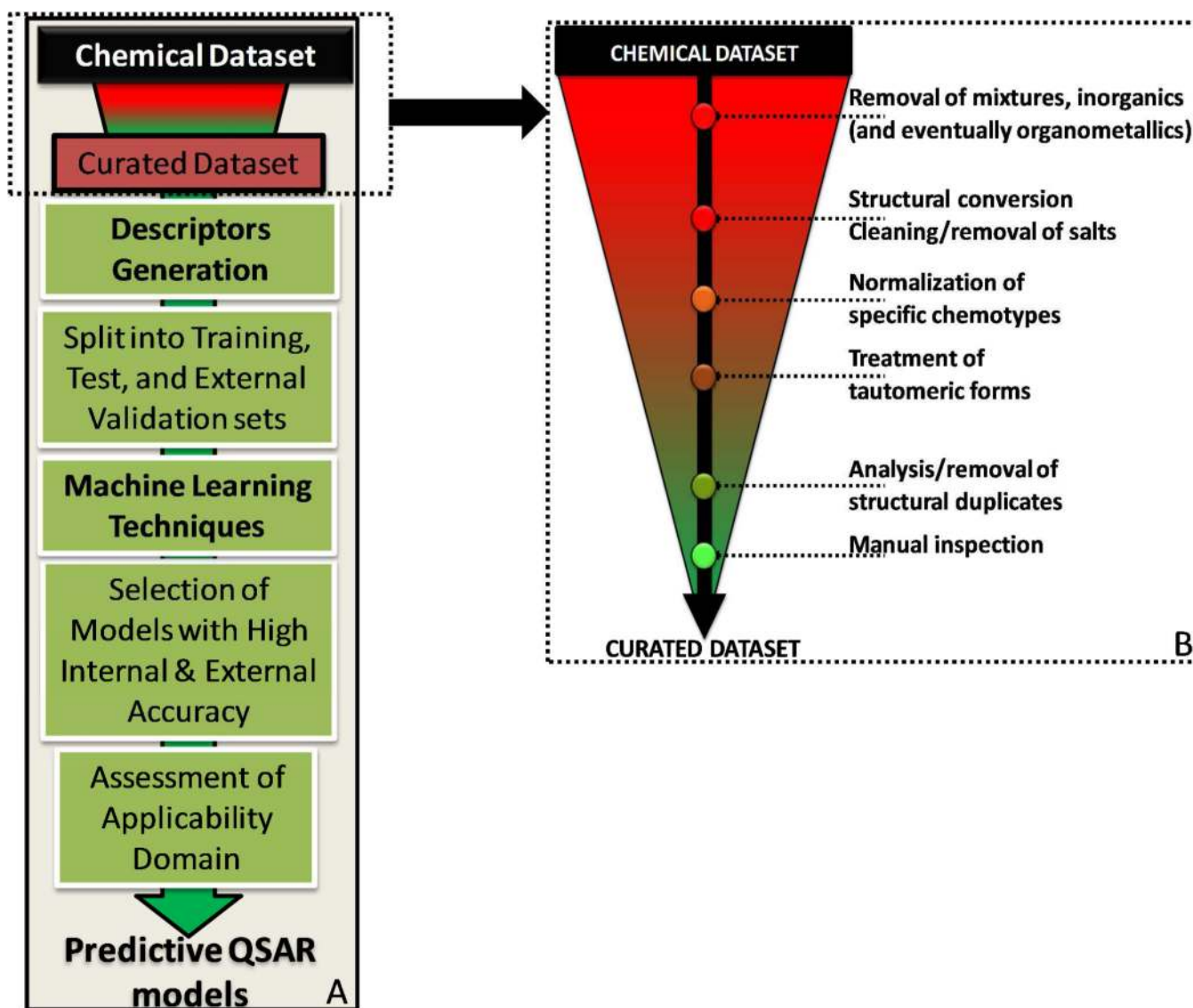
265. Thomas DG, Pappu RV, Baker NA. NanoParticle Ontology for Cancer Nanotechnology Research. *J. Biomed. Inform.* 2011; 44:59–74. [PubMed: 20211274]
266. Linkov I, Satterstrom FK, Corey LM. Nanotoxicology and Nanomedicine: Making Hard Decisions. *Nanomedicine.* 2008; 4:167–171. [PubMed: 18329962]
267. Garnett M, Kallinteri P. Nanomedicines and Nanotoxicology: Some Physiological Principles. *Occupat. Med.* 2006; 56:307–311.
268. Elder A. Nanotoxicology: How Do Nanotubes Suppress T Cells? *Nat.Nanotechnol.* 2009; 4:409–410. [PubMed: 19581890]
269. Zhao Y, Xing G, Chai Z. Nanotoxicology: Are Carbon Nanotubes Safe? *Nat. Nanotechnol.* 2008; 3:191–192. [PubMed: 18654501]
270. Oberdorster G. Safety Assessment for Nanotechnology and Nanomedicine: Concepts of Nanotoxicology. *J. Intern. Med.* 2010; 267:89–105. [PubMed: 20059646]
271. Dawson KA, Salvati A, Lynch I. Nanotoxicology: Nanoparticles Reconstruct Lipids. *Nat.Nanotechnol.* 2009; 4:84–85. [PubMed: 19197306]
272. Stone V, Donaldson K. Nanotoxicology: Signs of Stress. *Nat. Nanotechnol.* 2006; 1:23–24. [PubMed: 18654137]
273. Heister E, Lamprecht C, Neves V, Tilmaciu C, Datas L, Flahaut E, Soula B, Hinterdorfer P, Coley HM, Silva SRP, McFadden J. Higher Dispersion Efficacy of Functionalized Carbon Nanotubes in Chemical and Biological Environments. *ACS Nano.* 2010; 4:2615–2626. [PubMed: 20380453]
274. Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-Dantona E, Paik D, Pan S, Stafford GA, Freund ET, Klemm JD, Baker NA. ISA-TABNano: a Specification for Sharing Nanomaterial Research Data in Spreadsheet-Based Format. *BMC Biotechnol.* 2013; 13:2. [PubMed: 23311978]
275. Burello E, Worth A. Computational Nanotoxicology: Predicting Toxicity of Nanoparticles. *Nat. Nanotechnol.* 2011; 6:138–139. [PubMed: 21372838]
276. Meng H, Xia T, George S, Nel AE. A Predictive Toxicological Paradigm for the Safety Assessment of Nanomaterials. *ACS Nano.* 2009; 3:1620–1627. [PubMed: 21452863]
277. Liu J, Yang L, Hopfinger AJ. Affinity of Drugs and Small Biologically Active Molecules to Carbon Nanotubes: a Pharmacodynamics and Nanotoxicity Factor? *Mol.Pharm.* 2009; 6:873–882. [PubMed: 19281188]
278. Puzyn T, Leszczynska D, Leszczynski J. Toward the Development of “Nano-QSARs”: Advances and Challenges. *Small.* 2009; 5:2494–2509. [PubMed: 19787675]
279. Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, Hwang H-M, Toropov A, Leszczynska D, Leszczynski J. Using Nano-QSAR to Predict the Cytotoxicity of Metal Oxide Nanoparticles. *Nat. Nanotechnol.* 2011; 6:175–178. [PubMed: 21317892]
280. Fourches D, Pu D, Tassa C, Weissleder R, Shaw SY, Mumper RJ, Tropsha A. Quantitative Nanostructure-Activity Relationship Modeling. *ACS Nano.* 2010; 4:5703–5712. [PubMed: 20857979]
281. [(accessed Mar 13, 2013)] EPA Estimation Program Interface (EPI) Suite. <http://www.epa.gov/oppt/exposure/pubs/episuite.htm>
282. Tunkel J, Mayo K, Austin C, Hickerson A, Howard P. Practical Considerations on the Use of Predictive Models for Regulatory Purposes. *Environ. Sci. Technol.* 2005; 39:2188–2199. [PubMed: 15871254]
283. Madden JC, Enoch SJ, Hewitt M, Cronin MTD. Pharmaceuticals in the Environment: Good Practice in Predicting Acute Ecotoxicological Effects. *Toxicol. Lett.* 2009; 185:85–101. [PubMed: 19118609]
284. Hughes K, Paterson J, Meek ME. Tools for the Prioritization of Substances on the Domestic Substances List in Canada on the Basis of Hazard. *Regul. Toxicol. Pharmacol.* 2009; 55:382–393. [PubMed: 19766685]
285. Arvidson KB, Chanderbhan R, Muldoon-Jacobs K, Mayer J, Ogungbesan A. Regulatory Use of Computational Toxicology Tools and Databases at the United States Food and Drug Administration’s Office of Food Additive Safety. *Expert Opin. Drug Metab. Toxicol.* 2010; 6:793–796. [PubMed: 20491519]

286. REACH: Registration, Evaluation and Authorisation and Restriction of Chemicals. <http://europa.eu.int/comm/environment/chemicals/reach.htm>
287. JRC QSAR Model Database. <http://qsardb.jrc.it> [(accessed Mar 13, 2013)]
288. Maggiora GM. On Outliers and Activity Cliffs--Why QSAR Often Disappoints. *J. Chem. Inf. Model.* 2006; 46:1535. [PubMed: 16859285]
289. Doweiko AM. Is QSAR Relevant to Drug Discovery? *IDrugs.* 2008; 11:894–899. [PubMed: 19051151]
290. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko I. V Combinatorial QSAR Modeling of Chemical Toxicants Tested Against *Tetrahymena Pyriformis*. *J. Chem. Inf. Model.* 2008; 48:766–784. [PubMed: 18311912]
291. Golbraikh A, Muratov EN, Fourches D, Tropsha A. Dataset Modelability by QSAR. *J. Chem. Inf. Model.* 2013 [Online early access], DOI: 10.1021/ci400572x. Published Online: October 19, 2013.
292. Guha R, Van Drie JH. Structure--Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* 2008; 48:646–658. [PubMed: 18303878]
293. Seebeck B, Wagener M, Rarey M. From Activity Cliffs to Target-Specific Scoring Models and Pharmacophore Hypotheses. *ChemMedChem.* 2011; 6:1630–1639. [PubMed: 21751401]
294. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* 2013 [on-line early access], DOI: 10.1021/jm401120g. Published Online: August 27, 2013.
295. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* 2013 [Online early access], DOI: 10.1021/jm401411z. Published Online: October 23, 2013.
296. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR Applicability Domain Estimation by Projection of the Training Set Descriptor Space: a Review. *Altern.Lab Anim.* 2005; 33:445–459. [PubMed: 16268757]
297. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A. Critical Assessment of QSAR Models of Environmental Toxicity Against *Tetrahymena Pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J.Chem.Inf.Model.* 2008; 48:1733–1746. [PubMed: 18729318]
298. Jorgensen WL. QSAR/QSPR and Proprietary Data. *J Chem.Inf.Model.* 2006; 46:937.
299. Bajorath J. Computational Chemistry in Pharmaceutical Research: At the Crossroads. *J. Comput. Aided. Mol. Des.* 2012; 26:11–12. [PubMed: 22083841]
300. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD. Chem-Bioinformatics: Comparative QSAR at the Interface Between Chemistry and Biology. *Chem. Rev.* 2002; 102:783–812. [PubMed: 11890757]

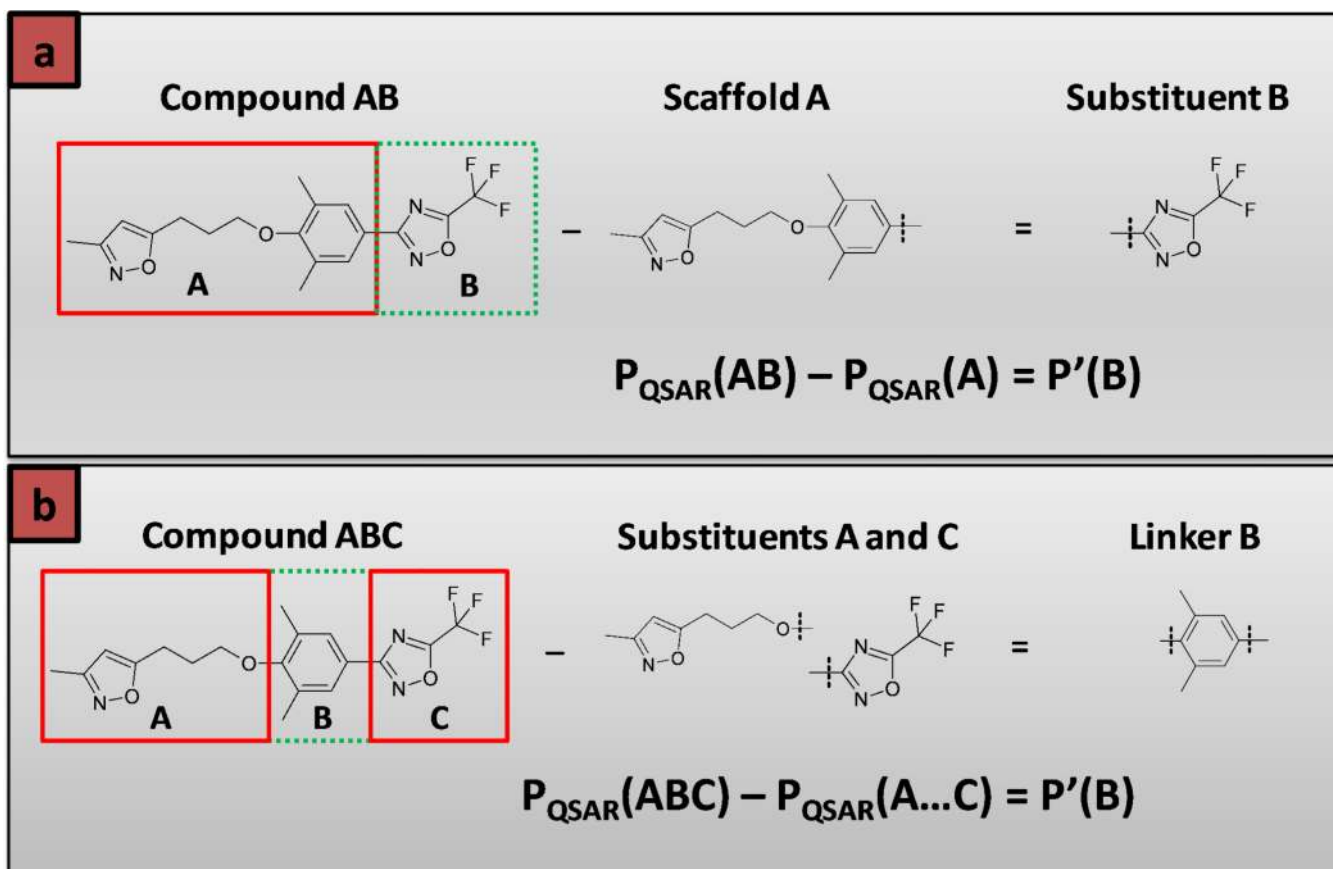


**Figure 1.**

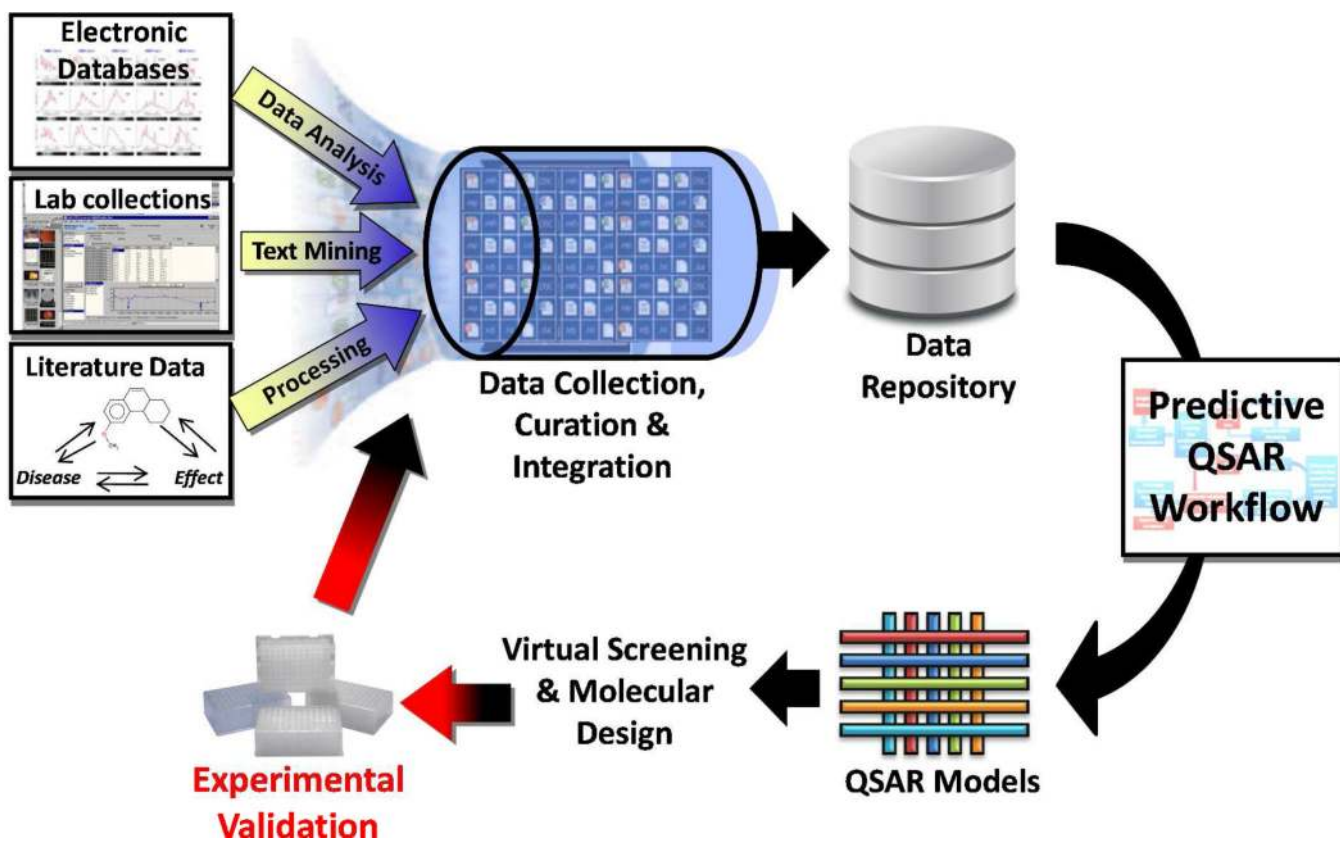
The growth of QSAR modeling is caused by the growth of experimental data. Chart is generated by Google Ngram Viewer (<http://books.google.com/ngrams>); Y-axis – percentage among all books in the Google Ngram database, X-axis – years.



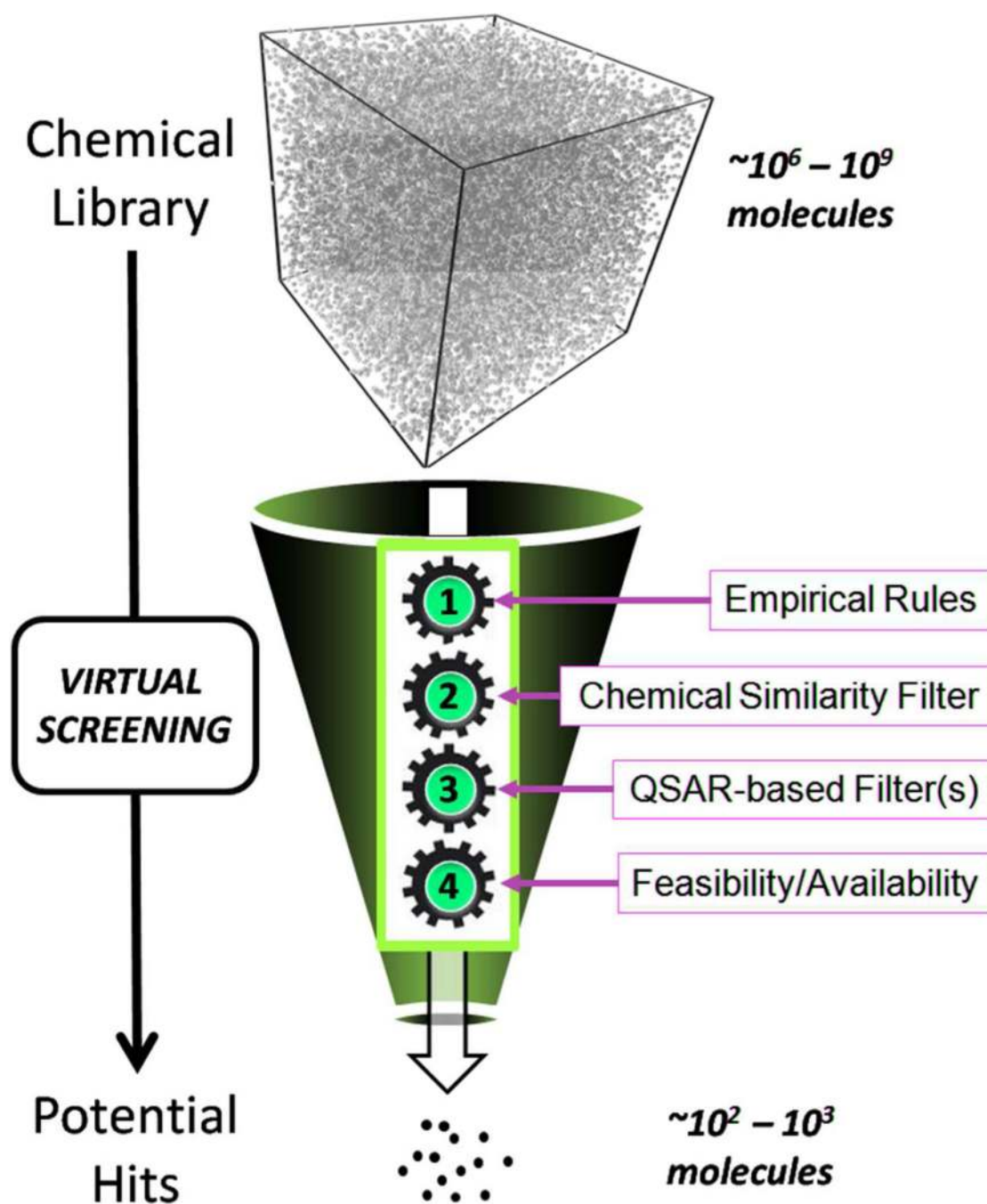
**Figure 2.** Workflow for predictive QSAR modeling (A) incorporating a critical step of data curation (within the dotted rectangle) that relies on its own special workflow (B).

**Figure 3.**

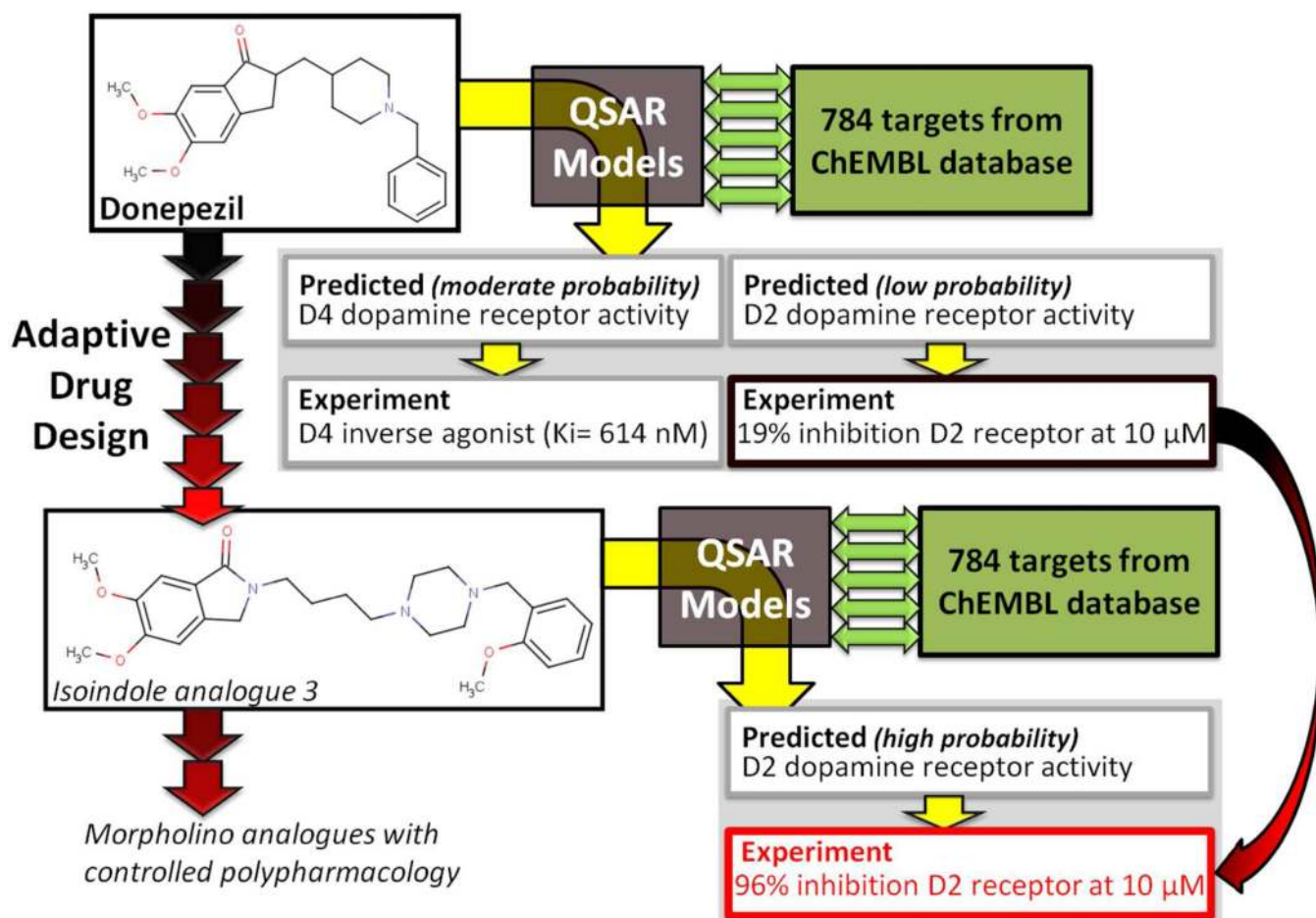
Estimation of terminal (a) and central (b) fragment's contributions to activity.  $P_{\text{QSAR}}$  – contribution estimated by developed QSAR model;  $P'$  – contribution estimated by the universal interpretation approach.



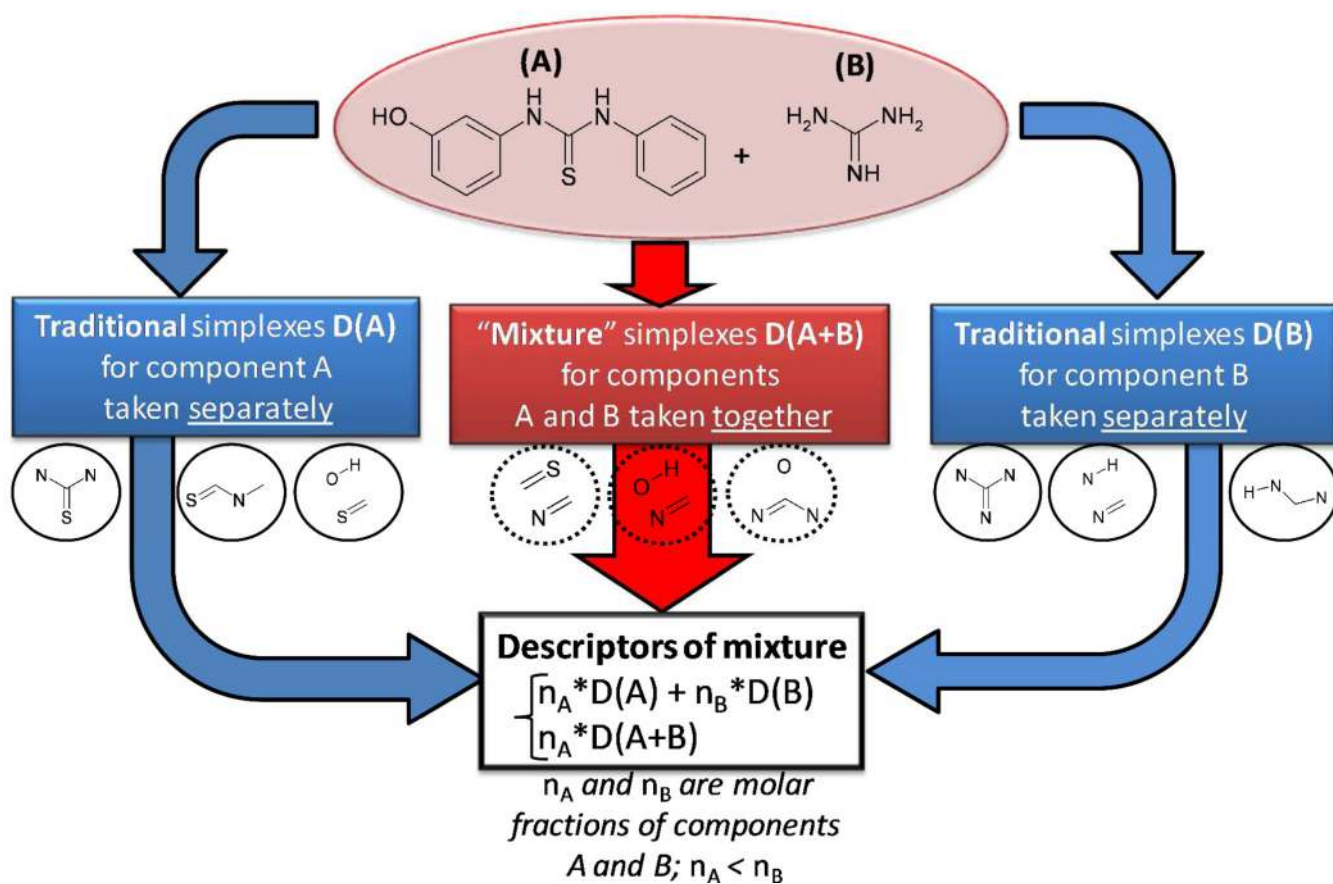
**Figure 4.**  
Overall study design of a QSAR-guided drug discovery project.



**Figure 5.**  
General workflow for screening chemical libraries using empirical and QSAR-based filtering.



**Figure 6.** Adaptive drug design for computer-aided generation of compounds with controlled polypharmacology (see earlier work).<sup>192</sup>



**Figure 7.**  
Generation of simplex descriptors for mixtures.