4.07 Predictive Quantitative Structure–Activity Relationship Modeling

A Tropsha, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

4.07.1 Introduction

4.07.1.1 Quantitative Structure–Activity Relationship (QSAR) Modeling in Modern Medicinal Chemistry

At the beginning of its over 40 years of existence as an independent area of research, quantitative structure–activity relationship (QSAR) modeling was viewed strictly as analytical physical chemical approach applicable only to small congeneric series of molecules. The technique was first introduced by Hansch *et al.*[3] on the basis of implications from linear free-energy relationships in general and the Hammett equation in particular.[4] It is based upon the assumption that differences in physicochemical properties account for the differences in biological activities of compounds. According to this approach, the changes in physicochemical properties that affect the biological activities of a set of congeners are of three major types: electronic, steric, and hydrophobic.[5] These structural properties are often described by Hammett electronic constants,[5] Verloop STERIMOL parameters,[6] hydrophobic constants,[7] etc. The quantitative relationships between biological activity (or chemical property) and the structural parameters could be conventionally obtained using multiple linear regression (MLR) analysis. The fundamentals and applications of this method in chemistry and biology have been summarized by Hansch and Leo.[5] This traditional QSAR approach has generated many useful and, in some cases, predictive QSAR equations and led to several documented drug discoveries.[8–10]

Many years of active research in QSAR have dramatically changed the breadth and the depth of this field in all its components including the diversity of target properties, descriptor types, data modeling approaches, and applications. QSAR modeling as an integral part of modern medicinal chemistry is experiencing one of the most exciting periods in its history. The changes have been brought about by an extraordinarily rapid expansion of available biomolecular databases and the growing use of advanced data mining technologies for the analysis of experimental medicinal chemistry data. Today, QSAR researchers are actively expanding the areas of application of QSAR approaches and concepts, with recent examples provided by the use of QSAR approaches as applied to the results of molecular dynamics simulations,[11] or protein sequence or structure classification,[12,13] or scoring functions for protein ligand docking.[14]

These recent developments have dramatically altered our approaches to the analysis of a relationship between chemical structure and biological action. Modern approaches to drug design and discovery are characterized by computational tool integration and a paradigm shift from the analysis of small congeneric series of molecules to the analysis of broad groups of biologically active molecules with applications to the design and discovery of drug-like molecules with optimal absorption, distribution, metabolism, and excretion (ADME), and toxicity properties. The most important changes in QSAR deal with a substantial increase in the size of data sets available for the analysis and an increasing use of QSAR models as virtual screening tools to discover biologically active molecules in chemical databases and virtual chemical libraries.

One of the most characteristic features of the modern age of QSAR as an integral part of drug design and discovery is an unprecedented growth of biomolecular databases, which contain data on chemical structure and in many cases, biological activity (or other relevant drug properties such as toxicity or mutagenicity) of chemicals. Naturally, the growth of molecular databases has been concurrent with the acceleration of the drug discovery process. According to an excellent, historical account of drug discovery,[15] due to high-throughput screening (HTS) technologies, the amount of raw data points obtained by a large pharmaceutical company per year has increased from approximately 200 000 at the beginning of last decade to around 50 million today. The total number of drugs used worldwide is approximately 80 000, which have fewer than 500 characterized molecular targets.[15] Recent estimates suggest that the number of potential targets lies between 5000 and 10 000, approximately 10-fold greater than the number of targets currently pursued (*see* 4.17 Chemogenomics in Drug Discovery – The Druggable Genome and Target Class Properties).[15]

The US National Institutes of Health (NIH) recently initiated an unprecedented public effort (Roadmap[107]) that is poised to boost all aspects of medicinal chemistry. One of the chief stated objectives of the Molecular Libraries Initiative (MLI), a component of Roadmap, is 'to develop and test new algorithms for computational chemistry and virtual screening' to facilitate the discovery of 'chemical probes to study the functions of genes, cells, and biochemical pathways.'[107] The funded Molecular Libraries Screening Center Network (MLSCN) includes 10 centers that are charged 'to screen a minimum of 100 000 compounds in 20 assays that have been adapted for HTS within each center per year' by the end of the pilot 3-year period. A simple estimate suggests that in 3 years, the MLSCN is likely to start depositing no fewer than 12 million screening data points per year! It is hard to evaluate the anticipated dimension of the cumulative library assay data matrix that will result from the MLI. However, given the current plans of the Roadmap to establish an initial collection of 500 000 diverse small molecules that will be eventually screened against hundreds of targets implemented in the MLSCN, it is safe to assume that the initial two-dimensional (2D) data matrix is likely to include at least 500 000 compound rows and at least 120 assay columns. Thus, the Molecular Libraries Roadmap will soon establish an unprecedented, at least in the public sector, collection of biological profiles of chemical compounds. The issues related to storing and accessing this collection will be addressed by the intramural PubChem[16] project. Enormous challenges remain, however, in converting these data into knowledge that will guide future library and compound design efforts and will ultimately have measurable benefits to public health in agreement with the NIH Roadmap goals.

While traditional QSAR modeling has typically been limited to dealing with a maximum of several dozen compounds at a time, rapid generation of large quantities of data requires new methodologies for data analysis. New approaches need to be developed to establish QSAR models for hundreds, if not thousands, of molecules. These new methods should be robust, yet sufficiently computationally efficient to compete with the data generation and the analytical requirements of experimental techniques, such as combinatorial chemistry and HTS.

It is practically impossible to review all, even relatively recent, developments in the field of QSAR in a single chapter. Many reviews discussing different QSAR modeling methodologies have been published,[17,18] and the reader is referred to this collection of general references and publications cited therein for additional in-depth information, and to Chapters 4.03 (Quantitative Structure–Activity Relationship – A Historical Perspective and the Future) and 4.23 (Three-Dimensional Quantitative Structure–Activity Relationship: The State of the Art).

The present chapter concentrates on current trends and developments in QSAR methodology, which are characterized by the growing size of the data sets subjected to the QSAR analysis, use of multiple descriptors of chemical structure, application of both linear and especially nonlinear, optimization algorithms applicable to multidimensional modeling, growing emphasis on rigorous model validation, and application of QSAR models as virtual screening tools in database mining and chemical library design. We begin by establishing general principles of QSAR modeling, emphasizing the common aspects of various QSAR methodologies. We then consider some popular approaches to the derivation of molecular descriptors and optimization algorithms, in the context of three key components of any QSAR investigation: model development, model validation, and model application. We conclude with several remarks on present status and future developments in this exciting research discipline.

4.07.1.2 Key Quantitative Structure–Activity Relationship Concepts

An inexperienced user or sometimes even an avid practitioner of QSAR could be easily confused by the diversity of methodologies and naming conventions used in QSAR studies. 2D or three-dimensional (3D) QSAR, variable selection or artificial neural network (ANN) methods, Comparative molecular field analysis (CoMFA), or binary QSAR present examples of various terms that may appear to describe totally independent approaches, which cannot be generalized or even compared to each other. In fact, any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form $P_i = \hat{k}(D_1, D_2, \ldots D_n)$, where P_i are biological activities (or other properties of interest) of molecules, D_1, $D_2, \ldots$, D_n are calculated (or, sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and $\hat{k}$ is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules. The relationship between values of descriptors D and target properties P can be linear (e.g., MLR as in the Hansch QSAR approach), where target property can be predicted directly from the descriptor values, or nonlinear (such as ANNs or classification QSAR methods) where descriptor values are used in characterizing chemical similarity between molecules, which in turn is used to predict compound activity. In general, each compound can be represented by a point in a multidimensional space, in which descriptors D_1, $D_2, \ldots$, D_n serve as independent coordinates of the compound. The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity. All QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities. This implies that points representing compounds with similar activities in multidimensional descriptor space should be geometrically close to each other, and vice versa.

Despite formal differences between various methodologies, any QSAR method is based on a QSAR table, which can be generalized as shown in Table 1. To initiate a QSAR study, this table must include some identifiers of chemical structures (e.g., company id numbers, first column of Table 1), reliably measured values of biological activity (or any other target property of interest, e.g., solubility, metabolic transformation rate, etc., second column of Table 1), and

Table 1 Generalized QSAR table

Structure id	*Target property (EC$_{50}$, K$_i$, etc.)*	*Structural properties (descriptors)*			
Compound 1	$P1$	$D11$	$D12$	⋯	$D1n$
Compound 1	$P2$	$D21$	$D22$	⋯	$D2n$
…	…	〃	〃	〃	〃
Compound m	Pm	$Dm1$	$Dm2$	⋯	Dmn

⬇

$$\{P\} = \hat{K}\{D\}$$

calculated values of molecular descriptors in all remaining columns (sometimes, experimentally determined physical properties of compounds could be used as descriptors as well).

The differences in various QSAR methodologies can be understood in terms of the types of target property values, descriptors, and optimization algorithms used to relate descriptors to the target properties and generate statistically significant models. Target properties (regarded as dependent variables in statistical data modeling sense) can generally be of three types: (1) continuous, i.e., real values covering certain range, e.g., IC_{50} values, or binding constants; (2) categorical related, classes of target properties covering certain range of values, e.g., active and inactive compounds, frequently encoded numerically for the purpose of the subsequent analysis as 1 (for active) or 0 (for inactive), or adjacent classes of metabolic stability such as unstable, moderately stable, stable; and (3) categorical unrelated, classes of target properties that do not relate to each other in any continuum, e.g., compounds that belong to different pharmacological classes, or compounds that are classified as drugs versus nondrugs. As simple as it appears, understanding this classification is actually very important since the choice of descriptor types as well as modeling techniques (see below) is often dictated by the type of the target properties. Thus, in general the latter two types require classification modeling approaches whereas the first type of target properties allows the use of linear regression modeling. The corresponding methods of data analysis are referred to as classification or continuous property QSAR.

Chemical descriptors (or independent variables in terms of statistical data modeling) can be typically classified into two types: continuous (i.e., range of real values, e.g., as simple as molecular weight or many molecular connectivity indices); or categorical related (i.e., classes corresponding to adjacent ranges of real values, e.g., counts of functional groups or binary descriptors indicating presence or absence of a chemical functional group or an atom in a molecule). Descriptors can be generated from various representations of molecules, e.g., 2D chemical graphs or 3D molecular geometries, giving rise to the terms of 2D or 3D QSAR, respectively. Understanding these types of descriptors is also important for understanding basic principles of QSAR modeling since as stated above any modeling implies establishing the correlation between chemical similarity between compounds and similarity between their target properties. Chemical similarity is calculated in the descriptors space using various similarity metrics (see excellent reviews by Willett[19]); thus the choice of the metric is dictated in many cases by the descriptor type. For instance, in case of continuous descriptor variables the Euclidean distance in descriptor space is a reasonable choice of the similarity metric whereas in case of binary variables metrics such as the Tanimoto coefficient or Manhattan distance would appear more appropriate.

Finally, correlation methods (which can be used either with or without variable selection) can be classified into two major categories, i.e., linear (e.g., linear regression (LR), or principal component regression (PCR), or partial least squares (PLS)) or nonlinear (e.g., *k* nearest neighbor (*k*NN), recursive partitioning (RP), ANNs, or support vector machines (SVMs). Most of QSAR researchers practice their preferred modeling techniques, and the choice of the technique is frequently coupled with the choice of descriptor types. However, there are recent attempts (discussed in more detail below) to combine various modeling techniques and descriptor types as applied to individual data sets.[20]

In some cases, the types of biological data, the choice of descriptors, and the class of optimization methods are closely related and mutually inclusive. For instance, MLR can only be applied when a relatively small number of molecular descriptors are used (at least five to six times less than the total number of compounds) and the target property is characterized by a continuous range of values. The use of multiple descriptors makes it impossible to use MLR due to a high chance of spurious correlation[21] and requires the use of PLS or nonlinear optimization techniques. However, in general, for any given data set a user can choose between various types of descriptors and various optimization schemes, combining them in a practically mix-and-match mode, to arrive at statistically significant QSAR models in a variety of ways. This situation is in essence analogous to molecular mechanics[22] calculations where different force fields and differently derived parameters are developed in different groups, but the common goal is to compute optimized energies and geometries of molecules from their chemical composition and coordinates of all atoms. Thus in general, all QSAR models can be universally compared in terms of their statistical significance, and, most importantly, their ability to predict accurately biological activities (or other target properties) of molecules not included in the training set (cf. molecular mechanics where different methods are ultimately compared by their ability to reproduce experimental molecular geometries). This concept of the predictive ability as a universal characteristic of QSAR modeling independent of the particulars of individual approaches should be kept in mind as we consider examples of QSAR tools, their applications and pitfalls in the subsequent sections of this chapter.

4.07.2 Molecular Descriptors

It has been said frequently that there are three keys to the success of any QSAR model building exercise: descriptors, descriptors, and descriptors. Many different molecular representations have been proposed, including Hansch-type parameters, topological indices,[23,24] quantum mechanical descriptors,[25] molecular shapes,[26] molecular fields,[27] atomic counts,[28] 2D fragments,[29] 3D fragments,[30] etc. A recent review by Livingstone[31] provides an excellent survey of various 2D and 3D descriptors, along with some associated diversity and similarity functions. Various physicochemical parameters such as the partition coefficient, molar refractivity, and quantum mechanical quantities such as highest occupied molecular orbital (HOMO) and lowest occupied molecular orbital (LOMO) energies have been used to represent molecular identities in early QSAR studies using linear and MLR. However, these descriptors are not suited for the analysis of large numbers of molecules either because of the lack of physicochemical parameters for compounds yet to be synthesized, or because of the computational expenses required by quantum mechanical methods. Recent years have seen the application of various topological descriptors that are usually derived from either 2D or 3D molecular structural information based on the graph theory or molecular topology. These descriptors are generated on the basis of the molecular connectivity, 3D molecular topography, and molecular field properties. We discuss below most popular types of molecular descriptors used in QSAR studies.

4.07.2.1 Topological Descriptors

Two widely applied examples of 2D molecular descriptors are molecular connectivity indices (MCI) and atom pair (AP) descriptors, initially developed by Carhart *et al.*[29] Most 2D QSAR methods have been extensively studied by Randic,[32] and Kier and Hall [33–38] based on graph theoretic indices. Although the physicochemical meaning of these structural indices is unclear, they certainly represent different aspects of molecular structures. These topological indices have been successfully combined with MLR analysis.[39] They have been extensively applied to analytical chemistry, toxicity analysis, and other areas of biological activity prediction.[40–43]

A popular MolConnZ software[44] affords the computation of a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices, kappa molecular shape indices, topological and electro-topological state indices, differential connectivity indices, graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, and counts of paths and edges between different kinds of vertices.

Overall, MolConnZ produces over 400 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In a typical QSAR study, only about a half of all possible MolConnZ descriptors are eventually used after deleting descriptors with zero value or zero variance. Figure 1 provides a summary of these molecular descriptors and presents some algorithms used in their derivation.

The idea of using atom pairs as molecular features in structure–activity relationship (SAR) studies was first proposed by Carhart *et al.*[29] AP descriptors are defined by their atom types and topological distance bins. An AP is a substructure defined by two atom types and the shortest path separation (or graph distance) between the atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The general form of an atom pair descriptor is as follows:

$$\text{atom type } i \text{ - - - - - - (distance) - - - - - - atom type } j$$

where atom chemical types are typically defined by the user. For example, 15 atom types can be defined using SYBYL mol2 format as follows: (1) negative charge center, NCC; (2) positive charge center, PCC; (3) hydrogen bond acceptor, HA; (4) hydrogen bond donor, HD; (5) aromatic ring center, ARC; (6) nitrogen atoms, N; (7) oxygen atoms, O; (8) sulfur atoms, S; (9) phosphorous atoms, P; (10) fluorine atoms, FL; (11) chlorine, bromine, iodine atoms, HAL; (12) carbon atoms, C; (13) all other elements, OE; (14) triple bond center, TBC; (15) double bond center, DBC. Apparently, the total number of pairwise combinations of all 15 atom types is 120. Further, distance bins should be defined to discriminate between identical atom pairs separated by different graph distances and therefore representing different molecular substructures. Thus, 15 distance bins can be introduced in the interval between graph distance zero (i.e., zero atoms separating an atom pair) to 14 and greater. Thus, in this a total of 1800 (120 × 15) AP descriptors can be generated for any molecular structure. An example of an AP descriptor is shown in Figure 2. Frequently, as applied to particular data sets, many of the theoretically possible AP descriptors have zero value (implying that certain atom types or atom pairs are absent in molecular structures).

Dragon descriptors[45] include different groups: constitutional descriptors, topological indices, molecular walk counts, BCUT descriptors, Galvez topological charge indices, 2D autocorrelations, charge indices, aromaticity indices, Randic molecular profiles, geometrical descriptors, radial distribution junction (RDF) descriptors, 3D-MoRSE descriptors,

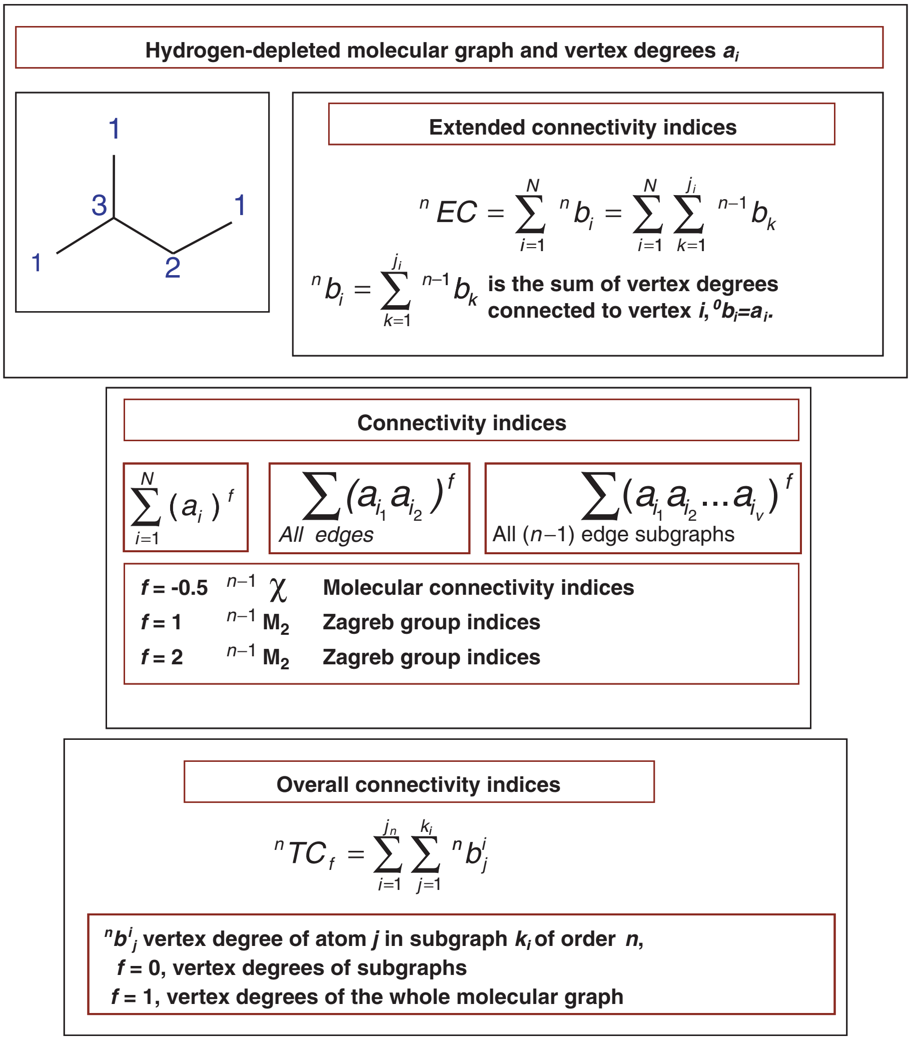
Hydrogen-depleted molecular graph and vertex degrees a_i
1
3
1
1
2
Extended connectivity indices
$^{n}EC = \sum_{i=1}^{N} {}^{n}b_i = \sum_{i=1}^{N}\sum_{k=1}^{j_i} {}^{n-1}b_k$
$^{n}b_i = \sum_{k=1}^{j_i} {}^{n-1}b_k$ is the sum of vertex degrees connected to vertex i, $^{0}b_i = a_i$.
Connectivity indices
$\sum_{i=1}^{N} (a_i)^f$
$\sum (a_{i_1} a_{i_2})^f$ All edges
$\sum (a_{i_1} a_{i_2} \ldots a_{i_v})^f$ All $(n-1)$ edge subgraphs
$f = -0.5$ $^{n-1}\chi$ Molecular connectivity indices
$f = 1$ $^{n-1}M_2$ Zagreb group indices
$f = 2$ $^{n-1}M_2$ Zagreb group indices
Overall connectivity indices
$^{n}TC_f = \sum_{i=1}^{j_n} \sum_{j=1}^{k_i} {}^{n}b_j^i$
$^{n}b_j^i$ vertex degree of atom j in subgraph k_i of order n,
$f = 0$, vertex degrees of subgraphs
$f = 1$, vertex degrees of the whole molecular graph

Figure 1 Examples of topological descriptors frequently used in QSAR studies.

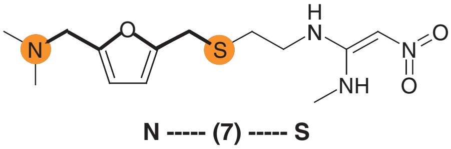

Figure 2 An example of an AP descriptor: two atom types, aliphatic nitrogen and aliphatic sulfur, are separated by the shortest chemical graph path of seven.

weighted holistic invariant molecular (WHIM) descriptors, empirical descriptors, GETAWAY descriptors, functional groups, atom-centered fragments, empirical descriptors, and properties.

Interesting recent developments in the area of molecular descriptors have dealt with addressing the inability of topological descriptors to distinguish stereoisomers. To address this deficiency, molecular chirality indices were proposed recently on the basis of 2D molecular graphs.[46] These descriptors afford different values for enantiomers and diastereomers. This work was recently extended to *cis–trans* isomerism.[47] Chirality descriptors have been successfully used in combination with nonchiral descriptors in quantitative structure–property relationship (QSPR) studies of several molecular data sets.[20,46] In all cases, highly predictive QSPR models have been obtained, having better or similar predictive abilities as compared to 3D QSAR methods.

4.07.2.2 Three-Dimensional Descriptors

With the rapid progress of 3D conformational searching of chemical structures, 3D QSAR approaches have been developed to address the problems of 2D QSAR techniques, such as their inability to distinguish stereoisomers.[48,49] These include molecular shape analysis (MSA),[50] distance geometry,[48,49] and Voronoi techniques.[51] The MSA method combined shape descriptors and MLR analysis, while the other two approaches applied atomic refractivity as structural descriptors and the solution of mathematical inequalities to obtain the quantitative relationships. CoMFA[52,53] perhaps is the most popular example of 3D QSAR. It has been widely used in medicinal chemistry and toxicity analysis by elegantly combining the power of molecular graphics and PLS technique.[54,55]

One of the most attractive features of the CoMFA and CoMFA-like methods is that due to the nature of molecular field descriptors these approaches yield models that are relatively easy to interpret in chemical terms. Famous CoMFA contour plots, which are obtained as a result of any successful CoMFA study, tell chemists in rather plain terms how the change in the compound's size or charge distribution as a result of chemical modification correlates with the binding constant or activity. These observations may immediately suggest to a chemist possible ways to modify their molecules in order to increase their potencies. However, as will be demonstrated in the next section, these predictions should be taken with caution only after sufficient work has been done to prove the statistical significance of the models.

VolSurf descriptors are obtained from 3D interaction energy grid maps.[56] They include size and shape descriptors, hydrophilic and hydrophobic regions descriptors, interaction energy moments, and other descriptors. The main advantage of VolSurf descriptors is that they are alignment free.

Molecular Operating Environment (MOE) descriptors[57] include both 2D and 3D molecular descriptors. 2D descriptors include physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors. 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors.

By analogy with 2D AP descriptors (**Figure 2**), 3D AP descriptors can also be defined using similar atom types and atom pairs and 3D molecular topography; in this case, a physical distance between atom types is used in place of chemical graph distance. The distance between two 'atoms' is measured and then assigned into one or two distance bins. Typically, the width of each distance bin is chosen as 1.0 Å. Since it is also designed to let the adjacent bins have 10% overlap with each other, the actual length of each distance bin is 1.2 Å. Any distance located in the overlap region is assigned to both bins. This 'fuzzy distance' concept is adopted to alleviate the possible unfavorable boundary effects of the distance bins. For example, with strict boundary conditions, a distance of 2.05 Å will be assigned only to bin No. 2, but it can be reasonably argued that it is almost as close to the upper half of bin No. 1 as to bin No. 2. With fuzzy boundary conditions, 2.05 Å belongs to both bin No. 1 and bin No. 2 allowing a possible match to either. All distances larger than 20 Å are assigned into the last bin.

4.07.3 Quantitative Structure–Activity Relationship Modeling Approaches

4.07.3.1 General Classification

Many different approaches to QSAR have been developed since Hansch's seminal work. As briefly discussed above, the major differences between these methods can be analyzed from two viewpoints: (1) the types of structural parameters that are used to characterize molecular identities starting from different representation of molecules, from simple chemical formulas to 3D conformations, and (2) the mathematical procedure that is employed to obtain the quantitative relationship between these structural parameters and biological activity. See Chapter 4.23.2.3.1 for an alignment (although not conformational) independent approach to 3D QSAR.

Based on the origin of molecular descriptors used in calculations, QSAR methods can be divided into three groups. One group is based on a relatively small number (usually many times smaller than the number of compounds in a data set) of physicochemical properties and parameters describing hydrophobic, steric, electrostatic, etc. effects. Usually, these descriptors are used as independent variables in multiple regression approaches. In the literature, these methods are typically referred to as Hansch analysis.

A more recent group of methods is based on quantitative characteristics of molecular graphs (molecular topological descriptors). Since molecular graphs or structural formulas are 'two-dimensional,' these methods are described as 2D QSAR. Most of the 2D QSAR methods are based on graph theoretical indices that are discussed above. Sometimes topological descriptors are also combined with physicochemical properties of molecules. Although these structural indices represent different aspects of molecular structures, and, what is important for QSAR, different structures provide numerically different values of indices, their physicochemical meaning is frequently unclear.

The third group of methods is based on descriptors derived from spatial (3D) representation of molecular structures. Correspondingly, these methods are referred to as 3D QSAR; they have become increasingly popular with the development of fast and accurate computational methods for generating 3D conformations and alignments of chemical structures. Perhaps the most popular example of 3D QSAR is CoMFA, developed by Cramer *et al.*,[58] which has combined the power of molecular graphics and PLS technique and has found wide applications in medicinal chemistry and toxicity analysis.[59] This method is one of the most recent developments in the area of ligand-based receptor modeling. This approach combines traditional QSAR analysis and 3D ligand alignment into a powerful 3D QSAR tool. CoMFA correlates 3D electrostatic and van der Waals fields around sample ligands typically overlapped in their pharmacophoric conformations with their biological activity. This approach has been successfully applied to many classes of ligands.[59]

CoMFA methodology is based on the assumption that, since, in most cases, the drug–receptor interactions are noncovalent, the changes in biological activity or binding constants of sample compounds correlate with changes in the electrostatic and van der Waals fields of these molecules. In order to initiate the CoMFA process, the test molecules should be structurally aligned in their pharmacophoric conformations. This makes the assumption that all bound ligands adopt the exact same conformation, which is unlikely considering accessory side chains may hinder or promote ligand binding depending on van der Waals and electrostatic interactions. After the alignment, steric and electrostatic fields of all molecules are sampled with a probe atom, usually sp3 carbon bearing a $+1$ charge, on a rectangular grid that encompasses structurally aligned molecules. The values of both van der Waals and electrostatic interaction between the probe atom and all atoms of each molecule are calculated in every lattice point on the grid using a force field equation and entered into the CoMFA QSAR table. This table thus contains thousands of columns, which makes it difficult to produce a statistically significant model when there are so many possible solutions. A cross-validated r^2 (q^2) that is obtained as a result of this analysis serves as a quantitative measure of the quality and internal predictive ability of the final CoMFA model. The statistical meaning of the q^2 is different from that of the conventional r^2; a q^2 value greater than 0.3 is considered significant.

4.07.3.2 Correlation Approaches

Both 2D and 3D QSAR studies have focused on the development of optimal QSAR models through variable selection. This implies that only a subset of available descriptors of chemical structures, which are the most meaningful and statistically significant in terms of correlation with biological activity, is selected. The optimum selection of variables was first achieved by combining stochastic search methods with correlation methods such as MLR, PLS analysis, or ANNs.[60–65] More specifically, these methods employ either generalized simulated annealing,[60] genetic algorithms,[61] or evolutionary algorithms,[62–65] as the stochastic optimization tool. It has been demonstrated that these algorithms combined with various chemometric tools have effectively improved the QSAR models compared to those without variable selection.

Most of the original QSAR techniques (both 2D and 3D) assumed the existence of a linear relationship between a biological activity and molecular descriptors. However, the assumption of linearity in the SAR may not hold true, especially when a large number of structurally diverse molecules are included in the analysis. Thus, several nonlinear QSAR methods have been proposed in recent years, such as ANNs[67] and *k* nearest neighbors.[68] Such applications, combined with variable selection, represent fast-growing trends in modern QSAR research.

Different QSAR methods have their own strengths and weaknesses. For example, 3D QSAR methods generally result in the diagrams of important molecular fields that can be easily interpreted in terms of specific steric and electrostatic interactions important for the ligand binding to their receptor. However, time-consuming and subjective alignment of molecular structures typically precludes the use of 3D QSAR techniques for the analysis of large data sets. On the other hand, 2D QSAR methods are much faster and more amenable to automation since they require no conformational search and structural alignment. Thus, 2D methods are best suited for the analysis of large numbers of compounds and computational screening of molecular databases; however, the interpretation of the resulting models in familiar chemical terms is frequently difficult if not impossible.

The generality of QSAR modeling approach as drug discovery tool irrespective of descriptor types or optimization algorithms can be best demonstrated in the context of inverse QSAR, which can be defined as designing or discovering molecular structures with a desired property on the basis of QSAR model. In practical terms, inverse QSAR also includes searching for molecules with a desired target property in chemical databases or virtual chemical libraries. These considerations emphasize the universal importance of establishing QSAR model robustness and predictive ability as opposed to concentrating on explanatory power, which has been characteristic feature of many traditional QSAR approaches.

4.07.4 Building Predictive Quantitative Structure–Activity Relationship Models: The Approaches to Model Validation

4.07.4.1 The Importance of Validation

The process of QSAR model development is divided into three key steps: (1) data preparation, (2) data analysis, and (3) model validation. The implementation and relative merit of these steps is generally determined by the researcher's interests and experience, and the availability of software. The resulting models are then frequently employed, at least in theory, to design new molecules based on chemical features or trends found to be statistically significant with respect to underlying biological activity.

The first stage includes the selection of a data set for QSAR studies and the calculation of molecular descriptors. The second stage deals with the selection of a statistical data analysis technique, either linear or nonlinear such as PLS or ANN. A variety of different algorithms and computer software are available for this purpose. In all approaches, descriptors are considered as independent variables, and biological activities as dependent variables.

Typically, the final part of QSAR model development is model validation,[1,69] in which estimates of the predictive power of the model are calculated. This predictive power is one of the most important characteristics of QSAR models. Ideally, it should be defined as the ability of the model to predict accurately the target property (e.g., biological activity) of compounds that were not used in model development. The typical problem of QSAR modeling is that at the time of model building a researcher has only the training set molecules, so predictive ability can be characterized only by statistical characteristics of the training set model, and not by true external validation.

Most QSAR modeling methods implement the leave-one-out (LOO), or leave-some-out, cross-validation procedure. The outcome from this procedure is a cross-validated correlation coefficient q^2, which is calculated according to the following formula:

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad [1]$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are the actual activities, the estimated activities by LOO cross-validation procedure, and the average activities, respectively. The summations in eqn [1] are performed over all compounds used to build a model (i.e., the training set). Frequently, q^2 is used as the criterion of both robustness and predictive ability of the model. Many authors consider high q^2 (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of a QSAR model. They do not test the models for their ability to predict the activity of compounds of an external test set (i.e., compounds that have not been used in the QSAR model development). For instance, in several publications[70–73] models were claimed to have high predictive ability in the absence of validation using an external test set. In other examples, models were validated using only one or two compounds that were not used in QSAR model development,[74,75] and the claim was made that these models were highly predictive.

Thus, it is still not common to test QSAR models characterized by a reasonably high q^2 for their ability to accurately predict biological activities of compounds not included in the training set. However, it has been shown[76,77] that various commonly accepted statistical characteristics of QSAR models derived for a training set are insufficient to establish and estimate the predictive power of QSAR models. Contrary to expectations, evidence would seem to indicate that no correlation exists between the LOO cross-validated q^2 and the correlation coefficient R^2 between the predicted and observed activities even when a test set of compounds with known biological activities is available for prediction (**Figure 3**). Furthermore, experience suggests,[1,78] that this phenomenon is characteristic of many data sets and is independent of the descriptor types and optimization techniques used to develop training set models. Several recent publications[69,76,77,79–81] suggest the only way to ensure the high predictive power of a QSAR model is to demonstrate a significant correlation between predicted and observed activities for a validation set of compounds that were not employed in model development.

4.07.4.2 *Y*-Randomization

The *Y*-randomization of response is another important validation approach that is widely used to establish model robustness.[82] This method consists of repeating the QSAR model derivation calculation procedure, but with randomized activities. The subsequent probability assessment of the resultant statistics is then used to gauge the robustness of the model developed with the actual activities. It is often used along with the cross-validation. In many cases, models based on the randomized data have high q^2 values, which can be explained by a chance correlation or structural redundancy.[83] If all QSAR models obtained in the *Y*-randomization test have relatively high R^2 and LOO q^2,

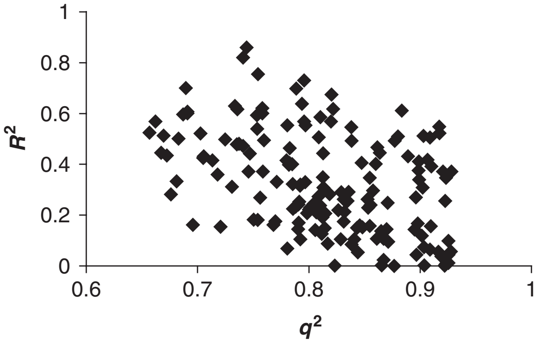

Figure 3 Beware of q^2! External R^2 (for the test set) shows no correlation with the 'predictive' LOO q^2 (for the training set). (Adapted from Golbraikh, A., Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.)

it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modeling method. A recent publication[20] provides examples of training set models that had high internal q^2 but were still unacceptable based on the *Y*-randomization test criteria.

4.07.4.3 Rational Division of Available Data Sets into Training and Test Sets

We should emphasize that both *Y*-randomization and external validation must be made a mandatory part of model development. This goal can be achieved by a division of an experimental SAR data set into the training and test sets, which are used for model development and validation, respectively. We believe that special approaches should be used to select a training set to ensure the highest significance, robustness, and predictive power of QSAR models.[1,78] Recent reviews and publications describe several algorithms that can be employed for such division.[76–78]

As follows from the above discussion, in order to estimate the true predictive power of a QSAR model, one needs to compare the predicted and observed activities of a sufficiently large external test set of compounds that were not used in the model development. One convenient parameter is an external q^2 defined as follows (similar to eqn [1] for the training set):

$$q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{\text{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{test}} (y_i - \bar{y}_{tr})^2} \qquad [2]$$

where y_i and $\hat{y}_i$ are the measured and predicted activities (over the test set), respectively, values of the dependent variable and $\bar{y}_{tr}$ is the averaged value of the dependent variable for the training set; the summations run over all compounds in the test set. Certainly, this formula is only meaningful when $\bar{y}_{tr}$ does not differ significantly from the similar value for the test set.[84] In principle, given the entire collection of compounds with known structure and activity, there is no particular reason to select one particular group of compounds as the training (or test) set; thus, the division of the data set into multiple training and test sets[78] or interchangeable definition of these sets[85] is recommended.

The use of the following statistical characteristics of the test set was also recommended[78]: correlation coefficient R^2 between the predicted and observed activities; coefficients of determination (predicted versus observed activities R_0^2, and observed versus predicted activities $R_0'^2$); slopes k and k' of the regression lines through the origin.

In summary, we consider a QSAR model predictive, if the following conditions are satisfied[78]:

$$q^2 > 0.5 \qquad [3]$$

$$R^2 > 0.6 \qquad [4]$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \qquad [5]$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15. \qquad [6]$$

It has been demonstrated[76,78] that all of the above criteria are indeed necessary to adequately assess the predictive ability of a QSAR model.

4.07.4.4 Applicability Domain of Quantitative Structure–Activity Relationship Models

It needs to be emphasized that no matter how robust, significant, and validated a QSAR model may be, it cannot be expected to be applicable to the entire universe of chemicals. Therefore, before any QSAR model is used to predict biological activity of any untested compound, its domain of application must be defined and predictions for only those chemicals that fall into this domain may be considered reliable. Described below are some approaches that aid in defining the applicability domain.

4.07.4.4.1 Extent of extrapolation

For a regression-like QSAR, a simple measure of a chemical being too far from the applicability domain of the model is its leverage, h_i,[86] which is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad (i = 1, \ldots, n) \tag{7}$$

where x_i is the descriptor row-vector of the query compound, and X is the $n \times k - 1$ matrix of k model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage h^* is, generally, fixed at $(3^* k)/n$, where n is the number of training compounds, and k is the number of model parameters. A leverage greater than the warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and, therefore, may not be reliable.[87,88]

4.07.4.4.2 Effective prediction domain

Similarly, for regression-like models, especially when the model descriptors are significantly correlated, Mandel[89] proposed the formulation of effective prediction domain, EPD. It has been demonstrated, with examples, that a regression model is justified inside and on the periphery of the EPD. Clearly, if a compound is determined to be too far from the EPD, its prediction from the model should not be considered reliable.

4.07.4.4.3 Residual standard deviation

Another important approach that can be used to evaluate the applicability domain is the degree-of-fit method developed originally by Lindberg *et al.*[90] and modified subsequently.[91] According to the original method, the predicted y values are considered to be reliable if the following condition is met:

$$s^2 < s_a^2(E_x)F \tag{8}$$

where s^2 is the residual standard deviation (RSD) of descriptor values generated for a test compound, $s_a^2(E_x)$ is the RSD of the X matrix after dimensions (components) a, and F is the F-statistic at the probability level α and $(p-a)/2$ and $(p-a)(n-a-1)/2$ degrees of freedom. The RSD of descriptor values generated for a test compound is calculated using the following equation:

$$s^2 = ||e||/(p-a) \tag{9}$$

where p is the number of x-variables, a is the number of components, and $||e||$ is the sum of squared residuals e_i expressed as

$$e_i = x_i - x_i BB' \tag{10}$$

where x_i is the ith x-variable, and B and B' represent the weight matrix and transposed weight matrix of x variables, respectively. Since the lowest possible value of F is 1.00 at $\alpha = 0.10$ (when both degrees of freedom are equal to infinity), the authors[91] decided to replace F with the degree-of-fit factor f to simplify the above condition. Thus, the modified degree-of-fit condition[91] is as follows: predicted y values are considered to be reliable if

$$s^2 < s_a^2(E_x)f \tag{11}$$

4.07.4.4.4 Similarity distance

Domain applicability can also be determined based on chemical similarity. Nonlinear methods such as k nearest neighbor (kNN) QSAR[92] employ models based on chemical similarity calculations. As such, a large similarity distance could signal that query compounds are too dissimilar to the training set compounds, and thus are not within the domain of applicability. A proposed[92] cutoff value, D_c (eqn [12]), defines a similarity distance threshold for external compounds:

$$D_c = Z\sigma + y \qquad [12]$$

Here y is the average and σ is the standard deviation of the Euclidean distances of the k nearest neighbors of each compound in the training set in the chemical descriptor space, and Z is an empirical parameter to control the significance level, with the default value of 0.5. If the distance from an external compound to its nearest neighbor in the training set is above D_c, we label its prediction unreliable.

4.07.4.5 Validated Quantitative Structure–Activity Relationship Modeling as an Empirical Data Modeling Approach: Combinatorial Quantitative Structure–Activity Relationship Modeling

We believe QSAR modeling is an empirical, exploratory research area where the models with the best-validated predictive power should be sought by a combinatorial exploration of various groupings of statistical data modeling techniques and different types of chemical descriptors followed by the consensus prediction of activities for external compounds by averaging the predicted activity values resulting from all validated models.[20] This strategy is driven by the concept that if an implicit SAR exists for a given data set, it can be formally manifested via a variety of QSAR models that use different descriptors and optimization protocols. We believe that multiple alternative QSAR models should be developed (as opposed to a single model using some favorite QSAR method) for each data set.

Several popular commercial and noncommercial QSAR software packages provide users with various descriptor types and data modeling capabilities. Practically, every package employs only one (or a few) type of descriptors and, typically, a single or a few molecular modeling techniques. Most commercially available programs provide a relatively easy-to-use interface and allow users to build single models with internal accuracy typically characterized by the q^2. As emphasized in the previous section, training-set-only modeling is insufficient to achieve models with validated predictive power; the QSAR model development process has to be modified to incorporate an independent model validation and applicability domain definition.[77,78] Since the process is relatively fast (and in principle, can be completely automated), these alternative models could be explored simultaneously when making predictions for external data sets. Consensus predictions of biological activity for novel compounds on the basis of several QSAR models, especially when predictions converge, provide more confidence in the activity estimates and better justification for the experimental validation of these compounds. This strategy is outlined in **Figure 4**.

The need to develop and employ the combinatorial QSAR approach is dictated by experience in QSAR modeling, suggesting that QSAR is still an experimental area of statistical data modeling. As such, it is impossible to decide a priori as to which particular QSAR modeling method will prove most successful. Every particular combination of descriptor sets and optimization techniques is likely to capture certain unique aspects of the SAR. Since the ultimate goal is to use the resulting models in database mining to discover diverse biologically active molecules, application of different combinations of modeling techniques and descriptor sets should increase the chances for success as demonstrated in recent publications.[20,93]

4.07.5 Validated Quantitative Structure–Activity Relationship Models as Virtual Screening Tools

Although combinatorial chemistry and HTS have offered medicinal chemists a much broader range of possibilities for lead discovery and optimization, the number of chemical compounds that can be synthesized and tested is still far beyond the capability of today's medicinal chemistry. Therefore, medicinal chemists continue to face the same problem as before: which compounds should be chosen for the next round of synthesis and testing? For chemoinformaticians, the task is to develop and utilize computational approaches to evaluate a very large number of chemical compounds and recommend the most promising ones to bench chemists.

Database mining associated with pharmacophore identification is a common and efficient approach for lead compound discovery. Pharmacophore identification refers to the computational approach to identifying the essential 3D structural features and configurations that are responsible for the biological activity of a series of compounds. Once a pharmacophore model has been developed for a particular set of biologically active molecules, it can be used to search databases of 3D structures with the aim of finding new, structurally different lead molecules with the desired biological activity.[94]

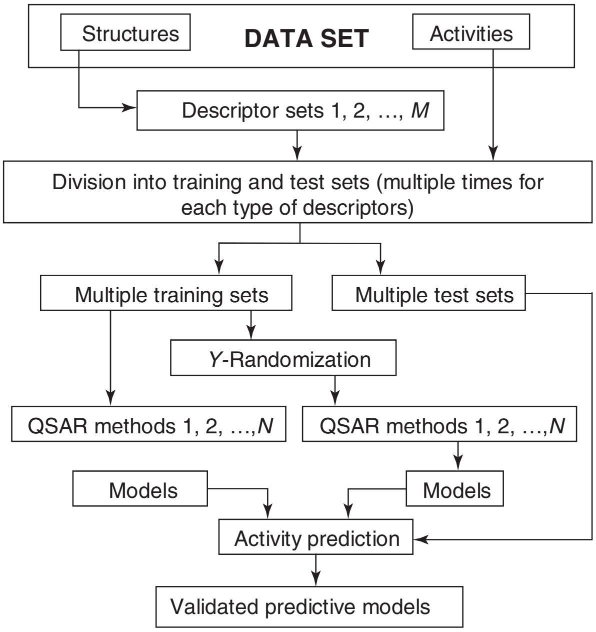

Figure 4 Flowchart of the combinatorial QSAR methodology.

An obvious parallel can be established between the search for pharmacophore elements, which are thought to describe the specificity of drug action, and the identification of a subset of descriptors contributing the most to the correlation with biological activity in a variable-selection QSAR model. Thus, the selection of specific pharmacophore features responsible for biological activity is directly analogous to the selection of specific chemical descriptors that contribute the most to an explanatory QSAR model. It is convenient to establish a concept of the descriptor pharmacophore in the context of variable selection QSAR modeling. Thus, by analogy with the conventional definition of pharmacophores, the descriptor pharmacophore can be defined as a set of descriptor variables implicated in highly statistically significant and predictive QSAR models. It has been demonstrated that QSAR models can be used in database mining, i.e., finding molecular structures that are similar in their activity to the probe molecules or even predicting the activities for the compounds in a database.[95–97] First, a preconstructed QSAR model can be used as a means of screening compounds from existing databases (or virtual libraries) for high-predicted biological activity. Alternatively, variables selected by QSAR optimization can be used for similarity searches to improve the performance of the database mining methods.

It should be noted that despite formal similarity between the common definition of pharmacophores and descriptor pharmacophores, there is also a significant difference in the procedure as well as expected outcome of virtual screening. As mentioned above, traditional approaches to database mining are based on chemical fragment or subfragment-based similarity searches. While this is an efficient approach that has enjoyed certain successes, it limits the chemical diversity of selected compounds to those that are similar to existing ligands. Search methodologies are based on chemical similarity estimated by Euclidean distance (or any other similarity measure) in multidimensional descriptor space (where descriptors are selected from the entire initial space in the process of variable selection model development) combined with quantitative predictions from combinatorial QSPR models. Due to the nature of the descriptors (e.g., whole molecule-based descriptors as opposed to fragments), such searches are more likely to result in accurate prediction of target properties for diverse novel compounds than traditional fragment-based search methodologies. This strategy was successfully tested in recent studies of anticonvulsant agents[2,98] and on the Ames Genotoxicity data set.[99] The approach is outlined in Figure 5. It is important to stress that the output of these studies is not models with their statistical characteristics as is typical for most QSAR studies. Rather, the modeling results are the predictions of the target properties for all database or virtual library compounds, which allows for immediate compound prioritization for subsequent experimental verification. Another advantage of using QSAR models for database mining is that this approach affords not only the identification of compounds of interest but also quantitative prediction of compounds' potency. For illustration, we shall discuss recent successes in developing validated predictive models of anticonvulsants[98] and their application to the discovery of novel potent compounds by the means of database mining.[2]

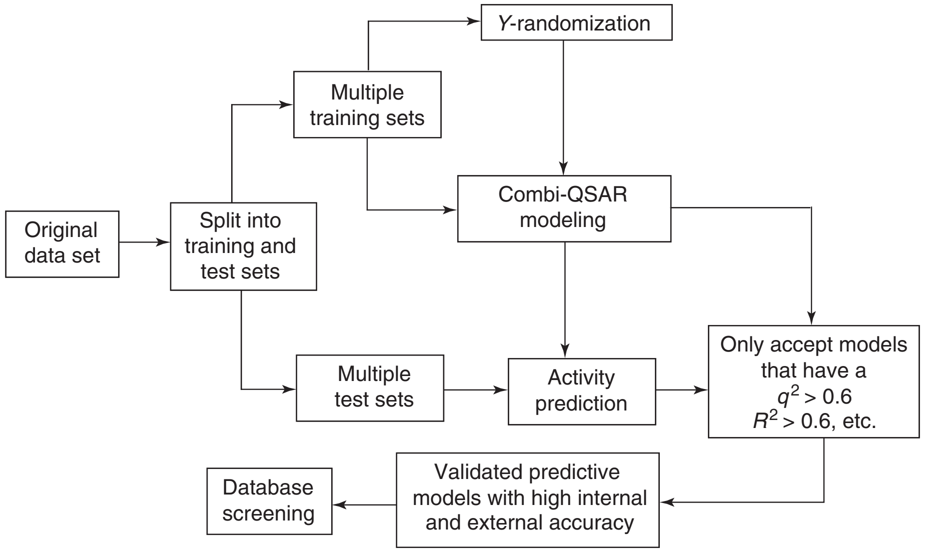
Y-randomization
Multiple training sets
Combi-QSAR modeling
Original data set
Split into training and test sets
Multiple test sets
Activity prediction
Only accept models that have a $q^2 > 0.6$ $R^2 > 0.6$, etc.
Database screening
Validated predictive models with high internal and external accuracy

Figure 5 Flowchart of predictive QSAR workflow based on validated combi-QSAR models.

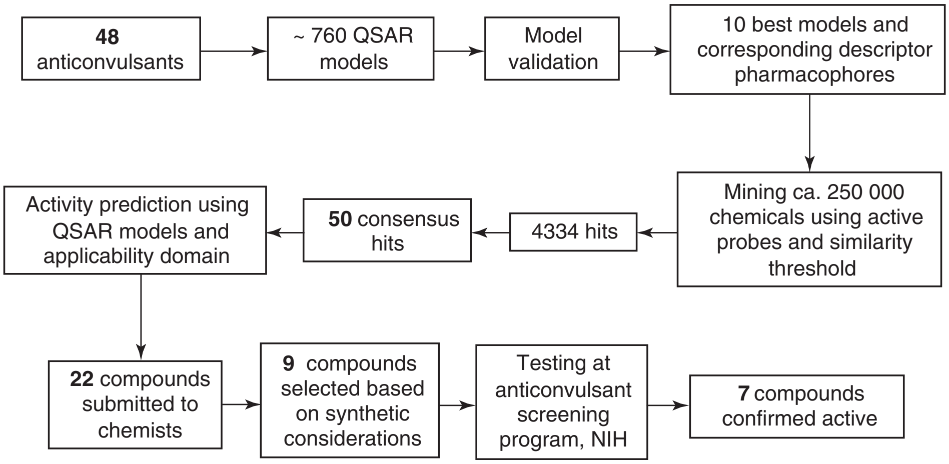
48 anticonvulsants
~ 760 QSAR models
Model validation
10 best models and corresponding descriptor pharmacophores
Mining ca. 250 000 chemicals using active probes and similarity threshold
4334 hits
50 consensus hits
Activity prediction using QSAR models and applicability domain
22 compounds submitted to chemists
9 compounds selected based on synthetic considerations
Testing at anticonvulsant screening program, NIH
7 compounds confirmed active

Figure 6 Computer-aided drug discovery workflow based on combination of QSAR modeling and consensus database mining as applied to the discovery of novel anticonvulsants.[2] The workflow emphasizes the importance of model validation and applicability domain in ensuring high hit rates as a result of database mining with predictive QSAR models.

Figure 6 summarizes the approach to using validated QSAR models for virtual screening as applied to the anticonvulsant dataset. Initially, the authors applied variable selection *k*NN and simulated-annealing PLS (SA-PLS) QSAR approaches to a data set of 48 chemically diverse functionalized amino acids (FAA) with anticonvulsant activity to develop validated QSAR models.[98] Both methods used multiple descriptors such as molecular connectivity indices or atom pair descriptors, which are derived from 2D molecular topology. QSAR models with high internal accuracy were generated, with leave-one-out cross-validated R^2 (q^2) values ranging between 0.6 and 0.8. The q^2 values for the actual data set were significantly higher than those obtained for the same data set with randomly shuffled activity values, indicating that models were robust. The original data set was further divided into several training and test sets, and highly predictive models were obtained with q^2 values for the training sets greater than 0.5 and R^2 values for the test sets greater than 0.6.

In the second stage of this process, the validated QSAR models and descriptor pharmacophore concepts were applied[2] to mining of available chemical databases for new lead anticonvulsant agents (Figure 6). Two databases have been explored: the National Cancer Institute[100] and the Maybridge[101] databases, including 237 771 and 55 273 chemical structures, respectively. Database mining was performed independently using the 10 best-validated QSAR models with the highest values of both q^2 and R^2. First, chemical similarity searches were performed between the training set compounds and database molecules using descriptor pharmacophores only (i.e., Euclidean similarity was calculated using only descriptors implicated in the 10 best validated *k*NN QSAR models) and over 4300 compounds found within the same similarity threshold in all 10 independent searches were selected as consensus hits (cf. Figure 2). Their activities then were predicted using individual QSAR models and the consensus hits with the highest predicted anticonvulsant activity were further explored experimentally.[2]

This study[2,98] presents a practical example of the drug discovery workflow that can be generalized for any data set where sufficient data to develop reliable QSAR models is available. These results certainly appear very promising and reassuring in terms of computational strategies, which emphasize that rigorous validation of QSAR models as well as conservative extrapolation are responsible for a very high hit rate.

4.07.6 Conclusions

A QSAR model describes a mathematical relationship between structural attributes and a property of a set of chemicals. The use of such mathematical relationships to predict the target property of interest for a variety of chemicals prior to, or in lieu of, expensive and labor-intensive experimental measurements has naturally been very enticing. The potential promise of using QSAR models for screening of chemical databases or virtual libraries before their synthesis appears equally attractive to chemical manufacturers, pharmaceutical companies, and government agencies, particularly in times of shrinking resources. Given the growing sizes of chemical databases resulting from combinatorial synthesis and the regulatory and social pressures for timely assessment of health and environmental risks of chemicals, the need for reliable QSAR models is imperative. For instance, environmental agencies in both Europe and the USA require reliable data on the environmental effects and the fate of all industrial chemicals. Traditionally, biological and environmental testing has provided such data; that are available for only a fraction of industrial chemicals, and thousands of industrial chemicals exist that will continue to go untested. Recently, Walker *et al.*[102–106] have addressed this problem by providing a set of guidelines for developing and using QSAR models for environmental risk assessment.

General guiding principles for building robust QSAR models have been described recently as well.[77] Thus, in order to be reliable and predictive, QSAR models should: (1) be statistically significant and robust, (2) be validated by making accurate predictions for external data sets that were not used in the model development, and (3) have their application boundaries defined so they may serve as effective database screening tools. The true power of QSAR results as we have emphasized in this chapter comes from their statistical significance and the ability of the model to predict accurately biological properties of chemical compounds in the training and most importantly test sets. Understanding and practicing these principles in QSAR modeling should help medicinal chemists to prioritize their experimental effort and significantly increase the experimental hit rates.

References

1. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
2. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2004**, *47*, 2356–2364.
3. Hansch, C.; Muir, R.; Fujita, T.; Maloney, P.; Geiger, E.; Streich, M. *J. Am. Chem. Soc.* **1963**, *85*, 2817–2824.
4. Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125–136.
5. Hansch, C.; Leo, A. Exploring QSAR. In *Fundamentals and Applications in Chemistry and Biology*; Hellen, S., Ed.; American Chemical Society: Washington, DC, 1995; Vol. 1, 580pp.
6. Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976, pp 165–207.
7. Testa, B.; Seiler, P. *Arzneimittelforschung* **1981**, *31*, 1053–1058.
8. Boyd, D. B. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991, pp 355–371.
9. Norinder, U.; Hogberg, T. *Acta Pharm. Nord.* **1992**, *4*, 73–78.
10. Van de Waterbeemd, W. H.; el Tayar, N.; Testa, B.; Wikstrom, H.; Largent, B. *J. Med. Chem.* **1987**, *30*, 2175–2181.
11. Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155–1158.
12. Agrafiotis, D. K. *Protein Sci.* **1997**, *6*, 287–293.
13. Giuliani, A.; Benigni, R.; Zbilut, J. P.; Webber, C. L., Jr.; Sirabella, P.; Colosimo, A. *Chem. Rev.* **2002**, *102*, 1471–1492.
14. Deng, W.; Breneman, C.; Embrechts, M. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 699–703.
15. Drews, J. *Science* **2000**, *287*, 1960–1964.
16. PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed Aug 2006).
17. Tropsha, A. Recent Trends in Quantitative Structure–Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*; Abraham, D., Ed.; John Wiley: New York, 2003, pp 49–77.
18. T. I. Oprea, 3D-QSAR Modeling in Drug Design. In *Computational Medicinal Chemistry and Drug Discovery*; Tollenaere, J., De Winter, H., Langenaeker, W., Bultinck, P., Eds.; Marcel Dekker: New York, 2004, pp 571–616.
19. Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D., Eds.; VCH: New York, 1996, pp 1–65.
20. Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582–595.

21. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238–1244.
22. Burkert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.
23. Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. *J. Pharm. Sci.* **1975**, *64*, 1971–1974.
24. Kier, L. B.; Murray, W. J.; Randic, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226–1230.
25. Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; Hansch, C. *J. Med. Chem.* **1991**, *34*, 786–797.
26. Jain, A. N.; Koile, K.; Chapman, D. *J. Med. Chem.* **1994**, *37*, 2315–2327.
27. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
28. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183–3187.
29. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
30. Sheridan, R. P.; SanFeliciano, S. G.; Kearsley, S. K. *J. Mol. Graph. Model.* **2000**, *18*, 320–334, 525.
31. Livingstone, D. J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
32. Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
33. Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513–520.
34. Kier, L. B.; Hall, L. H. *Quant. Struct.-Act. Relat.* **1993**, *12*, 383–388.
35. Hall, L. H.; Mohney, B.; Kier, L. B. *Quant. Struct.–Act. Relat.* **1991**, *10*, 43–51.
36. Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
37. Hall, L. H.; Kier, L. B.; Brown, B. B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.
38. Hall, L. H.; Kier, L. B. *J. Mol. Graph. Model.* **2001**, *20*, 4–18.
39. Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure–Property Modeling. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991, pp 367–422.
40. Anker, L. S.; Jurs, P. C.; Edwards, P. A. *Anal. Chem*. **1990**, *62*, 2676–2684.
41. Jurs, P. C.; Ball, J. W.; Anker, L. S.; Friedman, T. L. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272–278.
42. Nelson, T. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601–609.
43. Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109–115.
44. MolConnZ. MolConnZ. [4.05]. 2002. Hall Associates Consulting, Quincy, MA.
45. DRAGON. http://www.disat.unimib.it/chm/Dragon.htm (accessed Aug 2006).
46. Golbraikh, A.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 144–154.
47. Golbraikh, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 769–787.
48. Crippen, G. M. *J. Med. Chem.* **1980**, *23*, 599–606.
49. Crippen, G. M. *Mol. Pharmacol.* **1982**, *22*, 11–19.
50. Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
51. Boulu, L. G.; Crippen, G. M. *J. Comb. Chem.* **1989**, *10*, 673–682.
52. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
53. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *Prog. Clin. Biol. Res.* **1989**, *291*, 161–165.
54. Kubinyi, H.; Folkers, G.; Martin, Y. C. *Perspect. Drug Disc.* **1998**, *12*, V–VII.
55. Kubinyi, H.; Folkers, G.; Martin, Y. C. *Perspect. Drug Disc.* **1998**, *9–11*, V–VII.
56. Cruciani, G.; Pastor, M.; Guba, W. *Eur. J Pharm. Sci.* **2000**, *11*, S29–S39.
57. MOE. http://www.chemcomp.com/fdept/prodinfo.htm#Cheminformatics. 2005 (accessed Aug 2006).
58. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
59. Kubinyi, H.; Folkers, G.; Martin, Y. C. Eds. 3D QSAR in Drug Design. In *Recent Advances*, Kluwer Academic Publishers: Dordrecht, The Netherlands, 1998, Vol. 3, 368pp.
60. Sutter, J. M.; Dixon, S. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.
61. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
62. Kubinyi, H. *Quant. Struct.–Act. Relat.* **1994**, *13*, 285–294.
63. Kubinyi, H. *Quant. Struct.–Act. Relat.* **1994**, *13*, 393–401.
64. Luke, B. T. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279–1287.
65. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521–1530.
67. Andrea, T. A.; Kalayeh, H. *J. Med. Chem.* **1991**, *34*, 2824–2836.
68. Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
69. Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. *J. Med. Chem.* **1998**, *41*, 2553–2564.
70. Girones, X.; Gallegos, A.; Carbo-Dorca, R. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400–1407.
71. Bordas, B.; Komives, T.; Szanto, Z.; Lopata, A. *J. Agric. Food Chem.* **2000**, *48*, 926–931.
72. Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. *J. Med. Chem.* **2001**, *44*, 3254–3263.
73. Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718–726.
74. Recanatini, M.; Cavalli, A.; Belluti, F.; Piazzi, L.; Rampa, A.; Bisi, A.; Gobbi, S.; Valenti, P.; Andrisano, V.; Bartolini, M. et al. *J. Med. Chem.* **2000**, *43*, 2007–2018.
75. Moron, J. A.; Campillo, M.; Perez, V.; Unzeta, M.; Pardo, L. *J. Med. Chem.* **2000**, *43*, 1684–1691.
76. Golbraikh, A.; Tropsha, A. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357–369.
77. Tropsha, A.; Gramatica, P.; Gombar, V. K. *Quant. Struct.–Act. Relat. Comb. Sci.* **2003**, *22*, 69–77.
78. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.
79. Novellino, E.; Fattorusso, C.; Greco, G. *Pharm. Acta Helv.* **1995**, *70*, 149–154.
80. Norinder, U. *J. Chemomet.* **1996**, *10*, 95–105.
81. Zefirov, N. S.; Palyulin, V. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1022–1027.
82. Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van der Waterbeemd, H., Ed.; VCH: New York, 1995, pp 309–318.
83. Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating Models Based on Large Dataset. In *Rational Approaches to Drug Design*, Proceedings of the 13th European Symposium on Quantitative Structure–Activity Relationship, Aug 27–Sept 1; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Düsseldorf, Germany, 2001, pp 475–485.
84. Oprea, T. I.; Garcia, A. E. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186–200.
85. Oprea, T. I. *SAR QSAR Environ. Res.* **2001**, *12*, 129–141.

86. Atkinson, A. C. *Plots, Transformations and Regression*; Clarendon Press: Oxford, UK, 1985.
87. Gramatica, P.; Papa, E. *Quant. Struct.–Act. Relat.* **2003**, *22*, 374–385.
88. Gramatica, P.; Pilutti, P.; Papa, E. *Quant. Struct.–Act. Relat.* **2003**, *22*, 364–373.
89. Mandel, J. *J. Res. Nat. Bur. Stand.* **1985**, *90*, 465–476.
90. Lindberg, W.; Persson, J.-A.; Wold, S. *Anal. Chem.* **1983**, *55*, 643–648.
91. Cho, S. J.; Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259–268.
92. Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
93. de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Tropsha, A. *J. Med. Chem.* **2006**, *46*, 1245–1254.
94. *Pharmacophore Perception, Development, and Use in Drug Design*; IUL: La Jolla, CA, 2000.
95. Tropsha, A.; Cho, S. J.; Zheng, W. "New Tricks for an Old Dog": Development and Application of Novel QSAR Methods for Rational Design of Combinatorial Chemical Libraries and Database Mining. In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, A. L., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1999, pp 198–211.
96. Tropsha, A.; Zheng, W. *Curr. Pharm. Des.* **2001**, 7, 599–612.
97. Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. *J. Med. Chem.* **2000**, *43*, 4151–4159.
98. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2002**, *45*, 2811–2823.
99. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. *Mutagenesis* **2004**, *19*, 365–377.
100. NCI. http://dtp.nci.nih.gov (accessed Aug 2006).
101. Maybridge. http://www.daylight.com (accessed Aug 2006).
102. Walker, J. *Handbook on Quantitative Structure Activity Relationships (QSARs) for Pollution Prevention, Toxicity Screening, Risk Assessment and World Wide Web Applications*; SETAC Press: Pensacola, FL, 2002.
103. Walker, J. *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Physical Properties, Bioconcentration Potential and Environmental Fate of Chemicals*; SETAC Press: Pensacola, FL, 2002.
104. Walker, J. *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Endocrine Disruption Potential of Chemicals*; SETAC Press: Pensacola, FL, 2002.
105. Walker, J. *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Effects of Chemicals on Environmental-Human Health Interactions*; SETAC Press: Pensacola, FL, 2002.
106. Walker, J. *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Ecological Effects of Chemicals*; SETAC Press: Pensacola, FL, 2002.
107. National Institutes of Health. RoadMap. http://nihroadmap.nih.gov/ (accessed Aug 2006).

Biography

Alexander Tropsha is professor and Chair of the Division of Medicinal Chemistry and Natural Products in the School of Pharmacy, UNC-Chapel Hill. He received his MS degree in chemical enzymology in 1982 and his PhD in biochemistry and pharmacology in 1986, both from Moscow State University. Dr Tropsha is a member of several editorial boards, including the *Journal of Chemical Information and Modeling*. He is a permanent member of the National Institutes of Health Biodata Management and Analysis Study Section, and is an elected member of the Board and Vice-Chair of the International QSAR and Cheminformatics Society. His research interests are in the areas of computer-assisted drug design, cheminformatics, and structural bioinformatics. He has authored or co-authored more than 100 peer-reviewed publications and book chapters. His research is supported by grants from the National Institutes of Health, National Science Foundation, Environmental Protection Agency, and industry.

Comprehensive Medicinal Chemistry II
ISBN (set): 0-08-044513-6

ISBN (Volume 4) 0-08-044517-9; pp. 149–165