Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction

Alan R. Katritzky,* Minati Kuanar, Svetoslav Slavov, and C. Dennis Hall

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611

Mati Karelson,*,[‡] Iiris Kahn,[‡] and Dimitar A. Dobchev[‡,§]

Institute of Chemistry, Tallinn University of Technology, Akadeemia tee 15, Tallinn 19086, Estonia, and MolCode, Ltd., Soola 8, Tartu 51013, Estonia

Received July 2, 2009

Contents

* To whom correspondence should be addressed. E-mail: A.R.K., katritzky@chem.ufl.edu; M.K., mati.karelson@ttu.ee.
[‡] Tallinn University of Technology.
[§] MolCode, Ltd.

10.1021/cr900238d
Published on Web 08/23/2010

Alan Katritzky is Kenan Professor of Chemistry and Director of the Center of Heterocyclic Compounds at the University of Florida. He was based in the U.K. at the Universities of Oxford, Cambridge, and East Anglia before crossing the Atlantic to take up his present post in 1980. He has taught, researched, and consulted in many areas of organic and physical-organic chemistry, including structure property and activity relationships since 1990. His distinctions include 14 honorary doctorates from 12 European and Asian countries and membership of five National Academies. He has traveled widely and published extensively in the primary and secondary literature (*h* index of 77).

Minati Kuanar was born in the village of Astak, Orissa, India, in 1970. She obtained her M.Sc. (1991) and M.Phil. (1993) Degrees in Organic Chemistry from Sambalpur University, India. She received her Ph.D. degree in Physical Organic Chemistry in 1999 under the direction of Prof. Bijay K. Mishra at Sambalpur University India. She was a CSIR (Government of India) Fellow. Her work included synthesis of some small organic molecules, and solvent and substituent effects in various chemical processes. Further, she was awarded CSIR Associateship in 2000 and worked in the same department. In 2003 she joined as postdoctoral research fellow with Prof. Alan R. Katritzky at the University of Florida, Gainesville, USA. She is involved in applications of quantitative structure activity and property relationship (QSAR/QSPR) studies. Her current research interest is computer aided drug design.

Svetoslav H. Slavov, born in 1974 in Bulgaria, received his M.S. degree in Chemistry and Physics at Sofia University in 1997. As an assistant professor to Prof. B. Galabov at the same university, he led laboratory classes in molecular modeling, QSAR, and UV and IR spectroscopy. He obtained his Ph.D. in 2007 in molecular modeling at University of Tartu, Estonia. His research is focused on the improvement of the QSAR methodology and the application of the 3D-QSAR and the docking methods to drug design related problems.

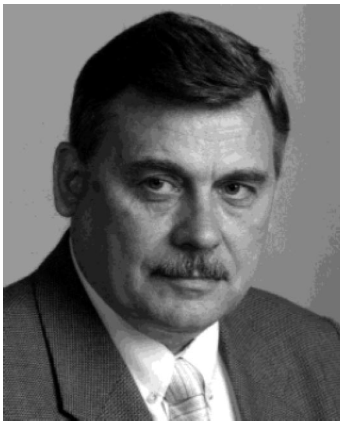
After retiring from his academic position at King's College, London, in 1999, Dennis Hall joined Alan Katritzky's research group at the University of Florida, where he acts as a group leader, Administrator for the on-line journal *Arkivoc* and co-organizer of the Florida Heterocyclic/Synthesis conferences (Flohet). Since joining the Katritzky team he has coauthored some 30 papers in the fields of heterocyclic chemistry, QSAR, insect control, and synthetic ion channels.

1. Introduction

1.1. Overview of QSPR Studies

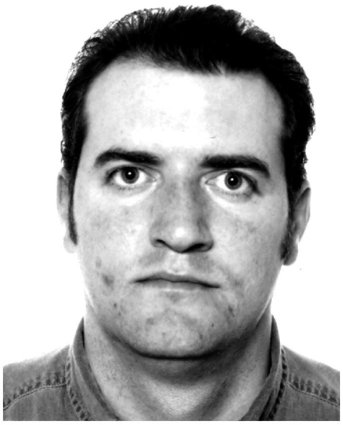
All properties of organic molecules—physical, chemical, biological, and technological—depend on their chemical structure and vary with it in a systematic way. The establishment of quantitative correlations between diverse molecular properties and chemical structure is now of great importance to society in assessing and improving environmental, medicinal, and technological aspects of life. These are expressed as quantitative structure—property relationships (QSPR) that relate physical, chemical, or physicochemical properties of compounds to their structures. Aside from the historically important QSPR models, the need of keeping the present review to a reasonable length limited our selection only to models of sufficient statistical quality published recently.

A major goal of the QSPR studies is to find a mathematical relationship between the property of interest and one or more descriptive parameters (descriptors) derived from the structure of the molecule. The descriptors used in the study may be empirical, i.e. experimental properties or properties derived from readily available experimental characteristics of the structure, or may be computed based on the structure. Classical physical organic chemistry has long been concerned with the correlation of chemical properties in terms of structures. The pioneering work of Hammet[1,2] and Taft[3−6] on the development of linear free energy relationships (LFERs) contributed considerable insight into organic reaction mechanisms. In 1947, the first structural descriptors (Wiener index and Platt number)[7−9] were developed for the correlations of thermochemical properties of paraffin hydrocarbons; nevertheless, throughout the 1950s, most of the correlation studies reported the use of empirical descriptors.



Mati Karelson (born in 1948) is Professor and Head of Theoretical Chemistry at the University of Tartu, Estonia. He received his Ph.D. in physical organic chemistry in 1975. His research has dealt with the theory of solvent effects, the foundations of QSAR/QSPR, and the development of the corresponding computer software. He is actively engaged in teaching, industrial consulting, and scientific management. He is also Courtesy Professor in Chemistry at the University of Florida.

Iiris Kahn received her M.Sc. (2003) and Ph.D. (2007) degrees in Molecular Engineering from the University of Tartu, Tartu, Estonia. She joined the group of Prof. Mati Karelson at the same university, working as a computational chemist (2000–2006) on the QSPR modeling of several environmentally relevant properties such as soil sorption coefficients and aquatic toxicity to the fish *P. promelas* and the ciliate *T. pyriformis* in the framework of the IMAGETOX programme. Together with a part of the research group, in 2006, she moved to the Tallinn University of Technology, Tallinn, Estonia, to work as a researcher on QSAR of cancer related targets within the CancerGrid project. Her research interests include the development of QSARs for use in environmental assessment and computer assisted drug design.



Dimitar A. Dobchev, born in 1977 in Bulgaria, received his M.S. degree in Theoretical and Atomic Physics at Sofia University in 2001. He obtained his Ph.D. in 2006 in molecular modeling at University of Tartu, Estonia. His research interests are related to various applications of QSAR/QSPR, chemometrics, mathematical chemistry and programming of chemical software.

Until the 1970s, most QSPR equations correlated spectroscopic, chromatographic, or other analytical properties of compounds. More recently, the QSPR approach has expanded to widely diverse areas of industrial and environmental chemistry. Initially, empirical molecular descriptors were obtained from experimental data for the development of QSPR equations (for example, Hammett substituent constants, σ; octanol/water partition coefficients, log P; Ostwald solubility coefficients, log L). However, many empirical descriptors reflect a complex combination of different physical interactions and, in addition, are not available for compounds yet to be synthesized.

Many efforts have been made to develop alternative molecular descriptors which can be derived using only the information encoded in the chemical structure. Much attention has been concentrated on "topological indices" and molecular descriptors derived from the connectivity and composition of a molecule which have made significant contributions in QSPR studies.[10−28] Nowadays, QSPR is used to correlate many diverse physicochemical properties of compounds with their molecular structures, through a wide variety of descriptors. The basic strategy of QSPR is to find an optimum quantitative relationship, which can be used for the prediction of the properties of compounds, including those unmeasured. QSPR studies became more prevalent with the development of new software tools, which allowed the understanding of how molecular structure influences properties and, significantly, afforded composition of structures with desired properties as a reverse task. The development of molecular descriptors based on structure is described in detail elsewhere.[29,30] QSPR has received significant contributions from various research schools.[31−37]

1.2. Scope of the Review

In the past two decades, QSPR models have gained extensive recognition in the correlation and prediction of physical, chemical, analytical, and technological properties of compounds. A major factor driving the widespread use of QSPR models is their aid in rational determination of properties of new compounds without the need to synthesize and test them. With the advancement of software technology, several computer programs have become available commercially and academically which enable the rapid calculation of thousands of structural descriptors for a compound in a fraction of a second. In order to process all these molecular descriptors at the same time and to build optimal structure−property models, multivariative statistical methods, such as multiple linear regression (MLR), principal component regression (PCR), and partial least-squares regression (PLS), are often used. Although thousands of molecular descriptors are already available for the QSPR modeling, the search for the best descriptors suitable to model a property is a major task. Experimentally determined values of many fundamental properties are unavailable in the literature, and their measurement is costly and time-consuming. Many QSPR models have been developed for the prediction of a wide range of properties, such as boiling and melting points, molar heat capacities, heats of vaporization, densities, aqueous solubilities, octanol−water partition coefficients, etc. Many reviews (including several from our group) have

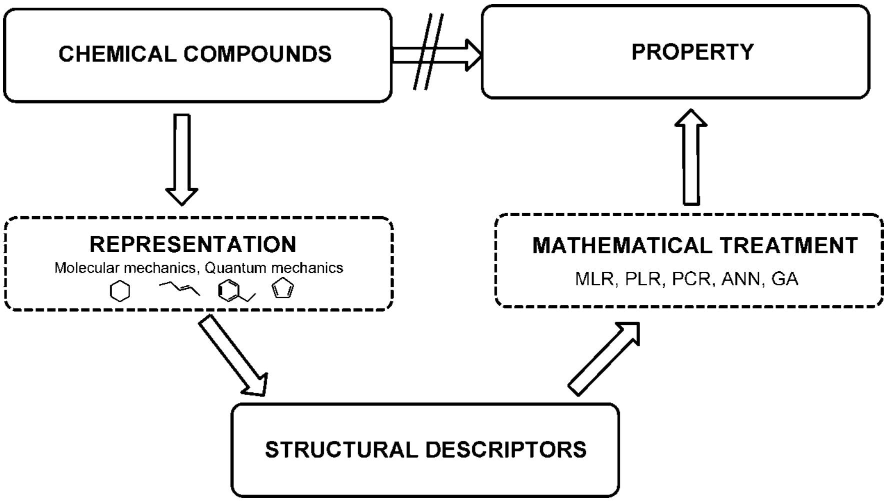


Figure 1. Flow chart of a QSPR problem.

appeared during the past decade on the QSPR applications based on structural descriptors.[22,36−46] The following are illustrative of the contributions from many other groups: (i) selection of molecular descriptors for quantitative structure−activity relationships (QSAR);[47] (ii) ANNs in molecular structure−property studies;[48] (iii) uniform-length molecular descriptors, QSPRs, and QSAR: classification studies and similarity searching.[49−51]

The present review summarizes recent QSPR research methods and applications. The main focus is placed on QSPR based on structural descriptors derived solely from chemical structure for the correlation and prediction of various physical, chemical, and physicochemical properties of compounds.

1.3. QSPR Approach

The main objective of the QSPR methodology is to quantify and relate determining factors for a particular measured property with molecular features of a given system of chemical compound(s). To achieve this purpose, one usually employs a mathematical model (F) that connects experimental property values with a set of features (molecular descriptors) derived from the molecular structures (eq 1):

$$\text{property} = F(\text{molecular descriptors}) \qquad (1)$$

The descriptors in eq 1 are numerical values which represent approximately the experimentally measured property in a space defined by the nature of the chemical compounds. Hence, the correct building of a model with relevant and consistent descriptors could provide insights into various underlying chemical, biological, or pharmacological processes. Also, a reliable model should be able to perform accurate prediction of the property values of other compounds not used in deriving eq 1; this is the second major goal of the QSPR methodology.

Building a QSPR model is an inductive process that depends on the set of compounds with predetermined properties. Therefore, there is no direct general model which can be applied for any compound. In practice, the QSPR methodology is applied in an indirect way that can be divided into two main stages relating the chemical compounds with their properties via structural descriptors and mathematical relations, i.e. (i) representation of the chemical objects and (ii) mathematical/statistical treatment. Figure 1 shows schematically the indirect approach to the QSPR problem. The representation is the process of applying the fundamental principles of chemical knowledge (molecular mechanics, quantum chemistry, etc) on the chemical compounds represented by their molecular structures. The connecting chain between (i) and (ii) comprises the structural molecular descriptors obtained on the basis of the representation and is regarded as part of (i). The mathematical treatment involves development of mathematical equations by taking into account both the molecular descriptors and the property in order to quantify their relationship.

2. Data Input for QSPR Modeling

2.1. Data Set Selection

The key step in developing comprehensive QSPR models "is the selection of an informative and representative data set", i.e. training set data.[52−56] QSPR models are only valid within their respective domains, being determined by the parameters associated with the chemicals in the training set, i.e., those chemicals used to develop QSPR models.

The input data preparation is a part of the representation stage (see Figure 1), and as such it is related to (i) the selection of the desired compounds used as objects to develop QSPRs and (ii) their predetermined (experimental) properties. The two points are up to the researcher and the task to be solved. The former point is connected with the chemical space of the compounds that the desired model would take into account. Mathematically, the chemical space for a given composition of matter can be defined as a set of all connection tables based on the molecular formula.[57] Clearly, the more atoms in the formula, the larger the number of possible variations of different compounds present in the parent formula. In practice, this leads to the existence of two classes of chemical sets, namely, homogeneous and diverse compound sets. Usually, the QSPRs developed on homogeneous data lead to better models compared to models developed on diverse sets. However, the applicability (predictions, analysis) of the models on homogeneous data sets is limited to compounds similar to those used to build the model.

Experimentally measured properties are either extracted from chemical databases or collected from the literature. The essential criteria for a satisfactory QSPR model are the availability of a set of experimental property data (i) of sufficient size and diversity and (ii) measured under the same (or similar) conditions with satisfactory reproducibility and accuracy. Several commercial databases of chemical compounds have been developed, but few of them are publicly available on the Web.[58] It is advisable to analyze the data set prior to building the QSPR models in order to confirm the basic requirements of the mathematical methods to be used to develop the quantitative relationship.

The reliability of the experimental property chosen for QSPR modeling is an important issue, since it determines the stability and predictability of the models. If the experimental errors are large, then building precise and reliable models is meaningless. In addition, many QSPR models are based on MLR techniques where normal distribution of the experimental property values is of vital importance for assessment of the model predictability and reliability. The closer the distribution function shape of the experimental data to normal, the more significant (accurate) are the statistical parameters assessing the models.[59] However, there are mathematico-statistical techniques that do not require normal distribution of the experimental data.[60]

If the original property values deviate greatly from the normal distribution, a transformation is desirable. The transformation functions frequently used are log(*Prop*), log(1/*Prop*), 1/*Prop*, and 1/(*Prop*)^2. Certain alternative transformations such as sine, tangent, or hyperbolic functions have been heavily criticized in the literature.[61]

Several empirically established criteria[62,63] concerning the data set selection are summarized below.

(i) Size of the data set: at least 25−30 compounds (frequently congeners) characterized by common structural features. Large and highly heterogeneous data sets are, however, typical for QSPR modeling of physicochemical properties.

(ii) Range of the property values: the property values should cover a range of at least 1 logarithmic unit.

(iii) Experimental error: the reliability of QSPR models decreases rapidly if the relative experimental error is higher than 15%.

(iv) The optimal ratio between the training and the test data sets should lie within the boundaries of 2:1 to 4:1.

2.2. Geometry Optimization

An important step in a QSPR study is the definition of the molecular structure. The (minimum) information required to specify a given molecule is its atomic composition and the manner in which those atoms are connected. The latter point requires the relative positions of all the atoms in space. Thus, the geometry optimization finds the coordinates of a molecular structure that represents a potential energy minimum (PEM). The stability of a compound is determined by its PEM and is usually related to the compounds that are the basic units of the pure substance. However, in reality, the calculations typically deal with equilibrium mixtures at nonzero temperatures. In this case, the measured properties reflect thermal averaging, possibly over multiple discrete conformers, stereoisomers, tautomers, etc. that are structurally different than the isolated compounds, and care is needed in making comparisons between theory and experiment.[59]

Obtaining the correct molecular structure is important for satisfactory descriptor calculations. Several geometry optimization methods are commonly used.[64,65] Such optimization methods combined with conformational search techniques lead to the PEM. The conformational search can be performed in a systematic or random manner with respect to small rotatable angles defining the step of each cycle. If many single bonds are present and/or the chosen rotational angle is small, the number of the conformers generated can be exceedingly large. In complex cases, it is desirable to have random sampling of the conformational space. As a result, a set of different conformers will be generated. The completeness of the set of conformations produced can be increased by augmenting the random searching cycles.[66]

The next stage requires full geometry optimization of the compounds studied. Recent progress in computational hardware and the development of efficient algorithms has assisted the routine development of molecular quantum mechanical calculations. The semiempirical methods supply realistic quantum chemical molecular quantities in a relatively short computational time frame. Quantum chemical calculations are thus an attractive source of new molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions. Indeed, many recent QSPR studies have employed quantum chemical descriptors alone or in combination with conventional descriptors.

In practice, the preparation of the molecules in computer format usually goes through drawing the molecular structure. There are many commercial and noncommercial programs having modules for two-dimensional (2D) drawing—Symyx Draw [www.symyx.com] (formerly ISIS Draw), Chem Draw [www.cambridgesoft.com], and Hyperchem [www.hyper.com]—or the drawing can be extracted from a chemical database in a certain file format. Three-dimensional (3D) structures are then generated using molecular mechanics (MM$^+$), quantum chemical methods, e.g. semiemprical AM1 (Austin Model 1) included in the MOPAC (MOlecular PACage) package, ab initio, etc.

2.3. Descriptor Calculation

Once the molecular structures are entered and the proper geometries are established, the next stage in the QSPR modeling is the generation of the descriptors. The molecular descriptors are precise mathematical values describing the physical and chemical properties of the molecules. Empirical indices (such as substituent constants and various electronegativity-related parameters) were most frequently used in early QSPR studies. Among them, electronegativity has remained a very popular and broadly employed descriptor.[67−73]

Substituent constants, however, are often disregarded by modern QSPR. With the increase of the computational power, quantum chemical, electronic, geometrical, constitutional, and topological descriptors are increasingly preferred. Quantum chemical, electronic, and geometrical descriptors (derived from empirical schemes or molecular orbital calculations) encode the molecule's ability to participate in polar or hydrogen bonding (donor, acceptor) interactions. The constitutional descriptors represent the chemical composition of a molecule and are independent of molecular connectivity and geometry. Examples of such descriptors are the numbers of particular atoms or bond types, numbers of particular ring systems, molecular weight, etc. They are fragment additive and reflect mostly the general properties of compounds related to their composition. Topological descriptors are calculated using data on the connectivity of atoms within a molecule. Consequently, these descriptors contain information about the constitution, size, shape, and branching, whereas bond length, bond angles, and torsion angles are neglected. Some of the most popular software packages capable of calculating an extensive list of descriptors include CODESSA PRO,[74] POLLY,[75] ADAPT,[76] OASIS,[77] Dragon,[78] Chem-X,[79] Tsar,[80] QSAR-Model,[81] and others.

However, due to mathematical and computational complexities, this seems unlikely to be realized in the foreseeable future. Thus, researchers need to rely on methods which, although approximate, have now become routine and have been demonstrated to provide results of real utility. Solving the Schrodinger equation for a moderate many-body particle system leads to a large number of quantities (both observable and nonobservable). For example, in the Hartree−Fock approximation (which is usually encoded in the semiempirical quantum-chemical methods; see below), the solution for the total (ground) energy of the system is obtained by averaging the Fock Hamiltonians. Several other parameters, including single electron energies, charges, potentials, etc., are also obtained. These variables can then be used to calculate diverse molecular descriptors, especially quantum-chemical descriptors. Therefore, the descriptors are direct outcomes from the theory and can include multiple quantum-chemical quantities in their definitions.

While the ab initio model Hamiltonian provides a complete representation of all nonrelativistic interactions between the nuclei and electrons in a molecule, available solutions of the respective Schrödinger equations are necessarily approximate and the computational time is proportional to a high exponential (N^4, N^5) of the number of electrons in the molecule, N; thus, practical ab initio calculations are still severely limited by the types of atoms and size of molecules.[82] However, even within these limits, molecules may be described by ab initio methods with some degree of reliability after a search of the potential energy surface(s) has been carried out at a lower level of theory. Most ab initio calculations have been based on the orbital approximation (Hartree−Fock method). In general, this method provides better results the larger the basis set (i.e., the number of atomic orbitals) employed, although, according to the variational principle, this is strictly valid only for the total electron energy of the molecule.

A wide variety of ab intio methods beyond Hartree−Fock have been developed and coded to account for the electron correlations in the molecule. These include configuration interactions (CI),[83,84] multiconfigurational self-consistent field (MC SCF),[85] correlated pair many-electron theory (CP-MET),[86] including its various coupled-cluster approximations, and perturbation theory (e.g., Møller−Plesset perturbation theory of various orders, MP2, MP3, MP4).[87,88] Most of these methods are extremely time-consuming and require large memory and fast CPUs. Therefore, they are impractical for the calculation of extended sets of relatively large molecules (i.e., more than 10 atoms).

As an alternative to ab initio methods, semiempirical quantum-chemical methods can be used for the calculation of molecular descriptors. These methods have been developed within the mathematical framework of the molecular orbital theory (SCF MO) but based on simplifications and approximations introduced into the computational procedure which dramatically reduce the computational time. Experimental data on atoms and prototype molecular systems have often been used to estimate the values of the parameters in these methods; therefore, they are widely known as semiempirical methods.[89] Different parametrizations, such as MNDO, AM1, or PM3, supply realistic quantum-chemical molecular quantities in a relatively short computational time. Thus, they are an attractive source of molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions. These semiempirical methods, however, are based on limited experimental parameters and in some practical cases fail to produce good results. Therefore, they are the subject of continual improvement.[90]

The methods used for analyzing the electron density (*charge* partitioning *schemes)* of molecular systems can be divided into three groups: (i) wave function based methods (Mulliken population analysis, natural population analysis); (ii) molecular electrostatic potential fitting based methods (such as the CHELPG and Merz−Singh−Kollman (MK) scheme), and (iii) electron density based methods (such as AIM). Due to the fundamental problem of deciding where atoms in a molecule actually start and where they end, no precise atomic charge can be defined. Nevertheless, the calculation of atomic charges can still be quite helpful, if only for use as an effective parameter in force field calculations or QSPR analysis of closely related (similar) systems.

Molecular modeling techniques enable the definition of a large number of molecular and local quantities characterizing the reactivity, shape, and binding properties of a complete molecule as well as of molecular fragments and substituents. Because of the large, well-defined physical information content encoded in many theoretical descriptors, their use in the design of the training set in a QSPR study presents two main advantages: (i) the compounds and their various fragments and substituents can be directly characterized on the basis of their molecular structure only, and (ii) the proposed mechanism of action can be directly accounted for in terms of the chemical reactivity of the compounds under study. Consequently, the derived QSPR models will include information regarding the nature of the intermolecular forces involved in determining the chemical, physical, or other property of the compounds in question.

3. Modeling Procedures

Most QSPR treatments utilize a program to calculate descriptors and then try to select a small number of significant descriptors in a purely empirical fashion to form an equation. The descriptors are calculated for a “training set” of compounds for which a property of interest has been measured. QSPR methodology has been aided by new software tools, which allow chemists to elucidate and to understand how molecular structure influences properties. Most importantly, this helps researchers to predict and prepare structures with optimum properties. Therefore, the software is also of great assistance for chemical and physical interpretation.

In the past ten years, our groups at the University of Florida and at Tartu/Tallinn in Estonia have developed multipurpose statistical analysis software in the form of the CODESSA (COmprehensive Descriptors for Structure and Statistical Analysis) program, later updated as the CODESSA PRO program.[74] The software includes robust multilinear regression (best multilinear and heuristic) algorithms which feature extraction, principal component regression (PCR), and partial least-squares (PLS) regression. In addition, it is being further updated with ANN, fragment molecular features, and genetic algorithms (GA) so that it can meet state-of-the-art methods for powerful QSPR modeling. CODESSA PRO also provides user-friendly tools for manipulation and extraction of the calculated descriptors for predefined data sets, which can be further treated by external software products such as Systat,[91] Statistica,[92] JMP,[93] etc.

3.1. Multivariate Approaches—Linear Aspect

3.1.1. Multilinear Regression

Multilinear regression (MLR) is a very widely used approach in QSPR. The simple linear regression model assumes that the response variable y is a straight-line function of a single explanatory variable x. Multiple linear regression is an extension of simple linear regression including more than one dependent variable. As with simple linear regression, the regression coefficients in MLR for each independent variable are determined so that the sum of the squares of the errors is minimized. Thus, the general multilinear equation can be written in the following form as in eq 2:

$$y = b_0 + b_1x_1 + b_2x_2 + \ ... \ + b_px_p \qquad (2)$$

where y is the response (or the dependent variable), x_1, x_2, ..., and x_p represent the explanatory (or independent) variables, b_1, b_2, ..., and b_p are the regression coefficients, and b_0 is the intercept.

A common method of handling a large number of x variables is to use a stepwise regression routine aimed to identify the “best” set of x variables, i.e., the set which explains the greater part of the data variance. This method adds, deletes, or both adds and deletes x variables (one at a time) to arrive at the “best” regression equation. At each step, the decision to add or exclude/skip an x variable is based on a test of whether that variable contributes significantly to the model. In the simplest implementation, an arbitrary point to cut off further variables from entering the equation “the p value” (level of significance) is used.

The squared correlation coefficients, R^2, squared cross-validated correlation coefficients, R^2_{CV}, Fisher criterion value, F, and standard deviation, s, all give information about the “goodness” of the model.

Usually a QSPR study deals with a large number of molecular descriptors. The search for the best MLR model in such a large descriptor space is not a trivial task. Various regression techniques coupled with variables selection procedures have been proposed for the selection of the “best set” of regression predictors, such as backward elimination, forward selection, and ridge regression.[94−97]

The best multilinear regression (BMLR) method implemented in CODESSA PRO is able to find the “best” regression in a short computational time in a descriptor space of hundreds of variables. The criteria are managed in a way that a certain chemical space can be explored more precisely for the best correlations. The BMLR method is based on the (i) selection of the orthogonal descriptor pairs and (ii) extension of the correlation (saved on the previous step), with the addition of new descriptors improving the statistical parameters of the model. A cutoff value determining the improvement of R^2 is used to limit the number of descriptors entering the equation. However, the number of descriptors in the equation should not exceed certain limits because it could lead to an overfitted model.[94] The problem of overfitting may be overcome by exploring the improvement of R^2 and R_{cv}^2 as a function of the number of descriptors in the model.

3.1.2. Principal Component Regression/Analysis and Partial Least Squares Analysis

Multiple linear regression suffers from several shortcomings:[98] (i) failure for highly intercorrelated data (i.e., descriptors); (ii) assumption that the data have no noise; (iii) ability to model only one y variable at a time; and (iv) requiring more observations than variables.

Principal component regression (combining PCA and MLR) and partial least-squares (PLS) are alternative methods that can be applied to address all of the issues i−iv.[99−101] PCA reduces statistically the dimensionality of the data while retaining most of the variation in the data set. This reduction is performed by identifying the directions (called principal components) along which the variation in the data is maximal. Theoretically, the number of components extracted in PCA is equal to the number of the observed variables. However, in most cases, the first few components alone already account for the majority of the variance, so only these first few components are retained, interpreted, and used.

The PCA components possess two very important characteristics: (i) each component accounts for the maximal amount of variance in the observed variables that was not accounted for by the preceding components, and (ii) no component is correlated with any preceding component.

The partial least-squares (PLS) method is constructed from the concept of PCA. Just as with PCA, in PLS the data analysis is simplified by projecting the data into a low dimensional “latent variable” space (the acronym PLS has also been taken to mean “projection to latent structure”). Components in PLS are constructed to maximize the covariance between the dependent variable y and the original independent variables x. However, (unlike PCA) PLS analysis also simultaneously calculates the latent variables for the two matrices—the matrix of independent (x) and the matrix of dependent (y) variables, together with the relationship between them. Thus, for PLS the new set of “latent variables” is a set of conjugant gradient vectors to the correlation matrix rather than a set of successive orthogonal directions that explain the largest variance in the data as in PCA.[102] The methodology and applications of PCA are described elsewhere.[103]

3.1.3. Chance Correlations

A QSPR model usually contains a small number of independent descriptors out of many evaluated. The descriptors selected for inclusion in such an equation are chosen so that the overall equation is highly significant by standard statistical criteria. However, these criteria relate to the individual variables in the final equation and do not take into account the number of descriptors actually screened for possible inclusion in the equation. When the number of possible independent variables considered becomes very large, it may become possible that a correlation will occur purely by chance. Because this factor is not reflected in the standard statistical criteria, it is important to consider the number of variables screened for possible inclusion in the equation.

Recently, various sets of published multiparameter QSPR models have been analyzed, in particular with attention to the possibility of "chance correlations" occurring in the published models.[104,105] However, we have demonstrated that the possibility of chance correlations can be minimized so as to be negligible by using appropriate procedures.[106] Most importantly, the collinearity of natural descriptor scales needs to be strictly controlled during the forward selection processing. Satisfactorily, the criteria used in the BMLR procedure have been proven to be sufficient for the elimination of chance correlations due to the nonorthogonality of the scales. A sufficiently large number of data points in the set give additional assurance for avoiding chance correlations. While tests with randomly generated scales might have possible significance, in such cases, the size of the space generated by these random scales must be compatible with the size of the actual descriptor space.

3.2. Multivariate Approaches—Nonlinear Approaches

The above-described chemometrics methods, MLR, PLS regression, and PCR, are widely used in the QSPR area. In principle, they give a multilinear expression of the property under study in a given descriptor (or principal component) space. However, nonlinear approaches, such as artificial neural networks (ANN) or support vector machines (SVM), can also be employed to derive flexible correlation models between the molecular structures and properties. They are able to "catch" hidden nonlinearities between the property and the descriptors which make them in most cases better predictors than the MLR models. However, these nonlinear methods are not as intuitively easy to interpret as the MLR models. Although nonlinear models are very useful, the real world is rarely "linear" and most QSAR/QSPR relationships are nonlinear in nature. Once a nonlinear relation has been found and validated, it can be a good predictor and indicator, such as, for instance, the famous J-shaped dose response dependence.[107]

3.2.1. Artificial Neural Networks

Artificial neural networks (ANN) have been applied in many diverse scientific endeavors, ranging from economics, engineering, physics, and chemistry to medical science.[108] These computational methods evolved from attempts to understand and emulate the brain's information processing capability. The brain consists of multimodal neural networks that extract and recombine relevant information received from their environments and render the brain capable of making decisions that satisfy the needs of the organism. These capabilities of the brain can be emulated with ANNs, which can conceive complex nonlinear input−output transformations. Their nonlinear feature extraction capability suggests their potential usefulness in QSPR.[103,109−111] There are numerous types of ANN, such as multilayer perception (MPL), Kohonen, probabilistic, radial basis, and entropy machines networks that differ by their ideology, topology, and optimization algorithms.[112]

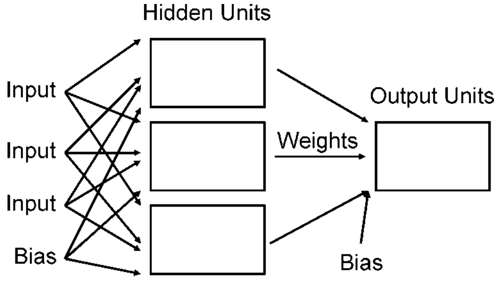


Figure 2. Typical ANN topology.

ANNs are typically used when a large number of observations are available and a nonlinear relationship is expected or when the problem is not understood well enough to apply other methods. The "architecture" of the ANN consists of a number of "neurons" that receive data from the outside, process the data using transformation functions, and produce a signal. The "neurons" actually act as nonlinear transformation functions. When more than one of these neurons is used, nonlinear models can be fitted. These networks have been applied to the modeling of numerous problems, including QSPR. Neural networks are known for their ability to model a wide set of functions without knowing the model a priori. The back-propagation network receives input signals which are then multiplied by each neuron's weights (Figure 2). For each neuron these products are summed and a nonlinear transfer/activation function is applied. The role of the bias is to shift the transfer function to the left or right. The resultant sums from the previous step are then multiplied by the output weights, transformed, and interpreted. Since a back-propagation network is a supervised method, the desired output must be known for each input vector. The difference between the desired output and the network's predicted output defines the ANN model error, which needs to be minimized. This error is then propagated backward through the network, adjusting the weights, so that, on the next cycle, the generated predictions will come closer to the desired output.

Four important factors must be considered when using neural networks:

(i) The design of the network is critical with respect to the number of hidden units involved—if too many hidden units are used, the network would overfit or "memorize" the data. Conversely, if too few hidden units are used, the network will fail to generalize and will become unstable.

(ii) The length of the training time must always be considered—if excessive training periods are used, the network might become overtrained and, thus, destabilized.

(iii) Appropriate test and training sets must be defined. The training set should adequately represent the entire data set and be sufficiently large in order to properly train the neural network.

(iv) The results obtained from ANNs can be difficult to interpret, particularly in application to drug design. The standard approach for interpretation is the analysis of the weight magnitudes.

In recent years, the literature concerning ANNs as applied to QSPR has grown dramatically, suggesting that the importance of applications of ANN in molecular modeling is a major driving force. Among numerous studies using ANN in QSPR, major contributions are due to the groups of Jurs,[103,113] Zupan and Gasteiger,[114] Zefirov,[111] and our group (Katritzky and Karelson).[115−117] The ANN models have frequently been shown to possess better predictive characteristics compared to the models using standard multilinear regressions. The flexibility of the ANN also exhibits a better ability to represent predictive models.

3.2.2. Genetic Algorithms

The genetic function approximation algorithm was originally conceived by taking inspiration from two seemingly disparate algorithms: Holland's genetic algorithm (GA) and

Friedman's multivariate adaptive regression splines (MARS) algorithm.[118,119] Today, this technique is widely used in the QSPR area as a tool for modeling of complex properties as well as for selecting meaningful variables for ANN.[120,121]

The genetic function approximation (GFA) algorithm is an alternative to standard regression analysis for constructing QSPR equations. The application of GFA leads to multiple models generated by evolving random initial models using a GA. Each cycle performs a crossover operation to recombine terms of better scoring models, thus improving the parameters of the model. The method is efficient for generating QSPR equations from a large number of descriptors.

GFA works well only for preselected smaller subsets of descriptors; otherwise, it might be trapped in local optima, and thus, an important descriptor or combination of descriptors could be lost during the crossover process. One of the main pitfalls of the method is that once an important descriptor is dropped during the crossover process, it can never be recovered.[122]

3.2.3. Support Vector Machines (SVMs)

An elegant alternative to the ANN approach was developed to avoid the existence of many local minima and the uncertainty about the number of neurons needed for a given task.[123] The so-called "support vector machine" methods are designed around the computation of an optimal separating hyperplane which provides the minimum expected generalization error in a multidimensional space called "future space".[124]

In this *m*-dimensional space, each compound is represented by a point which may be thought of as a vector of *m* numbers (descriptors). The support vector machine can actually locate the hyperplane without ever representing the future space explicitly, simply by defining a function, called a kernel function.

The main advantages of SVM are as follows: (i) stable, reproducible results are produced, independent of the optimization algorithm; (ii) the optimum solution (global minima) is guaranteed; (iii) a few parameters need to be adjusted—the regularization parameter (*C*) and the nature and the parameters of the kernel function.

Despite huge advantages over ANNs, SVMs have two major drawbacks in that they are even slower than ANNs and provide only "black-box" solutions.

3.3. Expert Systems

Expert systems were originally introduced by Edward Feigenbaum and were the first truly successful form of artificial intelligence software. Expert systems seek to provide an answer to a problem or clarify uncertainties without the need of consultation of human experts. All expert systems so far designed provide answers only in a specific narrow problem domain, but with the increase of the computational power, it is expected that in the near future more general systems will appear.

The most common methods to simulate the performance of a human expert are (i) the creation of a so-called "knowledge base" which uses the knowledge representation formalism to capture the subject matter expert's (SME) knowledge and (ii) a process of gathering that knowledge from the SME and codifying it according to the formalism, which is called knowledge engineering. It is not necessary for an expert system to have a learning component. However, its efficacy can be proven by placing it in the same real world problem solving situation as human experts, typically as an aid to human workers or as a supplement to an information system.[125−127]

Some of the distinctive characteristics of an expert system are as follows:

(i) it contains dynamically synthesized step sequences needed to reach a conclusion for each new case, which were not explicitly programmed when the system was built;

(ii) it allows processing of multiple values for any problem parameter and, thus, permits more than one line of reasoning to be followed and the results of incomplete (not fully determined) reasoning to be presented;

(iii) problem solving is achieved by applying specific knowledge rather than a specific technique. This characteristic reflects the belief that human experts do not process their knowledge differently from others, but they do possess different knowledge.

Compared to human operators, expert systems possess several advantages: (i) they always ask questions that a human might forget to ask; (ii) they maintain hold and maintain significant levels of information, unifying the knowledge of many human experts; (iii) they can work uninterruptedly; and (iv) they can assist more than one person at a time. The disadvantages of expert systems include the following: (i) lack of the common sense needed in some decision making; (ii) inability to respond creatively in unusual circumstances; (iii) errors in the knowledge base may lead to wrong decisions; and (iv) inadaptive to changing environments, unless the knowledge base is changed.

Most expert systems with applications in chemistry and biochemistry are designed to solve complex problems and predict properties such as toxicity, bioactivity, drug efficacy, etc. However, in some cases these systems are also able to predict widely used physicochemical parameters, such as pK_a, log *P*, retention times, etc. A few examples of expert systems providing solutions to QSPR problems are given below.

The CRIPES (chromatographic retention index prediction expert system) expert system was developed to predict the retention time properties of organic molecules in reversed-phase HPLC using indices based on an alkyl-aryl-ketone scale derived from empirical quadratic expressions.[128]

The SOL expert system[129] estimates the quality of experimental aqueous solubility data and is also able to screen out reported erratic values, such as the 1.29×10^{-8} and 6.34×10^{-8} values for the solubility of PCB 101.

The CAMEO (computer assisted mechanistic evaluation of organic reactions) is a modular expert system that predicts the outcome of organic reactions given starting materials, reagents, and reaction conditions.[130] This expert system is also able to predict the pK_a values of a wide range of organic compounds using a fragment based approach with an error not exceeding 2 pK_a units.[131]

SPARC (spark performs automated reasoning in chemistry) is an expert system for the estimation of chemical and physical reactivity.[132] In general, SPARC utilizes linear free energy theory (LFET) to compute thermodynamic properties and perturbed molecular orbital (PMO) theory to describe quantum effects such as delocalization energies or polarizabilities of icelectrons. Some of the properties which can be predicted are as follows: (i) equilibrium constants for complex speciation (ionization and tautomerization) and interphase distribution (gas/liquid, liquid/liquid, solubilities) and (ii) rate constants for reactivity (solvolysis and redox).

SPARC is especially accurate when predicting the pK_a values at 25 °C (R2 = 0.994 and RMSE = 0.37).

The HPLC-METABOLEXPERT expert system developed by Valko et al.[133] predicts the retention indices of metabolites. For this purpose, the system requires retention data and the octanol−water partition coefficient for the parent drug molecule. The log P is calculated according to the Rekker fragment system and by determining the contribution of the structural differences between the parent compound and the metabolite to the octanol−water partition coefficient, and then relating this contribution to reversed-phase retention data, the retention data of a metabolite can be predicted.

Szepesi and Valko[134] also developed the EluEx expert system, which can calculate the log P values on the basis of the structure of the compounds. Equations and pass−fail criteria (the capacity factors and asymmetry factors should fall within certain limits) are used to predict the final mobile-phase composition. According to the authors, the system may fail for some compounds, such as quaternary ammonium salts and very lipophilic compounds.

3.4. Model Selection

In a QSPR study, one usually develops several equations, among which the best should be chosen. The following general steps are used for the selection of an equation:

(i) Outlier identification: an outlier is an atypical value not belonging to the distribution of the rest of the values in the data set. Data points with deviations at least twice greater than the standard deviation of the data are usually considered outliers. This definition is correct only when the distribution is unimodal and symmetrical (most cases). For skewed data, the median is a better indicator of the central location than the mean. In such cases, observations lying more than 1.5 interquartile distances away from the closest quartile can be considered outliers. Extreme outliers are those which lie in more than 3 interquartile distances from the closest quartile. Outliers that cause a poor fit degrade the predictive value of the model; however, this has to be balanced with loss of generalizability if they are removed. Among all possible equations generated, those characterized with few and/or explainable outliers should be selected as reliable and potentially useful. In the case of univariate statistics, the outliers may be identified before fitting a model, but multivariate outliers, if present, can be identified only when the model of the best fit is obtained. Multivariate outliers having (i) high leverage and low discrepancy do not affect the regression line but tend to increase R^2 and reduce the standard error; (ii) low leverage and high discrepancy tend to influence the intercept but not the slope of the regression or R^2, while usually inflating the standard error; and (iii) both a high leverage and a high discrepancy influence the slope, the intercept, and the R^2 value. Parameters such as Mahalanobis distances,[135] leverages, Q^2 residuals, T-Hotelling,[136] etc. are commonly utilized for multivariate outlier identification.

(ii) The 5:1 rule of thumb: given enough parameters any data set can be fitted to a regression line. As a consequence, regression analysis generally requires significantly more data points than parameters. A useful rule of thumb is that the ratio between the objects and the variables should be at least five to one for the MLR analysis—otherwise there is a high risk of "by chance" correlation.[137]

(iii) Principle of parsimony (Occam's Razor): the principle postulated by William of Occam states that, among a set of equally good explanations for a given phenomenon, the simplest one is the most probable. It is called Occam's razor because he was "trimming down" his explanations to the bare minimum. In QSPR modeling, the principle of parsimony means that (a) models should have as few parameters as possible, (b) models should be pared down until they are minimally adequate, and (c) simple explanations are better alternatives than those more complex.

The process of model simplification is an integral part of hypothesis testing. In general, a variable is retained in the model only if its removal causes a significant decrease of the statistical parameters compared to those of the current model. However, when simplifying the model, one should be careful not to lose the essential parts. This situation is reflected in Einstein's clever addition to the Occam's razor: "A model should be as simple as possible. But no simpler."

3.5. Model Validation

Validation of the models developed is an important aspect of any QSPR study. Once a model is obtained, it is important to determine its reliability and statistical significance. Several procedures are available to assist in this. These can be used to check whether the number of parameters is appropriate for the data available, as well as to provide some estimate of how well the model can predict the property for new molecules. In order to be reliable and predictive, QSPR models should (1) be statistically significant and robust, (2) be validated by making accurate predictions for external data sets not used in the model development, and (3) have a defined domain of application. There are various ways to express the performance of regression models which are widely used in QSAR. The most common parameters are the "explained variance" for the response variable y, denoted R^2, and the residual standard deviation (RSD, s^2). The term R^2 is often referred to as "model fitness" and should preferably be as close to unity as possible, while the RSD should be kept small. For judging a model's predictive power, meaning how well it performs in forecasting, techniques such as cross-validation, bootsrapping, external validations, and permutations of the data (scrambles) are often used. The most widely used validation procedures can be classified as follows.

3.5.1. Internal Validation

The idea behind the internal validation is to predict the property value for a compound or a group of compounds using the regression equation calculated from the data for all remaining compounds of the QSPR model. Variations of this technique are leave-one-out (LOO) and leave-many-out (LMO).[52] In evaluating the prediction ability of regression models, the criterion most used is the leave-one-out cross-validated R_{loo}^2 (Q2,q2), defined as follows:

$$q^2 = 1 - \sum (y - y')^2 / \sum (y - y'')^2 \qquad \text{(A)}$$

In eq A, y, y', and y'' are the measured, predicted, and averaged (over the training set) values of the property. From a practical point of view, it is considered that if Q^2 is greater than 0.5, it is an indicator for the high predictivity power of the model. However, it has often been claimed that the use of only this criterion is often too optimistic for model validation,[53] because models so validated in some cases turn out to be not predictive if more severe validation is applied.

On analyzing such models, chance correlations, noisy variables, and too predictor collinearity are frequently the cause of their lack of predictivity.

In a typical LMO (in this case, leave-$^1/_3$-out) validation,[138] the parent data points are sorted in order of their property values and divided into three subsets (A, B, C) as follows: the first, fourth, seventh, etc. data points comprise the first subset (A); the second, fifth, eighth, etc. comprise the second subset (B); and the third, sixth, ninth, etc. comprise the third subset (C). Three training sets are then prepared as combinations of two subsets (A and B), (A and C), and (B and C). For each training set, a correlation equation is derived using the complete descriptor pool. The equations obtained are then used to predict property values for the compounds of the remaining sets (A, B, or C, respectively). The efficiency of the QSPR models to predict property values is assessed by the squared correlation coefficients (R^2) and standard deviations (s) between experimental and predicted data for each test set (A, B, or C). The descriptors found "effective" for each of the submodels are further used to form a reduced descriptor space from which the final model for the whole set (A + B + C) will be constructed.

The possibility of so-called chance correlations is an important issue related to the validation of the QSAR models. Typically, a pool of independent variables (descriptors) of possible relevance to important physicochemical parameters relating to the series of compounds under discussion is evaluated by multiple-regression analysis for correlation with the activity values. The correlation equations emerging from such an analysis generally contain a small number of independent descriptors from the large pool evaluated. The descriptors selected for inclusion in such equations are chosen so that the overall relationships are highly significant by standard statistical criteria. However, these criteria relate to the individual variables in the final equation and do not take into account the number of descriptors actually screened for possible inclusion in the equation. Clearly, the larger the number of possible independent variables considered, the greater the possibility that a correlation will occur purely by chance. There are very useful criteria and analyses in the literature for tackling this drawback of the models. We now refer the reader to the most useful works on this topic.[54−56,106]

In addition, randomization tests can be used in conjunction with the LMO or LOO. These tests consist of repeated elaboration and random shuffling of the data for which the model equations are tested. Due to a factorial increase in time of permutations, the Monte Carlo method is often used for producing a randomization test.[139,140]

Another useful technique for the validation of QSAR models is so-called bootstrap. The main idea of this method is that many new data sets called bootstrap samples are created from the original data set by random replacements. By performing such resampling many times, a good estimate can be obtained of the distribution of the statistics of interest. Hence, the distributions can be seen as approximations to the true distributions of the estimators, and, thus, statistics of interest such as bias, standard deviation, and confidence intervals can be derived from them in the usual manner.[141]

3.5.2. External Validation

One of the most widely used methods of correlation testing involves the use of an external validation set. This so-called "test set" should be sufficiently large to give a reasonable estimation of the model quality, especially when a random selection was used for its construction. However, a large test set can be constructed only when the initial data set itself is large—in all other cases, a small-sized validation test set would be useless because the final result will be a random estimate and would not reflect the "true" predictive power of the model.[142] Thus, for small-sized data sets, the splitting of the data is not a suitable solution, since the size decrease of the "training set" may itself lead to poorly constructed models. In such cases, we recommend an extensive use of internal validation procedures as described in section 3.5.1.

The purpose of the external validation is to evaluate how well the equation generalizes the data. The difference between the test statistics of the training and external validation data sets is a measure of the reliability of the correlation. A valid model with high generalization ability has R^2 and s for the validation set similar to those of the model. The predictive power of the QSPR models is often quantified in terms the root-mean-square error (RMSE), residual standard deviation (RSD), or predictive squared correlation coefficient Q^2.[142]

3.6. Model Interpretation

Another important aspect of the QSPR modeling is the extraction of the structure−property relationship information encoded in the model. The development of a (multi)linear model involves selecting one or more descriptors that provide a statistical correlation with the experimental property values. A preconceived notion of the physicochemical interpretation of the descriptors involved could result in misinterpretation of the underlying structure−activity relationship. There are two important pieces of information one needs in order to generate a meaningful QSPR model: (i) the knowledge of what features of the structure are measured by a given descriptor and (ii) the knowledge of how structural changes influence the experimentally observed property. In the case of whole-molecule descriptors, e.g. molecular connectivity indices, the changes in structure measured by the descriptor may be occurring in several places. However, the important structural changes which affect the observed property may be localized at a particular position in the molecule.

It is well-known that a statistical correlation between an observed property and a descriptor does not necessarily mean causality. A credible QSPR model should describe a causal relationship between descriptors and observable properties. To establish the credibility of the model, it is crucial to rationalize the physicochemical/biochemical basis of the correlation. Unless the equation is very simple, a routine examination of the model may not allow such rationalization because (i) the coefficients of the equation usually represent a combination of two or more structural trends in the model, (ii) some structural descriptors are mathematical constructs that may lack a direct physicochemical interpretation, e.g. molecular connectivity indices,[143] and (iii) a descriptor may be acting as a surrogate measure for structural features which are not characterized accurately enough by another descriptor with a more intuitive physicochemical interpretation. For example, the topological changes in the molecules are related to the changes in their geometry and, thus, could act as a measure of the shape of these molecules. However, changes in branching can affect the electron distribution, which often changes the reactivity or polarity. Thus, descriptors that characterize the charge distribution can act as a better measure of branching, and shape, than do the typical topological descriptors.[144] As a result, emphasizing the type of features measured by a certain descriptor could be misleading.

In summary, to build a statistically valid QSPR model, one must know both the essential structural features that influence the studied property and the nature of the descriptors involved.

4. Simple Physical Properties Involving Single Molecular Species

Most physical properties of organic compounds depend functionally upon the number, kind, and structural arrangement of the atoms in the molecule. The number and kind of atoms are both constant in isomers, and hence, the differences in their physical properties are due to structural relationships. Experimentally determined values of many fundamental properties are unavailable in the literature, and their measurement is costly and time-consuming. As a result, accurate prediction of properties of compounds has become increasingly important and useful to the producers and consumers of organic chemicals.

4.1. Boiling Points

Boiling point (conventionally at 760 mmHg pressure) is important for the characterization and identification of a compound. It also provides an indication of the volatility of a compound. Other physical properties, such as critical temperatures,[145] flash points,[146] and enthalpies of vaporization,[147] can be predicted or estimated from boiling points. With the increased need for reliable data for optimization of industrial processes, it is important to develop reliable QSPR models for the estimation of normal boiling points for compounds not yet synthesized or whose boiling points are unknown.

Many methods have been developed for the estimation of the normal boiling points of compounds, and numerous QSPR correlations have been reported. Early attempts were made to correlate boiling points of homologous hydrocarbons with the number of carbon atoms or molecular weight.[148] Later methods employed physical parameters such as parachor and molar refractivity.[149] Earlier methods for the estimation of boiling points have been summarized by Rechsteiner[147] and Horvath.[150] Efforts were made to estimate boiling points by group contribution additivity (GCA)[147,151] based on the assumption that cohesion forces in liquids are predominantly short-range[152] and proceed from the division of a molecule into predefined structural groups, each of which adds a constant increment to the value of the property.[153] Group contribution methods provide good prediction of boiling points,[154,155] with an average absolute error of 15.5 K, for small and nonpolar molecules. However, GCA methods are limited to molecules containing groups present in the calibration set of molecules, and some group contribution schemes are not comprehensive enough to cover multiple substitutions of functional groups.

Aside from simple correlations of boiling points with the carbon number or molecular weight for homologous series of compounds, Wiener was the first to correlate boiling points with structurally based topological descriptors.[7] Wiener introduced two structural parameters, path number W (named later the Wiener index), defined as the sum of the distances between any two carbon atoms in the molecule,[7] and Wiener polarity index (P), defined as the number of unordered pairs of vertices for which the distance between any two verteces is equal to 3. Based on these indices, he predicted the boiling points of paraffins with an average error of 1 °C.[7] Other topological indices, including the Randić,[11] and Kier and Hall[15] molecular connectivity indices, have been successful in correlating the boiling points of alkanes and amines.[15] For more than four decades, the correlation of boiling points of hydrocarbons with chemical structure has been of considerable interest. However, for better predictability of a property in the form of a general model, a search for better descriptors has been a focal point of QSPR research.

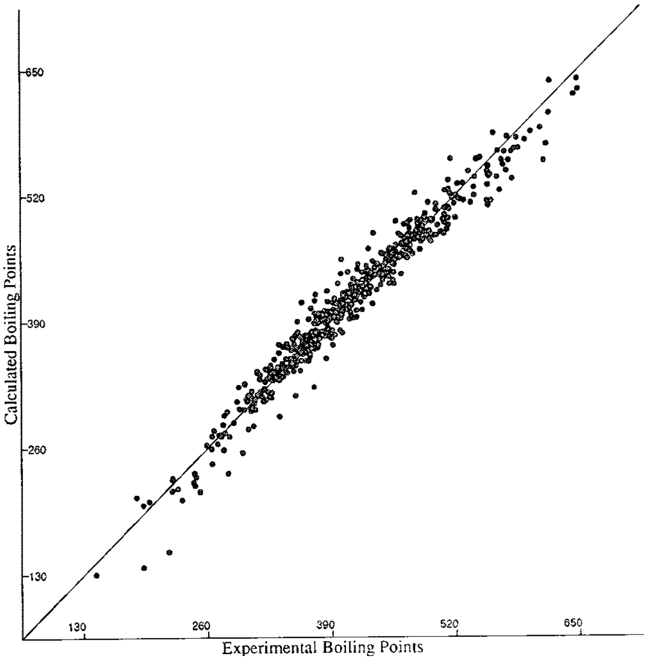


Figure 3. Plot of calculated vs experimental boiling points of 584 compounds using the 8-parameter model. Reprinted with permission from ref 157. Copyright 1998 American Chemical Society.

At present, a large number of QSPR models have been developed for the correlation and prediction of boiling points of diverse classes of organic compounds, such as hydrocarbons, halohydrocarbons, alcohols, carbonyl compounds, amines, nitriles, pyrans, furans, thiophenes, sulfides, ethers, and peroxides. Some QSPR models previously reported for predicting boiling points are summarized in Table 1.

Katritzky et al.[156] derived a two-parameter QSPR model for a training set of 298 diverse compounds (saturated and unsaturated hydrocarbons, halogenated compounds, and hydroxyl, cyano, amino, ester, ether, carbonyl, and carboxyl functionalities) using CODESSA (Figure 3). The two descriptor linear equation showed R^2 of 0.954 and is robust, as shown by the statistically significant squared cross-validation coefficient R^2_{CV} of 0.953 with standard error s of 16 K. Importantly, the two parameters selected by the descriptor forward selection procedure, the cubic root of the gravitation index ($GI^{1/3}$) and the hydrogen donor charged surface area (HDSA-2), are understandable in physical terms. The gravitation index describes the distribution of the mass of a molecule about its center of gravity and is associated with dispersion and cavity-formation effects in liquids. The hydrogen donor charged surface area is a measure of the propensity of a compound to form hydrogen bonds. The two-parameter QSPR equation reflects quantitatively the well-known fact that the boiling point of a compound depends on the mass of its molecules and their tendency to stick together, and it is equally well-known that the most important

Table 1. QSPR Models Developed for Prediction of Boiling Points

type of compd	*N*	molecular descriptor	QSPR method[a]	R^2	*s*	ref
alkanes	94	TIs (W, P)	MLR		0.97 (avg)	Wiener[7]
olefins	123	TIs (8)	MLR	0.998	1.78	Stanton et al.[158]
alkanes	74	TIs (5)	MLR	0.999	1.86	Needham et al.[159]
		ad hoc (5)		0.998	2.0	
alkanes (C_2-C_7)	21	TI (Kier connectivity index $\chi^{1/3}$)	RA	0.999	2.825	Randić et al.[160]
furans and tetrahydrofurans	209	CPSA	MLR	0.968	11.2	Stanton et al.[158]
furans, tetrahydrofurans, and thiophenes	209	TIs, electronic, geometrical	RA	0.969	11.2	Stanton et al.[161]
alkanes (C_2-C_7)	72	TIs (LOVI's)	RA	0.994	3.9	Balaban et al.[162]
alkanes	150	TIs (distance matrix)	MLR(2)	0.981	5.93	Mihalić et al.[163]
halogenated alkanes (C_1-C_4)	532	constitutional and connectivity (6)	MLR	0.970	10.94	Balaban et al.[164]
acyclic ethers, peroxides, acetals, and their sulfur analogues, aromatic compounds	185	TIS(3)	MLR	0.971	8.2	Balaban et al.[165]
furans/THFs, thiophenes, and pyrans)	299	TIs, electronic	RA	0.962	11.8	Stanton et al.[166]
pyrans	178 (pyrans)	geometrical, CPSA(11)		0.954	13.5	
pyrroles	278 (pyrrole)	(7)		0.962	12.3	
pyridines	291	TIs, electronic, CPSA	RA	0.933	15.0	Egolf et al.[167]
haloalkanes C_1-C_4	276	constitutional and topological	NN(5−10−1)	0.982	8.5	Balaban et al.[168]
(straight-chain, branched, cyclic hydrocarbons) (halogen, alcohol, cyano, amino, ester, ether, carbonyl, and carboxylic acid functionalities)	268	constitutional, topological, and CPSA (8)	RA	0.976	11.85	Egolf et al.[169]
hydrocarbons from DIPPR Database	296	electronic, CPSA, topological, mol. weight	MLR(6)	0.994	rms = 6.3	Wessel et al.[170]
	267		NN(6:5:1)		rms = 5.7	
O, S, and halogen containing compounds	248	constitutional, topological, and CPSA (10)	RA	0.982	11.6	Wessel et al.[171]
N containing compounds	90	constitutional, topological, geometrical, and CPSA (10)	RA	0.980	10.7	Wessel et al.[171]
alcohols, alkanes	245	atom-type E	MLR		8 K	Hall et al.[172]
diverse organic compounds	298	constitutional, topological, geometrical, and CPSA (4)	MLR	0.973(4)	12.4 K	Katritzky et al.[156]
				0.954(2)	16.15 K	
diverse organic compounds	298	atom type electrotopological indices	ANN	0.995	5.30	Hall et al.[173]
C_2-C_9	74	constitutional and topological	PCA (4)	0.980		Kuanar et al.[174]
C_2-C_9	74	hierarchical orthogonalized partially ordered molecular descriptors (7)	least square fit	0.977		Klein et al.[175]
alkanes and cycloalkanes	76	detour indices, *w* (1)	QR	0.961	12.1	Trinajstić et al.[176]
		W*w*(1)		0.990	6.2	
C_3-C_9	73	topological (1)	MRA	0.960		Kuanar et al.[177]
compounds containing C, H, N, O, S, F, Cl, Br, and I	584	constitutional, topological, geometrical, and CPSA (8)	MLR	0.965	15.5	Katritzky et al.[157]
compounds containing halogen, O, and S	185	constitutional, topological, geometrical, and CPSA(6)	MLR	0.984	6.3	Ivanciuc et al.[178]
halogenated alkanes (C_1-C_4)	534			0.990(5)	9.0(5)	
benzenoid hydrocarbons	22	distance based TIs (2)	MLR	0.992	12.2	Plavšić et al.[179]
acyclic compounds containing O or S atoms	185	constitutional, atom and bond weighted	MLR(5)(4)	0.978	7.19	Ivanciuc et al.[180]
		TIs based on electronegatitivity and covalent radius		0.973	7.98	
chlorosilanes	74	constitutional, topological, geometrical, and CPSA (8)	MLR	0.996	6.09	Bunz et al.[181]
alcohols	58	TIs (weighted path number)	MLR(3)	0.994	3.91	Randić et al.[182]
manohaloalkanes	45	connectivity TIs	MLR	0.965	10.00	Balaban et al.[183]
acyclic carbonyl compounds	200	TIs	MLR	0.964	6.93	Balaban et al.[184]
diverse compounds	241	constitutional, tolopological, geometrical, and CPSA (8)	NN(8−3−1)		7.75	Goll et al.[185]
sulfides	21	connectivity TIs	RA	0.992	2.61	Randić et al.[186]
			QR	0.996	1.83	
alkanes	21	variable connectivity based TIs	RA	0.998	2.481	Randić[187]
acyclic and cyclic hydrocarbons	180	distance related TIs(4)	MRA	0.985	5.31	Lučić et al.[188]
alcohols	58	weighted path related TIs	MRA	0.978	3.64	Randić et al.[189]
cycloalkanes and alkylcycloalkanes	42	variable connectivity based TIs	RA	0.997	3.029	Randić et al.[190]
acyclic carbonyl compounds	200	constitutional, topological, geometrical, and CPSA	MLR	0.976	5.6 °C	Ivanciuc et al.[191]
heterogeneous organic compounds	1168	connectivity indices, kappa shape index, dipole moment, sum of atomic number	fuzzy, ARTMAP, BPNN		2.09%	Espinosa et al.[192]
					10.3%	
hydrocarbons, aldehydes, ketones, thiols, and alkoxy silicon chlorides		TIs	MLR	0.990	<0.5	Zhou et al.[193]

[a] MLR, multilinear regression; RA, regression analysis; ANN, artificial neural network; PCA, principal component analysis; QR, quadratic regression; TI, topological inices.

attractive force between molecules is hydrogen bonding. The four parameter model showed improved statistical results ($R^2 = 0.973$, $s = 12$ K) and includes additional descriptors: the most negative atomic partial charge and the number of chlorine atoms. Katritzky et al.[157] extended their previous work and obtained a general model for boiling points for a large set of 612 diverse compounds consisting of C, H, N, O, S, F, Cl, Br, and I atoms based on molecular descriptors calculated solely from structure using CODESSA PRO. QSPR models were developed for various classes of compounds and also for the combined set. An eight-parameter MLR model gave $R^2 = 0.965$ and $s = 15.5$ K for 584 compounds. The descriptors include the previously used $GI^{1/3}$ and HDSA-2 and an additional six descriptors (N_F/N, N_{CN}, HASA-1, T_I, CSA-2_H, and CSA-2_{Cl}). The validity of the model was tested for 28 polyfunctional compounds, and except for three compounds, the predicted values were within the standard error. This correlation covers a larger diversity as compared to other QSPR models having a standard prediction error of 15.5 K and uses descriptors calculated solely from the molecular structures.

The applicability of several novel descriptors has been investigated for developing QSPR models of boiling points of diverse sets.[193−195] In a recent paper, Zhou et al.[194] used a novel topological index N_t based on equilibrium electronegativity and relative bond length, and path numbers P_2 and P_3 to develop ANN models for the boiling points of alkanes, aldehydes, ketones, and mercaptans. The authors obtained QSPR models with correlation coefficients, $R^2 = 0.99$, and standard error, $s < 0.5$.

4.2. Melting Points

Melting point is a fundamental physical property specifying the transition temperature between solid and liquid phases. It has been used as a criterion of purity of a compound and has also been used for the prediction of other physical properties such as aqueous solubility[196−198] and liquid viscosity.[199] Melting point has also been successfully used as a descriptor in correlations with the aqueous solubility of chlorophenols[200,201] and skin corrosivity of organic acids, bases, and phenols.[202] Since melting point affects the solubility of a compound, techniques for the estimation of the melting point of organic compounds would significantly assist medicinal chemists in designing new drugs with a specified range of melting point and solubility. Melting point also affects the toxicity of a compound through its solubility.[203]

The melting points of organic molecules in general depend on the arrangement of atoms in the crystal lattice as well as upon the strength of the pairwise group interactions.[204,205] Melting point is determined by the strength of the crystal lattice, which, in turn, is controlled primarily by three factors: intermolecular forces, molecular symmetry, and conformational degrees of freedom of a molecule.[203] Further, molecular motion in crystals affects melting point, which depends on the size and shape of the molecules, on their orientation in the crystal, and on temperature.[206] Importantly, melting point is often not unambiguous. Many compounds crystallize in more than one form, each with a different melting point, and hence exhibit the phenomenon of polymorphism. Phase transitions are complicated by polymorphism; molecules that exist in different crystal forms have their own distinct properties, including heat capacity and melting point. Additionally, measurements of melting points are affected by the purity of the compound and by experimental error.

Despite the large amount of melting point data available and knowledge of melting point transition, the correlation and prediction of melting points of diverse sets of compounds is still very difficult. Various QSPR methods, such as the property−property relationship (PPR),[207] and group contributions[208−210] have been used for the prediction of melting point. However, the group contribution methods generally have difficulty in providing reliable estimates of melting points, because they depend heavily on the nonadditive structural features, such as intermolecular interactions and molecular symmetry.[210] A comprehensive review has appeared of the relationship of melting points with chemical structures.[211]

Successful predictions of melting points have been achieved for 24 normal alkanes ($R^2 = 0.998$, $s = 0.51$) using topological indices such as the carbon number, Wiener index, and the Balaban distance sum connectivity index.[212] However, QSPR models developed by Needham et al.[159] using structural parameters show poor predictability ($R^2 = 0.570$, $s = 23.8$) for 56 normal and branched alkanes.

Abramowitz and Yalkowsky[213] correlated the melting points of 85 rigid, non-hydrogen bonding compounds with their boiling points and symmetry numbers to study the effect of symmetry on the melting point of organic compounds. The authors found a multiparameter correlation ($R^2 = 0.880$ and $s = 22.8$) of melting point with four variables such as boiling point (bp), logarithm of symmetry number (SIGMAL), the cube of eccentricity of the compound (EXPAN), and the number of groups that are in an ortho position to another group (ORTHO).

Dearden[203] developed a five-parameter model ($R^2 = 0.885$, $s = 24.6$ K) for the prediction of melting points of a series of 42 anilines using descriptors based on hydrogen bond donor ability (α), the hydrophobic substituent constant (π), the molar refractivity (MR), the STERIMOL width parameter (B_2), and the indicator variable of meta substitution (I_3).

Charton and Charton studied the correlation of melting points of a combined set of 366 branched and normal substituted alkanes using variables capable of accounting for the packing energy contribution of the alkyl group and found better predictability ($R^2 = 0.9185$, $s = 17.9$).[214]

Six-descriptor QSPR models ($R^2 = 0.931$, $R^2_{CV} = 0.816$) of 141 pyridines and piperidines,[215] and ($R^2 = 0.857$, $R^2_{CV} = 0.843$, $s = 36.1$) for pyridines and substituted pyridines of 140 compounds[216] were obtained for the prediction of melting points. The descriptors contributing to the above models are related to the hydrogen bonding ability of the compound, intermolecular interactions in condensed media, crystal lattice packing, the band gap between solid insulators, and the valence band and the unoccupied band.

A comparative study on the prediction of physical properties of aldehydes ($n = 27$, $R^2 = 0.833$), amines ($n = 48$, $R^2 = 0.795$), and ketones ($n = 30$, $R^2 = 0.865$) using five different molecular descriptors (topological, electronic, and geometrical) has given moderate correlations between the structures and melting point.[22]

The melting point of a large data set of 443 mono- and disubstituted benzenes has been correlated with a set of structural parameters, and a nine-parameter model with ($R^2 = 0.837$, $s = 30.19$ K) was obtained as shown in Figure 4.[217] Six-parameter equations were used to describe each of the individual ortho-, meta-, and para-substituted benzene subsets. In this QSPR investigation the descriptors related

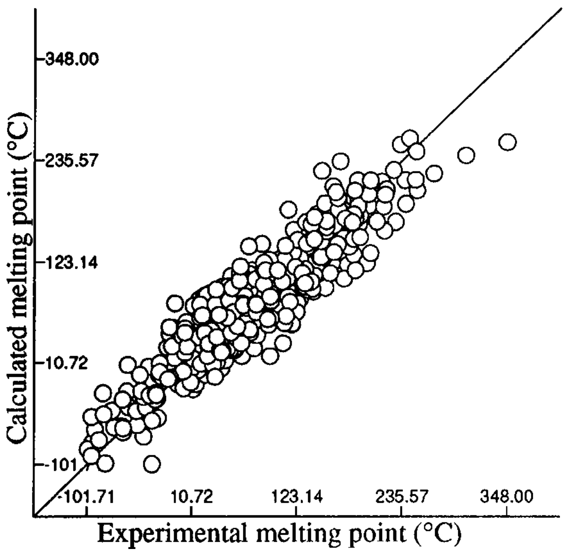


Figure 4. Plot of experimental vs calculated melting points of 443 compounds using the 9-parameter model. Reprinted with permission from ref 217. Copyright 1997 American Chemical Society.

to hydrogen bonding ability, molecular packing in crystals (effects from molecular shape, size, and symmetry), and other intermolecular interactions such as charge transfer and dipole−dipole interactions contributed to the prediction of melting point.

Molecular connectivity parameters including molecular ID numbers, atomic ID numbers, and variable linear combinations as descriptors have shown promising correlations in predicting various properties.[218−221]A QSPR model (R^2 = 0.856, s = 16.79 °C) was developed for prediction of the melting point of a series of 1,2,3-diazaborine compounds (n = 72) based on the electronic and topological descriptors from molecular structures.[222] The most important molecular descriptors describing this physicochemical property were the sum of the atomic charges for the heteroatoms, the sum of the Randić connectivity indexes ($^0\chi$, $^0\chi^1$, and $^0\chi^2$), the total number of atoms in the molecule, and the volume of the box in which the molecule fits. In addition to the regression techniques, a back-propagation neural network was used to predict the melting points successfully. The authors noted that the melting points of 1,2,3,-diazaborine compounds can be described by electrostatic interactions mediated by atomic charges and steric properties.

Melting point models have been reported previously for smaller alkanes. Burch et al.[223] recently developed multiparameter models to predict melting points of alkanes having 10−20 carbon atoms and only one methyl group, which are of special interest to petroleum engineers manufacturing synthetic diesel fuel. A nonlinear regression model (n = 69) with satisfactory predictability was obtained based on the number of carbon atoms, the Wiener path numbers, the mean Wiener index, and the methyl locant index.

Gramatica et al.[224] used weighted holistic invariant molecular descriptors (WHIM), 3D molecular descriptors based on the size, shape, symmetry, and atom distribution of the molecules, for the correlation of melting points of polychlorinated biphenyls (PCBs). A four-parameter model (R^2 = 0.82, s = 21.25) was obtained for melting points ranging from 16.5 to 310 °C for 82 PCBs out of 66 WHIM descriptors. The model found that the melting points depended on the size variables (both directional, A_m and T_u, and nondirectional, L_{2V}) and on symmetry variables G_{1p}.

The increasing importance of ionic liquids[225,226] underlines the significance of understanding melting behavior. A six-descriptor QSPR model (R^2 = 0.788) for the prediction of melting points of 126 structurally diverse pyridinium bromides in the temperature range 30−200 °C was obtained using molecular descriptors calculated by the CODESSA-PRO program.[227] The model obtained was based mainly on the descriptors such as information content indices, total entropy, and the average nucleophilic reactivity index for the nitrogen atom. The melting points of 104 substituted imidazolium bromide based ionic liquids were correlated with molecular descriptors.[228] The data set was divided on the basis of the N-substituents into three subsets: A, B, and C consisting of 57, 29, and 18 compounds. Another set D was formed consisting of 48 benzimidazolium bromides. Five-parameter correlations were obtained for set A (R^2 = 0.744), set B (R^2 = 0.752), and set D (R^2 = 0.690), while set C was correlated with a three-parameter equation with R^2 = 0.943. The descriptors involved in the correlations reflect both the intermolecular interactions and the influence of intramolecular electronic effects on those interactions. QSPR models were also developed for the melting point data of 126 pyridinium bromides based on molecular descriptors calculated by the DRAGON software.[229] Regression trees were initially built for the variable selection, and by use of a counterpropagation neural network (CP-NN) approach, a reasonable result was achieved (R^2 = 0.748). The authors obtained qualitative predictions for a new set of nine compounds, all with low melting points being recognized by several methods: the decision tree, the ensemble of trees, and the CP-NNs.

Bergstrom et al. investigated the role of calculated 2D and 3D molecular descriptors in predicting melting points of druglike compounds and classification of solid drugs.[230] The melting points of 277 structurally diverse druglike compounds were taken from the *Merck Index*. The normal distribution of data showed the majority of compounds displaying melting points between 140 and 160 °C. Correlations between the calculated descriptors and the melting point values were established with the PLS projection to latent variables using training and test sets. Three different descriptor matrixes were used for consensus modeling. The calculated properties were shown to explain 63% of the melting point variations of the druglike molecules, and the descriptors generated from the 2D representation of the molecule were more successful in the prediction of melting points than descriptors generated from the 3D configuration. Descriptors for hydrophilicity, polarity, partial atomic charge, and molecular rigidity were found to increase the melting point, whereas nonpolar descriptors and descriptors for molecular flexibility lowered the melting point. The authors achieved a qualitative classification of the compounds separated into groups of low, intermediate, or high melting points.

QSPR models for melting points of druglike compounds were developed based on three different software packages (CODESSA, DRAGON, and Tsar) for molecular descriptor generation and a combined set of all descriptors.[231] The melting point data of 323 druglike organic compounds were used for the study. Two QSPR models with reasonable statistical results were obtained with the combined set of descriptors based on stepwise regression (R^2 = 0.673, s = 40.4 °C) and genetic algorithms (R^2 = 0.660, s = 41.1 °C) descriptor selection methods. The authors analyzed the descriptors from three different software packages and noted the difficulty in predicting melting points of druglike compounds. In contrast to the simple case of hydrocarbons, interpretation of effective molecular features of complex compounds is a difficult task because of the various entropic parameters involved in the melting process. An eight-descriptor regression model for the training set with $R^2 = 0.66$ and $s = 40.9$ ($s = 42$ °C for the test) was obtained using genetic algorithms.

QSPR models for the prediction of melting points of polyamides have been recently reported[232] based on the descriptors calculated from molecular structures using the B3LYP/6-31G(d) basis set.[233,234] A four-descriptor MLR model was obtained for a training set of 41 polyamides for melting points (T_m (K)) with good statistical characteristics (eq 3).

$$T_m\ (\mathrm{K}) = 300.1545 - 23.5783\mathrm{PMA} + 55.9903\mathrm{LB} + 53.0969Q_O - 0.3119E_t \quad (3)$$

$$n = 41, R^2 = 0.900, R^2_{CV} = 0.931, F = 104.5, s = 9.98$$

The descriptors involved in the model for polyamide melting point temperatures T_m (eq 3) are the proportion of methylene to acylamino in the backbone chain (PMA), the value of benzene rings in the backbone chain (LB), the atomic charge for oxygen in acylamino (Q_O), and total molecular energy (E_t).

A BPANN (back-propagation artificial neural network) was developed using the descriptors selected by the MLR with a four-two-one network model.[232a] The training set data for 41 polyamides showed a good correlation coefficient ($R^2 = 0.931$) and prediction ($R^2 = 0.882$) for the test set of 39 polyamides. The authors noted that the four-descriptor model predicted T_m successfully for polyamides and that MLR and BPANN are practical methods for building a QSPR model. Furthermore, descriptors selected for the T_m are representative, and the value of T_m is governed mainly by the molecular rigidity and polarity.

The application of semiempirical quantum chemical descriptors calculated by the CODESSA program has enabled the development of robust QSPR models for chain melting temperatures (T_m). A predictive, chemically meaningful QSPR for phosphatidylcholines provided T_m values that agreed with the experimental values to within experimental error.[232b]

4.3. Viscosities

Viscosity (η) is one of the most important physical properties for understanding many processes in the chemical and petroleum industries. With the increased need of reliable data for optimization of the industrial process, it is important to develop an effective method to predict the viscosities of compounds for which a measured value is unavailable. Numerous predictive methods have been reported for the estimation of viscosities as reviewed elsewhere.[151,235] Suzuki and co-workers[236,237] used QSAR/QSPR and PLS techniques to estimate the liquid viscosity of 116 and 361 diverse organic compounds based on experimental physicochemical property data. The MLR model with $R^2 = 0.870$ obtained for 116 compounds includes four key physical properties: molar refraction, critical temperature, molar magnetic susceptibility, and cohesive energy. The authors also developed a five-component PLS model based on a cross-validation method that resulted in $R^2 = 0.867$ for the 116 compounds. MLR and two-layer ANN modeling with back-propagation were applied to derive predictive models for the liquid viscosity of 361 organic compounds.[237] A nine-descriptor QSPR model with $R^2 = 0.92$ and rms error of 0.17 log units was obtained for MLR with the predicted set of 124 compounds, resulting in $R^2 = 0.93$ and a rms error of 0.16 log units. The derived models allow predictive applications with expected uncertainty factors for η of 1.5 (MLR) and 1.4 (ANN), respectively, which is reasonably accurate for the wide range of chemical structures with η values covering 4 orders of magnitude. The use of experimental properties as independent variables in the QSPR model for liquid viscosity as proposed by the Suzuki research group makes their application difficult for a significant set of organic compounds whose properties (independent variables in the model equation) have not been determined.

Ivanciuc et al.[238] developed a QSPR model for the prediction of liquid viscosities of a diverse set of organic compounds based on molecular descriptors solely calculated from structures using CODESSA PRO. A five descriptor MLR model with $R^2 = 0.846$ and $s = 0.37$ was obtained for the prediction of η of 337 organic compounds. The five descriptors relating to the QSPR model for η are molecular weight, Randić connectivity index of order 3, hydrogen-donor charged surface area (*HDCA*-2), maximum electrophilic reactivity index for a C atom, and maximum atomic orbital electronic population. The predictive ability of the linear model was tested by the leave-20%-out cross-validation method that showed stability. However, this model is limited to compounds with polar groups.

Our group, in collaboration with others,[239] obtained a five-descriptor QSPR model for the liquid viscosity of 361 organic compounds containing C, H, N, O, S, and/or halogens with a statistically significant R^2 of 0.854 and s of 0.22 log units.

$$\eta = -10.3 + 1.77\mathrm{HDCA\text{-}2} + 0.000557G_I + 2.78N_{\mathrm{rings}} + 20.2\mathrm{FPSA\text{-}3} + 0.0897E_{\min}(\mathrm{C})$$
$$n = 361, R^2 = 0.854, R^2_{CV} = 0.840, F = 414.1, s = 0.22 \quad (4)$$

The most important descriptor in eq 4 was found to be the HDCA-2, which indicates that hydrogen bonding is a key factor for liquid viscosity.

Furthermore, a five-descriptor nonlinear multiple regression model was obtained for the liquid viscosity of the same data set of 361 compounds with $R^2 = 0.908$ and $s = 0.175$ based on the CROMRsel method.[240] The most important descriptor involved in the model is the gravitational index, which reflects the effective mass distribution in the molecule and describes intermolecular dispersion forces in the bulk liquid media (i.e., accounts simultaneously both for the atomic masses and for their distribution within the molecular space).[239] The three electrostatic descriptors involved in the model also reflect the bonding properties of the molecules, i.e. their capabilities to create hydrogen bonds. In summary, the key descriptors involved relate to the mass, size, and shape as well as hydrogen bonding abilities of the molecules.

4.4. Refractive Indices

Refractive index (n) is an important optical property and is used to indicate purity in material science and thus to evaluate the applicability of materials for various purposes.

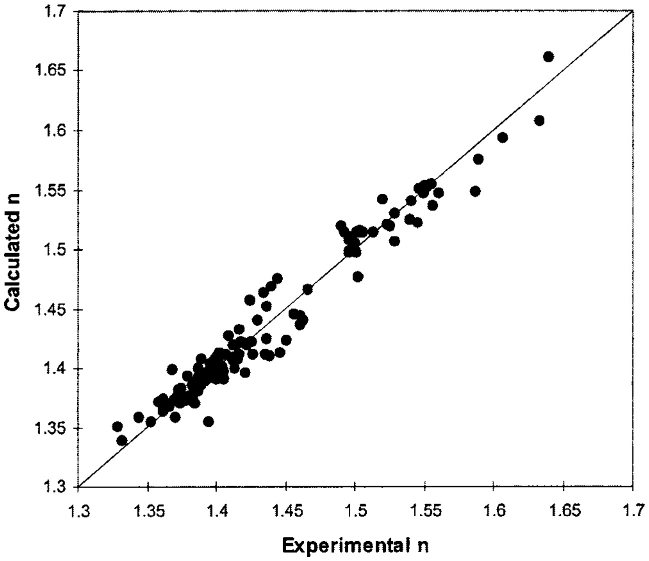


Figure 5. Calculated vs observed refractive index values by using a 5-parameter model. Reprinted with permission from ref 241. Copyright 1998 American Chemical Society.

It is related to other physicochemical properties such as polarizability, critical temperature, surface tension, density, and boiling point. Unlike the molar refraction, the refractive index was not used in many QSPR studies before 1990. In an early work, in 1993, our group correlated refractive indices of three different data sets of aldehydes, amines, and ketones with molecular descriptors.[22] QSPR models were developed for each class of compounds with correlation coefficients > 0.90. The topological descriptor Randić index is related to the refractive indices for the three diverse sets of compounds: aldehydes, amines, and ketones. Furthermore, we reported a five-parameter QSPR model ($R^2 = 0.945$, $s = 0.0155$) for the refractive index of a structurally diverse set of 125 compounds (hydroxyl amino, ether, ester, carbonyl, cyano, and carboxylic functionalities, halogenated, saturated and unsaturated hydrocarbons).[241] The five parameters included the HOMO−LUMO energy gap, minimum electron−nuclear attraction for a C atom, PPSA-2 [Zefirov's PC], HDSA [semi-MO PC], and gravitation index on all bonds. The calculated versus observed plot is shown in Figure 5. An estimated average error of 0.8% was achieved for the predicted values and can be used for the prediction of refractive indices with a high degree of confidence.

In an extension to this work, we correlated refractive indices of a set of linear polymers consisting of homochain polymers (only carbon atoms in the main chain) and polyoxides, and also a few polyamides and polycarbonates.[242] In the calculation of descriptors for polymeric molecules, the methods used for small molecules cannot be applied. However, in the case of linear polymers, we used the repeating unit to calculate appropriate descriptors for 95 compounds.

A four-descriptor QSPR model ($R^2 = 0.929$, $s = 0.0175$) was obtained for a set of 121 linear polymers using simple molecular descriptors: the sum of valence degrees (SVDe), the degree of unsaturation (DU), the relative number of halogen atoms (RNH), and the electrostatic attraction and the hydrogen bond between the main chains ($Q_{\pm}$).[243] The statistical characteristics of the general QSPR model using four descriptors are given by eq 5.

$$N(298\text{ K}) = 1.476 - (5.202 \times 10^{-4})\text{SVDe} + (2.337 \times 10^{-2})\text{DU} - 0.187\text{RNH} - 0.547Q_{\pm}$$
$$n = 121, R^2 = 0.929, R^2_{\text{CV}} = 0.926, s = 0.0175, F = 378.1 \qquad (5)$$

The four simple descriptors used in this model illustrate the molecular size and the intermolecular forces of polymers through structural analysis on the polymers. This model was found to have an average prediction error of 0.87%, comparable with the QSPR model obtained previously for 95 compounds, including mostly quantum-chemical descriptors.

Refractive indices of a series of organic solvents of the structural formula X-Y have been correlated with molecular descriptors by using CODESSA.[244] A comparative study has been reported based on the heuristic and best multilinear regression techniques included in CODESSA and with the multivariate PLS/GOLPE method. The best correlation for the refractive index was obtained using the GOLPE procedure ($R^2 = 0.9501$, $\text{SDEP}_{i,\text{LOO}} = 0.0159$, and $\text{SDEP}_\text{e} = 0.0180$).

A five-descriptor QSPR model ($R^2 = 0.902$, $s = 0.0055$) was obtained for refractive indices of 149 alkanes (C_2 to C_{20}) using molecular descriptors based on the polarizability index (PEI) calculated from eigenvalues of the bond adjacency matrix (eq 6).[245]

$$n_\text{D} = 1.1848 + 0.0224SX_{1\text{CH}} - 0.0374SX_{1\text{CC}} + 0.00018SV_{ij} + 0.00047\text{PEI} + 0.1127N^{2/3}$$
$$n = 149, R^2 = 0.902, s = 0.0055, F = 264.4 \qquad (6)$$

The predictive ability of the model was tested using the cross-validation procedure, and good statistical results were obtained ($q^2 = 0.879$, PRESS = 0.0060, $F = 207.3$). The authors designated a few outliers in the above QSPR model and suggested that the descriptor set could not be interpreted well for alkanes with branched structure.

The refractive indices of 180 diverse phosphates and diphosphates, comprising various types of structures (normal and branched aliphatic, or alicyclic and aromatic) for different temperatures in the range of 20−25 °C, were successfully predicted using an ANN trained with the back-propagation procedure[246] based on molecular descriptors. The best ANN model (40:2:1) showed good predictive ability with the average prediction error of 0.24% and R^2_{CV} equal to 0.99.

QSPR models were developed for refractive indices of 186 saturated compounds, 200 aromatic compounds, and the combined set of 386 compounds based on molecular descriptors.[247] The statistical characteristics of the three models are ($n = 186$, $R^2 = 0.9921$, $s = 0.004$ K, $F = 3054$), ($n = 200$, $R^2 = 0.9902$, $s = 0.005$ K, $F = 2052$), and ($n = 386$, $R^2 = 0.9881$, $s = 0.008$ K, $F = 34774$), respectively. However, the standard deviation, s, for the combined data set increased from 0.004 to 0.008 compared to the saturated compounds and 0.005 compared to the aromatics. Consequently, there is a need for further investigation to build a general QSPR model with higher precision.

Linear and nonlinear QSPR models for the prediction of refractive indices of polymers were developed based on a diverse data set of 120 polymers by using MLR analysis and feed-forward ANNs.[248a] A linear model was obtained with $R^2 = 0.943$ and $s = 0.016$ for a training set of 100 compounds. The mean relative error (MRE) of 0.79% was obtained for the whole data set based on the trained model. The nonlinear model showed better statistical results ($R^2 = 0.973$, $s = 0.0118$) for the training set and ($R^2 = 0.961$, $s = 0.0144$) for the test set, respectively.

Good QSPR models have been developed for the prediction and rationalization of the refractive indices of a wide variety of simple organic/organosilicon compounds (3 parameters, $R^2 = 0.924$)[248b] and polymer matrices (2 parameters, $R^2 = 0.924$)[248c] using SAM1 and AM1 CODESSA descriptors.

4.5. Densities of Organic Liquids

The normal density (i.e., the density at 1 atm and 20 °C) is a major physicochemical property for characterization and identification of compounds. Density can be estimated from molecular weight (M_w) and molar volume (V_m) by the simple formula $d = M_w/V_m$. In addition, it can be used to predict or estimate properties such as critical pressure, viscosity, thermal conductivity, diffusion coefficients, and surface tension.[249] Kier and Hall[15] showed a correlation with $R^2 = 0.815$ for the density of a set of 82 alkanes as an inverse relationship with the Randić index of order 1 ($^1\chi$). QSPR models for density of a set of aldehydes, amines, and ketones were also developed using molecular descriptors of topological, electronic, and geometrical types.[22] The three-parameter model had a R^2 value of 0.922 and $R^2_{CV} = 0.896$, and it included the connectivity descriptors, $^1\chi^V$, together with two information theoretic indices (^{0}IC and ^{0}SIC) for a set 59 aldehydes. Further, the four- and five-descriptor models constructed resulted in the improvement of the regression coefficients ($R^2 = 0.935$, $R^2_{CV} = 0.901$) and ($R^2 = 0.941$, $R^2_{CV} = 0.915$), respectively. The four-parameter QSPR model for the density of 109 amines had $R^2 = 0.940$ and $R^2_{CV} = 0.931$ and included three topological descriptors, the Randić index, $^1\chi$, the shape index, ^1k, the information index, ^{0}SIC, and the electronic descriptor, FNSA-1. A four-parameter QSPR model for the densities of a set of 60 different ketones with excellent fit and high stability ($R^2 = 0.955$, $R^2_{CV} = 0.947$) included two descriptors ($^3\chi$ and ^{1}SEPD) in their squared form in combination with two information indices of the zeroth order (^{0}EPD and ^{0}SEPD).

Gakh et al.[250] devised a computational method to predict the densities of organic compounds based on their molecular structure, which used graph theory to encode the structural information in numerical form included as input for the ANN model. The ANN model was trained using a data set of 109 saturated hydrocarbons (C_6-C_{10}), and it gave an average error of 0.60% for the test set of 25 compounds. Zhang et al.[251] developed a nonlinear ANN model by using a set of five molecular descriptors (W, P, w, p, s) for the prediction of densities of 85 alkenes (C_4-C_{20}). The five parameters are as follows: W based on the distance matrix of a molecule; P, the polarity number; w, representing the absolute contribution of a double bond to the whole size of a molecule; p, indicating the absolute contribution of a double bond to the shape of the molecule; and s, representing enantiomers of alkenes, respectively. Their ANN model showed the average RSD (relative standard deviation) of 0.44% for the test set of 16 alkenes.

A general two-parameter correlation model developed for the prediction of densities of 303 diverse organic compounds (containing C, H, N, O, S, F, Cl, Br, and I) gave promising results with $R^2 = 0.975$ and $s = 0.046$ g/cm^3.[252] The two parameters involved in the correlation are the intrinsic density values calculated as the ratio of the molecular mass over the theoretically calculated van der Waals molecular volume, and the total molecular electrostatic interaction per atom in the molecule (analogous to the Madelung energy in ionic crystals). Further correlation equations were developed for densities of various subsets of organic compounds that included one to four parameters having standard errors, s, ranging from 0.027 for hydrocarbons to 0.085 g/cm^3 for halogenated compounds.

An excellent 1-parameter correlation ($R^2 = 0.929$, SDEC $= 0.094$) for a training set of 61 compounds and a standard error of prediction (SDEP$_e$ $= 0.090$) for 28 compounds was obtained with the relative molecular weight.[244] The molecular weight and density correlated well within a homologous series of compounds, taking into account the structural heterogeneity of the training set. By the addition of another parameter, the minimum value of the net atomic charge for the variable molecular fragment (f-Min net atomic charge), both the model statistics ($R^2 = 0.960$, SDEC $= 0.071$) and the predictive capability of the regression model (SDEP$_e$ $=$ 0.072) increased. Further, the GOLPE multivariate analysis was applied for the correlation of density that included four parameters and three PLS components resulting in $R^2 = 0.948$ and SDEC $= 0.079$, and a relative loss in predictive ability with SDEP$_e$ $= 0.093$. Furthermore, the PLS pseudoregression coefficient of the relative molecular weight presents the highest value among those of the four selected descriptors and confirms convergence between this approach and the MLR methods.

Toropov et al.[253−258] estimated the predictive potential of the OCWLGI (optimization of correlation weights of local graph invariants) based on Morgan extended connectivity of LHFGs and GAO (graphs of atomic orbitals) in modeling density and other physicochemical properties. The statistical characteristics of the QSPR model ($R^2 = 0.984$, $s = 3.602$) and ($R^2 = 0.976$, $s = 3.790$) were obtained for the densities of a training set of 66 compounds and a test set of 67 compounds, respectively, based on the molecular descriptor $^0X_{CW}$(GAO,EC1), comprising 26 local invariants.[253]

Multivariate regression models have been developed to predict the densities of alkanes and monosubstituted alkanes based on the molecular descriptors calculated from the eigenvalues of the bonding orbital-connecting matrix, polarizability effect index (PEI) of alkyl, and Pauling's electronegativity concept.[259] A five-descriptor QSPR model was obtained for 213 compounds with a significant value of $R^2 = 0.992$, the rms error 0.0208 g/cm^3, the average absolute error 0.017 g/cm^3, and the average relative error 1.85% between experimental and predicted values.

4.6. Dielectric Constants

Dielectric constants measure the ability of a liquid to solvate a charged molecular species. The dielectric constant is frequently used as a practical parameter to characterize the polarity of organic solvents.

A large body of theory has been developed for the calculation of dielectric constants from properties such as dipole moment and polarizability.[260] The value of the dielectric constant is strongly related to the chemical structure of a molecule, intermolecular interactions, and external conditions (temperature, pressure, etc.) Several methods have been used for the calculation of dielectric constants.[261] Significant progress is made through the use of the Onsager equation. However, this does not take into account significant

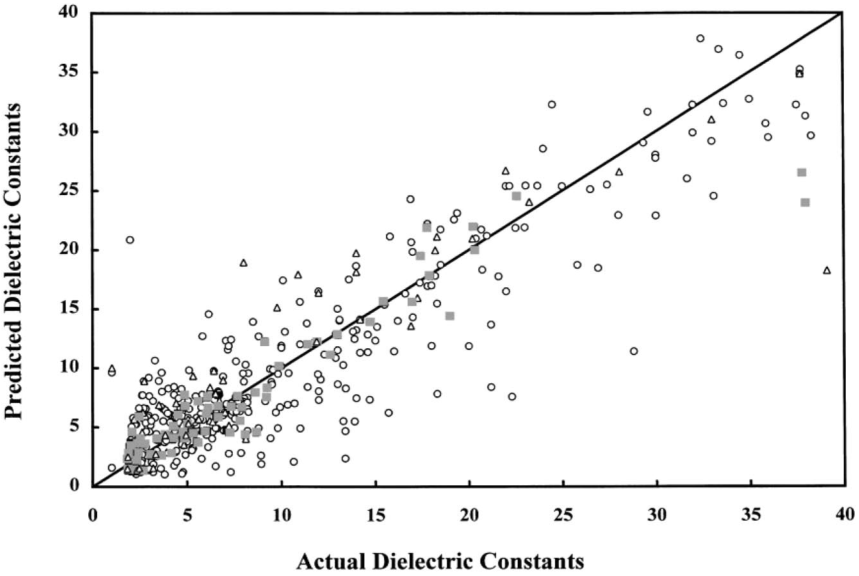


Figure 6. Calculated vs observed dielectric constant values using the 10-parameter nonlinear model. Reprinted with permission from ref 262. Copyright 1999 American Chemical Society.

intermolecular interactions, especially for hydrogen-bonding liquids. Computational efforts involving computer simulation and molecular dynamics are also limited for the hydrogen-bonding solvents, due to the sensitivity of the dielectric constant to the long-range intermolecular interactions. The earlier available theoretical approaches simply do not allow their use as general purpose tools for calculating dielectric constants of a wide variety of compounds.

The QSPR approach is an alternative and mathematically simpler option for the prediction of the dielectric constant. Cocchi et al. used MLR analysis and the multivariate PLS method to develop QSPR models for dielectric constants of organic solvents[244] based on molecular descriptors calculated using CODESSA. A three-descriptor MLR model was obtained for the dielectric constants of a training set of 23 compounds with $R^2 = 0.9564$. Dielectric constant values ranging from 2 to 41 were used in the study. The standard deviations, s, of the training ($n = 23$) and test sets ($n = 20$) were 2.262 and 4.650, respectively. The authors obtained a 15 variable PLS model using the GOLPE procedure with $R^2 = 0.974$, and the standard errors for the training and test sets were 1.576 and 3.213, respectively. However, this model seems to be overfitted.

A large diverse data set of dielectric constants for 497 compounds (ranging from 1 to 40) was used by Schweitzer and Morris for QSPR modeling.[262] They used molecular descriptors such as the dipole moment, polarizability, counts of elemental types, an indicator of hydrogen bonding capability, charged partial surface area (CPSA) descriptors, and molecular connectivity. The authors obtained a 10-parameter nonlinear QSPR model based on the Broyden−Fletcher−Goldfarb−Shanno (BFGS) training algorithm with training set error (rms) 3.77 and test set error 2.33, respectively. The data set was divided into three: 350 compounds for the training set, 50 compounds for the monitoring set, and the remaining 97 compounds for the test set. The ANN model obtained is shown in Figure 6.

The analysis of the figure reveals that although there are a number of outliers, many of the compounds have accurately predicted dielectric constant values. Thus, 86% of the compounds have an absolute error less than 2.0. The molecular connectivity and CPSA descriptors were found to make important contributions to the model. The authors developed improved QSPR models for the prediction of dielectric constants for various subsets of compounds.[263] Their full data set consisted of 454 compounds with dielectric constants ranging from 1 to 40. The authors divided the full data set into eight subsets, and for each subset, molecular descriptors were calculated as in their previous paper and nonlinear models were constructed. The resulting combined mean test set error for the eight local models of 1.31 is significantly better than the mean test set error of 1.85 for the general model. The authors have shown that the division of the set into subsets based on functional groups allows the development of local targeted models that result in much more accurate predictions of a test set of compounds than the general model.

MLR and ANN methods were used to develop QSPR models for the prediction of the dielectric constants for the training set of 155 diverse compounds by using molecular descriptors based on the electronic properties of the molecules and the intermolecular interactions between the molecules.[116] A six-parameter MLR model was obtained for the dielectric constant with a $R^2 = 0.945$ for 155 training set compounds. The rms errors of the training set ($n = 155$) and prediction set ($n = 46$) are 2.368 and 3.743, respectively. A nonlinear model was obtained with $R^2 = 0.948$, without noticeable improvement over the MLR model.

4.7. Polarizabilities

The polarizability of a molecule (α) is a significant electronic property that measures the distortion of a molecule in an external electric field. In principle, the polarizability of molecules is governed by the strength of the attractive interaction between electrons and atomic nuclei. Polarizability is determined experimentally from the molar refraction (MR_D) values, which are calculated using the refractive index, n_D, the density, ρ, and the molecular weight, MW, using the Lorentz−Lorenz equation (eq 7):

$$MR_D = [n_D^2 - 1/n_D^2 + 2]MW/\rho = {}^4/_3\pi N_0\alpha \quad (7)$$

where $\pi = 3.14...$, N_0 is the Avogadro constant, and α is the polarizability.

Polarizability has been shown to play an important role in chemical−biological interactions. The first attempt to apply molecular refractivity in terms of the Lorentz−Lorenz equation to biological processes was made by Pauling and Pressman.[264] Molar refractivity is approximately additive. A number of additive models for calculations of MR have been proposed using both the atomic and bond increments.[265,266] Leo calculated the MR from the fragments based on a method similar to that developed for octanol/water partition coefficients.[267] MR for a variety of molecules has also been determined by Vogel's group.[268−270] Many QSPR models have been reported for the calculation of the molar refractivity of compounds using molecular descriptors.[159,174,177,271]

Many semiempirical methods of differing accuracy have been proposed for calculating molecular polarizabilities.[272−277] Molecular polarizability influences several other physical properties, including electronegativity,[278,279] dipole moment,[280] and ionization potential.[281,282] A quantitative relationship between polarizability, hardness,[283] and size of different systems, such as atoms, molecules, metal clusters, and carbon clusters, has been demonstrated.[284,285] Polarizability has also been related to the structural parameters, such as the bond length alteration (BLA) and π-electron bond order alteration (BOA).[286] Polarizability has been used as an independent variable for the prediction of vapor pressure and octanol−air partitioning coefficients.[287] Polarizability fields derived from semiempirically determined atomic polarizabilities have been used in three-dimensional quantitative structure−activity relationships (3D-QSAR).[288]

Bosque and Sales[289] developed a MLR for the prediction of polarizabilities of a large set of solvents comprising 426 compounds based on the atomic composition of the molecules. The authors obtained a very good correlation ($R^2 = 0.9943$ and $R^2_{CV} = 0.9938$) between the polarizability and the number of atoms of each type present in the molecules of the training set of 340 solvents. The ten different types of atoms present in the data set were used in the correlations. Using the same set of parameters for the prediction set consisting of 86 solvents, a very good correlation was obtained. The average absolute relative errors for the training and prediction sets were 2.31% and 1.93%, respectively. The authors also obtained average atomic polarizabilities and provided a comparison with the previously reported values. Other estimates of polarizability using the semiempirical methods AM1, PM3, and MNDO for 426 solvents resulted in high average absolute relative errors of 34.7%, 39.1%, and 36.4%, respectively. The calculated polarizability values were lower than the experimental ones. Contrary to the classic additive methods, the existence of structural units such as the numbers of double and triple bonds and the number and size of the rings present were not considered.

A QSPR model developed by Zefirov et al. for the prediction of molecular polarizability of a set of 613 compounds was based on atomic composition and fragment descriptors.[290] A very good correlation ($n = 613$, $R^2 = 0.9898$, $s = 0.613$) was obtained using average atomic polarizability and the additivity model proposed by Bosque and Sales.[289] Zefirov's group included the numbers of C, H, N, O, S, P, F, Cl, Br, and I atoms, the numbers of double and triple bonds, and the number of aromatic bonds (aromatic systems) as independent variables to obtain very good results ($R^2 = 0.9967$, $s = 0.38$) for a training set of 552 compounds. The test set of 61 compounds gave an rms error of 0.75.

QSPR models were developed for the prediction of polarizability of organic compounds, including halogenated compounds based on theoretical and pseudoconnectivity descriptors.[291] A four-descriptor regression model based on a linear combination of the basis indices (${}^0\chi^V, {}^1\chi, D^V, \chi_t^V, U_0$) was obtained for a set of 54 organic compounds with $R^2 = 0.964$ and $s = 0.75$.

Hansch et al.[292] correlated polarizability (eq 8) with the number of valence electrons (NVE) for 37 compounds and reported R^2 of 0.924.

$$\alpha(0)\ [\text{Å}^3/\text{molecule}] = 0.27(\pm 0.011)\text{NVE} \qquad (8)$$

Agin et al.[293] used electronic polarizability to correlate the narcotic activity of a group of 39 compounds with $R^2 = 0.973$. The correlation of the narcotic activity of 37 compounds with MR had $R^2 = 0.969$. Hansch et al. reported several QSAR correlations based on MR and NVE.[292,294−296] Based on the importance of the molecular polarizability parameter in chemico−biological interactions, the correlation and prediction of polarizabilities of 219 diverse organic compounds were studied using calculated descriptors.[297] The authors developed MLR models for the logarithm of polarizability values with $R^2 = 0.941$, based on the descriptors related to the charge distribution within a molecule and the energies of the HOMO and LUMO.

A recent successful QSPR model ($R^2 = 0.9845$) was developed[298] for polarizability of a data set of 40 polyaromatic hydrocarbons (PAH) and fullerenes. The model involved just one descriptor: the total molecular two-center exchange energy. The model was externally validated, and the results were in good agreement with both the ab initio calculated and experimental property values.

4.8. Vapor Pressures

Vapor pressure (VP) plays an important role in the transport, distribution, and fate of environmental pollutants in the atmosphere.[299] Vapor pressure is used in designing various chemical processes, and additionally, VP can be used in the estimation of other physicochemical properties, such as liquid viscosity, enthalpy of vaporization, air−water partition coefficient, and flash points.[300] Vapor pressure determines the volatility of a chemical. It governs the exchange rate of a chemical across the air−water interface through Henry's Law. Vapor pressures are not determined for an ever increasing number of chemicals due to the lack of resources. The greatest difficulty and uncertainty arises in the determination of the vapor pressure of low volatile chemicals. Experimental VP data are abundant for low molecular weight hydrocarbons but scarce for most compounds with boiling points over 200 °C.

As a complement to the experimental data, numerous correlations for estimating VP have been proposed. Many vapor pressure estimation equations are either empirical or based on equations of state or on the Clausius−Clapeyron equation. Several equations are available based on the use of other physicochemical properties.[287,301−303] Several group contribution methods were applied for the prediction of VP of organic compounds.[304−306] An ANN was applied by Kühne et al.[307] to a training set of 1200 compounds and a prediction set of 638 compounds based on 23 parameters calculated from chemical structure, system temperature, and melting point of compounds, resulting in $R^2 = 0.995$ and 0.990 and absolute average errors of 0.08 and 0.13 logarithmic units for the training and prediction sets, respectively. A generalized model was obtained by Godavarthy et al.[308] for the prediction of VP of a diverse data set consisting of 1121 molecules, including 73 classes of chemicals taken from the DIPPR database based on a scaled variable reduced coordinates (SVRC) model equation and QSPR methodology. The authors developed a 10−12−1 backpropagation ANN model with average errors of less than 0.5%, based on triple point and critical point data plus structural descriptors.

The QSPR approach is a highly promising alternative for the estimation of vapor pressures from descriptors derived solely from the chemical structure. Basak et al.[309] used the hierarchical quantitative structure−activity relationship (HiQSAR) approach for the prediction of vapor pressures based on structural descriptors using topostructural and topochemical parameters and an additional parameter (HB_1) related to intermolecular interactions for the prediction of VP measured at 25 °C for 476 diverse chemicals taken from the ASTER (assessment tools for the evaluation of risk) database and obtained a ten-parameter model with $R^2 = 84.3\%$ and $s = 0.29$. Three linear regression methodologies—ridge regression (RR), principal component regression (PCR), and partial least-squares (PLS)—were used to develop HiQSAR models for a VP data set of 469 chemicals based on topological descriptors.[310] The results indicated that the RR outperforms PCR and PLS.

Liang and Gallagher[311] obtained a seven-parameter MLR model for the prediction of vapor pressure (log P_L) at 25 °C for 479 compounds using polarizability and polar functional group counts as descriptors (eq 9).

$$\log P_L = -0.432\alpha - 1.382(\mathrm{OH}) - 0.482(\mathrm{C{=}O}) - 0.416(\mathrm{NH}) - 2.197(\mathrm{COOH}) - 1.383(\mathrm{NO_2}) - 1.101(\mathrm{CN}) + 4.610 \quad (9)$$

$$n = 479, R^2 = 0.960, R^2_{\mathrm{CV}} = 0.957, s = 0.534$$

The correlation coefficient with a single descriptor, polarizability (α), gave $R^2 = 0.920$, and the addition of the six polar functional group counts increased the R^2 to 0.957. An ANN model (7−5−1) reported by the authors gave approximately the same results ($R^2 = 0.973$, $R^2_{\mathrm{CV}} = 0.960$, $s = 0.522$) as the MLR model.

Katritzky et al.[312] developed a five-descriptor MLR model for the prediction of vapor pressure, log (VP) of 411 compounds with a large structural diversity (eq 10).

$$\log(\mathrm{VP}) = (2.30 \pm 0.06) - (0.00618 \pm 0.00008)G_{\mathrm{I}} - (4.02 \pm 0.10)\mathrm{HDCA\text{-}2} + (0.129 \pm 0.006)\mathrm{SA\text{-}2(F)} + (6.02 \pm 0.574)\mathrm{MNAC(C1)} - (0.0143 \pm 0.0017)\mathrm{SA(N)} \quad (10)$$

$$n = 411, R^2 = 0.949, R^2_{\mathrm{CV}} = 0.947, s = 0.331$$

Two important descriptors (G_{I} and HDCA-2) used in the model represent the forces of intermolecular attraction; G_{I} is connected with the dispersion and cavity-formation effects in liquids, and HDCA-2 is connected with the hydrogen bonding ability of compounds. Three additional descriptors used in the model are as follows: the sum of the surface area of fluorine atoms SA-2(F), the maximum net atomic charge for a chlorine atom MNAC(Cl), and the surface area of nitrogen atoms SA(N). The cross-validated correlation coefficient $R^2_{\mathrm{CV}} = 0.947$, when compared to the $R^2 = 0.949$, indicates high stability of the regression equation, and the standard error $s = 0.331$ is less than that of the Liang and Gallagher model ($s = 0.534$).[311]

Goll and Jurs[313] reported a 7:3:1 computational neural network (CNN) model for the prediction of vapor pressure (log VP) for a data set comprised of 352 hydrocarbons and halohydrocarbons. The rms errors associated with the training ($n = 270$), cross-validation ($n = 30$), and prediction ($n = 52$) set compounds used for this CNN model were 0.163, 0.163, and 0.209 log units, respectively. The less diverse the types of compounds in the data set, the better the model appeared to be. McClelland and Jurs[314] developed an eight-descriptor CNN model, for the prediction of log VP at 25 °C of 420 diverse organic compounds, and they obtained a rms error of 0.37 log units for 65 compounds of an external prediction set based only on topological descriptors. The authors also obtained a ten-descriptor CNN model with an improved prediction set rms error of 0.33 log units within a descriptor range from topological, electronic, and geometrical to hybrid types.

Cash[315] developed a QSPR model for the prediction of VP of a large data set of 1676 compounds by employing the K-nearest-neighbors (KNN) method in a Euclidian space based on electrotopological state indices ($n = 1676$, $R^2 = 0.697$, $s = 0.560$). A modification of the KNN method using PCA to minimize the vector sum of the vectors from the test structure to the neighbors gave better results ($n = 1676$, $R^2 = 0.733$, $s = 0.526$).

Vapor pressure is highly temperature dependent, and thus temperature-dependent prediction methods are desirable. Chalk et al.[316] developed a temperature dependent model for VP based on a feed-forward ANN (27:15:1) and quantum chemical descriptors. The VP values ranged from −8.63 to 5.47 log-(Torr) units with temperatures ranging from 76 to 800 K. The training set of 7681 and validation data set of 861 compounds gave QSPR models with $R^2 = 0.9762$ and $s = 0.322$, and $R^2 = 0.9758$ and $s = 0.326$, respectively. A comparison with the McClelland and Jurs[314] results was made. Yaffe and Cohen[317] developed a back-propagation ANN QSPR model for the prediction of VP as a function of temperature. The vapor pressure−temperature behavior of the hydrocarbons (C_4−C_{12}) based on valence molecular connectivity indices, molecular weight, and temperature was predicted. The database of the Design Institute for Physical Property Data (DIPPR) containing VP data for a total of 7613 homogeneous compounds was used. The average absolute errors (see Figure 7a) and standard deviations were (11.6%, 8.0%) for the training set ($n = 5330$) and (8.2%, 5.9%) for the test set ($n = 1529$), (9.2%, 7.8%) for the validation set ($n = 754$), and (10.7%, 7.8%) for the overall data set ($n = 7613$). Finally, Yaffe and Cohen selected a small data set of 274 hydrocarbons, analyzed the overall performance of the 7−29−1 ANN model, and compared their model with those of Katritzky et al.,[312] Goll and Jurs,[313] and McClelland and Jurs.[314] Yaffe and co-workers obtained lower average and maximum pressure estimation errors from all reported models: 120 compounds from Goll and Jurs,[313] 108 compounds from Basak and Mills,[318] and 45 compounds from Katritzky et al.[312] were common to their data set. Katritzky et al.,[312] Goll and Jurs,[313] and Basak and Mills[318] reported QSPR models for heterogeneous data sets, whereas Yaffe and Cohen[317] built their model for a homogeneous data set;

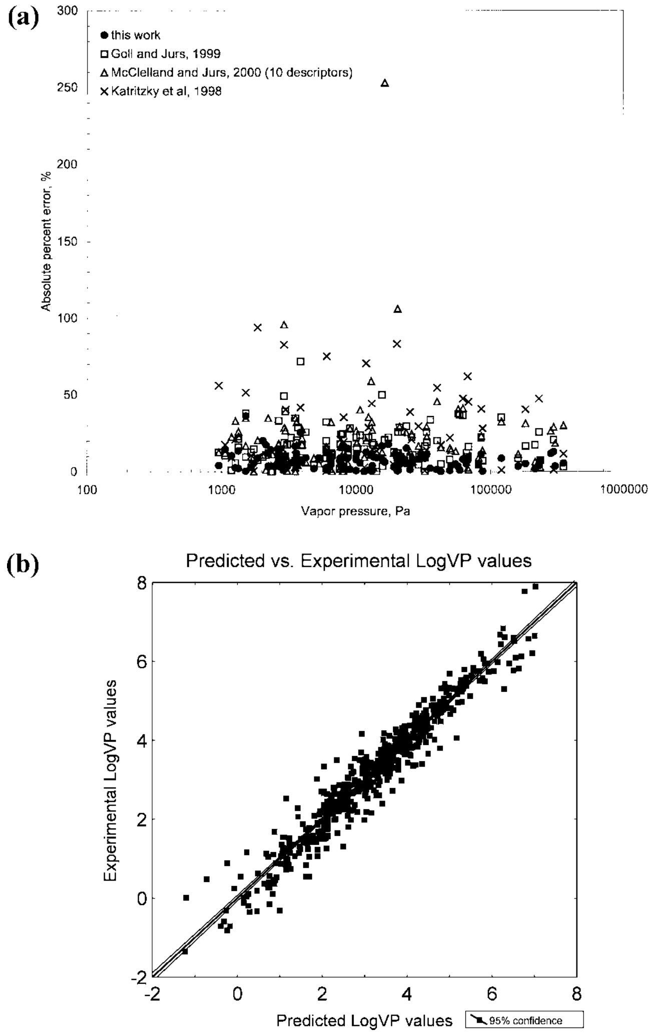


Figure 7. (a) Comparison of the absolute errors for the predicted hydrocarbon vapor pressures at 25 °C. Reprinted with permission from ref 317. Copyright 2001 American Chemical Society. (b) Scatter plot of the calculated vs experimental log(VP) values at 25 °C for 645 compounds. Reprinted with permission from ref 319. Copyright 2007 Elsevier B. V.

their greater accuracy may be due to the homogeneity of their data set.

Katritzky et al.[319] developed a general QSPR model for the log VP at 25 °C for 645 diverse compounds taken from their previous study[312] and from that of McClleland and Jurs.[314] A four-parameter MLR model was obtained with R^2 = 0.937 and s = 0.366 based on molecular descriptors calculated using CODESSA PRO. The four descriptors included in the model—the gravitation index (all bonds), the number of F atoms, HA dependent HDCA-1 (Zefirov PC), and FNSA-2 Fractional PNSA (PNSA-2/TMSA) (MOPAC PC)—gave significant contributions to the modeling of vapor pressure. The validity of the model was tested for the subsets of data by using the same set of four descriptors. The authors compared their data with QSPR models previously reported by Liang and Gallagher,[311] Goll and Jurs,[313] and McClleland and Jurs.[314] In their QSPR model Katritzky used fewer descriptors and a larger number of diverse compounds to obtain predicted values in good agreement with experimental data with minimum standard error values (see Figure 7b).

4.9. Surface Tension

Surface tension (ST) is an important physical property which reflects the intermolecular interaction of molecules. Many generalized statements have been made which associate polar and hydrogen-bonding intermolecular interactions with increased surface tension of pure liquids.[320,321] Surface tension has been shown to increase as a function of molecular weight for a set of congeners[322] and has been used to predict other physicochemical properties.[323,324] However, several QSPR models involving ST have been published.

Multiparameter regression models were reported by Needham et al.[159] to predict surface tension at 20 °C for a set of 68 alkanes having $R^2 = 0.986$, $s = 0.2$ and $R^2 = 0.989$, $s = 0.2$ with connectivity and ad hoc descriptors, respectively. Stanton and Jurs correlated the surface tension of 31 diverse organic compounds with structural descriptors including charged partial surface area (CPSA) descriptors which combined solvent accessible surface areas with partial atomic charges. They found a six-parameter model with R^2 = 0.908 and s = 2.32.[158] The authors also obtained good MLR models for surface tension of organic compounds containing 95 alkanes, 56 alkyl esters, and 35 alkyl alcohols by employing a wide variety of topological, geometrical, and electronic descriptors.[144] Finally, a 10-descriptor QSPR model for the combined set of data including 146 training compounds and 20 external validation compounds was obtained. The statistical results for the training and test sets were (R^2 = 0.983, s = 0.4) and (R^2 = 0.983, s = 0.7), respectively: the removal of an outlier from the test compounds lowered the error value to that of the training set, and that molecular surface area provided better results in modeling surface tension than the van der Waals or solvent-accessible surface area. Several different theoretical approaches have been applied based on molecular descriptors for the prediction of surface tension of homogeneous data sets of alkanes[175,325−327] and also, for diverse classes of compounds.[328,329] MLR and CNN were employed by Kauffman and Jurs[330] for the prediction of surface tension of 199 solvents. An eight-descriptor MLR model was obtained with $R^2 = 0.835$ and s = 3.37 for the training set and (R^2 = 0.837, s = 3.37) for the prediction set. The CNN model showed average percent

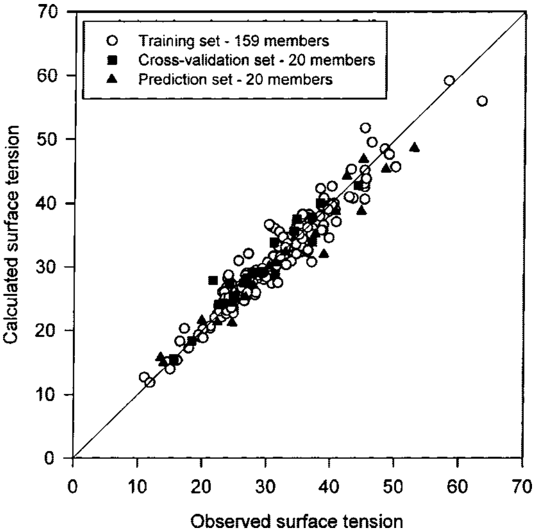


Figure 8. Plot of calculated versus observed surface tension for the training, cross-validation, and prediction set compounds using the CNN model. Reprinted with permission from ref 330. Copyright 2001 American Chemical Society.

errors of the predicted values of 5.3% for the training set, 6.1% for the cross-validation set, and 6.4% for the prediction set (see Figure 8). The authors reported a general CNN model for predicting surface tension, viscosity, and thermal conductivity of compounds at a greater level of accuracy.

A QSPR model was developed for the prediction of the surface tension of nonionic surfactants including a topological descriptor, the Kier and Hall index of zeroth order ($^0\chi$) of the hydrophobic segment of the surfactant, and a quantum chemical descriptor, the heat of formation of the surfactant molecules.[331] The QSPR model obtained between the surface tension and the descriptors produced $R^2 = 0.987$ for the studied 30 nonionic surfactants.

4.10. Critical Temperatures

Critical temperature is an important property determined by intermolecular interactions between molecules in the liquid state. Earlier estimates of the critical temperature were made from measured quantities such as boiling points, parachor, and molar refraction.[149] Among several different approaches, group contribution methods were successfully employed for the estimation of critical temperatures from molecular structures.[154,209] Grigoras applied a novel approach based on computation of the molecular surface interactions (MSI) to estimate the critical temperatures together with various other properties.[332] Several approaches based on molecular descriptors have also been employed for the prediction of the critical temperatures of acyclic hydrocarbons.[159,174,175] An eight-parameter MLR model was obtained by Jurs[169] which included three CPSA, two topological descriptors, and three constitutional descriptors, for the prediction of critical temperature of 147 diverse organic compounds with $R^2 = 0.978$ and $s = 11.9$ K. A three-descriptor regression model which included experimental boiling points as one of the parameters gave a correlation of critical temperature for 147 compounds with $R^2 = 0.988$ and $s = 8.48$. An improved eight-descriptor MLR model was developed by excluding two outliers (quinoline and hexanitrile) with an rms error of 9.16 K, and $R^2 = 0.986$ and by removing another outlier (acetone) an rms error of 9.13 was achieved.[333] The authors also developed a nonlinear 8−4−1 CNN model with an observed rms error of 7.3 K for the 132 training, 7.7 K for 15 cross-validation, and 9.9 K for the 18 prediction set compounds, respectively.

Hall and Story[173] applied an ANN based on 19 atom type electrotopological state indices for the prediction of critical temperatures of 165 compounds with an overall relative error ranging from 0.97% to 1.17% and a mean absolute error of 4.52 K.

One- and three-parameter QSPR models were developed by Katritzky et al.[334] to correlate the critical temperatures of the sets of 76 hydrocarbons and 165 structurally diverse compounds. The one-parameter model utilizing the cube root of the gravitation index allowed the prediction of critical temperatures for 76 hydrocarbons with $R^2 = 0.953$ and $s = 18.9$ while the three-parameter model for 165 diverse compounds gave $R^2 = 0.955$ and $s = 16.8$ K.

Bonchev[335,336] applied an overall connectivity, a topological complexity, and an overall Wiener index for the correlation of critical temperatures of alkane isomers. The line graph parameters[337] and the modified Harary index[338] also correlated well with the critical temperatures of alkane isomers. A fuzzy ARTMAP-based ANN QSPR model predicted the critical temperatures of 530 compounds with an absolute mean error of 1.4 K (0.24%) by using molecular descriptors which included the sum of atomic numbers, valence connectivity indices, the second-order Kappa shape index, and the dipole moment.[192] Yao et al.[339] utilized a MLR and radial basis function neural network (RBFNN) approach for the prediction of critical temperatures of 856 organic compounds based on molecular descriptors calculated solely from structure. A ten-descriptor linear model was obtained with the rms error of 16.21 K and $R^2 = 0.974$ for the 733 training set compounds. The rms error was 16.4 K for 123 external test compounds. A 10−33−1 RBFNN model for the critical temperatures of 856 organic compounds gave an rms error of 14.0 and 12.3 K for the 733 training and 14.2 K for the 123 test set compounds, respectively.

Simple topological molecular parameters based on atomic coordination numbers for different atoms and the corresponding identity of the chemical bonds were used for the correlation of critical temperatures of 164 diverse compounds.[340,341] QSPR models were obtained for the correlation of critical temperatures of 61 and 74 alkanes using the molecular descriptors based on the eigenvalues of the bond adjacency matrix[245] with $R^2 = 0.998$, $s = 4.0$ and the atom adjacency matrix[342] with the mean absolute error of 5.3 °C. A five-parameter MLR model was obtained for 139 alkanes with $R^2 = 0.996$ and $s = 16.1$ °C based on vertex, edge, ring, and distance related topological indices. Shacham et al.[343] applied a molecular similarity approach to various physicochemical properties of unmeasured data, and they obtained an average prediction error of 0.91 for the critical temperatures of 18 compounds. Many different topological parameters calculated from structures have been applied to predict the critical properties of compounds.[344,345] Charton[346] has reviewed the nature of topological parameters in the prediction of physicochemical properties of compounds, noting that topological parameters are composites representing counts of the numbers of atoms, bonds, electrons, and branching. Recently, Dobchev and Karelson[90] calculated novel parameters by the reparameterization of the AM1 based on a nonlinear optimization technique and obtained a two parameter QSPR model with $R^2 = 0.902$ and $s = 25.04$ K.

Godavarthy et al.[347] developed QSPR models for critical temperatures of a diverse data set containing over 1230 organic compounds based on molecular descriptors calculated solely from structure using CODESSA PRO. Several

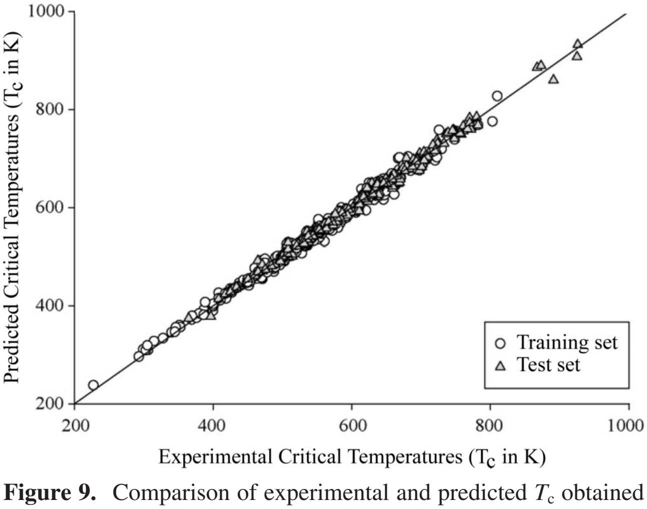


Figure 9. Comparison of experimental and predicted T_c obtained from the nonlinear BPNN model based on weighted mean square error (WMSE). Reprinted with permission from ref 347. Copyright 2007 Elsevier B. V.

approaches—including linear, nonlinear, and genetic algorithms (GA)—were employed in the model development. The statistical characterstics of various QSPR models obtained are as follows: MLR ($n = 1230$, $R^2 = 0.913$, AAD = 16.1), PLS ($n = 1230$, $R^2 = 0.935$, AAD = 13.8), nonlinear analysis (NLA) with linear descriptor reductions ($n = 1230$, $R^2 = 0.972$, AAD = 8.6), NLA with sum of squared errors for error propagation ($n = 1230$, $R^2 = 0.992$, AAD = 4.5), and NLA with GA based on WMSE (weighted mean square error) ($n = 1230$, $R^2 = 0.995$, AAD = 3.7). The resultant nonlinear QSPR models are capable of giving excellent predictions of the critical temperatures of diverse compounds (see Figure 9).

4.11. Critical Pressures

The critical pressure of a substance is the pressure required to liquefy a gas at its critical temperature. Critical temperature and pressure values are of great importance in chemical engineering for the calculation of the equation of state of thermodynamic and transport properties used in high-pressure phase equilibrium processes, such as oil recovery and supercritical fluid extraction.[348] However, the experimental determination of critical pressure is expensive and time-consuming, often involving uncertainty due to the impurities and/or decomposition. Consequently, many different approaches have been made to calculate critical property values (see ref 347), and several QSPR models have also been reported for the correlation and prediction of critical properties based on experimental measurements. However, most of the QSPR models use molecular descriptors calculated from structure and focus on a homologues series of hydrocarbons.[159,175,245,336,338,343−345]

Jurs and co-workers[333] obtained QSPR models for critical pressure ranging from 12 to 55 atm for 165 diverse organic compounds taken from the Design Institute for Physical Property (DIPPR) database using molecular descriptors calculated by ADAPT. An eight-descriptor MLR model was obtained with $R^2 = 0.929$ and rms error = 2.0 atm for 147 training set compounds and a rms error of 2.8 atm for 18 prediction set compounds. The authors found that the majority of errors in their model were in the prediction of compounds with high critical pressures. They also developed an 8−5−1 CNN model which gave rms errors of 1.51, 1.35, and 2.39 atm for 132 training set, 15 cross-validation set, and 18 prediction set compounds, respectively.

Duchowicz and Castro[340] applied simple constitutional descriptors, derived from atoms and classical bonds, as basic variables to predict the critical pressures of the Jurs set of 164 diverse compounds.[333] The authors obtained R^2 ranging from 0.8015 to 0.8941 through C_1-C_4 calculation schemes. The same descriptors were also implemented in the prediction of critical property values of 43 normal and 9 branched alkanes with $R^2 = 0.99$.[341]

Espinosa et al.[192] investigated the applicability of fuzzy ARTMAP QSPR models to estimate the critical pressures of 463 diverse compounds ranging from 8.95 to 1.02 MPa. Their fuzzy ARTMAP models with eight descriptors as input (sum of atomic numbers, five valence connectivity indices, second order kappa shape index, and dipole moment) showed the best predictive (absolute mean errors of 0.02 MPa) and extrapolation capabilities compared to optimal back-propagation models and group contribution methods (see Figure 10a). However, the inclusion of the dipole moment in the model showed a significantly smaller effect on the critical properties.

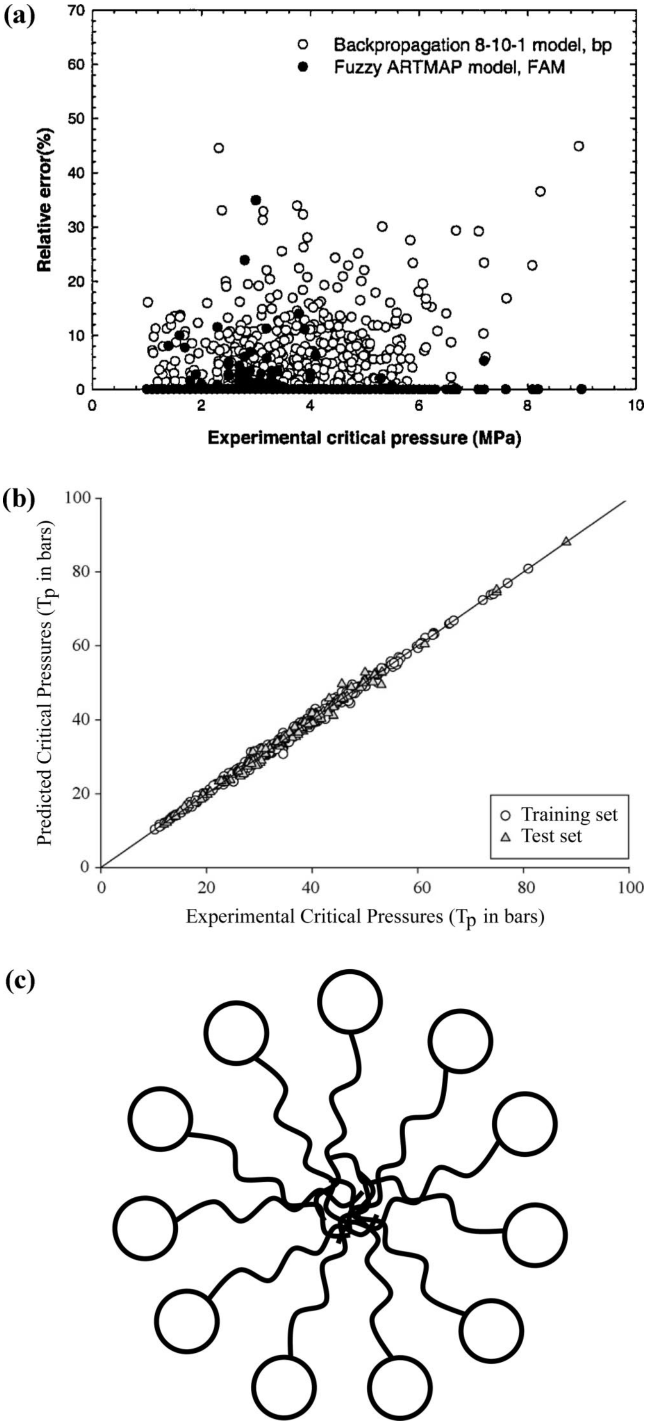


Figure 10. (a) Comparison of the relative errors of the critical pressures from the fuzzy ARTMAP model with 8−10−1 back-propagation architecture for the complete data set. Reprinted with permission from ref 192. Copyright 2001 American Chemical Society. (b) Comparison of experimental and predicted P_C obtained from the nonlinear BPNN model based on weighted mean square errors (WMSE). Reprinted with permission from ref 347. Copyright 2007 Elsevier B. V. (c) Schematic structure of micelle structure in aqueous solution.

In a recent paper, Sola et al.[349] obtained a QSPR model for the critical pressures of a set of 121 diverse compounds taken from the DECHEMA database based on CODESSA PRO methodologies. Their final eight-descriptor QSPR model showed a significantly higher accuracy ($R^2 = 0.9209$) with respect to the best available group-contribution method. Comparable results were also obtained with respect to other

Table 2. Summary of the QSPR Models of Critical Pressures

type of compd	N	molecular descriptors (n_d)	QSPR method	R^2	s	ref
normal alkanes	43	carbon number and its functional power (3)	MLR	0.9889	0.0179	Krenkel et al.[341]
branched alkanes	41	carbon number and its functional power (10)	MLR	0.9942	0.0007	Krenkel et al.[341]
alkanes (C_1-C_{20})	60	topological parameters (5)	MLR	0.9970	0.0771	Cao et al.[245]
diverse compounds	463	atomic numbers, valence connectivity, kappa shape index, dipole moment (8)	fuzzy ARTMAP neural network		0.08	Espinosa et al.[192]
diverse compounds	463	N, $\chi(0-4)$, k^2 (7)	fuzzy ARTMAP neural network		0.09	Espinosa et al.[192]
diverse compounds	463	N, μ, $\chi(0-4)$, k^2 (8)	backpropagation ANN (8-10-1)		0.39	Espinosa et al.[192]
diverse compounds	132	topological, electronic, geometrical (8)	MLR	0.9293	2.03 (atm)	Turner et al.[333]
diverse compounds	132	topological, electronic, geometrical (8)	feed forward ANN (8-5-1)		1.51 (atm)	Turner et al.[333]
alkanes	74	topological indices (5)	MLR	0.9769	0.655 (atm)	Lučić et al.[338]
alkanes	82	topological indices (2)	MLR	0.9130		Ni et al.[345]
hydrocarbons	129	topological indices (4)	MLR	0.9712	0.188 MPa	Yuan et al.[350]
alkanes	74	topological indices (2)	MLR	0.9131	1.25	Shamispur et al.[344]
alkanes	74	topological indices (5)	PCR	0.9760	0.67	Shamispur et al.[344]
alkanes	74	connectivity indices (5)	MLR	0.9810	0.60	Needham et al.[159]
alkanes	74	orthogonalized descriptors (7)		0.8845		Klein et al.[175]
alkanes	38	topological indices (5)	MLR	0.9905	0.40	Bonchev[336]

QSPR models[333] despite the different composition of the database, confirming the versatility and robustness of the QSPR method.

Godavarthy et al.[347] obtained QSPR models for the critical pressures of a large data set of 1230 diverse organic compounds with an average absolute percent deviation (%AAD) of 1.5 based on molecular descriptors calculated by using CODESSA PRO. The authors[347] compared their model with those previously reported in the literature. The various linear and nonlinear QSPR models obtained showed good statistical characteristics for MLR ($n = 1230$, $R^2 = 0.951$, AAD = 1.24), PLS ($n = 1230$, $R^2 = 0.971$, AAD = 0.95), nonlinear analysis (NLA) with linear descriptor reductions ($n = 1230$, $R^2 = 0.984$, AAD = 0.76), NLA with SSE ($n = 1230$, $R^2 = 0.991$, AAD = 0.52), and NLA with GA based on WMSE (weighted mean square error) ($n = 1230$, $R^2 = 0.992$, AAD = 0.49). The nonlinear QSPR models show promising results in predicting the critical temperatures of a diverse set of compounds (see Figure 10b) compared to previously published results. Some important QSPR models developed for the critical pressure together with the statistical parameters are summarized in Table 2.

4.12. Heats of Vaporization

The heat of vaporization, ΔH_V, is the energy required to transform a given quantity of a substance into a gas. Heat of vaporization is thus the energy required to overcome the intermolecular interactions in a liquid or solid, and it measures the strength of intermolecular forces. Practical applications of heats of vaporization include the understanding of distillation and vapor pressure. Distillation is ubiquitous for the separation and purification of compounds. The heat of vaporization is the fundamental quantity that determines the experimental conditions at which an industrial or laboratory-scale distillation should be run. The heat of vaporization of a liquid allows the calculation of vapor pressure at any temperature and permits the control of vapor pressure by setting the temperature of the liquid being vaporized, hence being an important tool.

Estimation of enthalpies of vaporization, ΔH_V, has been the subject of numerous papers. However, ΔH_V values depend not only on the molecular weight and composition of substances but also on their structure. The corresponding technique for additive calculations has been developed in detail and gives good results for hydrocarbons. For compounds containing heteroatoms, however, it is less developed and significant deviations from experimental data are found. An additivity calculation of ΔH_V for 295 diverse *n*-alkenyl compounds[351] was reported based on the ECN (effective carbon number) characterizing the functional group, the number of carbon atoms, and the position of the double bond. Their model explained 99.8% of the variance in the data, with the mean absolute deviation of 0.5%. An additive calculation of ΔH_V at 298.15 K for a large set of isomeric ketones C_5-C_{15} was carried out by Emel'yanenko and Roganov.[352] One of the probable causes of the poor correlation is the lack of sufficiently diverse sets of reliable data on ΔH_V. Makitra and Polyuzhin[353] correlated ΔH_V of a series of isomeric ketones C_5-C_9 with the Hammet-Taft equation. Several two-parameter (σ^* and E_s) regression equations were obtained for different sets of ketones with good correlation ($R^2 > 0.80$). The U.S. EPA group estimated heats of vaporization of a large number of compounds from the EPA database based on the SPARC physical process calculator method[354] with a rms error of the predicted values close to the intralaboratory experimental errors. A 3D QSPR model was developed by Puri et al.[355] for the correlation of ΔH_V of a set of polychlorinated biphenyls (PCBs) at 298.15 K with CoMFA (comparative molecular field analysis)[356a] based physicochemical parameters. Their model yielded $R^2 = 0.996$ and $R^2_{CV} = 0.852$ with the atom fit alignment and Gasteiger-Marsili charges, which were used for the prediction of the entire set of 209 PCB congeners.

Predictive QSPR models correlating experimental solubility parameters and enthalpies of vaporization have been derived from QM calculated descriptors.[356b] A four-descriptor Hildebrand total solubility parameter regression equation was developed with $R^2 = 0.97$, $R^2_{cv} = 0.97$, $F = 461.5$, $s^2 = 0.53$, and root mean square error (RMSE) = 0.69. A four-descriptor QSPR model for the prediction of enthalpies of vaporization (ΔH_{vap}) had $R^2 = 0.96$, $R^2_{cv} = 0.96$, $F = 230.3$, $s^2 = 4.75$, and RMSE = 2.04.

Several 2D structural descriptors have been developed time to time by numerous authors. In 1947 Wiener[8] calculated ΔH_V of linear and branched chain hydrocarbons by using structural parameters, such as the path number, *w*,

Table 3. Summary of the QSPR Models on Heats of Vaporization (ΔH_V)

no.	type of compd	N	molecular descriptors	QSPR methods	R^2	s	ref
1	alkanes	134	structural descriptors based on information theory	MLR	0.9896	0.63	Ivanciuc et al.[359]
2	C_3 to C_8 alkanes	38	TIs (overall Wiener indices)	MLR	0.9950	0.67	Bonchev[336]
3	organofluorine compounds	9	fragment descriptors	MLR	0.9628	rms = 0.128	Golovanov et al.[360]
4	alkanes	134	TIs based on reverse Wiener matrices	MLR	0.9777	0.65	Ivanciuc et al.[361]
5	C_2-C_9 alkanes	68	atom-type topological indices	MLR	>0.9900	0.34	Ren[362]
6	alkanes	69	TIs based on adjacency matrix	MLR	0.9980	0.34	Ponce[342]
7	alkanes		TIs based on distance complement matrix	MLR			Ivanciuc et al.[363]
8	alkanes, alcohols, polyols, ethers	32	descriptors based on molecular mechanics and quantum chemical calculations	MLR	0.9817	0.311 (kcal/mol)	Dyekjaer et al.[364]
9	alkanes	57	topological indices	MLR	0.9934	0.66 (kJ/mol)	Cao et al.[245]
10	hydrocarbons	159	topological indices	MLR	0.9910	1.34 (kJ/mol)	Yuan et al.[350]
11	C_4 to C_{12} aliphatic ketones	39	electrotopological indices	MLR			Marino et al.[365]
12	alkanes	69	topological indices	PCA	0.9921	0.50	Shamsipur et al.[344]
13	alkylbenzenes	47	topological indices	SVR		rms = 0.699	Yang et al.[342]
14	saturated hydrocarbons	66	fragment based topological descriptors	MLR	0.9878	σ = 0.942 (kcal/mol)	Tsygankova et al.[366]

and the polarity number, p, calculated solely from chemical structure. Needham et al.[159] in their study reported the correlation equation for the prediction of heats of vaporization of 69 normal and branched chain alkanes with an $R^2 = 0.99$ based on structural parameters. Numerous QSPR correlations have appeared in the literature for the prediction of ΔH_V of a homologous series of compounds, mostly hydrocarbons, by the use of molecular descriptors calculated from structure.[174,175,357,358] Some QSPR models developed for ΔH_V are summarized in Table 3.

4.13. Heats of Formation

The heat (or enthalpy) of formation ($\Delta H°_f$) is a fundamental thermodynamic property for predicting the chemical characteristics of compounds. Thus, heats of formation are important in the investigation of bond energies, resonance energies, the nature of chemical bonds, the calculation of equilibrium constants of reactions, etc.[367] Considerable effort has been directed toward the determination of the $\Delta H°_f$ in the past.[368−370] Various estimation methods for the calculation of the $\Delta H°_f$ are introduced based on isodesmic and homodesmic reactions atom group equivalents, transferability and additivity of Fock matrix elements, etc.[371−379] Chen and co-workers[380] applied first-principles methods, density functional theory (DFT),[381−383] and Hartree−Fock (HF) to calculate $\Delta H°_f$ of 180 organic molecules which showed large deviations. Duan et al.[384] implemented an ANN approach based on the descriptors obtained from natural bond orbital analysis, and an enlarged training set of 350 diverse compounds showed improved results as compared to the earlier calculation methods.[380−383] Upon ANN correction, the rms deviations for the 350 molecules were reduced from 11.2 to 4.4 and from 15.2 to 3.5 kcal/mol for B3LYP/6-31G(d) and B3LYP/6-311G(2d,d,p) methods, respectively. At the same time, the calculated $\Delta H°_f$ of the HF method improved greatly, and the rms deviations were reduced from 327.1 to 9.5 kcal/mol for the HF/6-31G(d) method.

Heats of formation were correlated with topological, quantum chemical, and various other descriptors calculated solely from structure, most successfully ($R^2 = 0.99$) for homologues series of alkanes. Various quantum chemical descriptors such as ionization potential (I), electron affinity (A), quantum chemical hardness index (η), softness index (S), electronegativity (χ), and electrophilicity (ω) correlated with $\Delta H°_f$ of a set of 39 C_2-C_8 alkanes.[385] The correlation of the $\Delta H°_f$ of alkanes with the indices I, η, and S showed a good linear fit with ($R^2 = 0.9274$, $s = 2.2$), ($R^2 = 0.893$, $s = 2.7$), and ($R^2 = 0.893$, $s = 2.7$), respectively.

Correlations of $\Delta H°_f$ with TIs based on overall Wiener indices[336] and the molecular electronegative distance vector (MEDV)[326] showed correlations ($n = 38$, $R^2 = 0.9984$, $s = 1.33$) and ($n = 54$, $R^2 = 0.9920$, rms = 2.58) for a set of alkanes, respectively. TIs derived from the distance and detour (maximum distance) matrix were applied to correlate the $\Delta H°_f$ of a set of 60 hydrocarbons.[386] The authors obtained regression models with two sets of five parameters which showed low average absolute deviations of 0.76 and 0.62 kcal/mol, respectively, as compared to the degree of uncertainty in the experimental determination around 2 kcal/mol. Also, TIs derived from the adjacency matrix were used in the correlation of polyhex polycyclic aromatic hydrocarbons.[387] G Gallegos and Girones[388] developed novel topological quantum similarity indices based on fitted densities from the atomic shell approximation procedure and used in the correlation of the $\Delta H°_f$ of alkanes. Their four-descriptor MLR model showed the best predictivity results ($R^2 = 0.990$ and q^2= 0.988) for $\Delta H°_f$ of 60 hydrocarbons. Topological indices derived by optimization of correlation weights of local graph invariants (OCWLGI) based on labeled hydrogen-filled graphs (LHFGs) and the graphs of atomic orbitals (GAOs) were applied for the correlation of the $\Delta H°_f$ for a set of alkanes.[253] Their QSPR model showed better results with ${}^0X_{CW}$ (GAO, EC1): $R^2 = 0.9984$, $s = 1.804$ for 66 training set compounds, and $R^2 = 0.9803$, $s = 1.791$ for 67 test set compounds compared to the previously reported results.[389] Distance connectivity based topological indices (Sh indices)[344] were correlated with $\Delta H°_f$ of 54 alkanes and gave $R^2 = 0.906$ and $s = 9.08$, and the bivariate regression with Sh_1 (first order Sh index) and n (number of carbon atoms) improved the correlation with $R^2 = 0.9711$, $s = 5.07$. The

Table 4. Summary of the QSPR Models of Heats of Formation ($\Delta H°_f$)

type of compd	N	molecular descriptors	QSPR methods	R^2	s	Reference
alkanes	46	graph theoretical descriptors	MLR	0.990	1.215	Garbalena et al.[357]
octanes	18	graph theoretical descriptors	MLR	0.936		Kuanar et al.[358]
alkanes	54	TIs based on molecular electronegative distance vector	MLR	0.992	rms = 2.58	Liu et al.[326]
alkanes	39	quantum chemical descriptors (ionization potential)	MLR	0.927	2.2	Thanikaivelan et al.[385]
aliphatic alcohols	20	fragment descriptors	MLR	0.993	6.227	Golovanov et al.[392]
alkanes	60	TIs based on distance and detour matrix		0.999	1.11	Mercader et al.[386,387]
				0.999	0.93	
alkanes	38	TIs based on overall Wiener indices	MLR	0.9984	1.33	Bonchev[336]
aliphatic ketones	39	electrotopological indices	MLR	0.99 (n = 22, train)	5.11	Marino et al.[365]
alkanes	133	TI derived from OCWLGI based on graphs of atomic orbitals (GAOs)	MLR	0.998 (n = 66, train)	1.804 (train)	Toropov et al.[253]
				0.980 (n = 67, test)	1.79 (test)	
alcohols and alkanes	88	topological indices (tau indices)	MLR	0.943	3.52	Roy et al.[391]
alkanes	54	distance-connectivity based topological indices	PCR	0.987	3.51	Shamsipur et al.[344]
alkanes	60	topological quantum similarity indices	MLR	0.999		Gallegos et al.[388]
nonaromatic polynitro compounds		physicochemical and topological indices	MLR			Sukhachev et al.[390]
diverse compounds	350	descriptors based on natural bond orbital analysis	HF/6-31G (d), ANN		rms = 9.5 kcal/mol	Duan et al.[384]

use of principal component variables in the QSPR model also showed improved results with $R^2 = 0.9873$ and $s = 3.51$.

Few articles report the QSPR correlations of the $\Delta H°_f$ of diverse sets. The heats of formation of a set of nonaromatic polynitro compounds were correlated with physicochemical and topological descriptors based on the QSPR approach.[390] QSPR models were obtained for $\Delta H°_f$ for a set of diverse functional acyclic compounds based on molecular connectivity indices (MCI), molecular negentropy (MN), and topochemically arrived unique (TAU) indices.[391] TAU indices developed in a VEM (valence electron mobile) environment showed better results for $\Delta H°_f$ of a set of 21 alcohols and 67 alkanes. The QSPR models of $\Delta H°_f$ for a set of 21 alcohols, 67 alkanes, and the combined set of alkanes and alcohols showed promising statistical results with ($R^2 = 0.982$, $s = 1.471$), ($R^2 = 0.992$, $s = 1.083$), and ($R^2 = 0.943$, $s = 3.520$), respectively, using TAU indices rather than MCI and MN indices. TAU indices unravel specific contributions of molecular bulk (size), functionality, branchedness, and shape parameters to the molecular thermochemical properties of diverse functional compounds. Some important QSPR models are summarized in Table 4.

4.14. Entropies

Entropy measures thermal energy per unit temperature of a system unavailable for doing useful work. The standard entropy is a highly important thermodynamic parameter of a substance used in physicochemical processes. In some cases, the measurement of this property involves experimental difficulties and the standard methods have substantial restrictions.[393,394] Previously, the molecular group contribution approach[395,396] and quantum chemical methods[397] were used to estimate entropy. These methods showed large absolute errors. Consequently, QSPR models were made based on the availability of experimentally measured data to be able to predict entropy values. Kuanar et al.[358] correlated the entropy values of 18 octane isomers with line graph parameters derived from molecular structure and obtained a single-parameter equation with $R^2 = 0.919$. Topological indices derived on the basis of optimized correlation weights of linear graph invariants were used for the prediction of entropy values of 40 acyclic and aromatic compounds.[398] The authors obtained a MLR model with $R^2 = 0.973$ and $s = 2.36$. Golovanov et al. used a simple approach based on a mathematical equation using only the number of C atoms in the alkyl radical, n, as the molecular descriptor for the calculation of entropy values as well as numerous other properties of a set of alcohols[392] and an improved model using 10 parameters for 117 saturated hydrocarbons,[399] giving highly accurate estimates for various properties.

Total entropies of melting for 370 pharmaceuticals and environmentally relevant compounds were predicted using two descriptors: molecular rotational symmetry number and molecular flexibility number with an average error of 21%.[400] In combination with the two descriptors mentioned, Johnson and Yalkowsky[401] used two novel structural parameters, eccentricity (ε) and spirality (μ), for the prediction of entropy of melting (ΔS_m) and obtained a regression equation with $R^2 = 0.90$ for a set of 117 aliphatic and aromatic hydrocarbons. Eccentricity was defined as the ratio of the volume of a box around a rigid molecule to the cubed radius of a sphere containing the same molecular van der Waals volume. Spirality was defined as the number of benzo[c]phenanthrene regions present in the molecule which results in repulsion and out of plane twisting to maximize the distance between hydrogens.

4.15. Rotational Activation Energies for Amides

Prediction of rotational barriers about the amide bond has been of substantial interest, particularly in relation to the conformational studies of peptides and polyamides. Semiempirical quantum-chemical methods[402,403] were used for the prediction of potential energy surfaces of amide bond rotation which underestimate the experimental activation energies. Ab initio calculations were made by Wiberg et al.,[404] and although values were closer to experimental, the calculations were very time-consuming and hence impractical for larger systems.

The main deficiency of the semiempirical calculations has been related to the inaccurate presentation of the lone pair interactions in the compounds containing nitrogen.[405] Consequently, the errors made in calculations of both the energy of the ground state and the energy of the rotational transition state result in unpredictable errors in activation energy. Moreover, the attempts to correlate the experimental free energies of activation with the calculated activation energies of rotation follow the assumption that the entropy change during the rotation is small and nearly equal for all amides. The applicability of such an assumption in the case of more variable structures is, however, questionable. For instance, the experimental activation entropy for various *N*,*N*-dialkylamides ranges between -3.7 and 1.5 cal K^{-1} mol^{-1} in the gas phase.[406,407] The variation of this activation entropy is even more substantial in the liquid phase.[407] Another source of errors leading to poor correlation between the quantum-chemically calculated activation energies and the respective experimental values emerges from the variation in experimental conditions (e.g., in media, concentration, or temperature) and the precision of measurements. Nevertheless, for a series of similar (homologous) compounds, the AM1 calculated activation energies were successfully correlated with the respective experimental conformational transition energies.[408]

In general, the origin of the rotational barrier about the N−C(O) bond in amides is related to the decoupling of mesomeric interaction within the amide group in the rotational transition state. Thus, the respective rotational activation energy should depend on the magnitude of the resonance stabilization in the planar rotational ground state. On the other hand, the ab initio investigations by Wiberg et al[409,410] and Bader et al.[411] revealed that the internal rotational barriers of isolated molecules can be described by the change in the attractive and repulsive interactions during the rotation. It has also been found that the respective energy changes are in good accord with the model considering the rehybridization of the nitrogen atom during the N−C(O) bond rotation. This rehybridization model confirms the view that the main difference in rotational barriers of different amides is caused by the different extent of nitrogen hybridization in the rotational ground state, whereas it is similar (i.e., sp^3-hybridized) in the transition state.

Leis and Karelson[412] first reported QSPR models of rotational barriers using general theoretical molecular descriptors and the relatively simple application of such descriptors is quite attractive for this purpose. A three-parameter QSPR model (eq 11) was obtained with $R^2 = 0.982$ for the free energies of activation for the amide bond rotation for a set of 24 *N*,*N*-dialkylamides using the CODESSA program.

$$\Delta G^{\ddagger}_{\text{gas}} = (12.13 \pm 0.44)E_{\text{CH,max}} - (1.75 \pm 0.11)E_{\text{C,min}} + (2.19 \pm 0.32)\text{HACA}_2 + (146.19 \pm 9.96) \quad (11)$$

The three parameters used in eq 11 are maximum Coulombic interaction for a C−H bond ($E_{\text{CH,max}}$), minimum atomic state energy for a C-atom ($E_{\text{C,min}}$), and charged surface area of hydrogen acceptor atoms (HACA_2). It was shown that the gas-phase rotational barriers are primarily determined by the electronic properties of molecules in the rotational ground state. The proposed model is suggested as an alternative to the commonly used potential energy surface (PES) calculations for the quantitative prediction of amide bond rotation barriers.

5. Complex Physical Properties Involving Interactions between Different Molecules

5.1. Solubilities

Solubility is defined as the amount of solute dissolved in a saturated solution under equilibrium conditions. It is an important molecular property, which plays a large role in the behavior of compounds and is of interest in diverse areas of pharmaceutical, material, physical, and environmental research. In particular, in the design of drugs, it is essential to consider aqueous solubility, which strongly influences pharmacokinetic properties such as absorption, distribution, metabolism, and excretion. Also, knowledge of solubilities is required for the prediction of the environmental fate of pollutants, soil adsorption coefficients, bioconcentration factors for nonionic pesticides, and the suitability of gaseous anesthetics, blood substitutes, and oxygen carriers.[413,414]

Significant effort has been invested in the prediction of the solubility of small organic compounds and environmentally important chemicals. There is a large amount of experimental solubility data available for small organic compounds, but only limited data are available for drugs and druglike compounds. The experimental conditions of the measurements, such as pH and temperature, affect the solubility of a compound. Thus, differences in experimental conditions and protocols lead to variations between laboratories in the measurement of solubility. The accuracy of the QSPR models based on available experimental solubility data is limited by the accuracy of the experimental measurements.

5.1.1. Solubility of Liquids and Solids

Yalkowsky and Banerjee have summarized the various methods used to develop solubility models.[196] The methods may be classified into three categories: (i) correlations with experimentally determined physicochemical properties such as partition coefficients, chromatographic retention time, melting point, boiling point, molar volume (derived from liquid density), or parachor (derived from density and surface tension); (ii) group contribution models, which are based on compilations of relevant structural features of the molecules; (iii) correlations with the parameters calculated only from molecular structure, such as molecular volume, and topological indices. Yalkowsky and Valvani[415,416] proposed the general solubility equation (GSE), which uses only two parameters, the octanol−water partition coefficient and melting point. Jain and Yalkowsky[417] revised the GSE based on complete miscibility as $X_0 = 0.50$. The revised GSE has been applied on diverse data sets and was found to provide a more accurate estimation of the aqueous solubility of the same set of data than the original GSE.[198,418] Meylan et al.[197] expanded this method to include molecular weight (MW). The approaches based on experimental properties are only suitable for compounds for which the measured values are available, and they are not applicable for compounds not yet synthesized or isolated. The group contribution method requires numerous parameters (up to 200) to achieve a good predictive model.[419−421] A model, developed for the prediction of solubility based on the fragmentation method, included the experimental melting point as a term to account for the entropy of fusion of solids.[422] Ruelle and Kesselring applied the mobile order thermodynamics method to compounds with no hydrogen bond donor capacity.[423] The group contribution method is unsuitable for the prediction of the solubility of compounds unless neighboring groups and conformation are taken into consideration. Hence, the QSPR models developed using calculated molecular descriptors are more reliable for prediction of solubility, and numerous linear and nonlinear models have been developed. Several reviews have been published about the prediction of solubility based on QSPR equations by using MLR, PLS, and ANN approaches.[316,424,425] We have listed in Table 5 some of the QSPR models developed for the prediction of solubility using calculated molecular descriptors.

Katritzky et al.[451] reported a QSPR prediction of free energies of solvation of single solutes in a series of solvents and specified solutes in ranges of solvents.[452] They developed 69 QSPR equations for various solvents in a series of solutes based on the molecular descriptors calculated from structure using CODESSA PRO.[451] The models showed the R^2 and s ranging from 0.837 and 0.32 for 2-pyrrolidone to 0.998 and 0.14 for di-*n*-propyl ether, respectively.[451] Subsequently, the authors correlated the free energies of solvation of 80 organic solutes in a range of 15 to 82 solvents with molecular descriptors calculated by CODESSA PRO.[452] In another study, the same authors reported the intrinsic characteristics of the solute−solvent interactions based on a solubility database of 4540 compounds.[453] These studies were complemented by developing QSPR models describing the solubility of PAHs and fullerene (C_{60}) in two different condensed media: 1-octanol and *n*-heptane.[454] Statistically good QSPR models were obtained by using forward selection techniques from a large group of theoretical molecular descriptors.

Recently, the structural similarity method was applied by Schüürmann and co-workers[455] to the water solubility of a data set of 1876 organic compounds. The similarity analysis was carried out on atom-centered fragments (ACFs) in accord with a k nearest neighbor procedure in 2D-structural space. In another recent study, the implementation of a data visualization technique assisted in the extraction of meaningful information from a large scale solubility database, which established that C log P and the molecular weight were critical factors in determining aqueous solubility.[456]

Water solubility of PAHs was modeled by Lu et al.[457] using quantum chemical descriptors computed at the B3LYP/6-31G(d) level, and PLS. Two optimized models with high correlation coefficients ($R^2 = 0.966$ and 0.970) were obtained for estimating the logarithmic mass and molar concentration of water solubility, respectively. The PLS analysis indicated that PAHs with larger electronic spatial extent and lower total energy values tend to be less soluble.

Recently, aqueous solubility has been modeled using the scores of extended connectivity fingerprint as molecular descriptors on a data set of 1302 compounds;[458] solubility of fullerene C_{60} in organic solvents has been modeled using multiplicative SMILES-based optimal descriptors;[459] and solubility of 29 anthraquinone, anthrone, and xanthone derivatives in supercritical carbon dioxide (SCF-CO_2) was modeled using structure based molecular descriptors via linear and nonlinear methods.[460]

Solubility is a key physicochemical factor in drug development. Accordingly, reliable models for prediction of druglike compounds are urgently needed, and specifically oriented studies have emerged. Duchowicz et al.[461] have developed a generally applicable linear QSPR based on 147 druglike compounds containing three molecular descriptors. Kim et al.[462] have correlated the water solubility of poorly soluble drugs, such as ursodeoxycholic acid, diphenyl hydrantoin, and dimethyl biphenyldicarboxylate. Three data sets of 50 compounds were extracted from the literature data according to their structural similarity with each drug. Fast and predictive QSPR models ($R^2 > 0.90$) were developed and validated ($R^2 > 0.85$). Huuskonen et al.[463] extracted a training set of 191 druglike compounds from the AQUASOL database to correlate aqueous solubility by a model of five parameters (C log P, molecular weight, indicator variable for aliphatic amine groups, number of rotatable bonds, and number of aromatic rings) with statistics of $R^2 = 0.87$ and $s = 0.51$. The model was applied to a test set of 174 druglike compounds with $R^2 = 0.80$ and $s = 0.68$. The results of this study suggest that increasing molecular size, rigidity, and lipophilicity decrease solubility whereas increasing conformational flexibility and the presence of a nonconjugated amine group increase the solubility of druglike compounds. Du-Cuny et al.[464] aimed at modeling the aqueous solubility of druglike compounds in congeneric series. Lipophilicity (C log P) combined with structural fragment information, fragmental based correction factors, and congeneric series indices were used as descriptors for a PCA followed by multivariate PLS regression. The resulting general model ($R^2 = 0.84$ and rms $= 0.51$) was based on an in-house data set of 1515 druglike compounds, and solubility of the test set of 958 compounds was predicted with a high degree of accuracy, $R^2 = 0.81$ and $s = 0.42$. In the course of model development, rules were derived which can be used by medicinal chemists or interested scientists as a rough guideline on the contribution of structural fragments to solubility.

5.1.2. Aqueous Solubility of Gases and Vapors

Due to the technical difficulties with accurate analytical determination of the solubility of gases and vapors, computational methods for their prediction are of great practical importance. The solubility of gases and vapors is denoted as L and is known as the Ostwald solubility coefficient. It is defined as the ratio of the concentration of a compound in a solution and in the gas phase at equilibrium. Another commonly used equilibrium parameter is the Henry's law constant (H), which is essentially an air−water partition coefficient (L_w^{-1}). Water−air partition coefficients can be estimated from the vapor pressure (VP) and aqueous solubility (S_w) of a compound.[465−467] Hine and Mookerjee reported the first empirically based group and bond contribution schemes[468] and estimated the solubilities of 292 compounds with a standard error of 0.12 log units based on 69 empirical

Table 5. List of Some QSPR Models of Aqueous Solubility (log *S*)

no.	compd	*N*	descriptors	approach	model statistics	Ref
1	alky-halo-substituted aromatics	38	$^1\chi^V$ and ϕ	MLR	$R^2 = 0.922$, $s = 0.200$	Nirmalakhandan and Speece[426]
	alcohols	50	$^1\chi$, $^1\chi^V$, and $^3\chi_P^V$	MLR	$R^2 = 0.961$, $s = 0.110$	
	alky-halo-substituted alkanes/alkenes/ aromatics and alcohols	145	$^0\chi$, $^0\chi^V$, and ϕ	MLR	$R^2 = 0.926$, $s = 0.318$	
2	miscellaneous compounds such as PCBs, PNAs, PCDDs, phenols, etc.	470	$^0\chi$, $^0\chi^V$, and ϕ	MLR	$R^2 = 0.980$, $s = 0.332$	Nirmalakhandan and Speece[427]
3	miscellaneous compounds	123	NO, NC, WTPT1, WRPT2, QSUM, SAAA1, SAAA2, FNSA3, GEOH	MLR	$R^2 = 0.998$, $s = 0.227$	Sutter and Jurs[428]
	miscellaneous compounds	123	NO, NC, WTPT1, WRPT2, QSUM, SAAA1, SAAA2, FNSA3, GEOH	CNN(9:3:1)	$s = 0.217$ (test)	
					$s = 0.282$ (cv, $n = 11$)	
4	miscellaneous compounds	295	SHDW2, SHDW5, MOLC3, PPSA1, DPSA3, WNSA1, SAAA3, CHAA2, EHBB	MLR	$R^2 = 0.931$, $s = 0.638$	Mitchell and Jurs[429]
		265 (trn)	SHDW3, GRAV3, ALLP3, WTPT4, 2SP3, QNEG, PPSA1, FPSA3, WPSA3	CNN(9:6:1)	$s = 0.394$ (test)	
					$s = 0.358$ (cv)	
5	hydrocarbons and halogenated hydrocarbons	241	MV, ^{0}BIC, PNSA	MLR	$R^2 = 0.959$	Hubiers and Katritzky[430]
6	diverse compounds	411	$Q_{\min}$, N_{el}, FHDSA$_2$, ABO(N), ^{0}SIC, RNCS	MLR	$R^2 = 0.879$, $s = 0.573$	Katritzky et al.[312]
7	drugs	160	topological (9), atom type electrotopological indices (24)	ANN(23-5-1)	$R^2 = 0.90$, $s = 0.46$	Huuskonen et al.[431]
		50			$R^2 = 0.86$, $s = 0.53$	
8	drug/organic	884	topological (6), *E*-state indices (24)	MLR	$R^2 = 0.89$, $s = 0.67$	Huuskonen[432]
		413		MLR	$R^2 = 0.88$, $s = 0.71$	
		884		ANN(30-12-1)	$R^2 = 0.94$, $s = 0.47$	
		413		ANN(30-12-1)	$R^2 = 0.92$, $s = 0.60$	
9	organic compounds	150	molecular descriptors, Monte Carlo simulations (5)	MLR	$R^2 = 0.88$, $s = 0.72$	Jorgensen and Duffy[433]
10	diverse compounds	500	descriptors calcd using PM3 and topological	FUZZY ARTMAP	$s = 0.14$ (validn) ($s =$ 0.28 validn for 11-13-1 BNN)	Yaffe[434]
11	drug/organic	1038 (lit.)	topological, geometrical, charge	Bayesian ANN, (ARD-automatic relevance procedure)	$R^2 = 0.95$, rmse = 0.50	Bruneau[435]
		673 (test)			rms = 0.84	
		522 (Astra-Zeneca)			$R^2 = 0.64$, rms = 0.67	
		261 (test)			rms = 0.78	
		1560 (all)			$R^2 = 0.94$, rms = 0.53	
		934 (test)			rms = 0.81	
12	drug/organic	1168	descriptors based on fragment atom scheme (118)	group contribution method	$R^2 = 0.95$, $s = 0.50$	Klopman and Zhu[421]
13	drug/organics	879	*E*-state indices (31)	MLR	$R^2 = 0.86$, $s = 0.75$ (trn)	Tetko[436]
		412		MLR	$R^2 = 0.85$, $s = 0.81$ (test)	
		879		ANN	$R^2 = 0.92$, $s = 0.56$ (trn)	
				ANN	$R^2 = 0.89$, $s = 0.68$ (test)	
14	drug/organic	1033	1D, 2D descriptors	ANN(7:2:1)	$R^2 = 0.86$, $s = 0.70$	Liu and So[437]
		258			$R^2 = 0.86$, $s = 0.71$	
15	chlorinated hydrocarbons	50	shadow XY, WNSA-3	MLR	$R^2 = 0.965$, $s = 0.45$	Delgado[438]
16	druglike compounds	930	2D and 3D descriptors (24)	MLR	$R^2 = 0.92$, rmse = 0.53	Gao et al.[439]
	druglike compounds	249		MLR	$R^2 = 0.91$, rmse = 0.49	
17	drug/organic	150 (trn)	quantum-chemical (3)	MLR	$R^2 = 0.90$, $s = 0.66$	Klamt et al.[440]
	pesticides	1078 (test)		MLR	$s = 0.61$	
18	drugs/organic	3042	topological, hydrogen bonding, lipophilic, 1D- and 2D-descriptors (63)	ANN	$R^2 = 0.91$, $s = 0.84$ (trn)	Engkvost and Wrede[441]
		309			$R^2 = 0.89$, $s = 0.87$ (test)	

Table 5. Continued

no.	compd	*N*	descriptors	approach	model statistics	Ref
		307			$R^2 = 0.86$, $s = 0.80$ (test)	
	drugs	80 (test)	$^0\chi^v$, $^3\chi_{ac}^v$, $^3\chi_c^v$	MLR	$R^2 = 0.91$, $s = 0.769$	
19	small organic molecules, drugs, druglike species	775 (trn)	*A* log P98, HBD*HBA, HBD, Rotlbonds, Wiener, Zagreb' S_aaaC, S_sOH	GA/PLS	$R^2 = 0.84$, rms = 0.87	Cheng and Merz[442]
		1665 (test)			$s = 1.01$	
20	diverse organic compounds	787	α, ΣCa, ΣCd, *n*(Cycl), *I*(alk), *I*(CX3), *I*(RCOOH), I (Hbintra) log 1/fui	RA	$R^2 = 0.935$, $s = 0.467$	Schaper et al.[443]
		569	α, ΣCa, ΣCd, log 1/fui	RA	$R^2 = 0.89$, $s = 0.49$	
21	aliphatics	50	weighted path numbers based on van der Waals volumes	MLR	$R^2 = 0.94$, $s = 0.38$	Nohair and Zakarya[444]
		50		ANN(4-4-1)	$R^2 = 0.98$, $s = 0.11$	
22	organic compounds	741 (trn)	18 descriptors: topological, HBD, HBA, indicator	MLR	$R^2 = 0.84$, $s = 0.78$	Yan and Gasteiger[445]
		552 (test)		MLR	$R^2 = 0.89$, $s = 0.68$	
		741 (trn)		BPANN(18-10-1)	$R^2 = 0.92$, $s = 0.51$	
		552 (test)		BPANN(18-10-1)	$R^2 = 0.94$, $s = 0.52$	
23	drug/organic	797 (trn)	RDF code 3D descriptors (32)	MLR	$R^2 = 0.79$, $s = 0.93$	Yan and Gasteiger[446]
		496 (test)		MLR	$R^2 = 0.82$, $s = 0.79$	
		797 (trn)		BPNN	$R^2 = 0.93$, $s = 0.50$	
		496 (test)		BPNN	$R^2 = 0.92$, $s = 0.59$	
24	alcohols log(1/*S*)	63	local graph invariants	least square regression	$R^2 = 0.986$, $s = 0.12$	Duchowicz et al.[447]
25	aromatic compounds	3343 (trn)	topological (47−67)	ANN, PLS-GA, MLR-GA	($R^2 = 0.88$, mae = 0.51) ($R^2 = 0.79$, mae = 0.71) ($R^2 = 0.77$, mae = 0.75)	Votano et al.[448]
	aromatic compounds	772 (test)			($R^2 = 0.77$, mae = 0.62) ($R^2 = 0.72$, mae = 0.78) ($R^2 = 0.72$, mae = 0.76)	
	nonaromatics	1674 (trn)	topological (35−52)	ANN, PLS-GA, MLR-GA	($R^2 = 0.88$, mae = 0.44) ($R^2 = 0.79$, mae = 0.61) ($R^2 = 0.76$, mae = 0.63)	
		166 (test)			($R^2 = 0.84$, mae = 0.56) ($R^2 = 0.78$, mae = 0.68) ($R^2 = 0.76$, mae = 0.66)	
26	miscellaneous compounds	930 (trn)	(topological, hydrophobicity, partial charge, polarizability) (22 MOE, 65 ISIS keys)	linear PLS	(rmse = 0.468) ($R^2 = 0.911$, rmse = 0.475)	Catana et al.[449]
		177 (test)				
		800 (trn)	41 descriptors	MLP (ANN)	($R^2 = 0.897$, rmse = 0.584) ($R^2 = 0.846$, rmse = 0.608)	
		177 (test)				
		800 (trn)	60 descriptors	linear NN	($R^2 = 0.93$, rmse = 0.483) ($R^2 = 0.903$, rmse = 0.501)	
		177 (test)				
27	PCDD/PCDFs and phthalate ester	35	topological (CRI), E_{HOMO}, E_{LUMO}, μ	MLR	($R^2 = 0.97$, $s = 0.347$)	Sacan et al.[450]

group contribution factors. Their bond contribution scheme reproduced the solubilities of 263 solutes with a standard error of 0.42 log units using 34 bond contributions. Another scheme based on 28 group contributions developed by Cabani et al.[469] was implemented in the correlation of 209 log L_w values of diverse compounds and a standard error of 0.09 log units. However, due to the large number of fitted parameters involved in these schemes, neither the group contribution nor the bond contribution method conveys much understanding of the physical nature of the relationship between molecular structure and intermolecular interactions, and hence, the solubility of gases in water. Moreover, the solubilities of compounds containing structural functionality not included in the training set were unsuitable for prediction.

Russell, Dixon, and Jurs correlated the logarithms of Henry's law constant, log *H*, of a small data set of 63 diverse gases in water, using five theoretically calculated descriptors.[470] Their MLR model had a R^2 of 0.956 and s = 0.375 log units. Based on the model descriptors, the authors suggested that the factors influencing the solubility of gases in water were related to the solute bulk, lipophilicity, and polarizability.

Abraham et al. correlated the solubility of 408 diverse gases in water with five linear solvation energy relationship (LSER) descriptors,[466] including the excess molar refraction (calculated from the experimental molar refraction), the experimentally determined dipolarity/polarizability π_2^H, the effective hydrogen-bond acidity $\sum\alpha_2^H$ and basicity $\sum\beta_2^H$, and the McGowan characteristic volume V_x (calculated from some tabulated atomic increments). The model had a correlation coefficient of 0.998, a standard deviation of 0.151 log unit, and an *F*-value of 16810. Although four of these descriptors were determined experimentally, this correlation equation can be interpreted term-by-term using well-established chemical principles, which has been a motivation to develop computational methods for obtaining the LSER descriptors in order to make a priori predictions. Absolv[471] is an excellent example of such a tool that enables calculation of solvation associated properties from Abraham-type equations and predicts calculation parameters necessary for those calculations.

The partitioning of two sets of organic gases and vapors between water and air (L_w) has been studied using the CODESSA program.[472] For the first set of 95 alkanes, cycloalkanes, alkylarenes, and alkynes, excellent predictions were obtained with a two-parameter correlation equation ($R^2 = 0.977$, $R_{cv}^2 = 0.975$, $s = 0.20$). The two descriptors involved—the gravitation index and the complementary information content—reflect the effective mass distribution and the degree of branching of the hydrocarbon molecule, and they adequately represent the effective dispersion and cavity formation effects for the solvation of nonpolar solutes in water. For the second set of 406 structurally diverse organic compounds (including structures containing N, O, S, and halogen atoms), a successful 5-parameter correlation equation ($R^2 = 0.941$, $R_{cv}^2 = 0.939$, $s = 0.53$) was also reported. The descriptors from this equation (which were completely different from those for the set of 95 nonpolar solutes) comprised the partial charge weighted normalized hydrogen-bonding donor surface area, counts of oxygen and nitrogen atoms, the HOMO−LUMO energy gap, the most negative partial charge weighted topological electronic index, and the number of rings. The descriptors account for the dispersion energy of polar solutes in solution, the electrostatic part of the solute−solvent interaction, and hydrogen-bonding interactions in liquids. In a related study, water−air partition coefficients were estimated from vapor pressure and aqueous solubility values predicted by the QSPR models derived from a set of 411 compounds.[312] The mean standard error of the predicted gas solubilities was found to be very similar to the standard error of the L_w, predicted using the equation derived directly from the experimental values of L_w[472]. The solubilities of 87 gases and vapors in methanol ($R^2 = 0.945$, R^2_{cv}) and 61 gases in ethanol ($R^2 = 0.969$, $R^2_{cv} = 0.964$) have also been reported.[473]

5.1.3. Activity Coefficients at Infinite Dilution

The infinite dilution activity constant, γ^∞, indicates how the solvent medium differs from the pure solute, measuring the interactions between solute and solvent in the absence of solute−solute interactions. The γ^∞ of aqueous solutions is important in environmental engineering as well as in industrial applications. Compared to solubility, the activity constants depend weakly on the temperature and to a lesser extent on the solute configuration. Several correlations between structural features and γ^∞ in aqueous solutions have been reported. Pierotti et al.[474] developed a scheme in which log γ^∞ is estimated from the contributions of individual interactions between the solute and solvent structural groups. The size of the structural groups was represented by the carbon number, and the nature of the groups was incorporated by the use of coefficients. Tsonopoulos and Prausnitz[475] correlated the activity coefficients of 147 aromatic solutes in dilute aqueous solutions with the number of carbon atoms and the types of groups present in the aromatic compound. The group contributions were found to be generally additive. Mackay and Shiu[476] correlated the hydrocarbon infinite dilution coefficient with the carbon number using a parabolic equation. Medir and Giralt[477] correlated ln γ^∞ for aliphatic and aromatic hydrocarbons using molecular descriptors that included the first-order molecular connectivity index, surface area, dipole moment, number of carbon atoms, total electronic energy, and acentric factor. Tochigi et al.[478] improved the ASOG (analytical solution of groups)—a group contribution method confirmed by the prediction of 442 γ^∞ values. Hansen et al.[479] revised the parameters of a group contribution method UNIFAC[480] based on binary group interaction parameters together with volume and surface area parameters capable of calculating γ^∞ in various solvents. Another method, based on the modified separation of cohesive energy density (MOSCED) model, for nonaqueous systems was developed by Thomas et al.[481] Using only pure component parameters, the method produced results comparable with the aforementioned approaches. According to Sherman et al.,[482] ASOG and UNIFAC cannot be extended to predict γ^∞ of aqueous solutions because of the strong nonidealities compared to other solvents due to the water's hydrogen-bonding capabilities and the small size of the molecule. Therefore, water exhibits large variation in γ^∞ and needs an individual approach. The authors evaluated data from different kinds of experiments and created a water database of 336 compounds at 298.15 K, and developed an LSER model that fitted the data to within an average absolute deviation of 0.294 ln unit.

Mitchell and Jurs[483] developed QSPR models for the ln γ^∞ of 321 organic compounds (in the range of −2.41 for dimethyl sulfoxide to 23.3 for 1-octadecanol) in aqueous solutions with predictive ability within the range of the experimental error of the measurements. The molecular structures were represented by calculated topological, geometric, and electronic descriptors at the PM3 and MNDO levels of theory. Genetic algorithm (GA) and simulated annealing (SA) routines were used to select subsets of 12 descriptors for the MLR and CNN models. The best model was obtained combining the GA and the CNN $s_{test} = 0.376$ and $s_{pred} = 0.434$ for the sets of 271 and 25 compounds, respectively.

Rani and Dutt[484] modeled γ^∞ values of 19 halocarbons in water and 18 organic compounds in five hydrofluoroparaffins as solvents over a temperature range of 291− 333 K with an ANN trained with 351 data points, with an average absolute deviation of 11.8% on the basis of γ^∞, compared to 94.3% obtained by MLR. The input variables included in the network were temperature, dipole moment, molar refraction, and critical pressure of the solute and solvent.

He and Zhong[485] conducted a QSPR study on ln γ^∞ for organic compounds in water at 298.15 K. Correlations based on 3...6 molecular connectivity indices were proposed for hydrocarbons ($n = 105$, $R^2 = 0.968$, $s = 0.478$), oxygen containing organic compounds ($n = 108$, $R^2 = 0.992$, $s = 0.339$), and halogenated hydrocarbons ($n = 70$, $R^2 = 0.834$, $s = 0.773$). Estrada et al.[486] used quantum-connectivity indices to model ln γ^∞ of the same data for hydrocarbons and oxygen containing compounds. Quantum-connectivity indices are defined by using a combination of topological invariants, such as interatomic connectivity, and quantum chemical information, such as atomic charges and bond orders. For 103 hydrocarbons, just two descriptors provided a correlation with $R^2 = 0.93$, while a single descriptor (the path bond-order-based bond connectivity index of order 1) described the 108 oxygen containing compounds with $R^2 = 0.98$.

Giralt et al.[487] used self-organizing maps for the extraction of relevant molecular features which were then used to identify 13 chemical classes and their characteristics within a data set of 325 organic compounds. A fuzzy-ARTMAP QSPR model with 11 topological and quantum-chemical descriptors was reported. Average absolute errors of 0.02 (0.36%) and 0.52 (6.64%) ln γ^∞ units were obtained for the training (280 compounds) and test sets (45 compounds), respectively.

Xu et al.[488] performed geometrical optimization and electrostatic potential calculations for a series of halogenated hydrocarbons at the HF/Gen-6d level. A number of electrostatic potentials and statistically based structural descriptors derived from these electrostatic potentials were obtained. MLR analysis and ANN were employed simultaneously. The results showed that the parameters derived from electrostatic potentials, σ2tot, V(s), and $\sum V_s(+)$, together with the molecular volume (V_{mc}), could be used to express the QSPR of γ^∞ of halogenated hydrocarbons in water. Validation of the model using an external test set demonstrated that the model obtained by using the BFGS quasi-Newton NN method had much better predictive ability than that from MLR.

In recent years, room-temperature ionic liquids (ILs) that are organic salts composed entirely of ions have gained great importance as media for reactions and extraction due to their unique physical properties. Particularly appealing is their low vapor pressure, which makes them essentially nonvolatile. Information on how solutes interact with these solvents is crucial in assessing their usefulness. Eike et al.[489] successfully modeled infinite dilution activity coefficients (ln γ_i^∞ at 298 K) for 38 solutes in three ionic liquids—1-butyl-4-methylpyridinium tetrafluoroborate ([bmpyr][BF4]), 1-methyl-3-ethylimidazolium bis(trifluoromethylsulfonyl)amide ([emim][Tf$_2$N]), and 1,2-dimethyl-3-ethylimidazolium bis(trifluoromethylsulfonyl)amide ([emmim][Tf$_2$N])—using QSPR methodology. Constant-temperature and temperature-dependent correlations were created with R^2 ranging from 0.90 to 0.99. In all three ionic liquids, log K_{OW} was the most significant property followed by solute aromaticity and charge distribution on solute−solvent interactions. In temperature-dependent correlations the hydrogen bonding (Hbonds) became the most influential interaction.

A similar set of 38 organic compounds with infinite dilution activity coefficients (ln γ_i^∞ at 313 and 343 K) in ionic liquids such as 1-methyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide, 1,2-dimethyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide, and 4-methyl-N-butylpyridinium tetrafluoroborate were studied by Tämm and Burk[490] using the CODESSA PRO program.[74] Three theoretical molecular descriptors correlated satisfactorily with the activity coefficients with R^2 ranging from 0.943 to 0.966. The complementary information content, the fractional partial negative surface area, and the count of hydrogen donor sites descriptors could be related to the nature of the dilution process in ILs. More recently, Wang et al.[491] showed that infinite dilution activity coefficients of molecular solutes in ILs can be represented by the tradional UNIFAC[480] model using a novel group segmentation method. Molecular solutes including alkanes, alkenes, aromatics, alcohols, ketones, and water in six ILs were well correlated within 9% of rmsd. Katritzky et al.[492] correlated the solubilities of 90 organic solutes measured as Ostwald solubility coefficients (log L) in eight ILs with $R^2 > 0.91$. The solute interactions with the ILs were most often described by descriptors reflecting the hydrogen donor/acceptor ability of the solutes and those reflecting size and shape effects. For more general screening for suitable solutes, a general QSPR representation of log L for all eight ILs as a group was created using four descriptors occurring most frequently in the previous models.

5.2. Partition Coefficients

5.2.1. Octanol−Water Partition Coefficient

The n-octanol/water partition coefficient is the ratio of the concentration of a chemical in n-octanol to that in water in the two-phase system at equilibrium. The logarithm of this partition coefficient, log P, is the parameter that determines the lipophilicity of a molecule, and it has found wide application in the prediction of biological activities, ADME, and toxicological end points. The partition coefficient has also been used in calculating numerous physical properties such as membrane transport and water solubility. Thus, a reliable computational model for the estimation of log P is of immense importance in drug discovery and design, since it is important to know the lipophilic properties of a compound before it is synthesized.

Several log P calculation methods from chemical structure have been developed. The methods fall into two classes: (i) the group contribution approach and (ii) the whole molecule approach. The group contribution approach includes “atom-based” and “fragment-based” methods, in which a molecule is divided into atoms or fragments and the log P values are calculated by summation of the contributions from the fragments or atoms present in the molecule. The whole molecule approach is based on molecular properties such as electrostatic potential, molecular surface area, and molecular volume.

Hansch and Fujita developed the first model for calculating log P in 1964[493] based on the π-system. The limitations of the π-system led Rekker[494] to develop the first “fragment-based” methods, and later Broto[495] reported the first “atom based” contribution methods for calculating the log P values. Since then several other calculation methods have been proposed. Mannhold et al. listed the addresses of the programs for the calculation of log P,[496] and they reported a comparative study of log P calculation methods.[497−499] In 2000, Leo summarized the various uses of log P as a descriptor in the prediction of biological activities of 3500 QSARs, assessment of the environmental hazard of organic chemicals, and biodegradation of chemicals.[500a] Katritzky et al.[41] and Holder et al.[500b] recently reported QSPR models involved in the prediction of log P. Various methods used for the prediction of the octanol−water partition coefficient and the advantages and limitations of the approaches were described by Livingstone.[501] A minireview has appeared describing recent methodologies for the calculation of log P and their use in the prediction of membrane transport of drugs,[502] which summarizes the methods of calculation of log P, and QSPR models developed using molecular descriptors, during the past decade.

The first computerized log P calculation model was developed by Leo and Hansch in 1982 based on the group contribution approach, and it was implemented in a computer program as ClogP.[503] In the latest, revised version of ClogP 4.0,[504] a new algorithm FRAGLAC, which is based on a set of around 600 dependably measured descriptors, was used to calculate the contribution values of the fragments having only aliphatic or aromatic bonds. The average deviation of the model is 0.31 log unit.[503] KlogP[505] uses an artificial

Table 6. List of Some Computer Programs Available for Calculation of log *P*

program	calculation method[a]	software released		ref
ClogP	fragmental-HL	BioLoom		www.biobyte.com
PCModels	fragmental-HL	ClogP	4.0	www.daylight.com
PrologP	fragmental-R	PrologP		www.compudrug.com
SYBYL	fragmental-R	ClogP		www.tripos.com
ACD/LogP	fragmental-A/F	ACD/LogP		www.acdlabs.com
LOGKOW	fragmental-A/F	KowWin		www.srcinc.com
KlogP	fragmental-C	KlogP		Klopman et al.[530]
ALOGPS	electrotopological indices	ALOGPS	2.1	www.vcclab.org
VLOGP	LFER and topological indices			Gombar and Enslein[518]
SLIPPER	polarizability and hydrogen bond acceptor strength	SLIPPER-2001		www.timtec.net

[a] The fragmental methods refer to the systems of Hansch and Leo (HL), Rekker (R), computer identified (C), and atom/fragment contributions (A/F).

intelligence system, Computer Automated Structure Evaluation (CASE) methodology for the development of the log *P* model. The log *P* model was derived by using 94 atomic and fragment based group descriptors based on a database of 1663 compounds and showed an R^2 value of 0.928 and standard error of 0.38 log units. The AlogP[506] model uses a pure atom-based contribution method for calculations of log *P* values based on a data set of 9920 organic compounds having R^2 of 0.918 and s of 0.68 log unit. In 2000, Viswanadhan et al.[507] developed a novel program, HlogP, which uses new fragment types, molecular holograms, as descriptors for building a PLS regression model. HlogP has been shown to have better predictability for druglike molecules compared to ClogP and AlogP models.[477] Manhold's $\sum f$-system[498,508] represents a log *P* model based on 169 hydrophobic fragmental descriptors and 13 correction factors. In the revised LOGKOW model, a new methodology, the experimental value adjusted (EVA) approach, was used to calculate the fragment contributions by comparison of closely related analogues.[509] This program is suitable for the calculation of log *P* values of unknown structures from a target compound on the learning set. Meylan et al. recently revised the LOGKOW model, which was derived from 150 atom-fragment and 250 correction factors.[509] They reported $R^2 = 0.943$ for calculating the log *P* values of the learning set of 10589 compounds. ACD/LogP v7.0 (Advanced Chemistry Development Inc., Toronto Ont., Canada, 2003) includes 500 basic fragment descriptors (*f*) and over 2000 correction factors (*F*) using fragmentation rules based on the definition of an isolated carbon (IC).[510] An XLOGP model was developed by Wang et al. in 1997 based on a pure atomic contribution approach.[511] The recently released version of the model XLOGP v 2.0[512] ($s = 0.35$ and $R^2 = 0.946$) includes 90 atom type descriptors and 10 additional correction factors in describing the log *P* for a database of 1853 organic compounds.

Property-Based Approaches. The partition coefficient log *P* is proportional to the molar Gibbs free energy of transfer between octanol and water, and hence, it should be dependent on the molecular properties that contribute to this free energy. In this approach various statistical methodologies (PLS, ANNs, etc) have been applied. An early attempt was made in 1969 by Rogers to correlate log *P* with molecular properties based on MO theory.[513] Klopman et al.[514] used quantum mechanical calculations based on the MINDO program and Huckel type calculations for the estimation of log *P*. The BlogP of Bodor and Huang,[515] QlogP of Bodor and Buchwald,[516,517] VLOGP of Gombar and Enslein,[518,519] MlogP model of Moriguchi et al.,[520] AUTOLOGP of Devillers,[521,522] HYBOT and SLIPPER-2001 of Raevsky and co-workers,[523,524] ScilogP and ALOGPS of Tetko and co-workers,[525,526] CLIP_logP of Gaillard et al.,[527] and HINT model of Kellogg[528,529] were reviewed by Klopman and Zhu[502] regarding the basic features of the programs. In Table 6 we have listed some recent computer programs available for calculation of log *P*.

Klopman et al.[530] reported a revised group contribution model for the calculation of log *P*. Their model includes 153 basic parameters, 41 extended parameters, and 14 molecular surface property descriptors based on a training database of 8320 chemicals. The model achieved significant improvement after modifying the traditional group contribution equation by using a 3D steric hindrance modulator. The predictability of the model was assessed by calculating the log *P* values of a test set of 1667 organic chemicals and 137 druglike chemicals. A structural analogue approach has been applied by Sedykh and Klopman[531] which includes 102 basic parameters and 36 correction factors whose coefficients are optimized on the basis of the sets of similarity pairs produced from the training set data of 8320 chemicals. The authors reported a comparison of the present similarity model with their previous model and other known models (ClogP, KowWin, AUTOQSAR/MLR, AUTOQSAR/PLS, AUTOQSAR/NN). Eros et al.[532] reviewed the reliability of log *P* predictions based on calculated molecular descriptors. They also assessed the reliability of their logP program called AutoQSAR (Auto-MLR, Auto-PLS, Auto-NN). A ALOGPS program was developed with 12908 molecules from the PHYSPROP database using 75 E-state indices. Sixty-four ANNs were trained using 50% of molecules selected by chance from the whole set. The log *P* prediction accuracy had rms = 0.35 and a standard mean error of 0.26 log unit.[525,526]

An ANN was applied by using atomic fragment descriptors included in the Pallas PrologP program (www.compudrug.com) based on Ghose-Crippen fragmentation[533] and additional correction terms to modify atomic contributions for the estimation of log *P*. The correlation statistics for the training set (8729 compounds) and the test sets 1 and 2 (2000 compounds) were ($R^2 = 0.94$, $s = 0.01$), ($R^2 = 0.91$, $s = 0.02$), and ($R^2 = 0.92$, $s = 0.02$), respectively.[534] Lombardo et al.[535] devised a RP-HPLC method, for the determination of log P_{oct} values of neutral drugs, which showed high accuracy for a set of 36 drug molecules. Linear free energy relationship (LFER) analysis, based on solvation parameters, showed that the method encodes the same information as obtained by a shake-flask log P_{oct} determination. Hawkins et al.[536] reported prediction of partition coefficients based on the geometry-dependent atomic surface tensions. No et al. revised the previously reported solvation free energy density (SFED)[537] model based on overestimation in the hydration free energies of the molecules having highly

Table 7. List of QSPR Models Developed for the Prediction of log *P*

compounds	*N*	descriptors	approach	model statistics	ref
polychlorinated biphenyls	133	electrophilicity index, E_{LUMO}, N_{Cl}	MLR	$R^2 = 0.914$, $s = 0.225$	Padmanabhan et al.[540]
organic compounds	69	correlation weights of local graph invariants	MLR	$R^2 = 0.995$, $s = 0.096$	Basak and Mills[541]
diverse compounds	136	topochemical	RR	$R^2_{\mathrm{CV}} = 0.570$, $s = 0.225$	Shamsipur et al.[542]
diverse organic compounds	379	distance–connectivity based TIs	PC-ANN	$R^2 = 0.963$, $s = 0.281$	Al-Fahemi et al.[543]
diverse compounds	76	molecular descriptors based on momentum-space (p-space) electron density	MLR	$\mathrm{R}^2 = 0.964$, $s = 0.256$	Lamarche et al.[544]
druglike compounds	79	descriptors calculated based on polarity, hydrogen bond acidity, basicity	MLR	$R^2 = 0.886$, rms = 0.43	Oliferenko et al.[545]
diverse organic compounds	90	basicity, acidity, polarizability, integral polarity	MLR	$R^2 = 0.970$, $s = 0.233$	Gao and Cao[546]
polychlorinated biphenyls (PCBs)	157	HOMO–LUMO interaction	MLR	$R^2 = 0.9235$, $s = 0.224$	Zou et al.[547]
disubstituted benzene	103	sum of the surface minima values of the electrostatic potential, molecular volume, PSA	MLR (6)	$R^2 = 0.925$, $s = 0.247$	Wegner et al.[548]
diverse organic compounds	1853	descriptors based on differential Shannon entropy (DSE)	GA-ANN	$R^2 = 0.846$, $s = 0.44$	Peruzzo et al.[549]
diverse industrial chemicals	76	correlation of local invariants of hydrogen filled graphs	MLR	$R^2 = 0.887$, $s = 0.51$	Padmanabhan et al.[540]

polarizable atoms.[538] The authors calculated log *P* from the free energy differences and from the log *P* density (LPD) based on the SFED model, which showed absolute mean errors of 0.34 and 0.32, respectively. A recent study by Machatha and Yalkowsky[539] showed ClogP to be a more accurate predictor of log *P* as compared to the predicted log *P* with the ACD/logP and KowWin programs. Their analysis showed the average absolute error (AAE) for KowWin 0.358, ACD/logP 0.386, and ClogP 0.329, respectively, for a set of 108 diverse compounds.

Some recent QSAR/QSPR models of log *P* developed using molecular descriptors are listed in Table 7.

To a lesser extent, partitioning coefficients in other aqueous systems have been studied, e.g the work of Leahy and coworkers, who employed Abraham LSER (linear solvation energy relationship) descriptors to model the respective log *P* values in four solvent–water systems: octanol (amphiprotic), alkane (inert), chloroform (proton donor), and propylene glycol dipelargonate (PGDP; proton acceptor)[550]—the "critical quartet.[551] An almost complete data matrix of 82 fragment values (*f*-values) for all four solvents resulted. It was suggested that the ”critical quartet” could be used as a model solvent system for membrane binding and transport characteristics that might have special relevance to biological selectivity.

In addition to the P_{oct} (octanol–water), three other partition coefficients of liquids and solids in different solvent systems, such as P_{16} (water–hexadecane), P_{alk} (water–alkane), and P_{cyc} (water–cyclohexane), were estimated by Khadikar et al.[552] using the PI (Padmakar–Ivan) index[553] and the widely used Wiener index (W). For *n*-alkanes the advantages of the PI index in terms of higher correlation coefficients with the studied properties compared to the W index were clearly shown.

Oliferenko et al.[545] have applied their newly established quantitative scales of hydrogen bond (HB) basicity and acidity to seven equilibrium partitioning data sets—octanol–water (see Table 7), hexadecane–water, and chloroform–water—as well as gas–water, gas–octanol, gas–hexadecane, and gas–chloroform partition coefficients. The hydrogen bond descriptors when supplemented by a cavity-forming term and a dipolarity term showed high performance in correlations of the partition coefficients of aliphatic compounds. These new HB descriptors can be used in studying hydrogen bonding and fluid phase equilibria as well as scoring functions in ligand docking and descriptors in ADME evaluations.

5.2.2. Aqueous Biphasic Partitioning

Aqueous biphasic systems (ABS) are formed by the addition of two (or more) water-soluble polymers or a polymer and salt to an aqueous solution above a certain critical concentration or temperature. ABSs are unique because each of the two nonmiscible phases is over 80% water on a molal basis and each possesses different solvent properties.[554,555] The distribution coefficient (*D*) is defined as the total concentration of a solute in the upper polyethylene glycol (PEG)-rich phase (C_{PEG}) divided by the concentration in the lower salt-rich phase (C_{salt}). Usually the logarithmic function (log *D*) is used for describing the distribution coefficients.

Due to its highly aqueous and hence mild nature, which is consonant with the maintenance of macromolecular structure, ABS has been employed for the separation of biological macromolecules for over 40 years[555,556] and for the evaluation of the relative hydrophobicity of organic compounds or biopolymers such as peptides and proteins.[557−560] Proteins and nucleic acids are prone to denaturation in alcohols, whereas suitable polymer/water compositions can more closely resemble the native living cell conditions. Eiteman and Gainer have studied the partitioning of amino acids, small peptides, and alcohols in ABS.[561,562] Gulyaeva et al.[563] studied the partitioning of 153 dihexapeptides in an aqueous dextran-PEG biphasic system and reported that the peptide bitterness threshold is quantitatively related to the relative hydrophobicity and lipophilicity (log D) of peptides, which are responsible for their biological activities.

However, due to variable composition, it seems that the QSPR analysis of ABS is more complex than that of the octanol−water system, which has a constant phase composition. Experimental investigations of the partitioning behavior of organic molecules in PEG/salt ABS have been reported by Rogers et al.,[564−566] who applied a linear solvation energy relationship (LSER) based on Abraham's generalized solvation equation, which enabled a direct comparison between the solvent properties of PEG/salt ABS and those of traditional solvent/water systems.[567,568] LFER studies concluded that the principal determinants which govern the partitioning in ABS arise from the size, basicity, and aromaticity or halogenicity of the solute. The first theoretical molecular model for the prediction of partitioning in ABS using descriptors solely calculated from structure was that of Katritzky et al.,[569] who employed structural descriptors included in CODESSA PRO for the prediction of log D of organic solutes in a PEG/aqueous biphasic system. The partitioning of 29 small organic probes in a PEG-2000/$(NH_4)_2SO_4$ ABS was satisfactorily described with a three-parameter QSPR model ($R^2 = 0.967$, $R^2_{cv} = 0.956$). All the descriptors involved were calculated solely from chemical structures and have definite physical meaning corresponding to different intermolecular interactions. A single-parameter model involving calculated log P (octanol/water) values as an independent variable also demonstrates high statistical quality ($R^2 = 0.89$). The results described in this paper should help to improve our understanding and prediction of partition coefficients in PEG/salt ABS for structurally diverse compounds.

log P refers to the neutral state of molecules. In the presence of a basic or acidic group, the ionization of a molecule provides an additional factor, since partition becomes pH-dependent. The pH-dependent distribution coefficient, log D, was shown to correlate with a number of biological properties, such as the effective permeability in human jejunum,[570] blood brain barrier (BBB) permeability,[571] plasma protein binding,[572] CYP 450 oxidation,[573] and volume of distribution (V_D).[574,575] The pH-dependent distribution coefficient, log D, is related to log P through pK_a. The problem of predicting log D is more complicated. As a rule, it is computed from log P and pK_a (eq 12), assuming that only the neutral form of the molecule will partition into the organic phase.[501,576]

$$\log D_{(\mathrm{pH})} = \log P - \log(1 + 10^{(\mathrm{pH}-\mathrm{p}K_a)\Delta i}) \qquad (12)$$

where $\Delta i = \{1, -1\}$ for acids and bases, respectively.

If several groups may be ionized, correction terms must be included in the equation for each of them. Thus, the log D prediction potentially suffers errors due to errors in both log P and pK_a predictions. Lombardo and co-workers[577] determined a robust method by using RP-HPLC of octanol−water distribution coefficients at pH 7.4, noted as Elog$D_{oct.}$ Xing and Glen[578] reported a three-parameter (polarizability and partial atomic charges on nitrogen and oxygen atoms) equation with $R^2 = 0.89$, for log P of 592 compounds, and an estimation of pK_a values in order to calculate log D values. Development of computational approaches is further complicated by the absence of large data sets with experimental log D values. Only a few programs are available to estimate log D values including PrologD,[579] ACD/logD [www.acdlabs.com], and SLIPPER [software.timtec.net]. Tetko and Bruneau[580] used ALOGPS 2.1 based on self-learning properties of associative neural networks and reported a rms of 0.7 for 2569 neutral log P, and a mean average error of 0.5 for 8122 pH-dependent log $D_{7.4}$, distribution coefficients from the AstraZeneca "in-house" database. Later on, Tetko and Poda[581] evaluated ALOGPS, ACD/log$D$, and PALLAS prologD software to calculate the log D distribution coefficients and reported a high rms of 1.0−1.5 log units for two in-house Pfizer's log D data sets of 17,861 and 640 compounds. The authors have demonstrated that the ANN-based ALOGPS is superior, compared to the ACD/LogD and PALLAS PrologD programs, which reduced rmse for log D prediction to 0.64 and 0.65 (compared to 1.17 and 1.33) for data sets of 17,341 and 640 compounds, respectively.

5.2.3. Gas to Olive Oil Partition Coefficients

In 1923, Meyer[582,583] measured gas to olive oil partition coefficients, K(olive), and demonstrated the use of gas to olive oil partition coefficients as a model for gaseous narcosis or anesthesia. K(olive) is defined as the concentration of a solute in olive oil to the concentration of the solute in the gas phase. K(olive) values have traditionally been related to anesthetic properties and used in empirical relationships for the prediction of gas to tissue partition coefficients.[584,585] Few correlations have been reported for the prediction of K(olive) values. Abraham and Weathersby[586] correlated gas to oil partition coefficients of 88 organic compounds based on the Abraham solvation equation with $R^2 = 0.997$ and $s = 0.082$. Abraham and Fuchs[587] later correlated log K(olive) for 52 compounds with $R^2 = 0.947$ and $s = 0.233$ by using three descriptors (volume, molar refraction, and dipole moment). Klopman et al.[588] used a group contribution approach for the prediction of 159 compounds with $R^2 = 0.938$, $s = 0.295$ from 24 fragments. However, their data set included some previously calculated log K(olive) values.

Katritzky et al.[589] employed calculated molecular descriptors using CODESSA PRO software for the correlation of log K(olive) values for 100 training set compounds and an independent test set of 33 compounds. A five descriptor MLR model with $R^2 = 0.922$ and $s = 0.232$ log units was obtained for the training set. The authors used the same equation for the prediction of log K(olive) values for the 33 test compounds with a fit characterized by $R^2 = 0.846$. In a recent study, Abraham and Ibrahim[590] obtained a QSPR model for log K(olive) with $R^2 = 0.981$ and $s = 0.196$ log unit for a data set of 215 compounds based on the Abraham's linear free energy equations. Their study showed that gas to biological phase partition can be described in an empirical way by a combination of gas to olive oil and gas to saline coefficients.

5.3. GC Retention Indices

Gas chromatography (GC) is one of the most widely employed analytical techniques due to its simplicity, rapidity of analysis, high sensitivity of detector systems, and efficiency of separations. Thus, GC has found wide application in pharmaceuticals, environmental studies, petroleum industries, clinical chemistry, analysis of pesticides, food preservatives, etc.[591] Identification of a compound is often accomplished on the basis of gas chromatographic peak comparisons with an authentic standard of the suspected material. However, it is not always possible to obtain samples of the pure standard material. Thus, it is desirable to develop methods for the prediction of retention characteristics of the unknown compound based on the structural features and chromatographic properties of other representative compounds. Retention is a phenomenon that is mainly dependent on the solute–sationary phase interactions. Ideally, each solute will exhibit unique retention characteristics based on its chemical, structural, and electronic properties. QSPR methodology is widely accepted in various areas of application, which relate the properties of a molecule with its structure. The process of relating chemical structure to chromatographic retention comprises a field of research known as quantitative structure–retention relationships (QSRR). Numerous publications have appeared in this area during the past two decades, including a book by Kaliszan[592] based on several aspects of the development of valid estimation models and the significance of model parameters. Numerous QSPR models developed for the prediction of the retention index are listed in Table 8.

Gas chromatographic retention times were related by Katritzky et al.[620] to chemical structures. Duvenbeck and Zinn reported a general method for fitting the GC retention index data using three topological indices and one electrotopological state (E-state) index in the so-called vertex and edge MLR models.[602] For a data set of 217 acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones, and ethers, their edge model gave a mean absolute error of 9 retention index units. However, application of this model to the prediction of retention indices for test compounds of the same classes gave prediction errors from 15 to 22 retention index units.[603] Jurs et al. correlated indices of substituted pyrazines,[621] polycyclic aromatic compounds,[622] stimulants and narcotics,[598] and anabolic steroids[599] with charged partial surface area (CPSA) and topological and geometrical descriptors. These MLR analysis descriptors encode information related to the interactions between the solute (and solvent) molecules in the stationary phase during the separation process. As the polarity of the stationary phase is changed, different descriptors become important, and therefore, each phase must be modeled separately.

Buydens, Massart, and Geerlings combined topological descriptors with quantum-chemical descriptors to predict the GC retention indices of mono- and bifunctional alcohols and ketones.[623] In a more general QSPR study, Katritzky et al. used a mixed set of topological and quantum-chemical descriptors to correlate GC retention times of 152 structures encompassing a wide cross section of organic compounds.[604] A forward procedure for the selection of molecular descriptors for MLR analysis in the CODESSA program gave a six-parameter model ($R^2 = 0.959$, $R_{cv}{}^2 = 0.955$, $s = 0.515$), in which the AM1-computed α-polarizability was the most important descriptor. These results were recently re-evaluated using improved CODESSA procedures and new methods for the efficient selection of variables in the MLR analysis.[606] Quantum-chemical descriptors were employed by Donovan and Famini in a theoretical linear solvation energy relationship (TLSER) investigation of the GC retention indices of 37 organosulfur compounds.[624] From each of the three semiempirical methods (MNDO, AM1, and PM3) used to compute the six TLSER descriptors, similar correlations were obtained: $R^2 = 0.88-0.92$. These results had a statistical quality similar to that of the previous study of the same compounds by Woloszyn and Jurs.[600] However, the TLSER approach[624] was also able to handle compounds containing sulfur–sulfur bonds, which were omitted in the four-parameter correlation with topological and CPSA descriptors obtained by Woloszyn and Jurs.[600]

Whereas most QSRR predictions of GC retention indices are based on multilinear regressions, Bruchmann, Zinn, and Haffer showed recently that ANNs can be trained using electrotopological indices of monofunctional compounds by the back-propagation technique to predict the corresponding retention index data.[625] Sutter, Peterson, and Jurs applied ANNs to predict retention indices of alkylbenzenes from their molecular structure.[605] They used the ADAPT software to calculate 182 descriptors and MLR analysis in combination with evolutionary optimization algorithms to select a subset of descriptors relevant to mapping retention indices. Six descriptors from their best MLR model (three topological, one geometrical, and two CPSA descriptors) were used with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method to train a 6:5:1 ANN, which improved the rms for both training and prediction sets from 18.0 and 21.8 to 13.4 and 17.6, respectively. The counter-propagation ANNs applied for the prediction of GC retention indices were shown to be inferior compared to the back-propagation ANN and MLR models.[626,627]

Zarei and Atabati[628] correlated the GC retention indices for 178 insect-produced methyl substituted alkanes using an ANN approach based on simple structural descriptors. The authors found a correlation with $R^2 = 0.978$ and RMSE of 3.1 by using a 9:8:1 ANN architecture.

5.4. GC Response Factors

Application of gas chromatography (GC) as a tool for quantitative estimation requires knowledge of the response factor (RF) for each compound under the GC experimental conditions employed. Since numerous compounds are unavailable as standards, the development of a theoretical method for estimating the RF is potentially useful.

During the last five decades, several books and articles have been published on the determination and explanation of the RF on various detector devices connected to a GC. For detailed information on response factors, the reader is referred to the literature.[591,629–633]

The flame ionization detector (FID) response is based on the ionization of carbon containing molecular fragments and is dependent only on the carbon content of the molecule in question. The number of carbon atoms per gram of the compound not bonded to one or more heteroatoms or halogen atoms is the so-called "effective carbon number" (ECN). Scanlon et al.[634] used the ECN approach to calculate FID response factors for compounds not available in pure form. Predictions of GC (FID) response factors for a diverse structural class of compounds were published for the first time by our laboratory in collaboration with Musumarra's group,[635] using a multivariate statistical PLS treatment. A three component PLS model was found which explains 84% of the variance in the RF data. In another paper[604] the QSPR treatment for the prediction of Dietz response factors for a large and widely diversified set of 152 organic compounds was established using CODESSA. A MLR model with cross-validated squared correlation coefficient $R^2_{CV} = 0.881$ was achieved using six structural descriptors, including the

Table 8. List of QSPR Models Reported on the Prediction of Retention Indices

compounds	*N*	descriptors	approach	model statistics	Reference
olefins (C_4-C_6)	86	physical properties (BP, log *P*) and geometrical descriptors and mol. wt	MLR (4)	$R^2 = 0.994$, $s = 7.79$	Rohrbaugh and Jurs[593]
polycyclic aromatic compounds (PACs), nitrated PACs	73	physical properties (BP, MR) geometrical descriptors	MLR (3)	$R^2 = 0.984$, $s = 6.87$	Rohrbaugh and Jurs[594]
polychlorinated biphenyls	209	fragment, geometrical	MLR (5)	$R^2 = 0.997$, $s = 0.01$	Hasan and Jurs[595]
diverse drug compounds	100 (trn)	mol. wt, fragment descriptors	MLR (5)	$R^2 = 0.941$, $s = 122$	Rohrbaugh and Jurs[596]
	44 (test)	mol. wt, fragment descriptors	MLR (5)	$R^2 = 0.902$	
pyrazines (OV-101)	107	topological, geometrical, electronic	MLR (6)	$R^2 = 0.994$, $s = 22.9$	
pyrazines (carbowax-20)	107	topological, geometrical, electronic		$R^2 = 0.986$, $s = 36.3$	
polyhalogenated biphenyls	53	geometrical, connectivity	MLR (5)	$R^2 = 0.989$, $s = 45$	Hasan and Jurs[597]
pyrazines (carbowax-20)	107	CPSA, geometrical, structural	MLR (6)	$R^2 = 0.988$, $s = 32.9$	Stanton and Jurs[158]
pyrazines (carbowax-20)	107	CPSA, geometrical, structural	MLR (9)	$R^2 = 0.994$, $s = 26.7$	
stimulants and narcotics	57	topological, electronic, fragment	MLR (6)	$R^2 = 0.982$, $s = 0.046$	Georgakopoulos et al.[598]
anabolic steroids	45	geometrical, topological	MLR (9)	$R^2 = 0.982$, $s = 0.027$	Georgakopoulos et al.[599]
sulfur vesicants (DB-1)	31	topological, geometrical, CPSA	MLR (4)	$R^2 = 0.996$, $s = 31.5$	Woloszyn and Jurs[600]
sulfur vesicants (DB-5)	31	topological, geometrical, CPSA	MLR (4)	$R^2 = 0.996$, $s = 33.1$	
sulfur vesicants (DB-1701)	30	topological, geometrical, CPSA	MLR (4)	$R^2 = 0.996$, $s = 43.8$	
sulfur vesicants (DB-1 and DB-5)	62	topological, geometrical, CPSA, indicator variable	MLR (5)	$R^2 = 0.996$, $s = 34.7$	
hydrocarbons separated from naphtha mixture, SE-30	67	mol. wt, topological, geometrical	MLR (4)	$R^2 = 0.966$, $s = 18.6$	Woloszyn and Jurs[601]
hydrocarbons separated from naphtha mixture, carbowax-20M	65	mol. wt, topological, geometrical, CPSA	MLR (5)	$R^2 = 0.933$, $s = 23.3$	
diverse compounds (alkanes, alkenes, alcohols, esters, ketones, and ethers)	217	topological, E-state index	MLR	mean absolute error (7...9)	Duvenbeck and Zinn[602,603]
diverse organic compounds	152	constitutional, thermodynamic, electronic	MLR (6)	$R^2_{CV} = 0.955$, $s = 0.503$	Katritzky et al.[604]
alkylbenzenes	150	topological, geometrical, electronic	MLR (6) based on evolutionary optimization technique	$R^2 = 0.982$, rms = 18.0 (train), rms = 21.8 (predn set)	Sutter et al.[605]
	150	topological, CPSA	ANN (6)	rmse = 11.7 (train), rms = 13.4 (predn set)	Sutter et al.[605]
polychlorinated biphenyls	209	3D descriptors (WHIM), GA selected descriptors	MLR (2)	$R^2 = 0.984$, SDEC = 0.023, SDEP = 0.023	Gramatica et al.[224]
diverse organic compounds	152	topological, geometric, electronic (variable selection)	nonlinear (6)	$R^2 = 0.977$, $s = 0.379$	Lučić et al.[606]
methyl alkanes	178	topological	MLR (4)	$R^2 = 0.959$, $s = 5.8$	Katritzky et al.[607]
aldehydes, ketones (HP-1)	31	topological (Xu and atom-type-based AI indices)	MLR (4)	$R^2 = 0.998$, $s = 7.73$	Ren[608]
aldehydes, ketones (HP-50)	31	topological (Xu and atom-type-based AI indices)	MLR (4)	$R^2 = 0.996$, $s = 9.19$	
aldehydes, ketones (DB-210)	31	topological (Xu and atom-type-based AI indices)	MLR (5)	$R^2 = 0.995$, $s = 12.03$	
aldehydes, ketones HP-Innowax	31	topological (Xu and atom-type-based AI indices)	MLR (5)	R^2=0.993, $s = 13.45$	
diverse organic compounds	632	semiempirical topological index (I_{ET})	LR	R^2=0.9997, $s = 17.7$	Junkes et al.[609]
	548	semiempirical topological index (I_{ET})	LR	$R^2 = 0.9997$, $s = 7.01$	
polycyclic aromatic hydrocarbons	44	mol. wt, Wiener index, polarizability, hardness, E_{LUMO}	PLS, PCR	$R^2 = 0.898$, $s = 13.45$	Alves de Lima Riberio and Ferreira[610]
alkanes	64	topological and others	MLR	$R^2 = 0.997$, $s = 8.09$	Cao et al.[245]
CNSagents (benzodiazpines, barbiturates, and phenytoin (DB-5))	37	topological and electronic	MLR	$R^2 = 0.976$, $s = 18.8$	Hodjmohammadi et al.[611]
CNSagents (benzodiazpines, barbiturates, and phenytoin (DB-17))	32	topological and electronic	MLR	$R^2 = 0.966$, $s = 49.8$	
hydrocarbons	207	topological	MLR (7)	$R^2 = 0.999$, $s = 3.49$	Hu et al.[612]
diverse organic compounds	846	topological, constitutional, electronic, fragmental (calculated by DRAGON*), stepwise selection	PLS	SEP = 79, SEC = 81	Garkani-Nejad et al.[613]
nitrogen containing polycyclic aromatic hydrocarbons	117	topological, fragmental	MLR	$R^2 = 0.985$, $s = 10.3$	Hu et al.[614]
methyl-substituted alkanes	178	fragment, indicator variables	ANN (9:8:1)	$R^2 = 0.978$, rms = 3.1 (calibration set) ($n = 30$, $R^2 = 0.939$, rms = 4.9), prediction set	Zarei et al.[615]

Table 8. Continued

compounds	N	descriptors	approach	model statistics	Reference
methyl alkanes	177	topological, fragment		$R^2 = 0.999$, SEC = 4.6 (SEP = 3.7, n = 30 test set)	Liu et al.[616]
carbazoles	49	topological, constitutional, electrostatic, quantum chemical	MLR (7)	$R^2 = 0.9966$, $s = 0.58$	Nakajima et al.[617]
halogenated hydrocarbons	23	topological (Lu index)	LR	$R^2 = 0.992$, $s = 21.5$	Lu et al.[618]
diverse organic compounds	22, 995	group increment values	MLR	average and median absolute deviations AAD = 70, MAD = 45	Stein et al.[619]

relative weight of "effective" carbon atoms and the total molecular one center one-electron repulsion energy in the molecule. A variable selection method implemented nonlinear cross-term multiregression (MR) model was developed[606] for the prediction of the RF with greater accuracy. This variable selection based MR approach enabled the selection of the best possible MR models from 10^{10} possibilities.

Huang et al.[636] determined FID relative weight response factors and FID relative carbon weight response factors for a variety of compounds: hydrocarbons, chlorohydrocarbons, bromohydrocarbons, and oxygenated hydrocarbons. They also determined the FID relative response factor and the FID relative carbon response factor for a variety of compounds.

Morvai et al.[637] determined FID response factors for 130 organic acid esters, such as ethyl, isopropyl, *n*-propyl, isobutyl, and *n*-Bu esters of C_1-C_{20} fatty acids, C_2-C_{12} aliphatic dicarboxylic acids, and benzoic and *o*-phthalic acids by using the ECN approach.

Jalali-Heravi and Fatemi[638] applied a ANN to develop a nonlinear model for the FID response factors of various classes of organic compounds.[604] They showed the higher predictive power of the ANN model over a MLR model. Jalali-Heravi and Fatemi[639] implemented the ANN approach for the prediction of thermal conductivity detection (TCD) response factors of a set of 110 organic compounds, including hydrocarbons, benzene derivatives, esters, alcohols, aldehydes, ketones, and heterocyclics. They also developed linear models using theoretical molecular descriptors for the prediction of TCD-RFs and applied those descriptors as inputs for the ANN. Saradhi et al.[640] studied the response mechanism of thermionic detection (TID) of a series of organophosphate esters. They found that the response to TID decreases with the increase of the alkyl chain length of the molecules studied.

The relative response factors (RRFs) of a flame ionization detection (FID) system were predicted for diverse organic compounds,[641] using molecular descriptors for the development of a MLR model and then employing those descriptors as inputs for the self-training ANNs. They then compared the two models and observed the superiority of ANNs over that of the MLR method.

The RRFs of an electron-capture detection (ECD) system were predicted for a set of 118 polychlorinated biphenyls (PCBs)[642] based on two different internal standards. The authors developed a MLR model for the prediction of RRFs by using two descriptors of molecular ion ionization potential (MIIP) and ionization potential of the molecule (IP) that are related to the affinity of the compounds. The descriptors employed were those used in the MLR model as inputs for developing the back-propagation ANNs. A QSPR model was developed for the gas chromatography thermionic detector (GC-TID) response factor of organophosphorous compounds taken from Saradhi et al.[640] using physicochemical and electronic descriptors.[643] The authors employed a combinatorial protocol in MLR, by using a "filter"-based variable selection procedure for the development of the model and achieved a four-descriptor correlation with $R^2 = 0.891$ for 28 organophosphorous compounds.

5.5. Gas Phase Homolysis

Homolysis is a simple elementary chemical reaction of bond fission generating two free radicals (eq 13):

$$A{:}B \rightarrow A^{\bullet} + B^{\bullet} \qquad (13)$$

Gas-phase homolysis reactions are unimolecular at low pressure[644] and thus characterized with a temperature-dependent first order rate constant which is usually described by the activation parameters of the respective Arrhenius equation. The rate constant of homolysis is the key parameter of thermal stability.

Theoretical prediction of the gas phase homolytic rate constants of 79 diverse nitro compounds by the QSPR approach was proposed by Sukhachev et al.[645] The radical cleavage of C−N and N−N bonds was the primary step in their thermolysis. About 3000 descriptors were computed for each structure, including topological and information indices, indices based on electronegativities of atoms, substructures, etc. The most stable of the regression models was constructed on 11 descriptors ($R^2 = 0.99$), nine of which were electrotopological states descriptors of atoms in different fragments, which are known to reflect the electron density distribution on atoms. The predictive power of the model was $R^2 = 0.96$, based on the test set of 10 compounds.

Hiob and Karelson[646,647] have modeled the rate constants of the gas-phase C−X and $C-CH_3$ bond homolysis. Initially, five-, four-, and three-parameter models were developed for the kinetics parameters of 287 different C−X bonds gas-phase homolyses at 891 K. A general model ($R^2 = 0.80$) included all 287 data points and the following five molecular descriptors: rotational entropy (300 K), relative number of C atoms, maximum $\sigma-\pi$ bond order, HOMO-1 energy − HOMO energy, and the relative number of I atoms. In the following study,[653] a six-parameter model was developed for the prediction of the log k (at 1047 K) for 58 different $C-CH_3$ bonds derived from information encoded in the chemical structure of compounds.

5.5.1. Gas Phase Ion Mobility Constants

Due to the high sensitivity and rapidity of the ion separation process (an order of milliseconds) combined with ease of use, gas-phase ion mobility constants have found wide analytical application in the detection of explosives, drugs, chemical weapons, and environmental pollutants by means of ion-mobility spectrometry. It is also a research tool for analysis of biological materials, especially in proteomics. The gas-phase mobility of an ion, K, is determined from the drift velocity, v_d, attained by the ion in a weak electric field, E, at atmospheric pressure by eq 14.

$$\nu_d = KE \qquad (14)$$

The reduced mobilities K_0 (normalized by pressure and temperature) increase linearly with the logarithm of molecular weight or ion mass for a homologous series[654] but not for a diverse set of compounds.[648]

Wessel et al.[649] have used the QSPR methodology to investigate gas-phase ion mobility. The values of gas-phase reduced ion mobility constants, K_0, were modeled for 70 organic compounds by MLR and ANN. The exclusion of three outliers (anthracene, *m*-toluidine, and *n*-butyl acetate) gave a good MLR ($R^2 = 0.98$ and $s = 0.047$) with five descriptors: the charge on the most negative atom, QNEG; the Kier path 3 shape index corrected for heteroatoms, KAPA-3A; the Wiener number, ALLP-5; the sum of all path weights starting from heteroatoms, WTPT-3; and the square root of the molecular weight, SQMW. This model performed well for the external test set of seven compounds: $s = 0.047$. The rms errors for the 5−3−1 ANN model including the same descriptors for the training set of 57 and validation set of 10 compounds were 0.041 and 0.039, respectively. The rms error for the external validation set was 0.039. The ANN enhanced the ability to predict K_0 values. The highest R^2 value of 0.89 between the SQMW descriptor and K_0 indicated the importance of the molecular weight. Wessel[650] attempted to model K_0, using a 6−4−1 ANN on a training set of 135 and a validation set of 15 compounds. The best model (rms $= 0.04$ K_0 units) was found with a feature selection routine which couples the genetic algorithm with MLR analysis. The model contained six molecular descriptors (charge on most negative atom, QNEG; Kier path 3 shape index, KAPA-3; path 1 valence connectivity, V1; number of oxygens, NO; sum of path weights from heteroatoms, WTPT-3; and number of secondary sp^2 C atoms, 2SP2) and was externally validated with a rms error of 0.038 for 18 compounds. The model can predict K_0 values of compounds for which there are no empirical K_0 data without the need for geometry optimization.

Agbonkonkon et al.[651] improved the data set of Wessel et al.[650] by including diverse ions but started out with the same descriptors. To meet the increased diversity within the data set, two of the molecular descriptors, namely KAPA-3 and WTPT were adjusted. The resulting MLR model for 162 compounds had $R^2 = 0.80$. Later on, this data set was modeled by Liu et al.[652] to develop linear and nonlinear models for predicting the gas-phase K_0 using MLR and projection pursuit regression. The molecular descriptors were generated and selected with the aid of the CODESSA software to be used as inputs for the models. The values of R^2 were 0.908 and 0.938, and the s values were 0.066 and 0.055, for the linear and nonlinear models, respectively, based on the whole data set of 159 compounds. The models were internally validated by splitting into training and test sets.

5.6. Soil Sorption Coefficients

The soil sorption coefficient measures the partitioning capacity of a compound between two phases, liquid (i.e., water) and solid (i.e., the soil components). The organic fraction of the soil is recognized as the part of the soil that is responsible for the sorption of contaminants. This has led to the normalization of the soil sorption coefficient by the organic carbon content, K_{OC}, which also makes comparable the experimental data from different soils. Soil sorption is considered of great environmental importance since it determines the distribution of chemicals and accordingly their availability to living organisms. Reference soils (the EUROSOILS[653]) have been determined by the European Commission for profound investigations of the sorption phenomena. More than 200 QSPRs on soil sorption have been reviewed by Gawlik et al.[654] It was shown that the log K_{OC} values were most frequently modeled with water solubility (S_w), *n*-octanol/water partition coefficient (K_{OW}), RP-HPLC capacity factor (k'), topological indices, or linear solvation energy parameters. It was concluded that log K_{OW} was most commonly used to describe soil sorption. Most of the QSPRs were chemical class- and soil-specific. More recent contributions are summarized in Table 9.

Due to the heterogeneity of the soil organic constituents, sorption can involve either nonspecific or specific mechanisms. Nonspecific interactions are exhibited by hydrophobic compounds. Topological indices have given good results in modeling homologous series of compounds. The size and branching, accounted for by the topological indices, may affect the mobility of the contaminant physically in the humic matrix. Specific interactions of soil sorption exhibited by polar compounds have been covered by the inclusion of constitutional, electrostatic, quantum-chemical, and weighted holistic invariant molecular (WHIM) descriptors. More general models resulted from incorporating descriptors for both mechanisms of sorption. In addition, chemical reactions may take place with the soil components, which decrease the mobility of these compounds.

In a different approach, Winget et al.[655] developed a set of quantum mechanical solvent descriptors using SM5 solvational parametrization to characterize the organic carbon component of soil. These descriptors were subsequently used to develop QSPR models to be applied in partitioning of solutes between soil and air. The combination of this set of effective solvent descriptors with solute atomic surface tension parameters developed for water/air and organic solvent/air partitioning allowed for prediction of the partitioning of solutes between soil and water.

5.7. Solvent Scales

Solvents form the basis of many chemical reactions and are of fundamental importance in chemistry. Solvents influence chemical and physical processes by solvating the substrate either through van der Waals interactions or hydrogen bonding or by providing solvent pockets or cages for encapsulating the substrate. The structure of both solute and solvent play important roles in the solvation phenomenon. Most solvent scales are based on a model system by recording the changes in a measurable parameter when the solvent is changed. The model processes represent different intermolecular interactions in the system, although no one scale can be universal and applicable to all properties. The empirical properties used to define solvent polarity scales include the following: (i) equilibrium and kinetic rate constants of chemical reactions of solutes, (ii) spectroscopic properties of solutes in different solvents comprising absorption and fluorescence energies or vibrational translational energies, (iii) solvation free energies of solutes, (iv) macroscopic properties such as dielectric constant, dipole moment, refractive index, molecular volume, polarizability index, etc., and (v) composite experimental parameters. Individual solvents are rarely represented in all common scales, and no scale covers all the common solvents. To date, more than a hundred quantitative solvent polarity scales have been developed based on diverse physicochemical properties

Table 9. Summary of Some Important QSPR Models on log K_{OC}[a]

no.	compounds[b]	N	model descriptors[c]	R^2	s	F	ref
1	nonpolar compounds	64	$^1\chi$	0.96	0.27	1371	Meylan et al.[656]
2	nonpolar and polar organic compounds	189	$^1\chi$, $\Sigma P_f N$ ($f = 26$)	0.96	0.23		Meylan et al.[656]
3	heterocyclic nitrogen compounds	12	$^1\chi$	0.88	0.38	74	Liao et al.[657]
4	heterocyclic nitrogen compounds	12	log S_w, $\Delta^1\chi^v$	0.91	0.32	54	Liao et al.[657]
5	heterocyclic nitrogen compounds	12	log K_{OW}, $\Delta^1\chi^v$	0.87	0.38	38	Liao et al.[657]
6	phenylthio, phenylsulfinyl, phenylsulfonyl	25	log k'_w	0.93	0.13	320	Hong et al.[658]
7	phenylthio, phenylsulfinyl, phenylsulfonyl	25	log K_{OW}	0.83	0.21	115	Hong et al.[658]
8	phenylthio, phenylsulfinyl, phenylsulfonyl	25	$^3\chi^v{}_c$(Ph), $^0\chi$(R$_4$), $^0\chi$(Ph), $^3\chi_c$(Ph)	0.91	0.15	63	Hong et al.[658]
9	diverse organic compounds	72	log K_{OW}	0.91			Baker et al.[659]
10	POPs (log K_{OW} > 5)	18	log K_{OW}(calc)	0.29	0.59		Baker et al.[660]
11	POPs (log K_{OW} > 5)	18	$^1\chi$, $^4\chi^v{}_C$, $^3\chi_C$	0.81	0.30	25	Baker et al.[661]
12	diverse organic compounds	66	log K_{OW}(calc), V$^+$, B_{MAX}	0.84	0.38		Müller[662]
14	diverse set of reference substances	21	log k'_w (cyanopropyl phase)	0.91			Szabo et al.[663]
15	diverse set of reference substances	21	log k'_w (humic acid phase)	0.93			Szabo et al.[663]
16	PCBs	48	RRT	0.92	0.16		Hansen et al.[664]
17	PCBs	48	TSA	0.92	0.17		Hansen et al.[664]
18	PCOCs	65	log K_{OW}	0.86	0.44	386	Dai et al.[665]
19	PCOCs	65	μ, E_{homo}, qH$^+$, q^-, TE	0.85	0.46	69	Dai et al.[665]
20	benzaldehydes (AM1)	14	μ, qH$^+$, $^3\chi_{pc}$, $^2\chi^v{}_p$	0.91	0.10	35	Dai et al.[666]
21	benzaldehydes (PM3)	14	μ, qH$^+$, $^3\chi_{pc}$, $^2\chi^v{}_p$	0.92	0.10	40	Dai et al.[666]
22	benzaldehydes (AM1)	14	μ, qH$^+$, q^-	0.86	0.13	28	Dai et al.[666]
23	benzaldehydes (PM3)	14	μ, qH$^+$, q^-	0.82	0.16	19	Dai et al.[666]
24	heterogeneous pesticides	143	MW, nBr, nNO, nHA, I_{CEN}, MAXDP	0.82	0.38	106	Gramatica et al.[667]
25	carbamates	29	nO, nX, nNO, ξ^C	0.95	0.17	110	Gramatica et al.[667]
26	organophosphates	28	$I^E{}_{deg}$, IC, MAXDP, η1u, Ts	0.89	0.23	35	Gramatica et al.[667]
27	phenylureas	43	MW, nCl, nCIT, λ1v, η2s	0.91	0.12	76	Gramatica et al.[667]
29	diverse organic compounds	592	74 fragment constants, 24 structural factors	0.97	0.37		Tao et al.[668]
30	diverse organic compounds	592	$^1\chi^v$, $^2\chi$, $^4\chi_c$, $^6\chi$, ΣP_j ($j = 17$)	0.77	0.44		Tao et al.[668]
31	diverse organic compounds	387	5σ moments	0.71	0.62[d]	189	Klamt et al.[669]
32	substituted aromatic compounds	28	log K_{OW}	0.61	0.22	43	Wu et al.[670]
33	substituted aromatic compounds	28	$^2\chi^v$, $\Delta^5\chi^v$	0.69	0.20	31	Wu et al.[670]
34	substituted aromatic compounds	28	MW, π^d, V$^-$, EN	0.95	0.08	128	Wu et al.[670]
35	substituted aromatic compounds	27	log K_{OW}	0.79	0.08	92	Wu et al.[671]
36	substituted aromatic compounds	27	log K_{OW}, $^3\chi_c$	0.88	0.06	91	Wu et al.[671]
37	substituted aromatic compounds	27	α, π^d, O	0.86	0.07	48	Wu et al.[671]
38	diverse organic pesticides	143	$^1\chi$, 11 E-state indices (S_i)	0.82	0.37	51	Huuskonen[672]
39	diverse organic compounds	403	log S (calc)	0.80	0.51	1622	Huuskonen[673]
40	diverse organic compounds	403	log S (calc), HBA, NAR, MW, I_{acid}	0.85	0.44	451	Huuskonen[673]
41	diverse organic compounds	403	log K_{OW} (calc)	0.79	0.52	1475	Huuskonen[673]
42	diverse organic compounds	403	log K_{OW} (calc), NAR, ROT, MW, I_{acid}	0.86	0.43	491	Huuskonen[673]
43	organic compounds containing C, H, N, O, S	82	N_ϕ, MW, N_N, N_O, N_S	0.94	0.33	228	Delgado et al.[674]
44	halogenated benzenes, anilines, and phenols	28	α, β, V_{CSE}, $^3\chi_C$, O_V	0.96	0.07	129	Wei et al.[675]
45	substituted anilines and phenols	42	log K_{OW}, E_{homo}, α, μ (MLR)	0.78	0.37	32	Liu and Yu[676]
46	substituted anilines and phenols	42	log K_{OW}, E_{homo}, E_{lumo}, q_N, q_O, MW, α, μ (ANN)	0.87	0.28		Liu and Yu[676]
47	diverse organic compounds	68	log K_{OW}, η [AM1]	0.76	0.44	101	Kahn et al.[677]
48	diverse organic compounds	344	log K_{OW}, PNSA-1, η, $P^{max}_{\pi-\pi}$	0.76	0.41	266	Kahn et al.[677]
49	POPs	32	Lu index, 5 DAI indices	0.90	0.23	40	Lu et al.[678]
50	heterogeneous pesticides	143	μ_{10}Dip, μ_{15}Dip, μ_4Dist, μ_1H, μ_5H, μ_7P Dip 10	0.84	0.37	117	Gonzalez et al.[679]
51	chlorinated phenols	19	log K_{OC} (exp) using cubic spline polynomials (in $x\alpha \rightarrow \beta$) on the oriented edges $\alpha \rightarrow \beta$ of the Hasse diagram	0.88	0.26		Ivanciuc et al.[680]
52	diverse organic compounds	550	VED1, nHAcc, MAXDP, CIC0	0.82	0.54		Gramatica et al.[681]
53	diverse organic compounds	550	consensus model[e]	0.82	0.42[f]		Gramatica et al.[681]
54	heterogeneous pesticides	143	NCONN, ATS2p, O-058, n_P, D_s, V_m	0.90	0.29	203	Duchowicz et al.[682]
55	polycyclic aromatic hydrocarbons	20	3 PLS PCs (using nine DFT level descriptors)	0.99	0.14	890	Lu et al.[683]

[a] N, number of compounds used to develop a model; R^2, correlation coefficient; s, standard deviation; F, F-test value. [b] PCOCs, polychlorinated organic compounds; POPs, persistent organic pollutants; PCBs, polychlorinated biphenyls. [c] $P_f N$ and P_j, structural fragment contribution factors (P_f) for polar structural fragments; N, number of times the fragment occurs in the structure; $^n\chi$, molecular connectivity indices; S_w, water solubility; $\Delta^1\chi^v$, nondisperse force factor; K_{OW}, n-octanol/water partition coefficient; V$^+$, potential of the positive atomic charges; B_{MAX}, maximum charge difference between connected atoms; k', HPLC capacity factor; S_{ester} and S_{alkyl}, group electrotopological indices; RRT, gas chromatographic relative retention time; TSA, molecular total surface area; μ, dipole moment; E_{homo}, energy of the highest occupied molecular orbital; qH$^+$, most positive net atomic charge on hydrogen atom; q^-, largest negative atomic charge on an atom; TE, total energy; MW, molecular weight; nBr, number of Br atoms; nNO, number of NO bonds or groups; nHA, number of hydrogen bond acceptor atoms; I_{CEN}, centric information index; MAXDP, maximum positive intrinsic state difference; Ts, global WHIM descriptor of molecular size; nO, nCl, and nX, numbers of O, Cl, and halogen atoms; ξ^C, eccentric connectivity index; $I^E{}_{deg}$, mean information content on vertex degree equality; IC, information content on multigraph; η1u, λ1v, and η2s, directional WHIM descriptors; nCIT, number of total rings; DELS, index, mainly related to total charge transfer in the molecule; σ moments, real solvents σ-moment descriptors; E-state indices (S_i), electrotopological-state indices; V$^-$, potential of the negative atomic charges; EN, electronegativity; α, polarizability; π*, α/Connolly accessible volume; O, ovality of a molecule; HBA, number of N and O atoms; NAR, number of aromatic rings; I_{acid}, indicator variable for ionization of carboxylic acids; ROT, number of rotational bonds; N$_\phi$, number of benzene rings; N$_N$, N$_O$, and N$_S$, numbers of N, O, and S atoms; α and β, hydrogen bond (acceptor and donor) terms; V_{CSE}, Connolly solvent-excluded volume (Å^3); O_V, ovality; q_N and q$_O$, net negative atomic charges on atoms N and O; NCONN, number of urea derivatives; ATS2p, Broto−Moreau autocorrelation of a topological structure—lag 2/weighted by atomic polarizabilities; O-058, number of O-fragments; n_P, number of phosphorous atoms; D_s, D total accessibility index/weighted by atomic electrotopological states; V_m, V total size index/weighted by atomic mass; VED1, eigenvector coefficient sum from distance matrix; CIC0, complementary information content index[46,47] with neighborhood symmetry of 0 order; μ_nX (X = Dip, Dist, H, P), where n is the order of the spectral moment and X is the type of bond weight; Dip, dipole moment; Dist, standard distance; H, hydrophobicity; P, polarizability; PLS PC, partial least squares principal components; DFT, density functional theory. [d] rms = root-mean-square. [e] Averages of predicted log K_{OC} values from 10 models. [f] Mean residual.

of solutes and solvent, including chemical reactivity, spectroscopic properties, directly measured energies, and free energies of solvation and others. Several analyses of solvent scales together with reviews and discussions on the subject have been published.[684−686]

In an early study, Katritzky et al.[687] obtained a multilinear QSPR model ($R^2 = 0.936$) for the nonspecific solvent polarity scale (S') containing 67 solvents on the basis of three calculated molecular descriptors. Their later study,[684] correlated 45 different solvent scales, which contained data for a total of 350 solvents with theoretical molecular descriptors. The resulting QSPR equations for the different scales gave considerable insight into both the nature of the scales and the nature of the solvents. A review on quantitative measures of solvent polarity includes an overview of solvent scales.[685] A recent study[686] reports the data for 127 solvent scale values containing data for a total of 774 different solvents. QSPR models were developed for all the 127 scales based on molecular descriptors calculated using CODESSA PRO software. The descriptors include constitutional, geometrical, topological, charge-related, quantum chemical, and thermodynamical types derived solely from molecular structure which do not require the knowledge of experimental data. The BMLR algorithm was used to build QSPR models with up to five descriptors, based on the size of the data set of each scale. The 127 solvent scales are categorized based on experimental techniques used for the measurement scales as follows: (i) spectroscopic (67), (ii) equilibrium (17), and (iii) kinetic measurements (4). The remaining 39 solvent scales were grouped into class (iv): other measurements. The QSPR model statistics with respect to the solvent scales are listed in Table 10.

The QSPR model equations developed for the 127 solvent scales contain a total of 168 different descriptors. The 168 individual descriptors included in the 127 QSPR models comprise (i) 10 constitutional (applied 14 times), (ii) 2 geometrical (applied 3 times), (iii) 29 quantum chemical (applied 135 times), (iv) 22 topological (applied 47 times), (v) 13 thermodynamical (applied 23 times), and (vi) 92 electrostatic and charged partial surface area descriptors (applied 203 times). Altogether, the molecular descriptors were applied 425 times.

As shown in Table 10, most of the QSPR models are characterized by statistically good correlation coefficients. The R^2 values for 127 models range from 0.726 to 0.999; 18 models have $R^2 < 0.800$. The ranges for the 127 solvent scales predicted using the proposed models indicate that 26 solvent scales have predicted values within the experimental range for 774 solvents. For 101 solvent scales, the predicted range of values is at most 20% outside the experimental range values.

A good fit of a model mostly depends on the quality of the experimental measurements used in the development of the solvent scales. The authors used solvents having wide structural variability, including molecules without carbon atoms (water, ammonia, hydrazine, and hydrogen sulfide) and molecules without hydrogen atoms such as carbon tetrachloride, and the overall statistical quality of QSPR models for different solvent scales showed results that ranged from satisfactory to excellent. The predicted value for water for solvent scale (Z) using model 29 is 94.8, which fits well with the experimental value of 94.6. A classification approach for these solvents and solvent scales based on the above models was attempted by using PCA.[686]

5.8. Surfactant Properties

5.8.1. Critical Micelle Concentrations

Surfactants are amphiphilic molecules, that is, that contain a nonpolar segment, “tail,” and a polar segment, “head” (see Figure 10c). Under specific conditions, the presence of these two substructural features causes aggregation: when the surfactant concentration is low, the molecules exist as individual entities, but when the concentration increases, the molecules tend to form aggregates.

In aqueous solution, the hydrophobic tails of the surfactant associate, leaving the hydrophilic heads exposed to the solvent. The simplest of such aggregates, having approximately spherical shape, are called micelles. In nonpolar solvents, the hydrophilic segments are usually poorly solvated. As a result, the heads will form the interior of the aggregates, while the hydrophobic segments surrounding the polar core will be responsible for the solubility.[688] The structures formed are therefore called “reverse micelles”.

The transition from premicellar to micellar solutions occurs at a concentration called the critical micelle concentration (CMC). It was found that, at the CMC, many important properties of the surfactant solution, such as surface tension, interfacial tension, conductivity, osmotic pressure, detergency, emulsification, foaming, and so on, change sharply.[689−691] Therefore, CMC can be regarded as one of the most useful quantities for characterizing surfactants and can be correlated with many industrially important properties. For example, to perform micellar electrokinetic chromatography (MEKC) and micellar liquid chromatography (MLC), a surfactant solution at a concentration higher than the CMC must be used as a separation solution; thus, the CMC is an essential factor in the experiment.[692]

The first attempts to determine the CMC theoretically occurred more than 50 years ago. Based on a vast amount of experimental data concerning the CMC of surfactants, many empirical equations relating the CMC to the various structural units in surfactants were obtained.[693] In 1976, Rosen reported a linear relationship (eq 15) between the logarithm of the CMC and the number of alkane carbon atoms, n, in a homologous series.[694]

$$\log \mathrm{CMC} = A - Bn \qquad (15)$$

where A and B are empirical regression coefficients. In 1984, Becher[695] published a similar relationship for a series of linear alkylethoxylate surfactants. It connects log CMC and the carbon number, n, on one hand, and the ethylene oxide number, m, on the other. Ravey et al.[696] improved the correlation by including a nonlinear term in the form of a product of the alkane carbon number and the ethylene oxide number, nm (eq 16):

$$\log \mathrm{CMC} = A + Bn + Cm + Dnm \qquad (16)$$

More recently, thermodynamic treatments have been used to describe the phase behavior of surfactant solutions in an attempt to predict the CMC.[697,698] The thermodynamic models most commonly used are the phase separation model and the mass action model.[699] The phase separation model represents micellization as an equilibrium between two pseudophases: the micelles and the monomers in solution. The CMC can be calculated through the standard free energy of micellization. This simple model allows qualitative

Table 10. QSPR Models for the Solvent Scales Obtained by Katritzky et al.[a]

no.	solvent scale	physical background of solvent scale	N	R^2	s
1	AN	acceptor number, derived from ^{31}P NMR of triethylphosphine oxide in different solvents	52	0.903	5.648
2	B	basicity from stretching frequency of CH_3OD in different solvents	71	0.808	32.66
3	BCo	ratio of the fluorescence intensities of bands I and III of the vibronic spectra of benzo(α)coronene	25	0.929	0.075
4	B_{KT}	calculated from the difference of the longest wavelength band in the UV−vis spectra measured for *p*-nitroaniline and *N*,*N*-diethyl-*p*-nitroaniline	44	0.788	0.135
5	BPe	relative band intensities(I/III) for benzo[*ghi*]perylene fluorescence spectra	25	0.902	0.103
6	Co	relative band intensities(I/III) for coronene fluorescence spectra	25	0.967	0.050
7	Cu-λ_{max}	maximum absorption band of Cu(tmen)(acac)(solv)	36	0.851	13.81
8	DCo	ratio of the fluorescence intensities of bands I and IV of the vibronic spectra of dibenzo[α,*j*]coronene	23	0.908	0.076
9	D_S	donor strength—decrease in symmetric stretching frequency of Hg_2Br_2, between the gas phase and solutions	56	0.768	7.481
10	$E_{(NR)}$	Nile red transition energy	82	0.829	1.424
11	E^*_{MLCT}	solvent dependence of the metal to ligand charge transfer absorption maxima of $W(CO)_4$ with 1,10-phenanthroline	33	0.941	3.569
12	$E_{CT(A)}$	CT spectra of $W(CO)_4$ complexes with TCNE: E^*_{MLCT} (cm^{-1}) = 3000 (cm^{-1}) × $E_{CT(\pi)}$ + 12360 (cm^{-1})	28	0.823	0.132
13	E_B^N	energy of $N \rightarrow \pi^*$ transition in the 2,2,6,6-tetramethylpiperidine *N*-oxyl spectrum	52	0.951	0.047
14	$E_{CT(A)}$	UV charge transfer absorption maxima of tetra-*n*-hexylammonium iodide trinitrobenzene	23	0.922	0.684
15	$E_T(30)$	molar electronic transition energy of dissolved negatively solvatochromic pyridinium *N*-phenolate betaine dye	335	0.826	2.917
16	$E_T(N)$	molar electronic transition energy of dissolved negatively solvatochromic pyridinium *N*-phenolate betaine dye	335	0.821	0.092
17	E_T^{SO}	UV/vis spectra of *N*,*N*-(dimethyl)thiobenzamide-*S*-oxide	35	0.965	0.544
18	G	infrared vibration shift of hydrogen bonding	21	0.774	11.49
19	$^2J_{119Sn-117Sn}$	tin−tin spin coupling constant $^2J(^{119}Sn-^{117}Sn)$ of the hexaorganodistannoxanes	18	0.933	4.011
20	K	equilibrium constants for the conformational mobile (+)-*trans*-α-chloro-5-methylcyclohexanone	25	0.837	16.38
21	NCo	ratio of the fluorescence intensities of bands I and III of the vibronic spectra of naphtha[2,3-α]coronene	25	0.890	0.152
22	Ov	relative band intensities(I/III) for ovalene fluorescence spectra	25	0.945	0.155
23	P_s	bathochromic UV/vis spectral shifts of λ_{max} of (α-perfluoroheptyl-β,β-dicyanovinyl)aminostyrenes	107	0.844	0.911
24	P_y	relative band intensities I_1/I_3 for pyrene fluorescence spectra	93	0.839	0.144
25	Qm	heat of mixing data, for mixtures of chloroform and solvents measured by infrared spectra	19	0.764	166.8
26	SA	solvent acidity, evaluated from UV/vis spectra of *o*-*tert*-butylstilbazolium betaine dye and its nonbasic homomorph *o*,*o*′-di-*tert*-butylstilbazolium betaine dye	121	0.849	0.071
27	SB	solvent basicity, evaluated from UV/vis spectra of 5-nitroindoline and its nonacid homomorph 1-Me-5-nitroindoline	200	0.828	0.126
28	SPP^N	calculated from the UV−vis spectra of 2-(dimethylamino)-7-nitrofluorene and its homomorph 2-fluoro-7-nitrofluorene	100	0.870	0.058
29	Z	transition energies for the charge transfer band of the complex from 1-ethyl-4-methoxycarbonylpyridinium iodide	60	0.906	2.730
30	α	solvatochromic parameter of solvent HBD (hydrogen-bond donor) acidity	184	0.773	0.204
31	β	solvatochromic parameter of solvent HBA (hydrogen-bond acceptor) basicity	184	0.756	0.147
32	π^*	solvatochromic parameter—index of solvent dipolarity/polarizability, which measures the ability of the solvent to stabilize a charge or a dipole by virtue of its dielectric effect	216	0.751	0.145
33	π^*_{azo}	bathochromic shifts of 6 azo merocyanine dyes	29	0.914	0.090
34	χ_R	transition energy of merocyanine dye (VII)	58	0.852	1.343
35	$\int^{C6H5F}$	^{19}F NMR shielding parameters of fluorobenzene in infinitely dilute solutions relative to a fixed external standard (20% *p*-difluorobenzene in CCl_4)	23	0.952	0.367
36	$\int_H^{P-NO2}$	^{19}F NMR shielding parameters in *p*-nitrofluorobenzenes	29	0.846	0.394
37	Δ	difference of ^{19}F nucleus shifts of *p*-fluorophenol between that in solvents relative to that in carbon tetrachloride	54	0.770	0.267
38	$\Delta\delta CHCl_3$	shift of pure chloroform relative to that of chloroform in dilute solution	28	0.820	0.234
39	$\Delta\upsilon_A$	perturbation of solvents on the C═O vibration band of acetophenone	27	0.856	1.388
40	$\Delta\upsilon_D$	perturbation of solvents on the O−D vibration band of methanol-*d*	92	0.842	27.62
41	θ_{1K}	polarity of solvent, based on the PCA combined with a cross-validation technique of solvatochromic shift data	80	0.780	0.806
42	θ_{2K}	polarizability of the solvent, based on the PCA combined with a cross-validation technique of solvatochromic shift data	80	0.772	0.118
43	Λ	maximum absorption of electronic spectra of heteroleptic molybdenum complexes	24	0.896	8.052
44	λ_F^{MHN12}	fluorescence band maxima for MHN12	20	0.861	1.517
45	π_{1^*}	calculated from the frequency shifts of the electronic absorption spectra of *N*,*N*-dimethyl-4-nitroaniline	95	0.873	0.134
46	π_{2^*}	calculated from the frequency shifts of the electronic absorption spectra of naphthalene	72	0.941	0.043
47	υ_{CE}	relative IR frequency shifts for chloroethane	22	0.881	1.506
48	$\Delta\nu_{CI}$	IR frequency shifts of iodine cyanide C−I bonds	66	0.726	11.46
49	$\Delta\nu_{OH}$	IR frequency shifts of the phenol hydroxyl group	66	0.794	80.40
50	C_p-SCS	substituent chemical shifts for the para carbon of *N*,*N*-dimethyl for the monosubstituted (n = 21) benzene	5	0.806	0.266
51	CTTS	absorption maxima of iodide ions	16	0.941	751.2
52	H	molar concentration of OH dipoles in (55.4 M) TEMPO	11	0.999	0.011
53	Kq^{MMA}	quenching rate constants for the deactivation of triplet thioxanthone by methyl methacrylate	12	0.906	8.112
54	log γK_c	fluorescence quenching rate constants for naphthonitrile-olefin and furan pairs	11	0.975	0.056
55	m*	NMR chemical shift of free base and protonated base	9	0.949	0.043

Table 10. Continued

no.	solvent scale	physical background of solvent scale	N	R^2	s
56	pK_{BH+}	NMR chemical shift of free base and protonated base	9	0.926	0.378
57	XX	solvent induced frequency shifts of SO_2	20	0.923	0.745
58	$\int_N^{pyrrole}$	nitrogen NMR shieldings of pyrrole referred to neat nitromethane	13	0.924	1.282
59	δ	chemical shifts of Li nucleus in different solvents	11	0.922	0.428
60	δ_0	^{23}Na chemical shifts of sodium iodide in different solvents	15	0.853	2.763
61	λ_A^{Na}	maximum absorption wavenumbers for charge transfer bands of $Na_3(acpy)Fe(CN)_5$	7	0.879	1.067
62	T	vibrational cooling times of azulene by picosecond spectral study	7	0.960	2.479
63	λ_A^{MS}	absorption band maxima for MS	13	0.912	0.555
64	Φ_f^{BBVB}	fluorescence quantum yield of BBVB	10	0.919	0.063
65	φ_f^{CEA}	fluorescence quantum yield of coelenteramide	8	0.917	0.037
66	Φ	position of the $n-\pi^*$ transition of a set of ketones	23	0.952	0.041
67	B-2	acid−base hydrogen bond formation induced shifts of the phenol OH group stretching frequency	113	0.798	74.36
68	C_B	susceptibility to covalent interaction of a base statistical from ΔH data of different bases and acids	65	0.801	0.665
69	D_H	ΔG of the transfer of Na^+ from solvent to reference solvent (1,2-dichloroethane) for hard acceptors	24	0.828	7.273
70	E_B	susceptibility to electrostatic interaction of a base statistical from ΔH data of different bases and acids	65	0.780	0.298
71	PA	calculated from equilibrium constants for various gaseous proton-transfer reactions with various solvents	20	0.954	2.744
72	$\Delta_{acid}H$	calculated by measuring the difference between the solvation enthalpies of *N*-methylimidazole and *N*-methylpyrrole along with SPP scale values	63	0.826	3.482
73	ΔH_v	experimental enthalpy of vaporization	22	0.923	707.3
74	$\Delta H°_{solv}$	linear combination of the $\Delta H°_{solv}$ for the four probes (pyrrole, *N*-methylpyrrole, benzene, and toluene)	35	0.845	2.430
75	$\varepsilon°$ (SVB)	average equilibrium and chromatographic distribution constants on Amberlite *XAD*-2, *SM*-2, and *XAD*-4	29	0.816	0.030
76	$-\Delta H°_{BF3}$	enthalpy of complexation of solvents with BF_3 in dichloromethane	76	0.812	12.00
77	μ	difference between the mean of the Gibbs free energies of transfer of sodium and potassium ions from water to a given solvent and the corresponding quantity for silver ions divided by 100	34	0.817	0.167
78	D1	$\Delta G°$ between *cis*- and *trans*-2-isopropyl-5-methoxy-1,3-dioxane	16	0.906	0.121
79	a_H	calculated $E_{(H\text{-bond})}$ values from the enthalpies of solvation for 7 solute−solvent systems	13	0.972	0.055
80	log K	stability constants of sodium complexes with DITHIA-18C6	6	0.892	0.396
81	Sp	solvophobic parameter, calculated from the energies of solute transfer from water to solvents	12	0.990	0.010
82	$-\Delta S_S°$	experimental values of entropy of solvation of electrolyte NaBr	8	0.985	1.852
83	X	solubility of *trans*-stilbene in organic nonelectrolyte solvents	28	0.923	0.002
84	DN	donor number, negative ΔH value for the 1:1 adduct formation between $SbCl_5$ and the solvent molecules in a dilute solution of 1,2-dichloroethane	110	0.763	6.267
85	D_π	second order rate constants for the reaction of DDM and TCNE	34	0.754	0.341
86	log k_{DC}	rate constants for decarboxylation of 3-carboxybenzisoxazoles	24	0.912	0.699
87	Rp	rate constants of pyridine-catalyzed decomposition of *tert*-butylperoxy formate in various solvents at 90 °C	19	0.967	7.242
88	A	anion solvating tendency	54	0.944	0.066
89	Ap	acidity parameter calculated from the data for the Gibbs solvation energy for the alkali metal cations and halide ions	18	0.956	1.794
90	BB′	cation solvating tendency	55	0.772	0.160
91	Bp	basicity parameter calculated from the data for the Gibbs solvation energy for the alkali metal cations and halide ions	18	0.847	0.417
92	d	dielectric constants	55	0.926	5.881
93	DC	calculated from thermodynamic model of protein denaturation	22	0.948	6.481
94	E	acidity derived from E_T and P and Y	84	0.920	1.462
95	J	expression of dielectric constant	57	0.846	0.090
96	log K	solvatochromic parameter α calculated from other solvent scales	27	0.931	0.131
97	log L^{16}	based on the logarithmic gas−liquid partition coefficient in *n*-hexadecane	167	0.969	0.223
98	log P	partition coefficient, calculated from hydrophobic fragmental constants	104	0.950	0.578
99	M	expression of refractive index	57	0.921	0.007
100	N	dielectric function	57	0.847	0.094
101	P′	chromatography strength	78	0.849	0.854
102	q−	electrostatic HBAB	28	0.840	0.055
103	q+	electrostatic HBDA	29	0.938	0.017
104	S	derived from Kosower's Z values, uses R for process sensitivity	46	0.896	0.042
105	S′	solvent polarity, derived from experimental observations $\Delta\chi = PS' + W$	46	0.901	0.164
106	V_{mc}	molecular volume	29	0.993	0.028
107	Xd^R	selectivity parameter: reflects a composite of solvent dipolarity−polarizability, hydrogen bond basicity, and hydrohen bond acidity	52	0.880	0.025
108	Xe^R	selectivity parameter: reflects a composite of solvent dipolarity and solvent acidity	52	0.849	0.042
109	Xn^R	selectivity parameter: reflects predominately solvent dipolarity, with small contributions from hydrogen bond basicity and acidity	52	0.842	0.026
110	X_d	proton donor index	72	0.819	0.028
111	X_e	proton acceptor index	72	0.822	0.041
112	X_n	strong dipole	72	0.810	0.029
113	ΔG_6 Å	calculated energy necessary to form a cavity of appropriate size for a solute which has a diameter 6 Å from the effective hard-sphere diameter of solvent	25	0.790	0.881
114	ΔH^{acid}	calculated from the enthalpies of solution of two probes: *N*-methylimidazole and *N*-methylpyrrole and relative permittivity	36	0.893	2.480
115	$-\Delta H_f$	heat of formation for the hydrogen-bonded complexes between *p*-fluorophenol and solvents which act as base	53	0.816	0.610
116	$\varepsilon°_{alumina}$	normal phase solvent eluotropic strength ($\varepsilon°$) using alumina adsorbent calculated from other solvent parameters (π^*, α, and β)	23	0.948	0.056

Table 10. Continued

no.	solvent scale	physical background of solvent scale	N	R^2	s
117	$\varepsilon°_{\text{silica}}$	normal phase solvent eluotropic strength ($\varepsilon°$) using silica adsorbent calculated from other solvent parameters (π*, α, and β)	19	0.947	0.047
118	ε_β	covalent HBAB	29	0.873	0.004
119	$\Theta(\in_B)$	expression of dielectric constants	39	0.918	0.035
120	μ_D	dipole moments	39	0.941	0.331
121	π_I	polarizability index	29	0.932	0.003
122	σ_1	calculated from the surface tension of hard sphere liquids	25	0.985	0.139
123	ε_α	covalent HBDA	29	0.991	0.002
124	γ_{SO2}	experimental infinite dilution activity coefficients of SO_2	17	0.835	0.061
125	δ_H	square root of cohesive energy density	30	0.928	2.038
126	Y	polarity expression of dielectric constant	66	0.921	0.023
127	P	polarizability expression of refractive index	66	0.961	0.004

[a] Reprinted with permission from ref 686. Copyright 2005 American Chemical Society. *N* is the number of data points, R^2 is the squared correlation coefficient, and *s* is the standard deviation.

understanding of the micellar solution, but it cannot provide information on the size of the micelle.

With increased computational power and the development of modern QSAR/QSPR approaches, powerful methods for the prediction of CMC have eventually become available. Their big advantage is that they do not require any experimentally determined values or empirical constants.

Employing the general QSPR approach, Huibers et al.[700] proposed a three-parameter QSPR model for a set of 77 nonionic surfactants ($R^2 = 0.983$, $F = 1433$, $s = 0.177$) using only topological descriptors calculated for the hydrophobic fragment of the surfactant molecule. The three descriptors represent contributions from the size of the hydrophobic group, the size of the hydrophilic group, and the structural complexity of the hydrophobic group.

In later studies, Huibers et al. reported a three-parameter QSPR model from 119 anionic surfactants (sulfates and sulfonates).[701] The log CMC values were found to correlate well ($R^2 = 0.940$, $F = 597$, $s = 0.217$) with the Kier and Hall index (zeroth order) calculated for the tails, the relative number of the carbon atoms in the head, and the total dipole of the molecule.

The CMC values of 46 octyl phenol and linear and branched alkyl chain oxyethylene derivatives with different numbers of carbon atoms in the hydrophobic groups were studied by Kuanar et al.[702] Only purely topological descriptors derived from the chemical graph theory were used, and a PCA model with $R^2 = 0.993$ and $s = 0.134$ was proposed to predict the CMC of nonionic surfactants.

Roberts correlated the CMC of 133 anionic surfactants including ether sulfates and ester sulfonates with good results ($R^2 = 0.976$, $F = 5360$, $s = 0.12$) using the hydrophobic parameter π_h for the hydrophobic domain of the surfactant and the length of the hydrophobe.[703]

The QSPR treatment of 40 anionic surfactants studied by Wang et al.[704] led to a six-descriptor model with $R^2 = 0.978$. The descriptors involved are as follows: the Kier and Hall index of zeroth order KH0 calculated for the hydrophobic fragment, the total molecular energy E_{total}, the heat of formation ΔH_f, the dipole moment D, and the energies of frontier orbitals—E_{LUMO} and E_{HOMO}—of the surfactants. The same strategy was later applied to predict the logarithm of the critical micelle concentration of 77 nonionic surfactants in aqueous solution.[705] The best QSPR model ($R^2 = 0.986$) involved seven molecular descriptors: ΔH_f, D, E_{LUMO} and E_{HOMO}, MW, the number of the oxygen and nitrogen atoms (*n*ON) of the hydrophilic fragment, and KH0 for the hydrophobic fragment.

Li et al.[706] developed a general QSPR model for 98 anionic surfactants using the RHF ab initio method and 6-31G(d) basis functions to optimize the molecular structures. They reported a three-parameter regression equation with good statistical characteristics ($R^2 = 0.980$, $R^2_{\text{cv}} = 0.978$, $F = 1505$, $s = 0.103$) which involves variables such as the total number of atoms in the hydrophobic−hydrophilic segment, the maximum atomic charge on the carbon atom, and the dipole moment.

Based on the simple harmonic vibration model of mechanic vibration theory, Ming-Hua et al.[707] calculated the inherent frequencies of 40 anionic surfactant molecules viewed as multifreedom spring-mass vibration systems. Based on the 2D representation of the molecular structure, fundamental frequency (ω_0), and sum-frequency ($\sum\omega_i$), QSPR models having the following general representation (eq 17) were proposed:

$$\log \text{CMC} = A_0 + A_1\omega_0 + A_2\sum\omega_1 \tag{17}$$

All the models were characterized by correlation coefficients $R^2 > 0.98$ and a mean relative error less than 0.0316.

Elshafie et al.[708] published a MLR model using a data set of 50 nonionic surfactants. The best five-descriptor model ($R^2 = 0.9889$, $F = 391.6$, $s = 0.486$) was selected. The descriptors involved are as follows: the molecular weight, hydrophobic/hydrophilic fragments molecular weight ratio, polarizability, log *P*, and the hydration energy.

Katritzky et al.[709] proposed a general QSPR model for a wide range of sodium salts, potassium alkanecarboxylates, and *p*-isooctylphenol ethoxylated phosphates. Correlation was studied using a data set of 181 anionic surfactants of CMC values measured at 40 °C with molecular descriptors calculated by CODESSA PRO. A five-parameter model was obtained involving descriptors calculated for the whole molecule and for the hydrophobic and hydrophilic fragments separately. The reported statistical parameters are as follows: $R^2 = 0.897$, $R^2_{\text{cv}} = 0.877$, $F = 303.7$, $s = 0.295$.

5.8.2. Cloud Points

The cloud point is an important property of nonionic surfactants. Below this temperature a single phase of molecular or micellar solution exists; above it the surfactant has reduced water solubility, and a cloudy dispersion results. A general MLR model ($R^2 = 0.937$, $s = 6.5$) has been developed for estimating the cloud point of pure nonionic surfactants of alkyl ethoxylates using only topological descriptors.[710] The set of 62 structures is composed of linear alkyl, branched alkyl, cyclic alkyl, and alkylphenyl ethoxylates. For this set the cloud points can be estimated to an accuracy of ±6.3 °C (3.7 °C median error) using the logarithm of the number of ethylene oxide residues and three topological descriptors that account for the hydrophobic domain variation. The topological descriptors model various aspects of the hydrophobic tail structure.

5.9. Cyclodextrin Complexation Free Energies

Cyclodextrins (CDs) are cyclic oligomers of α-D-glucose which result from the action of certain enzymes on starch. The family includes three well-known industrially produced members—α-CD (six glucose units), β-CD (seven units), and γ -CD (eight units)—as well as several other less well-known oligosaccharides. The α-, β-, and γ-CDs, commonly referred to as the native cyclodextrins, are crystalline, homogeneous, nonhygroscopic substances which form cylindrical or doughnut-shaped molecules with their OH groups on the outside of the molecule. CD molecules are shallow truncated cones rather than toruses. The primary hydroxyl rim of the cavity opening possesses a somewhat reduced diameter compared with the secondary hydroxyl rim.

The CD exterior, containing many OH groups, is fairly polar, whereas the interior of the cavity is nonpolar relative to water, which is the usual external environment.[711] In principle, in aqueous solution, the slightly apolar CD cavity is occupied by water molecules, which are energetically unfavored (polar−apolar interaction) and can therefore be readily substituted by appropriate ”guest molecules” which are less polar than water. The dissolved CD is the host molecule, and the driving force of the complex formation is the substitution of the high-enthalpy water molecules by an appropriate guest molecule. This host−guest property allows CDs to be used in numerous applications in industrial, pharmaceutical, agricultural, and other fields, including improving the solubility and stability of drugs and selectively binding materials that fit into the central cavity in affinity and chromatography purification methods.[712,713]

Katrizky et al.[714] correlated free energies of complexation of β-cyclodextrins with molecular descriptors calculated using CODESSA PRO, and fragment descriptors calculated by the TRAIL program.[715] A seven-parameter equation was obtained with $R^2 = 0.796$, $R_{cv}{}^2 = 0.779$, and $s = 1.542$. However, for a data set of 195 compounds (with exclusion of 23 compounds), better results were obtained ($R^2 = 0.943$, $R_{cv}{}^2 = 0.848$, and $s = 1.65$) by using 79 fragmental descriptors. The two approaches individually and in combination led to statistically stable and predictive QSPR models.

5.10. UV Spectral Intensities

High performance liquid chromatography (HPLC) combined with ultraviolet (UV) spectrophotometric detection is applied widely in organic chemistry for analyzing reaction products.[716] UV is also considered a nearly universal detector for druglike molecules: 85% of the structures in the MDDR (a database of drugs and candidate drugs) contain an aromatic group, and most of the remaining 15% contain another chromophore. A computational method for prediction of the relative response of organic molecules in the UV region of spectra would therefore be of benefit to researchers.

In recent years, however, several authors have reported the prediction of electronic absorption parameters using quantum theory.[717−726] A robust method for the calculation of UV spectra has been the ZINDO modification[727−729] of INDO (intermediate neglect of differential overlap), which works well with extended conjugation for many other organic systems, excluding the systems containing nonbonded electrons. Ab initio predictions of UV spectra have also been carried out by using highly correlated methods such as configuration interaction singles (CIS) or time dependent density functional theory (TD-DFT)[730] combined with high order basis sets (cc-pVTZ + sp and B3LYP,[731] respectively), but these calculations require powerful computing capacity and are very time-consuming for relatively large molecules. Therefore, fast QSPR approaches which could be extended to larger molecules consisting of hundreds of atoms would possess significant advantages. It is understood that correlation of UV spectra intensities in terms of extinction coefficients at a certain wavelength has no sound basis in theory, since such coefficients depend in complex ways on the location and intensities of spectral maxima.

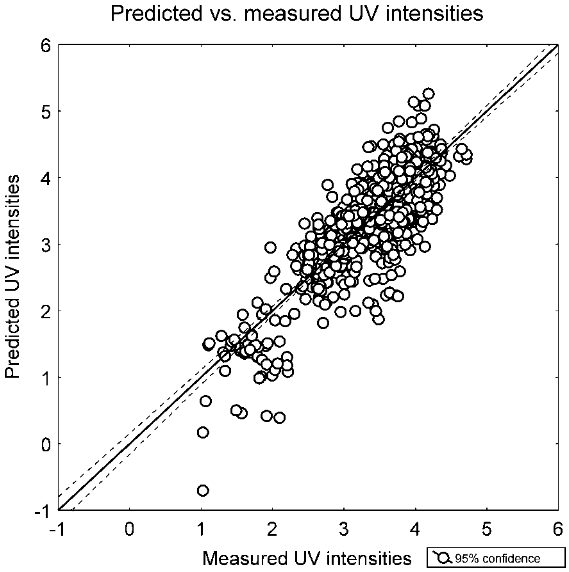


Figure 11. Predicted vs experimental UV absorption intensities at 260 nm in water. Reprinted with permission from ref 138. Copyright 2007 Springer.

Only a few QSPR treatments of UV intensities have been reported on sets of derivatized molecules based on common chromophores.[732] Fitch et al.[716] and Molnar and King[733] correlated UV intensities with structural descriptors.

Recently, MLR and backpropagation feed-forward ANN approaches allowed the correlation of UV absorbance at 260 nm of a lage data set of 805 organic compounds.[138] The authors reported a correlation with $R^2 = 0.692$ and $s = 0.537$ log unit for UV absorption intensities at 260 nm and 25 °C in water of a set of 805 organic compounds by using five structural descriptors calculated by CODESSA PRO software (Figure 11). Consequently, a corresponding nonlinear model was developed and validated using the external data. The descriptors (square root of partial surface area (MOPAC PC) for atom C, relative number of double bonds, HOMO−LUMO energy gap, moments of inertia C, and average information content order 1) involved are calculated solely from chemical structure and possess definite physical meaning related to the nature of the process.

6. Chemical Properties

6.1. Lithium Cation Basicities

The measurement of lithium cation basicities helps the understanding of fundamental interactions implied in analytical mass spectrometry, organic synthesis, catalysis, lithium battery electrochemistry, and cation transport through ion channels. In general, the activity of Li cations toward ligands controls the formation of adducts, or clusters, that can be considered as ions "solvated" by one or more ligands. The gas-phase lithium cation basicity (LCB) is defined as the Gibbs free energy associated with the thermodynamic equilibrium of eq 18.

$$B + Li^+ \overset{K_1}{\Leftrightarrow} [B - Li^+] \qquad (18)$$

where $\Delta G_{Li+} = -RT \ln K_1$ and $LCB = -\Delta G_{Li+}$.

Tämm et al.[734] have used QSPR to study gas-phase LCBs of 205 compounds. The BMLR (best multilinear regression) method implemented in CODESSA PRO was used to extract six theoretical descriptors, explaining most of the data variance: (i) minimum net atomic charge, (ii) relative number of S atoms, (iii) energy of the orbital lying below HOMO, (iv) total point-charge component of the molecular dipole, (v) total molecular surface area weighted positive surface area, and (vi) surface charge weighted area of hydrogen-bonding acceptor atoms. The internal leave-one-third-out procedure was used for validation. A statistically significant model with $R^2 = 0.801$, $R_{cv}{}^2 = 0.785$, $F = 133.1$, and $s = 2.96$ was reported. Charge related descriptors dominate, confirming the electrostatic nature of the Li cation−base interactions.

Jover et al.[735] developed linear and nonlinear QSPR models to relate the LCBs of 229 structurally diverse compounds to calculated molecular descriptors. The best model was obtained with ADAPT neural networks (NN). A genetic algorithm routine using a NN fitness evaluator was applied to a 7−5−1 architecture for the descriptor selection. A seven descriptor model (training set of 166 and test set of 19 compounds) involving the numbers of hydrogen, oxygen, and nitrogen atoms, the HOMO-1 energy, the total dipole of the molecule, the total molecular electrostatic interaction divided by the number of atoms, and the surface charge weighted area of the hydrogen-bonding donor atoms, HDCA-2, was reported. The statistical parameters for the training and the test set were $R^2 = 0.954$, RMSE = 6.54 and $R^2_{test} = 0.914$, RMSE = 8.61, respectively. Compared with high level ab initio and DFT results, for the same compounds, the QSPR approach showed better predictions, especially for the diverse set of compounds.

6.2. Stability Constants

Stability constants are associated with the formation of chemical complexes in equilibrium reactions. These constants are a measure of the stability of complex formation, usually obtained by the reaction $mA + nB \leftrightarrows [A]_m[B]_n$, and are functions of both the reactants and products. They are very often involved in computational models for various physicochemical properties of reaction products as well as experimental assessments of such properties. Their application in many areas of chemistry illustrates the need for stability constant values. However, these are not always experimentally available, and thus, QSPR can be useful for prediction of stability constants. There are numerous QSPR reports in the literature related to the assessment of stability constants, and several illustrative examples are discussed below.

Toropov and co-workers developed QSPRs for calculating stability constants using an optimization of correlation weights (OCW) of local graph invariants approach. Models were derived for data sets of transition metal complexes with ammonia or ethylene diamine[736] and for different sets of complexes of biometals (Mg^{2+}, Ca^{2+}, Mn^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, and Co^{2+}) with adenosine mono-, di-, and triphosphates.[737] This optimization used molecular graphs (MG), whose vertices were atoms and AO graphs (AOG) whose vertices were AOs (1s, 2s, 2p, and others). It was established that (i) the AOG-based models were more accurate than those based on MG and that (ii) the models obtained by OCW of the Morgan degrees of vertices of MG/AOG were more accurate than those based on normal degrees of vertices. Morgan degrees refer to the employment of the extended connectivity (EC) of Morgan[738] in calculation of the descriptors. The extended connectivity of an atom is specified as the sum of the connectivities of the neighboring atoms in an iterative procedure which ends when the same atom ordering results in two consecutive iterations.[739] The statistical characteristics of the best model (OCW of the first-order Morgan degrees of AOG) were $n = 20$, $R^2 = 0.971$, $s = 0.28$, and $F = 608$ (learning sample) and $n = 20$, $R^2 = 0.990$, $s = 0.196$, and $F = 1691$ (control sample). In subsequent work, a descriptor calculated from correlation weights of the main quantum number, orbital quantum number, number of electrons on the atomic orbital, and Morgan degrees of the second-order vertices in the graph of atomic orbitals (GAO) was proposed.[740] The quality of the models was verified by reference accesses. This approach allowed prediction of the stability constants of the Ca^{2+}, Cu^{2+}, and Zn^{2+} complexes with adenosine phosphate derivatives using teaching accesses containing no complexes of these metals. The hydrogen bond index (HBI), the global invariant of a molecular graph that equals the number of vertices representing hydrogen and nitrogen atoms, was considered as a measure of the capability of a complex to form hydrogen bonds. Together with the local graph invariants, HBI was used for the OCW QSPR modeling of the stability of 110 biometal M^{2+} complexes with α-amino acids and phosphate derivatives of adenosine.[741] The statistical parameters of the best model reported were $n = 55$, $R^2 = 0.984$, $s = 0.279$, and $F = 3328$ (learning sample) and $n = 55$, $R^2 = 0.986$, $s = 0.248$, and $F = 4027$ (control sample). Finally, the stabilities of 150 complexes containing adenosine derivatives, α-amino acids, and other biological ligands based on OCW of the nearest neighborhood codes (NNC), the HBI, and the cyclicity code (CC) were described.[742] The NNC is a local topochemical invariant of a vertex of the MG whose numerical value is a function of the total number and the composition of vertices adjacent to the given vertex. The CC is a global topological invariant of the graph equal to the number of rings present in the ligand structure. The statistical characteristics of the best model proposed were as follows: $n = 75$, $R^2 = 0.949$, $s = 0.457$, $F = 1337$ (training sample); $n = 75$, $R^2 = 0.960$, $s = 0.461$, $F = 1724$ (test sample).

The performances of several popular modeling techniques, associative neural networks (ANN), support vector machines (SVM), k nearest neighbors (kNN), maximal margin linear programming (MMLP), the radial basis function neural network (RBFNN), and MLR were compared by Tetko et al.[743] A QSPR of the stability constants, log K_1, for the 1:1 (metal/ligand) complexes and log$^{\beta}{}_2$ for the 1:2 complexes of the metal cations Ag^+ and Eu^{3+} with diverse sets of organic molecules in water at 298 K at ionic strength 0.1 M was obtained. The methods were tested on three types of descriptors: molecular descriptors including E-state indices, counts of atoms determined for E-state atom types, and substructural molecular fragments (SMF). The models were compared using a 5-fold external cross-validation procedure. The Wilcoxon signed-rank test was used to compare the performance of these methods. It is a useful nonparametric alternative to the paired *t* test, which is similar to the Fisher sign test. This test assumes that there is information in the magnitudes of the differences between paired observations, as well as the signs.[744] Estimating this test, the nonlinear QSPR methods demonstrated a significantly better performance than the models built using MLR analysis. However, the averaging of several MLR models based on SMF descriptors provided predictions as good as most of the efficient nonlinear techniques. SVM and ANN produced the largest number of significant models. Models based on fragments (SMF descriptors and E-state counts) demonstrated higher predictive ability than those based on E-state indices. The use of SMF descriptors and E-state counts provided similar results, whereas E-state indices lead to less significant models. The study illustrated the difficulties of quantitative comparison of different methods: conclusions based only on one data set without appropriate statistical tests could be wrong.

In a significant contribution, Solov'ev et al.[745] applied the molecular fragment contribution method to model the stability constants (log *K*) of the complexes of strontium(II) with organic ligands in water. In this work a data set of 130 ligands which were separated into different substructural fragments based on the ISIDA approach[746] were used. The models were utilized for the generation and screening of a combinatorial library of virtual ligands. Several good models were derived based on the fragment descriptors ranging from $R^2 = 0.91$ to 0.94. Based on these models the authors suggested the construction of new, potentially good binders. They concluded that O—C—C═O, N—C—C—N, N—C—C═O, and N—C—C—O fragments largely contribute to log *K*.

Recently, significant progress was reported by Ghasemi and Saaidpour[747] on the stability constants of 58 complexes of 1,4,7,10,13-pentaoxacyclopentadecane ethers. Their best model of five descriptors was able to correlate the experimental and predicted stability constants with $R^2 = 0.95$. In addition, 12 complexes were used as an external validation set for which the prediction gave $R^2 = 0.92$. The model descriptors were related to the specific charges of H atoms, showing their importance for the interpretation of the formation of the complexes. In addition, interactions between C−H and C−C atom pairs and charge distributions were indicated as important by the descriptors. The symmetry and shape of the complexes were also accounted for by the model parameters. The work of Ghasemi and Saaidpour clearly demonstrates that a QSPR equation can be developed for the interpretation of complexation processes. It should be emphasized that a comprehensive QSPR model requires consideration of conformational changes upon metal binding, solvation of the coordinated ligand molecule and side chain, or lariat effects.

There is considerable interest in the synthesis, structure, and luminescence or magnetic resonance spectral properties of novel binuclear compounds exhibiting electronic lanthanide coupling ($Ln^{3+}-Ln^{3+}$). For instance, the potential for such couplings to produce unusual tunable electronic behavior to generate sharper image contrasts in magnetic resonance (MRI) and fluorescence imaging continues to encourage interest in these compounds. Significant water solubility and stability of some binuclear lanthanide(III) compounds also make them attractive as biomedical agents. For example, free Gd(III) ion is extremely toxic at the concentrations needed for MRI studies. However, being administered in the form of stable complexes, the metal ion is not released while in the human body. The stabilities of complexes are also very important for the development of efficient separation methods for lanthanides, as separability depends on the stability constants of the complexes. The above-listed applications require the development of lanthanide chelates with carefully tailored chemical, structural, and spectroscopic (or magnetic) properties, which in turn can be found with the aid of QSPR studies.

Svetlitski and Karelson[748] developed a QSPR model for the stability constants, K_1, of complexes between 63 different organic ligands and 14 lanthanides with the BMLR method implemented in CODESSA. The stability constant, K_1, measured in aqueous solutions at the ionic strength $\mu = 0.1$ and temperature 25 °C, is defined as in eq 19:

$$K_1 = \frac{[Ln\,L^{n+3}]}{[Ln^{3+}][L^n]} \tag{19}$$

Models for the series involving a single metal were constructed using only theoretical descriptors for the ligands. QSPR models for the series involving a constant ligand were constructed using various physical properties of metals as descriptors. The models contained two to four descriptors from a variety of classes. The largest groups included hydrogen bonding descriptors, topological indices of the organic ligands, general electronic properties, and bonding interactions. In addition, descriptors reflecting the geometry and constitution of ligands and partial surface areas appeared in the QSPR models. The frequency of the descriptors in QSPR models indicated that bidentate complex formation with the lanthanide ions is predominantly determined by the hydrogen-bonding capabilities and the geometrical and even topological aspects of the ligands. The descriptors reflecting the charge distribution in the ligands and the related electrostatic interactions had smaller contributions. In the case of the correlations with the lanthanide (metal) descriptors, the most important contribution was given by the successive ionization potentials of the metals, that appeared altogether 42 times, of which 18 cases involved the ionization potential of the Ln^{3+} ion. Another group of descriptors of substantial importance included the heats of vaporization and fusion of the metals. In principle, these descriptors depend on the London forces between the metal atoms and may thus reflect similar noncovalent interactions in the complexes. Most of the models were characterized by $R^2 > 0.8$ and prediction of an external test set with $R^2{}_{ext} = 0.588$.

Further, a QSPR modeling of the distribution coefficient (log *D*) of uranyl cations extracted by phosphoryl-containing podands from water to 1,2-dichloroethane was reported by Katritzky et al.[749] Two different approaches were used: one based on classical physicochemical descriptors (implemented in the CODESSA PRO program) and another based on fragment descriptors (implemented in the TRAIL program[715]). Taking into account the conformational flexibility of podands, only conformationally invariant or weakly conformationally dependent descriptors were used in the CODESSA PRO calculations. Several robust models were obtained from CODESSA PRO which involved its "own" descriptors together with fragment descriptors generated by TRAIL. Using TRAIL alone, three statistically significant models involving sequences of atoms, bonds, or augmented atoms were developed. The QSPR models obtained were applied to the estimation of log *D* values for a virtual combinatorial library of 2024 podands generated with the CombiLib program. Eight of these hypothetical compounds which span the range of log *D* variation for experimentally studied molecules were then synthesized and tested experimentally. Comparison of calculated and new experimental results showed that the QSPR models successfully predicted log *D* values for 7 of the 8 compounds from the "blind test" set.

6.3. Rate Constants

6.3.1. Decarboxylation Rates

Decarboxylation involves cleavage of a C−C bond in a carboxylate ion to produce CO_2 and an organic residue containing an unshared pair of electrons. This organic product can be stabilized by delocalization of the electron pair. In some cases, as in the tetramethyl guanidinium salt of 3-carboxy-6-nitrobenzisoxazole (**I**), the carboxylation process is greatly influenced by the nature of the reaction environment. This phenomenon has attracted attention with respect to applications in biological and organic synthetic fields as well as for probing solvents and varied media such as micelles, bilayers, macrocyclic hosts, and polymers.[750] The authors derived equations for 24 pure solvent scales ($R^2 = 0.870$, $s = 0.73$) and for 60 pure and mixed solvents ($R^2 = 0.904$, $s = 0.60$) for decarboxylation rates (log *k*) of **I** using their own experimentally determined solvent scales. A year later the effects of solvents on the decarboxylation rates of **I** in 24 pure solvent scales was studied by Katritzky et al.[751] employing the CODESSA program and using theoretical descriptors related to the solvents. The three-parameter correlation ($R^2 = 0.909$, $R_{cv}{}^2 = 0.870$, $F = 66.21$, $s = 0.712$) relates the log *k* values of the rate of decarboxylation of 6-nitrobenzisoxazole-3-carboxylates to (i) the hydrogen acceptor accessible surface area, HASA, (ii) the structural information content (order 1), ^{1}SIC, and (iii) the image of the Onsager−Kirkwood solvation energy, SE_{OK}. According to these results, H-bonding interactions of the carboxylate with the solvent impede the decarboxylation reaction by stabilizing the ground state of the carboxylate ion. The rate of the reaction also depends on branching of the solvent, which affects the interactions of the solvent with the substrate and/or transition state and the polarity, as revealed by the contribution of the SE_{OK}. With an increase in the value of this descriptor, the energy of activation for the decarboxylation process diminishes.

6.3.2. Hydroxyl Radical Rate Constants

Volatile organic compounds (VOCs) are chemically transformed in the troposphere by reacting with photochemically generated oxidants. The lifetimes of organic chemicals can be calculated from the rate constant of their degradation reaction with OH radicals, k_{OH}, and ozone during the daytime and NO_3 radicals at night. The hydroxyl radical reacts with practically every organic compound in the troposphere and has been studied extensively, providing sufficient experimental data for QSPR modeling. These models could help in rapid recognition of safe or high risk organic chemicals and be useful in planning and development of new safer organic chemicals. Advances in the QSPR study of atmospheric degradation of chemicals are quite considerable, while modeling of biodegradability in water and soil has produced very modest results according to the comprehensive overview of the degradability of organic compounds by Sabljić et al.[752]

Hydroxyl radical reactions include the following: (i) hydrogen atom abstraction, (ii) addition to double and triple bonds, (iii) addition to aromatic rings, and (iv) reactions with nitrogen, sulfur, and phosphorus compounds. The environmental importance of these degradation pathways and the modeling methods has been reviewed and evaluated by Güsten et al.[753,754] Published prediction models on abiotic tropospheric degradation can be grouped into (i) empirical models, QSPRs using measured physicochemical properties, (ii) QSPRs based on semiempirical quantum-chemical descriptors, and (iii) ab initio MO calculations. Among these models only a few were generally applicable. Meylan and Howard[755] noted two general methods for the prediction of the tropospheric OH radical degradation rates. First, the Atkinson's group/fragment methodology combined with known reaction mechanisms is implemented in US EPA's AOPWIN estimation software[756,757] comprising 89 parameters derived by nonlinear least-squares analyses of the kinetic data. The total rate constant is the sum of the four reaction pathways (i)−(iv) mentioned above. Despite the fact that the method has no mechanistic background and has applicability limitations, it is quite successful and not class-specific. The other outstanding model, the so-called MOOH model, was made by Klamt,[758,759] who used the AM1 semiempirical SCF-MO method to derive six molecular descriptors combining MO energies and atomic charges on the reaction centers of the molecules. A nonlinear optimization procedure was performed to obtain regression coefficients. The obtained system of models covers reaction pathways (i)−(iii), including oxygen-containing compounds.[759]

An approach similar to Atkinson's was used by Neeb.[760] Additive group rate constants were determined for each structural group or sites of attack. For the classes of compounds considered in this study, only pathways (i) and (ii) were considered important.

Gramatica, Pilutti, and Papa developed models for the atmospheric degradation of VOCs,[761] using the GA-VSS (genetic algorithm-variable subset selection) strategy for the selection of significant variables from a large set of structural, topological, empirical, and WHIM (weighted holistic invariant molecular) descriptors followed by an OLS (ordinary least squares) method for model formation of OH radical reaction rate constants, −log k_{OH}. PCA and the Kohonen artificial neural network (K-ANN) were used to extract representative training (51 chemicals) and validation (94 chemicals) sets from the initial set of 326 chemicals. The best six-descriptor model was characterized by $R^2 = 0.85$ for the training set and $R^2 = 0.75$ for the validation set. Later, with the same variable selection methodology,[762] the authors proposed a MLR model for OH radical tropospheric degradation of 460 heterogeneous VOCs from theoretical molecular descriptors. D-optimal Experimental Design and K-ANN were applied to the original data set for splitting the data into training and validation sets. The resulting two MLR models involved the same descriptors (HOMO energy, number of halogen atoms, complementary information content index, and number of unsubstituted aromatic C atoms) with identical R^2 values of 0.828 ($n_{\text{training}} = 234$, $n_{\text{test}} = 226$). Although the performance of the models was very similar, the D-optimal design, where chemicals with higher diversity were selected for the training set, yielded superior models with predictive power exceeding that of the models developed on the basis of the training set selected by K-ANN. The authors emphasized the need for chemical domain determination and external validation for the development of successful predictive models.

Bakken and Jurs[763] developed QSPRs using computational artificial neural networks (ANNs) with structure-based descriptors for the set of 57 unsaturated hydrocarbons previously used by Medven.[764] A 5−2−1 CNN produced a rms error of 0.0705 log unit for the training set and 0.0639 log unit for an external prediction set. The residual sum of squares for all 57 compounds was a favorable 0.234 log unit. Additionally, a 10−7−1 ANN for a diverse set of 312 compounds produced a rms error of 0.229 log unit for the training set and 0.254 log unit for the external validation set. Accurate predictions over a wide range of functionalities were demonstrated.

The same data set[764] was also used by Pompe et al.,[765] who developed a CODESSA MLR model for the $-\log k_{\text{OH}}$ using only topological descriptors. A model with statistical parameters comparable with other QSPRs on the same data was obtained with six descriptors and a rms_{CV} error of 0.119 log unit. Additionally, a regression model using a variable connectivity index (${}^1\chi^{\text{f}}$) was developed. The variable connectivity index accounts for and adjusts the relative contributions of different atoms and bonds to best suit the property studied. The model provided worse cross-validation results with an rms_{CV} error of 0.16 log unit, but it enabled a mechanistic interpretation of the reaction. The largest contribution to the reactivity was given by the cyclic and acyclic sp^2-hybridized carbon atoms, in accord with experimental findings. Since all 58 compounds contained one or more C=C double bonds, the reactions were classified as OH radical addition to multiple bonds. Later, higher-order variable connectivity indices were introduced by Pompe and Randić[766] to account for the combination of positive as well as negative relative contributions of atoms and bonds in the construction of the QSPR/QSAR models. The superiority of the variable connectivity index ${}^1\chi^{\text{f}}$ compared to the simple and valence analogues ${}^1\chi$ and ${}^1\chi^{\text{v}}$ was clearly shown with the aid of a selected data set of 39 organic compounds containing carbon, oxygen, and chlorine atoms with known reaction rates with OH radicals, $-\log k_{\text{OH}}$. The respective rms errors of the models were as follows: 0.62, 0.53, and 0.35. The introduction of anticonnectivity further refined the result, rms = 0.34. The optimization of diagonal weights of the augmented adjacency matrix of the “anticonnectivity model” pointed out the significant enhancing effect of oxygen and the suppressive effect of chlorine on the overall atmospheric reactivity of organic compounds with OH radicals—a valuable result offering direct knowledge about the role of the individual structural components that influence the reactivity of the compounds.

Öberg[767] outlined how validation and domain definition can facilitate the modeling and prediction of the OH radical reaction rates for a large database. A set of 867 theoretical descriptors was generated from the 2D-molecular representation of the structures for compounds presented in the Syracuse Research Corporation’s PhysProp Database to give a QSPR model using PLS regression validated with an external test set. The main factors of variation were attributed to two reaction pathways, hydrogen atom abstraction and addition to double bonds or aromatic systems. When projected onto the PLSR model, 74% of 17,293 compounds with similar molecular weight fell inside the applicability domain determined by the chosen limits for the residual standard deviation and the leverage. The predicted hydroxyl reaction rates for 25% of these compounds were slow or negligible, with atmospheric half-lives in the range from days to years. The list of the persistent organic compounds was matched against the OECD list of high production volume chemicals (HPVC). Nearly 300 compounds were identified as both persistent and/or in high volume production.

More recently, a new method of developing QSPR models based on fuzzy “if−then” rules was demonstrated by Kumar et al.[768] The fundamental issues involved in QSPR studies related to modeling errors associated with the chosen descriptors and structure of the model were addressed. The construction of fuzzy mappings was based on a robust criterion that the maximum possible value of energy-gain from modeling errors to the identification errors was minimum. Such an identification method would guarantee that small modeling errors would not lead to large identification errors. Simulation studies and three QSAR modeling examples provided by the authors illustrated that, in the presence of modeling errors, the proposed fuzzy modeling was more suitable than the Bayesian regularized ANNs. For the example of predicting the rate constant for OH radical tropospheric degradation, $-\log k_{\text{OH}}$, the data set of 460 heterogeneous organic compounds and the model descriptors taken from Gramatica et al.[762] provided a model with rms errors of 0.43 and 0.37 for the training and test sets, respectively (Table 11). This result outperformed both reference models, the Bayesian ANNs and the model by Gramatica et al.

6.3.3. Methyl Radical Addition Rate Constants

Methyl radical reactions are important in many fields of chemistry. Environmentally, methyl radicals are formed in the atmosphere during the reaction of methane with OH radicals and subsequently determine the fate of other chemicals found in the atmosphere. Methane is considered the second most important gas after carbon dioxide that affects the ozone layer.

Methyl radical addition rate constants were modeled by Bakken and Jurs[769] using 191 small organic compounds. Topological, geometrical, electronic (using PM3 Hamiltonian), or combined descriptors were used to encode substrate information. The best results were achieved by nonlinear feature selection combined with CNN model development. Alkynes, allenes, and heterocycles were absent from the data set. A seven-descriptor CNN was built for 172 compounds. The reported rms error for the training set was 0.424 log unit, and the rms error for the prediction set was 0.409 log unit. The error of the predictions of the proposed model was on the order of the experimental error.

Table 11. QSPR Models for Prediction of the Gas-Phase OH Radical Rate Constants ($-\log k_{OH}$)[a]

no.	compounds	N[b]	methods	n_d, model descriptors	R^2	s	R_{valid}^2	s_{valid}	ref
1	update for oxygenated compounds: ketones, alcohols, ethers, carbonic acids, and aldehydes	93	nonlinear optimization procedure	AM1 semiempirical MO calculation parameters		1.6			Klamt[759]
2	unsaturated hydrocarbons	57	PLS	18 (8 empirical, 8 quantum-chemical (AM1), and 2 constitutional), 3 latent variables	0.86				Medven[764]
			MLR, stepwise	3 (E_{HOMO} (AM1), dipole moment (*Dip*), log *P*)	0.82	0.125			
3	unsaturated hydrocarbons	52 (5)	CNN 5-2-1	5 (topological)		0.071[c]		0.064[c]	Bakken and Jurs[763]
	diverse compounds	281 (31)	CNN 10-7-1	10 (8 topological, 2 electronic (PM3))		0.23[c]		0.25[c]	
4	VOCs	51 (94)	GA, OLS	6 (structural, topological, empirical, and WHIM descriptors)	0.85	0.47	0.75		Gramatica[761]
5	alkanes, alkenes, and oxygenated hydrocarbons	250	GAP	21 group rate constants					Neeb[760]
6	unsaturated acyclic and cyclic organic compounds ($C_3...C_{10}$)	53 (5)	MLR	6 (topological)	0.88	0.12[c]		0.097[c]	Pompe[765]
7	heterogeneous VOCs	234 (226)	GA, MLR	4 (HOMO energy, number of halogen atoms, complementary information content index, number of unsubstituted aromatic C atoms)	0.83	0.47	0.83	0.48	Gramatica[762]
8	diverse VOCs	495 (238)	PLSR	333 2D descriptors (using SMILES)	0.91	0.39	0.84	0.50	Öberg[767]
9	organic compounds containing C, O, and Cl atoms	39		${}^1\chi$	0.24	0.62[c]			Pompe[766]
				${}^1\chi^V$	0.45	0.53[c]			
				${}^1\chi^f$	0.76	0.35[c]			
				${}^1\chi^a$	0.77	0.34[c]			
			MLR	6 (topological and constitutional)	0.93	0.21[c]			
10	diverse VOCs (Gramatica 2004)	460	fuzzy mapping	"if-then" rules		0.43[c]		0.37[c]	Kumar[768]
			Bayesian NNs			0.45[c]		0.38[c]	

[a] VOCs, volatile organic compounds; OLS, ordinary least squares; WHIM, weighted holistic invariant molecular. [b] The number of validation set compounds is shown in parentheses. [c] rms error.

A reliable QSPR model for estimation of the rate constants of radical addition reactions was developed by Heberger and Borosy.[770] Carbon-centered radicals with very different features and vinyl-type alkenes with diverse substituents served as reactants. The data set of 178 compounds was split into training (114), monitoring (56), and validation (19) subsets. Linear and nonlinear methods were applied, and the rms error of prediction (RMSEP) was used to compare the predictive power of the methods used. The six important descriptors comprised the following: reaction heat (HR), singlet−triplet energy gap of alkenes (ETR), ionization potential of radicals (IPR), ionization potential of alkenes (IPA), electron affinity of radicals (EAR), and electron affinity of alkenes (EAA). The model exhibited strong nonlinearity: the RMSEP of 0.37 log unit for the logarithm of the rate constant achieved by ANN for the 19 members of the validation set was 70% lower than the RMSEP of the MLR and PCR linear methods.

6.4. Acid Dissociation Constants

The dissociation or acidity constant, K_a, is extremely important in organic chemistry, equilibrium studies, and drug design. K_a measures the propensity of a compound to donate a proton (eq 20):

$$K_a = \frac{[H_3O^+][A^-]}{[HA]} \qquad (20)$$

where $HA + H_2O \Leftrightarrow A^- + H_3O^+$

(acid + base < = > conjugate base + conjugate acid)

For convenience, the acidity scale is expressed as pK_a, where pK_a = −log K_a. At a pH above the pK_a of an acid, the conjugate base predominates, and inversely, at a pH below the pK_a, the conjugate acid predominates. Inductive and resonance effects, which are summarized in the Hammett equation, affect the pK_a's of organic acids. Structural effects, such as *cis*−*trans* isomerism, may also alter the stability of the conjugate acid. The pK_a value of a compound influences reactivity and spectral properties (color) and is of general importance in chemistry because ionization of a compound alters its physical behavior and macro properties such as solubility and lipophilicity. In biochemistry the pK_a values of proteins and amino acid side chains are of major importance for the activity of enzymes and the stability of proteins. Ionization increases solubility in water but decreases lipophilicity. In drug development, the concentration of a compound in the blood can be adjusted by the pK_a of an ionizable group. The affinity of a drug molecule to a target or the efficiency of RNA as an active transport carrier may be critically dependent on the degree of dissociation. Hence, pK_a is important to the activity and/or toxicity of drugs and it is useful to develop broadly applicable and accurate models for the prediction of the pK_a values in the early phases of drug design. Following are the main contributions for the QSPR modeling of pK_a of organic compounds, including druglike compounds, using theoretical molecular descriptors. The models' technical and statistical parameters are given in Table 12.

In 1981, Perrin et al.[771] published a book on pK_a prediction, which is widely used but is impractical for large systems, especially for high-throughput virtual screening applications. A number of useful fragment methods are available as

Table 12. QSPR Models for Prediction of pK_a

no.	compounds	N^a	methods	model descriptors, n_d	R^2	s (log unit)	R_{valid}^2	s_{valid} (log unit)	ref
1	pure extractants	15	MLR	2 molecular connectivity indices	0.933				Shan et al.[801]
2	carboxylic acids, substituted phenols, and alcohols	48	MLR	number of carbon atoms (N_C) and group philicity	0.991	0.49			Giri et al.[800]
3	diverse organic compounds incl. drugs	(123)	computer program SPARC	contributions of the structural components while based on perturbation models			0.92	0.78^b	Lee et al.[796]
	diverse organic compounds	(537)	computer program SPARC	contributions of the structural components while based on perturbation models			0.80	1.05^b	Lee et al.[796]
4	acidic nitrogen compounds	421	PLS	8 components	0.97	0.41^b			Milletti et al.[794]
	six-membered heteroaromatics	947	PLS	10 components	0.93	0.60^b			Milletti et al.[794]
	druglike compounds	(28)	PLS	14 models out of the system of 33 models			0.85	0.90^b	Milletti et al.[794]
5	carboxylic acids	31	MLR	variable anticonnectivity index of order one		0.47^b			Pompe and Randić[793]
6	phenols in 10 solvents	199 (55)	CNN	solute, 5 quantum chemical; solvent, H-bond donation ability and dipole moment	0.982	0.71^b	0.977	0.83^b	Jover et al.[798]
7	benzoic acids in 9 solvents	379 (98)	CNN	solute, 5 quantum chemical; solvent, H-bond donation ability and cohesive energy	0.998	0.21^b	0.998	0.21^b	Jover et al.[799]
8	aromatic acid derivatives	74 (33)	MLR	3 quantum chemical and geometrical			0.988	0.27^b	Ghasemi et al.[795]
9	phenols	106 (22)	MLR	6 molecular descriptors	0.913	0.523^b	0.895	0.562^b	Habibi-Yangjeh et al.[792]
	phenols	106 (22)	ANN	6 molecular descriptors	0.999	0.036^b	0.999	0.011^b	Habibi-Yangjeh et al.[792]
10	substituted imidazolines	23	PLS	2 latent variables	0.995				Popelier and Smith[791]
	imidazoles	15	PLS	3 latent variables	0.989				
11	neutral and basic drugs	59 (15)	HM, MLR	5 constitutional, topological, geometrical, electrostatic, quantum chemical	0.78	0.48^b	0.48	0.99^b	Luan et al.[789]
	neutral and basic drugs	59 (15)	HM, RBFNN	5 constitutional, topological, geometrical, electrostatic, quantum chemical	0.785	0.458^b	0.543	0.613^b	Luan et al.[789]
12	carboxylic acids	40	QTMS	descriptors of the AIM theory	0.920				Chaudry and Popelier[790]
	anilines	36		descriptors of the AIM theory	0.974				Chaudry and Popelier[790]
	phenols	19		descriptors of the AIM theory	0.952				Chaudry and Popelier[790]
13	imidazol-1-ylalcanoic acid derivatives	15 (3)	MLR	2 quantum chemical, 1 constitutional	0.978	0.1			Soriano et al.[788]
14	carboxylic acids	826	MLR	21 operational atomic contributions	0.941	0.104			Cherkasov et al.[787]
	protonated amines	802	MLR	19 operational atomic contributions	0.933	0.182			Cherkasov et al.[787]
15	diverse organic acids	645	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types	0.93				Xing et al.[578]
	diverse organic bases	384	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types	0.92				Xing et al.[578]
	organic compounds (validation)	(25)	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types			0.95	0.7	Xing et al.[578]
16	diverse organic acids	625	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types; 22 atom types and 11 group types	0.98	0.405			Xing et al.[785]
	diverse organic bases	412	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types; 22 atom types and 11 group types	0.99	0.302			Xing et al.[785]
	organic compounds (validation)	(25)	PLS	tree structured fingerprint describing the ionizing centers: 24 atom types and 9 group types; 22 atom types and 11 group types			0.99	0.40	Xing et al.[785]
17	phenols	175	MLR	4 semiempirical quantum mechanical descriptors derived from frontier electron theory	0.93	0.599			Tehan et al.[783]
	aromatic carboxylic acids	99	MLR	4 semiempirical quantum mechanical descriptors derived from frontier electron theory	0.87	0.357			Tehan et al.[783]
	aliphatic carboxylic acids	185	MLR	4 semiempirical quantum mechanical descriptors derived from frontier electron theory	0.69	0.564			Tehan et al.[783]
18	heterocyclic compounds	150	MLR	3 semiempirical quantum mechanical descriptors derived from frontier electron theory	0.72	1.168			Tehan et al.[784]

Table 12. Continued

no.	compounds	N^a	methods	model descriptors, n_d	R^2	s (log unit)	R_{valid}^2	s_{valid} (log unit)	ref
	anilines and amines	132	linear	1 (electrophilic superdelocalizability)	0.94	0.985			Tehan et al.[784]
19	anilines	36	linear	1 (Hammett constant)	0.940	0.310			Gross et al.[782]
	anilines	36	linear	1 (minimum molecular surface local ionization energy)	0.949	0.285			Gross et al.[782]
20	carboxylic acids	56 (9)	MLR	3 (partial charges on O and H atoms, O−H bond order)	0.84		0.95		Citra[776]
	benzoic acids	31 (10)	MLR	3 (partial charges on O and H atoms, O−H bond order)	0.89		0.85		Citra[776]
	phenols	101 (15)	MLR	3 (partial charges on O and H atoms, O−H bond order)	0.96		0.97		Citra[776]
	alcohols	27	MLR	3 (partial charges on O and H atoms, O−H bond order)	0.89				Citra[776]
21	phenols and aromatic and aliphatic carboxylic acids	190	MLR	4 MNDO and AM1 calculated descriptors	0.90				Grüber and Buss[781]

[a] Training set (prediction set): QTMS, quantum topological molecular similarity; AIM, atoms in molecules; HM, heuristic method. [b] rms error.

commercial software[772−774] but are limited in scope. Because every prediction is based on a congeneric parent structure, pK_a's can only be predicted reliably for compounds very similar to those in the training set, making it difficult to get good estimates for novel structures. The latter is especially true when predicting pK_a values for compounds of pharmaceutical interest. Ab initio and semiempirical quantum mechanics calculations have been used extensively,[775,776] and pK_a values can be calculated formally from statistical thermodynamics, based on numerical solutions of the Poisson−Boltzmann equation.[777−779] Methods have also been developed for the prediction of pK_a of amino acid residues in proteins in which the environmental effects are particularly important and difficult to estimate.[780]

Grüber and Buss[781] used MNDO and AM1 levels of QM theory on PCMODEL-optimized geometries to calculate the pK_a-values of some 190 phenols and aromatic and aliphatic carboxylic acids. The best correlation encompassing all compounds employed four descriptors and had $R^2 = 0.900$.

Citra[776] correlated partial atomic charges and bond orders with the pK_a of sets of phenols, carboxylic acids, and alcohols. A three-descriptor equation with $R^2 = 0.84$ for 56 acids was reported. Citra carried out a conformational analysis on the molecular structures and used descriptors that were averaged over all low energy conformations.

Gross et al.[782] have developed single-parameter correlations of five ab initio quantum chemical indices for anilines to study the effects of substituents on the dissociation constant, pK_a. Among the calculated quantities, the best representation of the aniline pK_a's was produced by the minimum average local ionization energy on the molecular surface. The good performance of the five calculated descriptors as compared to Hammett constants in their ability to estimate pK_a proved that quantum chemical parameters can be applied instead of empirical ones in modeling and in providing a fundamental understanding of the property variations. Tehan et al.[783] proposed QSPRs for the pK_a of molecules or fragments with relevance to the pharmaceutical industry (extracted from the Physprop database, http://www.syrres.com) using semiempirical quantum mechanical descriptors. These descriptors were calculated on and around the atoms of the functional group of interest. Electrophilic superdelocalizability (SE) was highly correlated with pK_a, and additionally, SE was able to distinguish between the *meta-/para-* or *ortho*-substituted acids or phenols. The pK_a values of the nitrogen containing functional groups of amines, anilines, and nitrogen containing heterocyclic compounds were also successfully modeled with the same descriptors.[784]

Xing et al.[578] predicted pK_a values for both acids and bases in water using a novel tree-structured fingerprint method describing the neighborhood of ionizing centers by constructing a count vector based on the total number of atoms and groups of each type at each level originating from the center. The results of the initial approach were considerably improved (from $R^2 = 0.93$ to 0.98 for acids, and from 0.92 to 0.99 for bases, respectively) by individual treatment of the chemical classes using the same approach.[785] Polanski et al.[786] predicted pK_a for benzoic and alkanoic acids by coupled ANN-PLS based on the comparison of molecular surfaces.

Cherkasov et al.[787] presented a new method to quantify the substituent effect, called "3D correlation analysis" (3D-CAN), based on empirical inductive and steric constants, taking into account the 3D structure of substituents. New formulas allowing calculation of pK_a values for 826 carboxylic acids and 802 protonated amines were established, and the possibility of interpretation of the physical nature of the substituent effects within the framework of 3D-CAN was presented.

Soriano et al.[788] estimated the pK_a values of different series of imidazol-1-ylalkanoic acid derivatives using combinations of semiempirical or ab initio methods and two semiempirical solvation models SM2 and SM5.4. None of these procedures was able to describe the zwitterionic structure of the carboxylic monoacid series as their most stable form determined experimentally in solution. A comparison of the theoretical and experimental pK_a values showed rms differences ranging from 1.43 to 3.04 pK_a units. As an alternative strategy, a QSPR model for pK_a determination is described based on two quantum chemical descriptors, the natural atomic charge on the N3 proton (q_H^+) and the frontier orbital energy (ε_L), and the number of ester groups in the molecule n. The model reproduced the experimental values of imidazol-1-ylalkanoic compounds within 0.1 pK_a unit.

Luan et al.[789] developed QSPR models to predict the pK_a values of a set of 74 neutral and basic drugs via linear and nonlinear methods. The CODESSA approach was used to derive descriptors and to build linear models; RBFNN (radial basis function neural networks) was used to generate the nonlinear models. Both models used the same descriptors selected by the heuristic method: the descriptors accounted for the relative nitrogen content and polarizability of the compounds related to the ease of protonation of the molecules. The results were rated as "fair" in view of the complexity and relatively large size of the drug molecules.

Multivariate models for three classes of compounds were developed by means of the quantum topological molecular similarity (QTMS) tool, using descriptors from the "atoms in molecules" (AIM) theory.[790] Correlations obtained outperformed the Hammett and other traditional parameters. The results of QTMS were demonstrated by the following R^2/q^2 values: 0.920/0.891 (acids), 0.974/0.953 (anilines), and 0.952/0.884 (phenols). Popelier and Smith[791] used the QTMS method based on quantum chemical topology (QCT) to define electronic descriptors drawn from modern ab initio wave functions of geometry-optimized molecules enabled by the present computing power. Seven data sets of medicinal interest were investigated including pK_a values for a set of substituted imidazolines and imidazoles. A PLS analysis in conjunction with a GA delivered excellent models that were also able to highlight the important bonds responsible for the observed property.

Habibi-Yangjeh et al.[792] used MLR and ANNs to model the pK_a values of 106 phenols with diverse chemical structures. Six molecular descriptors [the polarizability term (π_I), the most positive charge of the acidic hydrogen atom (q^+), the molecular weight (MW), the most negative charge of the phenolic oxygen atom (q^-), the hydrogen-bond accepting ability (ε_B), and the partial-charge weighted topological electronic (PCWTE) descriptor] of the MLR model were used as inputs. External validation of the models with 22 compounds produced R^2 0.895 and rms 0.562 for the MLR model compared with the values of 0.99996 and 0.0114, respectively, for the ANN model.

Pompe and Randić[793] optimized a variable anticonnectivity topological index for the modeling of pK_a values. The variable anticonnectivity index of order one showed superior modeling capabilities compared to the ordinary variable connectivity index of the same order because it accounts for the combination of positive and negative contributions in the molecular descriptor.

Milletti et al.[794] presented a new computational method for pK_a prediction of organic compounds using descriptors generated by the program GRID, based on molecular interaction fields precomputed on a set of molecular fragments. The new method was trained and cross-validated by using a diverse data set of 24,617 pK_a values. The results were presented for a class of 421 acidic nitrogen compounds (rms = 0.41, $R^2 = 0.97$, $q^2 = 0.87$) and for a class of 947 six-membered *N*-heterocyclic bases (rms = 0.60, $R^2 = 0.93$, $q^2 = 0.85$). External validation with 28 novel compounds with nine different ionizable groups and 39 experimentally determined pK_a values demonstrated good predictive ability ($R^2 = 0.85$, rms = 0.90). For the validation set of the 28 druglike compounds, the method gave better results when compared with the ACD/pK_a program ($R^2 = 0.76$, rms = 1.36).

A very simple, interpretable model, based on MLR and quantum chemical descriptors for pK_a's of aromatic acid derivatives, was developed by Ghasemi et al.[795] Three significant descriptors, related to the partial charges at each atom in the $O^{\delta-}-H^{\delta+}$ bond (*pchg*$H^{\delta+}$ and *pchg*$O^{\delta-}$) and the change in the bond length of O−H (bl(O−H)), were identified. A model with low prediction error and high correlation coefficient was obtained based on 74 molecules as a training set. The average relative error of the prediction set of 33 compounds was lower than 1% (rms = 0.27), and R^2 was 0.988. The pK_a values of aromatic acids generally decreased with increasing positive partial charges on the acidic hydrogen atom.

Lee et al.[796] used the computer program SPARC[797] (SPARC Performs Automated Reasoning in Chemistry) to predict the ionization state of drugs. This program has been developed based on the physical chemistry of reactivity models and applied successfully to predict numerous physical properties as well as chemical reactivity parameters. SPARC predicts both macroscopic and microscopic pK_a values strictly from molecular structure. A high correlation ($R^2 = 0.92$) between experimental and the SPARC calculated pK_a values was obtained with rms of 0.78 log unit for a set of 123 compounds, including many known drugs. A set of 537 compounds from the Pfizer internal data set gave $R^2 = 0.80$ and rms = 1.05.

Jover et al.[798] utilized CNNs to compose a multicomponent system to correlate the pK_a values of 94 phenols in protic (water, methanol, isopropanol, and *tert*-butanol) and aprotic (DMSO, *N*,*N*-dimethylformamide (DMF), acetonitrile, nitromethane, acetone, and *N*,*N*-dimethylacetamide (DMA)) solvents. The phenols were characterized by the CODESSA descriptors, and the solvents by several physical properties and the most used multiparametric polarity solvent scales. The final model contained seven descriptors: five of them belonging to the solutes and the remaining two to the solvents. RMSE and (R^2) of 0.71 (0.982) for the training, 0.83 (0.977) for the prediction, and 0.95 (0.975) for the validation sets were reported. The same methodology was used to derive a QSPR model for the pK_a prediction of benzoic acids in different solvents.[799] The system studied contained 519 pK_a values corresponding to 136 benzoic acids determined in water and 8 organic solvents. The training, prediction, and cross-validation sets all had the same R^2 (0.998) and RMSE (0.21). The descriptors of both models were clearly related to interactions playing a role in the dissociation process.

Giri et al.[800] used a simple descriptor, the number of carbon (N_C)/non-hydrogenic (N_{NH}) atoms present in a molecule, for the development of QSPR models for several useful properties, including pK_a values of carboxylic acids, phenols, alcohols, etc. High statistical parameters ($R^2 = 0.991$, $R^2_{cv} = 0.990$, $s = 0.490$, $n = 48$) suggest the significance of this descriptor, which improves the two-parameter QSPR models with electrophilicity or its local variant as an additional descriptor. The simplicity of this descriptor is seen as a great advantage of these models.

Shan et al.[801] established QSPR models for the pK_a of some pure extractants and the apparent basicity (p$K_{a,B}$) of three typical mixture solvents: trioctylamine (TOA)/hexane, TOA/1-octanol, and TOA/methyl isobutyl ketone (MIBK). The models include the concentration of extractant in the solvent and three kinds of molecular connectivity indices of extractant and diluent. The calculated values from the models of the pure extractant and mixture solvents showed good consistency with experimental values.

Despite the numerous advances in high-throughput measurements, in silico determination is still the fastest and cheapest way of obtaining an estimate of pK_a. This research demonstrates the high priority of developing prediction models for pK_a for applications in chemical technology and in medicinal chemistry.

6.5. Thermal Decomposition Temperatures of Nonlinear Optical (NLO) Chromophores

Organic second-order nonlinear optical (NLO) materials have potential applications in telecommunications, optical information processing, computing, and data storage. These materials are typically made from chromophores, small organic molecules, incorporated into polymer matrices, and poled with an electric or optical field to achieve a noncentrosymmetric dipole alignment. There are several qualities of the NLO chromophores that need to be optimized for producing a successful industrial material. To ensure the stability of the material during fabrication, the thermal decomposition temperature, T_d, should be over 573 K. QSPR model development can be helpful in prediction of the qualities of the potential NLO chromophores.

In attempts to predict T_d of compounds, Bicerano[802] developed a QSPR for a set of 140 polymers, with 21 descriptors involved. Using a molar thermal decomposition function Y_d ($Y_d = T_dM$, where M is the molecular weight) as a dependent variable, an R^2 value of 0.998 was reported. Xu et al.[803] studied QSPRs between theoretical descriptors representing the molecular structures and T_d for a diverse set of 90 s-order nonlinear optical (NLO) chromophores in the temperature range 473−685 K. A seven-parameter MLR model was developed for the molar thermal decomposition function Y_d, following the same methodology as in the previous work, with $R^2 = 0.964$ and SEE = 14.01 K. The mean relative error for the prediction of T_d was 4.46%. The model descriptors supported the physical origin of T_d, expressing size, shape, resonances, and transfers of intramolecular charge. Cross-validation of the model indicated good stability of the description of the property by the selected descriptors; however, no external validation was performed.

6.6. Chain Transfer Constants

A transfer constant is a dimensionless quantity defined as the ratio between the rate constant for the formation of the unreactive polymer and the rate constant for the propagation reaction. Understanding chain transfer clarifies our understanding of the microkinetic processes in polymerization reactions. During polymer synthesis, chain-transfer reactions modulate molecular weight and broaden the molecular weight distribution, which in turn determines polymer processability. Thus, control of these macromolecular features is required when high molecular weight polymers are not suitable for a given application. During the last 15 years, considerable interest has developed in the use of chain-transfer agents to produce "living polymers". Knowledge of chain-transfer rate constants which can be obtained with the aid of QSPR estimation models assists the industrial scale-up of polymerization processes using kinetic modeling techniques and reduces the number of iterative adjustments required to achieve optimum (co)polymerization.

Ignatz-Hoover et al.[804] deduced quantitative structure−reactivity relationships (QSRR) for kinetic chain-transfer constants, log C_X, for 90 agents for styrene polymerization at 60 °C. A five-parameter correlation with $R^2 = 0.818$, $R_{cv}^2 = 0.795$, and $s = 0.818$ logarithmic unit was derived. Despite the heterogeneity of the radical size within the systems studied and differences in the experimental testing conditions, a good correlation was obtained. The descriptors involved in the correlations were consistent with the proposed mechanism of chain-transfer reactions. The model allows the prediction of the transfer constants for a variety of additives (transfer agents) and helps in the theoretical understanding of free-radical polymerization kinetics.

6.7. Flash Points and Autoignition Temperatures

The flash point, T_f, is the lowest temperature at which the vapor of a volatile liquid can form an ignitable mixture with air. The combustible substance reacts with oxygen in the air in an exothermic oxidation reaction giving a momentary flash. The flash point of compounds is important in terms of both practical uses (i.e., combustion chemistry) and safety (i.e., handling and transporting of the compounds in bulk quantities). Numerous methods have been developed to estimate flash points for pure liquids as well as mixtures.[805] Many of these methods involve different mathematical equations using empirical parameters, such as boiling points, critical temperature, vapor pressure, and activity coefficients. QSPR models utilizing theoretical molecular descriptors have the advantage of not needing data measurements for prediction of the property. The most significant publications on QSPR modeling of the flash points are reviewed and summarized in Table 13 below.

Zhokhova et al.[806] constructed several MLR and ANN models for the T_f of different sets of diverse organic compounds using structural fragment descriptors. The best result was obtained with the ANNs using 25 fragmental descriptors on a set of 398 compounds: $R^2 = 0.959$ and $\text{rms}_{pr} = 14.6$ °C. The MLR model using the same data was comparable to the ANN model with $R^2 = 0.956$ and $\text{rms}_{pr} = 15.8$ °C.

Gramatica et al.[807] studied solvent properties in order to provide a tool for selecting a suitable solvent. A QSPR for the flash points of 136 organic solvents was developed starting from a large descriptor pool. Using the genetic algorithms-variable subset selection (GA-VSS) procedure, a six-parameter model with $R^2 = 0.813$ and $Q_{LOO}{}^2 = 78.7$ was obtained. A data set of 153 esters was studied for the prediction of the basic physicochemical properties.[808] The MLR approach was based on a variety of theoretical molecular descriptors, selected by the GA-VSS. The best linear QSPR models were internally and externally validated ($Q_{LMO50\%}{}^2 = 0.78−0.93$; $Q_{EXT}{}^2 = 0.88−0.94$). The leverage approach was used to define the model's domain of applicability. The predictions of the class-specific QSPR models were compared with the US-EPIWIN prediction and showed better performance.

Tetteh et al.[809] developed radial basis function (RBF) ANN models for the simultaneous estimation of flash (T_f) and boiling points (T_b) based on 25 molecular functional groups and their first-order molecular connectivity indices. The RBF networks were trained by the orthogonal least squares (OLS) learning algorithm. After dividing the initial set of 400 compounds into training (134), validation (133), and test (133) subsets, the average absolute errors were obtained and compared. For the validation and test sets, they ranged from 10 to 12 °C and 11 to 14 °C for T_f and T_b, respectively, in agreement with the experimental value of about 10 °C. This work is an extension of one of the authors' previous studies,[810] where a predictive MLR model with the same parameters on the whole data set of 400 compounds ($R^2 = 0.94$, $s = 13.5$ °C) was reported.

A QSPR study of the flash points of a diverse set of 271 compounds by Katritzky et al.[811] provided a general three-parameter QSPR model ($R^2 = 0.9020$, $R^2_{cv} = 0.8985$, $s =$

Table 13. QSPRs for the Prediction of Flash Points (T_f) of Volatile Compounds

no.	compounds	N^b	methods^a	model descriptors, n_d	R^2	s (°C)	R_{valid}^2	s_{valid} (°C)	ref
1	diverse compounds	400	MLR	25 atomic and group increments, first-order molecular connectivity index	0.94	13.5			Suzuki et al.[810]
2	diverse compounds	267 (133)	RBF-ANN	25 (atomic and group increments, first-order molecular connectivity index)	0.96^c	10.8	0.92	14.3	Tetteh et al.[809]
					0.96^d	10.1	0.92	14.0	Tetteh et al.[809]
3	diversely substituted pyridines	126	MLR	6 theoretical descriptors	0.76				Murugan et al.[215]
4	diversely substituted pyridines	121	MLR	6 theoretical descriptors	0.84	16.7 K			Katritzky et al.[216]
5	diverse compounds	271	MLR	3 (theoretical descriptors, AM1)	0.90	16.1 K			Katritzky et al.[811]
				3 (descrs and exp boiling point)	0.95	11.2 K			Katritzky et al.[811]
				3 (descrs and calc boiling point)	0.92	14.2 K			Katritzky et al.[811]
6	diverse compounds	398	MLR	25 (the number of atoms or molecular fragments)	0.96	11.4			Zhokhova et al.[806]
		398	NN		0.96				
7	organic solvents	136	GA-VSS MLR	6 (structural, empirical, topological, 3D-WHIM)	0.81				Gramatica et al.[807]
8	diverse organic compounds (−50 to 133.9 °C)	59 (∼600)	empirical equation	3 (normal boiling point, standard enthalpy of vaporization at 298.15 K, number of C atoms)^e		2.9^f		3.4	Catoire and Naudet[812]
9	diverse compounds including bioactive	418	group contribution	first- (104) and second-order group contributions	0.97	14.7 K			Stefanis et al.[813]
10	alkanes	92 (15)	BPANN	9 (group bond contributions)	0.98	3.8 K^f	0.98	4.8 K^f	Pan et al.[817]
			MLR		0.97	6.27 K		6.1 K^f	Pan et al.[817]
11	diverse organic compounds	758	MLR	4 (boiling point (calc), electrostatic, topological)	0.849	18.9 K			Katritzky et al.[818]
			ANN		0.878	12.6 K			Katritzky et al.[818]

[a] RBF-ANN, radial basis function artificial neural networks [b] In parentheses is the number of validation set compounds. [c] Single output values. [d] Double output values. [e] The normal boiling point and standard enthalpy of vaporization can be calculated using a number of theoretical models. [f] Mean absolute deviation.

16.1 K) using a large pool of 3D theoretical descriptors retrieved from AM1 calculations. The use of the experimental boiling point as a descriptor in the equation resulted in $R^2 = 0.9529$. When using calculated boiling points, the R^2 of the model was found to be 0.925. The other two parameters involved in the last equation were (i) the difference between the positively charged partial surface area and the negatively charged partial surface area, DPSA, and (ii) the minimum electron attraction for a C atom, $E_{e\text{-}n,C}$. DPSA is responsible for the polar interactions between molecules, whereas the quantum-chemical descriptor $E_{e\text{-}n,C}$ can be related to the reactivity of any carbon atom within the molecule in a combustion reaction. The results were validated by a leave-many-out procedure. In their earlier work,[216] a modest six-parameter correlation ($R^2 = 0.837$, $R_{cv}^2 = 0.832$, $s = 16.7$ K) was obtained for the flash points of 121 pyridines. The descriptors employed in this equation indicated the importance of molecular bulk and hydrogen-bonding effects in determining flash points.

Catoire and Naudet[812] developed a unique empirical equation for estimating T_f for most classes of organic liquids. Fifty-nine carbon-containing compounds were selected for the establishment of the equation considering the reliability of the measurements, wide range of temperature (−50 to 133.9 °C), and structural variations. Three parameters were used: the normal boiling point, the standard enthalpy of vaporization at 298.15 K, and the number of carbon atoms in the molecule. In the case of missing experimental data for the two empirical parameters, several accurate theoretical estimation methods were suggested. The developed equation reproduced T_f with a mean absolute deviation of 2.9 °C and a maximum absolute error of 7 °C. The equation was extensively tested on diverse organic compounds including those containing N, O, S, Si, P, Sn, Ni, B, Ge, and halogen atoms. Polyhalogenated compounds (that also include ignition inhibitors) were detected as outliers in this model.

Stefanis et al.[813] developed a group-contribution method that uses two kinds of groups: first-order groups (104) that describe the basic molecular structure of the compounds and second-order groups that are based on the theory of the conjugated operators and improve the accuracy of the predictions by providing more structural information to distinguish the isomers. New groups were defined to ensure that the molecular structure of any compound of biochemical interest, including complex aromatic, multiring, and heterocyclic compounds, could be described and the reliability of the predictions was enhanced. The flash points were estimated with $R^2 = 0.967$, $s = 14.7$ K, and a mean error of 3.27%.

A particular challenge is the estimation of the properties of mixtures because a simple mixing rule will not work when interactions among the mixture components are strong. A review of flash point estimates for mixtures is given in Vidal et al.[805] as well as the discussion of special cases in which the flash point of a mixture is below the flash points of the individual components.[814] Estimates for the binary mixtures, such as methanol−water and ethanol−water, were presented and compared with experimental values as well as those for the flammable mixtures of octane−ethanol and octane−1-butanol, which exhibit the minimum flash point behavior (MFPB). It was demonstrated that the UNIFAC group contribution method[480] can be used to reproduce the flash points of binary mixtures when the liquid mixture is nonideal. Catoire et al. extended their equation for pure compounds[812] to binary and tertiary mixtures,[815,816] and they were also able to reproduce the MFPB phenomenon measured experimentally.

Pan et al.[817] constructed models of the relationships between the structures and flash points of 92 alkanes by means of ANNs using the group bond contribution method. Group bonds which contain information of both the group property and the group connectivity in the molecules were used as molecular descriptors. The data set of 92 alkanes

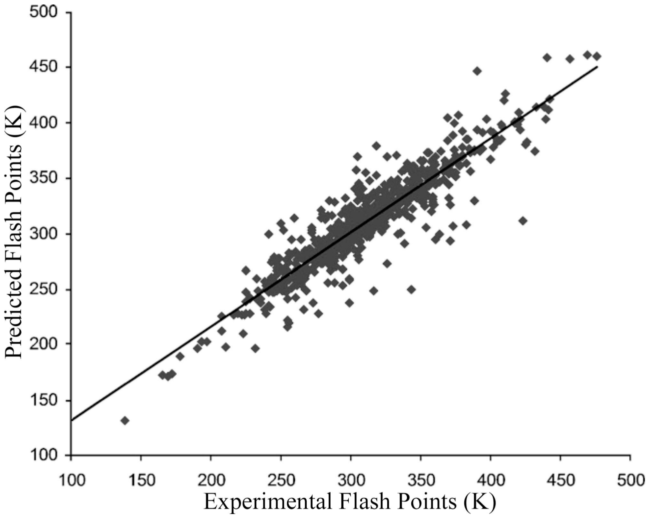


Figure 12. Experimental vs predicted flash points according to the MLR model. Reprinted with permission from ref 818. Copyright 2007 Elsevier B. V.

was randomly divided into a training set (62), a validation set (15), and a testing set (15). The optimal condition of the ANN was obtained by adjusting various parameters by trial-and-error. Simulated with the final optimum BP ANN [9−5−1], the results showed that the predicted flash points were in good agreement with the experimental data, with the average absolute deviation of 4.8 K and the rms error of 6.86, which were shown to be superior to those of the MLR method.

Katritzky et al.[818] published an update of their previous QSPR study of flash points[811] using an extended data set of 758 organic compounds collected from the literature published after 2004. Both MLR (see Figure 12) and ANN models were developed using geometrical, topological, quantum mechanical, and electronic descriptors calculated by the CODESSA PRO software. The best model obtained had a good representation of the property (with an average error of 13.9 K) with only four molecular descriptors: boiling point, BP (calculated from a QSPR model), HA dependent HDCA-1/TMSA (Zefirov PC), HASA-1/TMSA (Zefirov PC) (all), and the relative number of triple bonds. The descriptors appearing in this model were primarily related to electrostatic and hydrogen bonding interactions as well as to the molecular shape. The ANN model gave better statistical characteristics: $R^2 = 0.878$ and average error of 12.6 K based on only slightly different decriptors. The developed QSPR model could be used for the prediction of flash points for a wide range of organic compounds.

Autoignition temperature (AIT) is another important fire safety parameter in handling bulk chemicals. It is defined as the lowest temperature at which a substance in air will ignite in the absence of a spark or flame. Autoignition occurs when the rate of heat evolved by this reaction is greater than the rate at which heat is lost to the surroundings. AIT is also crucial to the performance of internal combustion engines through the phenomenon of engine knock. The autoignition mechanism proceeds by a free radical reaction and the stability of the free radical intermediates determines the ease of oxidation. The structural features that affect AIT are the chain length, degree of unsaturation, degree of branching, aromaticity, and the functional groups of the compounds. QSPR models of AIT based on calculated molecular descriptors are discussed below and summarized in Table 14.

Egolf and Jurs[819] correlated chemical structural features of 312 diverse hydrocarbons, alcohols, and esters to their AIT. The general model for the whole set displayed moderate statistical parameters ($R^2 = 0.726$), and therefore, the chemical classes were approached separately. They observed a shift of the regression line on the observed vs calculated AIT plot and suggested two different mechanisms for the AITs of hydrocarbons at high (633−848 K) and low (475−610 K) temperatures. The final models for the four groups defined had R^2 in the range 0.88−0.95, and s between 12 and 24 K. The above MLR models contained from four to eight molecular descriptors. Structural features such as radical stability, steric strain, and molecular rigidity were found to be important for modeling autoignition. The study was repeated with an enlarged pool of descriptors[820] on a data set consisting of 327 hydrocarbons, halohydrocarbons, and oxygen, sulfur, and nitrogen containing compounds. MLR and CNN were used to develop predictive AIT models. Both GA and SA routines were used to select subsets of descriptors. The models developed for several subsets of the data had predictive ability in the range of experimental error (rmse ≤35 °C).

Tetteh et al.[821] used both radial basis function (RBF) and back-propagation (BP) ANNs to model AIT with six descriptors, two empirical and four structural. The RBF and BP ANNs led to satisfactory models for the training set ($n = 85$, $R^2 = 0.953$ and 0.945, respectively) but performed only moderately for the validation set ($n = 148$, $R^2 = 0.834$ and 0.837 with average errors of prediction of 30.1 and 29.9 °C, respectively). In a subsequent study, Tetteh et al.[822] enlarged their data set to 232 organic compounds and used biharmonic spline interpolation to optimize both the spread parameter and the number of neurons in the hidden layer of the RBF ANNs. Comparable results with their previous study were reported for the training ($R^2 = 0.826$, error 30.2 °C, $n = 78$), validation ($R^2 = 0.833$, error 30.1 °C, $n = 77$), and test ($R^2 = 0.861$, error 32.9 °C, $n = 77$) sets, respectively.

Yoshida and Funatsu[823] proposed a new hybrid method that combines the genetic algorithm (GA) and the QPLS method (GA-QPLS) acting as a nonlinear analogue of PLS to model AIT using the training set data of 85 compounds of Tetteh et al.[821] The hybrid method led to a significant improvement over the conventional QPLS method.

Kim et al.[824] used the genetic functional approximation (GFA) to find the best MLR model for AIT within 72 molecular descriptors for a set of 200 diverse organic compounds. Nine topological, functional group counts, and AM1 based charge related descriptors for a training set of 157 data points were used. For the training set $R^2 = 0.920$, and the RMSE was 25.9 °C. The corresponding statistical parameters for the test set (43 data points) were $R^2 = 0.910$ and RMSE = 29.0 °C.

Albahri and George[825] used ANNs to identify structural groups within the framework of the structural group contribution (SGC) method that could best represent AIT for about 490 substances. The chosen 58 single and binary structural groups were derived from the Ambrose, Joback, and Chueh-Swanson definitions of group contributions and modified to account for the location of the functional groups in the molecule. The proposed method developed using 470 compounds performed excellently in the applicability range of the model, predicting the AIT of 20 pure components with an average error of 2.6% and R^2 of 0.98.

Table 14. QSPRs for the Prediction of Autoignition Temperatures (AIT) of Volatile Compounds

no.	compounds	N (n_{valid})[a]	methods[b]	model descriptors, n_d	R^2	s (°C)	R^2_{valid}	s_{valid} (°C)	ref
1	all compounds	312	MLR		0.73	59 K			Egolf and Jurs[819]
	hydrocarbons (low AIT)	58		8 (topological, AM1)	0.95	12 K			
	hydrocarbons (high AIT)	46		5 (topological, AM1)	0.88	16 K			
	alcohols	28		4 (topological)	0.94	24 K			
	esters	25		4 (topological, geom, AM1)	0.93	20 K			
2	all compounds	300	GA, SA, quasi-Newton BFGS NNs, MLR	11		58.5[c]			Mitchell and Jurs[820]
	hydrocarbons (low AIT)	47 (5)		5 (topol, electronic)		8.77[c]		5.11[c]	
	hydrocarbons (high AIT)	46 (5)		6 (topol, electronic)		18.5[c]		15.7[c]	
	nitrogen compounds	36 (4)		6 (topol, electronic)		34.9[c]		28.2[c]	
	oxygen/sulfur compounds	132 (4)		7 (topol, geom, electronic)		30.8[c]		32.5[c]	
	alcohol/ether compounds	67 (x)		6 (topol, geom)		19.6[c]		20.0[c]	
3	diverse organic compounds (AIT 170...630 °C)	250	MLR	6 (critical pressure, parachor, atomic charges, 0th order connectivity index, group indicators)	0.89	34.8[f]			Suzuki[828]
4	diverse organic compounds	85 (148)	RBF NN	6 (2 empirical, 2 theoretical, 2 indicator variables)	0.95	17[d]	0.83	30[d]	Tetteh et al.[821]
			BP NN		0.95	18[d]	0.84	30[d]	
5	diverse organic compounds	85	GA-QPLS	6 (2 empirical, 2 theoretical, 2 indicator variables)	0.95	23.7			Yoshida and Funatsu[823]
6	diverse organic compounds	232 (77)	RBF NN	6 (2 empirical, 2 theoretical, 2 indicator variables)	0.83	30.2	0.86	32.9	Tetteh et al.[822]
7	diverse organic compounds	157 (43)	GFA, MLR	9 (topological, spacial, AM1)	0.92	25.9	0.91[e]	29.0[e]	Kim et al.[824]
8	diverse organic compounds	470 (20)	SGC, ANN	58 single and binary structural groups	0.98	17.7[f]	0.98	17.8[f]	Albahri and George[825]
9	hydrocarbons	118 (42)	BP NN	16 atom-type E-state indices	0.974	17.5[c]	0.906	31.1[c]	Pan et al.[826]
10	alkanes	50 (10)	SVM	6 atom-type E-state indices	0.968	16.4[c]	0.968	17.7[c]	Pan et al.[827]
	organic compounds	142 (90)		6 (critical pressure, parachor, atomic charges, 0th order connectivity index, group indicators)	0.927	29.8[c]	0.908	31.0[c]	
11	diverse organic compounds	446 (90)	GA, SVM	9 (topol, indicator, charge)	0.901	33.2[c]	0.874	36.9[c]	Pan et al.[829]
	(AIT 170...680 °C)		MLR		0.869	38.0[c]	0.856	39.9[c]	

[a] N, number of all compounds. [b] GA, genetic algorithms; SA, simulated annealing; BFGS, Broyden−Fletcher−Goldfarb−Shanno; GFA, genetic functional approximation; SGC, structural group contribution; QPLS, quadratic partial least squares; SVM, support vector machine. [c] rms error. [d] Average error of prediction. [e] Validation set was used for choosing the model. [f] AAD, average absolute deviation.

More recently, Pan et al. performed a series of QSPR studies on modeling AIT. They constructed a BPNN [16−8−1] model to predict the AIT of 118 hydrocarbons using atom-type E-state indices as molecular descriptors which combine together both electronic and topological characteristics of the molecules.[826] The predicted AIT values were in good agreement with the experimental data, with the average absolute error being 21.6 and the rms error being 31.09 °C for the testing set. The same research team conducted another study for the development of QSPR models for predicting AIT of organic compounds.[827] In this study, the calibration and predictive ability of support vector machines (SVM) were investigated using two different data sets (from the latest Internet databases) and compared with those of MLR and BPNN. The first data set involved 50 saturated hydrocarbons, whose structural characteristics were encoded by atom-type E-state indices as molecular descriptors, while the second one comprised of a total of 142 organic compounds described by both the physicochemical parameters and molecular descriptors as previously employed by Suzuki.[828] The results showed that, for both data sets, the performances of the SVM models were comparable or superior to those of MLR and BPNN, especially in external predictive ability. In a following study, Pan et al.[829] combined SVM with GA-PLS as the variable selection method on a large pool of calculated molecular descriptors, including topological, charge, and geometric descriptors. The leave-one-out cross-validation was used to determine the optimal values for the SVM parameters. The resulting model showed the prediction ability, with the rms error being 36.86 for the external validation set of 90 compounds (20% of the data set), which is within the range of the experimental error of the AIT measurements. The information contained in the selected descriptors by the GA-PLS suggests that AIT of organic compounds can be reasonably explained by their electrostatic and steric effects.

The proposed model[829] together with that of Albahri et al.[825] can be pointed out as good alternatives to the experimental measurements of AIT, being applicable for a wide range of organic compounds using only the information which can be derived directly from the molecular structure.

6.8. Octane and Cetane Numbers

Octane number (ON) or octane rating (OR) is a figure representing the resistance of gasoline to premature autoignition when exposed to heat and pressure in the combustion chamber of an internal-combustion engine. Such autoignition is wasteful of energy in the fuel and potentially damaging to the engine, indicated by knocking noises that occur as the engine operates. ON is numerically equal to the percentage of isooctane by volume in a mixture of isooctane and normal heptane in the given gasoline. Research and motor octane numbers (RON and MON) constitute the main quality characteristics of gasoline, providing a sensitive indication of the antiknocking behavior of the fuel. The higher the octane number, the better the gasoline resists detonation and the smoother the engine runs. During the past decade, the increase in the compression ratio of motor vehicle engines led to higher requirements in the octane rating of the fuels. Additionally, restrictions in using octane number improvers encouraged the refineries to utilize algorithms for the prediction of the octane rating of gasoline blends.

Numerous studies have attempted to describe mathematically the ON as a function of the gasoline composition measured by gas chromatography. The blending of gasoline is nonlinear in nature, and in general, the developed models for the ON of the gasoline fuel are empirical in origin and recognize the nonlinear dependence of the blend octane number by modeling it with functions containing a linear part and a nonlinear correction term.

Myers[830] established a correlation between the octane number and readily measurable characteristics of 77 gasoline samples. A MLR model was developed to express the octane number as a linear combination of the isoparaffin index, the aromatic content (volume %), the lead content (g/gal), and the sulfur content (weight %). The alkyl lead concentration was shown to be the most important variable, closely followed by the isoparaffin index. The standard deviations encountered in engine testing were approximately 0.25 and 0.45 octane number for RON and MON, respectively. A standard deviation $s = 1.1$ ON was obtained for both calculated properties. The values of the index of determination (p^2) obtained in this analysis were 0.87 for RON and 0.90 for MON, indicating a satisfactory correlation with the variables selected. To adequately predict the octane numbers, the use of hydrocarbon mixtures should be limited in the range of 91−103 RON units.

Meusinger and Moros[831] determined the influence of molecular structure of 240 organic compounds on their knocking behavior using a nonbinary genetic algorithm (GA). The molecular structures of the potential gasoline components were divided into 16 different structural groups. The partial ONs for paraffines, naphthenes, olefins, aromatics, and oxygenates subclasses were calculated. The sum of the calculated partial octane numbers supplies the ON of the compound. The results obtained by GA were significantly better than those obtained by MLR: for paraffins $R^2_{\mathrm{GA}} = 0.976$ ($R^2_{\mathrm{MLR}} = 0.910$), naphthenes $R^2 = 0.950$ (0.769), olefins $R^2 = 0.968$ (0.920), aromatics $R^2 = 0.893$ (0.769), and oxygenates $R^2 = 0.929$ (0.845), respectively. The calculated partial ONs allow the quantitative determination of the influences of the structure modifications on the knocking characteristics of the gasoline components. In a further study, Meusinger and Moros[832] related the constitutions of more than 300 individual gasoline components to their knock rating (blending research octane number, BRON). ^{13}C NMR spectra of all compounds were classified into 28 chemical shift regions. The number of individual carbon signals of the nearly 2500 carbons was counted in each shift region and was combined with the information about the presence or absence of the following atoms or functional groups: oxygen, rings, aromatics, aliphatic chains, and olefins. These numbers were used as descriptors in the ANN model ($R^2 = 0.891$, $s = 15.7$). For the validation set of 50 individual chemicals from various organic classes consisting only of C, H, and O atoms, a good agreement with their experimentally determined BRON was found ($R^2 = 0.870$, $s = 20.2$).

Estrada and Gutierrez[833] used a generalization method of topological indices based on a vector−matrix−vector multiplication procedure to optimize the Balaban J index for describing the motor octane number (MON) of octane isomers. The reported correlation coefficient between the obtained optimal Balaban index J^{**} and MON was 0.983. A cubic model between MON and J^{**} produced an excellent correlation with $R^2 = 0.992$ and $s = 3.51$.

The Balaban index, Balaban-like topological indices (the complement Balaban index, the Harary−Balaban index, the quotient Balaban indices of first and second order), their variable counterparts, and vertex- and edge-connectivity indices were used in the comparative study of the structure-motor ON modeling by Nikolić et al.[834] The variable indices produced slightly better linear models than the fixed indices. The best models obtained were quadratic models with the Harary−Balaban index and the quotient Balaban index of second order.

Hosoya[835] studied the relationship between the ON of heptane and octane isomers and various topological indices. The single parameter correlation with the Balaban (B) and Wiener (w) indices showed satisfactory statistical results. The best model reported for the octane isomers was as follows: $\mathrm{ON} = 76.389 + 8.179\ (0.3B + 1.1p - 0.6Z)$ with $R^2 = 0.964$ and $s = 6.56$. It was concluded that the highly branched (small Z index of Hosoya, or low boiling point), spherically shaped (large), and compact (large polarity number, p, or high liquid density) gasoline isomers will have high ON or will burn without knocking, and based on these results, several candidates of nonane isomers whose ON is expected to be higher than 100 were suggested.

An analytical method was developed by Albahri et al.[836] to predict ONs of petroleum fuels. The minimum input data for the equations were the boiling point and the specific gravity; however, when the composition of the mixture in addition to the boiling point was known, better results were obtained. Average deviations of about 4−7 for the ON were observed when evaluated with a wide range of data sets. A structural group contribution method for modeling the ON of 200 pure hydrocarbon liquids (ON range of −20 to 120) was also reported by Albahri.[837] The method required knowledge of only the chemical structure of the molecule. The average deviations of the models for the RON and MON of pure hydrocarbon liquids were 4 and 5.7, and those for the external validation sets were 1.3 (9 compounds) and 1.5 (12 compounds), respectively. The results of two different sets of structural groups derived from the Joback group contribution approach were tested and compared.

Podlipnik et al.[838] introduced indirect evaluation of molecular shape similarity. As a first step of the molecular comparison, a conversion of the 3D-molecular structure to translational and rotational invariant RDF code was performed. Second, the similarity indices were computed based on the RDF code comparison for each pair of molecules. These similarity indices were then used as descriptors for generating QSAR/QSPR models. In a practical example, the approach was used to correlate the octane isomer structures to their ON. The results were comparable to those obtained by topological indices.

Ghosh et al.[839] presented a model that predicts the RON and MON of a wide variety of gasoline process streams and their blends including oxygenates based on detailed composition. The ON was correlated to a total of 57 hydrocarbon "lumps" measured by gas chromatography. The model is applicable to any gasoline fuel regardless of the original refining process. It is based on the analysis of 1471 gasoline fuels from different naphtha process streams such as reformates, cat-naphthas, alkylates, isomerates, straight runs, and various hydroprocessed naphthas. Blends of these individual process streams were also considered. The model predicts the ON within a standard error of 1 RON or MON unit and is applicable to a range of ONs between 30 and 120. Further improvements in the model predictions are demonstrated by a data reconciliation algorithm used in tandem with the predictive model.

ANN models have been developed by Pasadakis et al.[840] to determine the RON of gasoline blends produced in a Greek refinery. The ANN models used the volumetric concentrations of the seven most commonly used fractions in the gasoline production and their respective RON numbers as input variables. Additionally, the RON values of the first five fractions weighted by their concentrations in the blends were included as input variables. The model parameters (ANN weights) were presented in a way that enabled the model to be easily implemented. The predictive ability of the models yielded an rms error of prediction (RMSEP) less than 0.2 RON. Based on the ANN models, the effect of each gasoline constituent on the formation of the blend RON value was revealed.

A new method for solving QSPR tasks was proposed by Smolenskii et al.[841] based on transition from numerical values to topological equivalents (TEs) of physicochemical properties of chemical compounds. The TEs are unambiguously related to the corresponding properties; for n-alkanes, they are linear functions of the number of carbon atoms. Since the TE depends only on the corresponding physicochemical parameter, it can be calculated for any hydrocarbon using the same relationships as those known for n-alkanes. The optimal topological index (OTI) was constructed using the chemical structure matrix for TEs. For the ONs of alkanes and cycloalkanes, a model with impressive statistical characteristics [$R^2 = 0.999$ and $s = 0.829$ for the training set ($n = 41$), and $R^2 = 0.990$ and $s = 2.620$ for the test set ($n = 37$)] was derived as an example.

Cetane number (CN) is a measure of the combustibility of diesel fuel under compression. Like octane number for gasoline, CN for the diesel measures how quickly it autoignites under diesel engine conditions. It is rather difficult to measure the CN properly; therefore, for most practical purposes, the cetane index is used. It can be calculated on the basis of the density and distillation range of the oil. Diesel engines run well with a CN between 45 and 50. After 50, the performance of the fuel reaches a plateau. The alpha form of methylnaphthalene is given a standard value of 0, and cetane ($C_{16}H_{34}$) is given a standard value of 100. However, there is very little actual cetane in the diesel fuel. Some fuel additives used to raise the CN are alkyl nitrates and di-*tert*-butyl peroxide, while aromatic additives reduce ignition quality.

Like octane number, the CN also depends on the molecular composition of the fuel that can be measured by gas chromatography (GC) or inferred from different spectroscopic methods, such as Fourier transform infrared spectroscopy, nuclear magnetic resonance (^{1}H NMR, ^{13}C NMR), etc. Numerous attempts have been made in the past to correlate the CN with various physical and chemical attributes of the diesel fuel. These include correlations based on bulk properties such as API (American Petroleum Institute) gravity, boiling points, and aniline points. Ladommatos and Goacher[842] tested 22 empirical CN equations for their prediction ability using the data for more than 500 fuels collected from the literature. These equations were used routinely to monitor CN in activities such as fuel blending at the refinery. The equations proposed predicted the CN with $s < 2$. Untypical diesel fuels, such as vegetable oils and diesel blends containing alcohols, were predicted less accurately, indicating that they could be outside the domain of these equations. The measured CN of diesel fuels was found to correlate very well with their aniline point (AP). According to the authors, the most accurate equation (with $s = 1.56$ for 267 fuels) belonged to the Canadian General Standards Board containing AP, viscosity, various distillation temperatures, and the density of the fuel.

Yang et al.[843] used a backpropagation ANN to correlate and predict the CNs of 21 isoparaffins and 120 diesel fuels. For the isoparaffins, 10 branched paraffins were employed to train the ANN using the group additivity method to express the degree of branching. According to the branching positions in the molecular structure, four carbon groups, plus normal boiling point, were taken as input elements to train the network. The input values used (N_i/N_t%) were the fractions of the number of C atoms in an individual group (N_i) over the total number of C atoms of that specific isoparaffin (N_t). The R^2 for the test set was 0.97. For the diesel fuels, the best model was obtained with 8 parameters, viz. density, viscosity, aniline point, and distillation temperatures (IBP, 10%, 50%, 90%, and FBP), as inputs. The trained ANN model for the diesel fuels gave $R^2 = 0.86$ and $s = 1.62$. In a later study, Yang et al.[844] used ANNs to correlate and predict the CN and the density of 69 diesel fuels from chemical composition. The CN and density were correlated to 12 hydrocarbon groups in the diesel fuels determined by liquid chromatography (LQ) and gas chromatography−mass spectrometry (GC-MS). The best among the tested ANN architectures for correlating the CN was a general regression neural network (GRNN). The mean absolute error for the test set of 21 diesel fuels was 1.23 (CN). Predictive equations were also developed using the standard MLR method. It was found that the ANN approach provided better results for complex nonlinear problems such as the correlation of the CN with the hydrocarbon type.

Kapur et al.[845] developed MLR models for predicting seven essential physicochemical properties of diesel fuels using structural parameters as observed by ^{1}H NMR. About 60 commercial diesel samples were included in the study, and their properties were measured by standard methods. High quality models ($R^2 > 0.9$) for all studied properties were obtained. The validation with a separated test set of 20 samples gave equally high R^2 values except for the CN ($R^2 = 0.79$). The latter was related to the evident presence of the cetane improvers in the samples, which would not be detected by this method. The same group of scientists carried out an ANN study[846] on the same data set to improve the prediction of the CN. The NMR spectra were divided into 18 structural regions, which were reduced to 8 parameters using a primary ANN (18:12:8:12:18). The hidden layer containing eight nodes was used as an input. Six samples were identified as outliers and removed from the data set. Coefficient values of $R^2 = 0.91$ and 0.85 for the training ($n = 36$) and the validation ($n = 18$) sets were obtained. An ANN model for the cetane index (CI) showing a much higher correlation (>0.97) between the actual and predicted CI values was also developed. The CI, contrary to the CN, is independent of the presence of ignition improvers.

Ghosh and Jaffe[847] developed a simple composition-based model for predicting the CN of diesel fuels. The model can be applied to any diesel fuel regardless of the refining process it originates from, and it is used to support various product quality predictions for diesel fuels in ExxonMobil's refineries worldwide.[847] The CN was correlated to nine different hydrocarbon classes containing a total of 129 different hydrocarbon "lumps" determined by a combination of supercritical fluid chromatography, gas chromatography, and mass spectroscopic methods. A total of 203 diesel fuels derived from 45 diesel-range refinery process streams and their 158 commercial blends were considered. The experimental data set was split into training (90% of the samples) and test (10% of the samples) subsets. A constrained least-squares minimization problem was solved using the Levenberg−Marquardt algorithm in order to regress the parameters of the model. Across all the used data, the model predicted the CN with the standard error, SE = 1.25 units, which is comparable to the experimental error. This result is also superior to the predictions based on the ASTM D4737 model for CN with the SE = 3.32 units using this data.

Santana et al.[848] have conducted a study on the CN improvement strategies in diesel fuels. As an essential part of their research, a ANN QSPR model was created for the estimation of the CN of individual components of diesel fuel based on 147 hydrocarbons classified as n-paraffins, isoparaffins, cycloparaffins, olefins, and aromatics. Two separate correlations, one for the paraffins and the other for olefins and aromatics, were made using multilayer perception ANNs on the molecular descriptors calculated by MDLQSAR. Separation of the data set found justification from combustion chemistry. The R^2 and s were equal for the models with both data sets, being 0.89 and 8 CNs, respectively. A few CNs measured experimentally by the authors were used for external validation of the models.

Smolenskii et al.[849] derived QSPR models for the prediction of the cetane numbers of alkanes and cycloalkanes introducing a purely theoretical approach similar to their study of octane numbers[841] reviewed above. A recently proposed computational scheme[850] was used to obtain $R^2 = 0.99998$ and $s = 0.117$ for the training set of 27 compounds, and $R^2 = 0.99255$ and $s = 2.49$ for the 44 test set compounds based on optimal topological indices (OTI) from the chemical structure matrix designed ad hoc. The accuracy of the prediction results encouraged the authors[850] to apply the model for the estimation of the CNs of a collection of 180 unstudied hydrocarbons with the number of carbon atoms, $n \leq 10$.

The ON of gasoline and the CN of diesel fuels being in the center of focus by the refineries have benefited greatly from the use of prediction models. As an alternative to the QSPRs using various empirical properties, the more recent approaches incorporating calculated molecular descriptors have displayed relatively high accuracy in addition to the advantages of avoiding expenditure of experimental resources.

6.9. Rubber Vulcanization Rates

A major commercial interest in the rubber vulcanization process is the efficiency of heterocyclic sulfenamide and sulfenimide accelerators, which provide a delay interval before the onset of sulfur cross-linking. The delay is necessary in processing large rubber items such as tires. Morita[851] found that inductive constants, σ^*, correlated reasonably well with the vulcanization activity of substituted phenylthio- and anilino-benzothiazole compounds. Two linear relationships with opposite signs were obtained for *N*-substituted phenyl- and *N*-alkyl-sulfenamides. Longer scorch delays were observed for phenyl compounds with electron-withdrawing substituents and for compounds with sterically hindered alkyl groups. Amino derivatives of higher basicity were characterized by faster acceleration rates, higher cross-linking efficiencies, and longer scorch delays.

Together with colleagues at Flexsys, our group studied the kinetics of vulcanization of styrene–butadiene rubber.[852] QSPR modeling was done both on the parent molecular accelerators (12 sulfenamides, 11 sulfenimides) and also on zinc complexes of the accelerators with thiolate fragments. Experimental characteristics such as the time to scorch, t_s2, and the maximum rate of cure, MXR, measured at 426 K, were used for correlation with the accelerator structures. The structures were represented by a variety of descriptors, including those derived from AM1 semiempirical quantum chemical calculations. The R^2 of the models for both studied properties ranged from 0.925 to 0.967 while the model descriptors supported previously proposed mechanisms describing the origin of the delayed action of fast curing sulfenamide accelerators. In addition, the results supported a carbanionic concerted mechanism for the sulfurization and cross-linking reactions. Based on the gained knowledge, a new structure was proposed as the active sulfurating intermediate.

6.10. Glass Transition Temperatures

The glass transition temperature, T_g, is a fundamental characteristic of amorphous polymeric materials: plastics, glasses, rubber, and other amorphous materials such as organic light-emitting-diode (OLED) materials. Below the T_g, the material becomes rigid and brittle because of loss in relative mobility of its molecules. For cross-linked thermosetting plastics, this process is irreversible. Thus, rubber objects once brought below their T_g will shatter. The T_g depends on the chain mobility, that is affected by the molecular weight or length of the polymer molecules, the flexibility of the chain, and its interactions with other chains. In addition, T_g is also influenced by the presence of additives, fillers, and/or impurities. Due to the large and variable size of polymer molecules, their properties are modeled by extrapolation from their monomers or repeat units. Most works handle homopolymers but some copolymers and cross-linked polymers have also been introduced. The overview of the theoretical models for the prediction of T_g is summarized in Table 15.

By using the group additive property (GAP) theory, Van Krevelen[853] predicted several polymer properties. In the framework of this theory, the property under consideration is assumed to be a scalar sum of the corresponding properties averaged for the component chemical groups of compounds with available experimental data. A more universally applicable atomistic QSPR model was developed by Bicerano[802] with R^2 of 0.95 and s of 24.65 K for a data set of 320 polymers. The T_g was related to the solubility parameter and the weighted sum of 13 topological bond connectivity parameters of the monomer structures. An ANN model was developed by Sumpter and Noid[857] using topological indices used by Bicerano for a data set of 320 compounds (T_g values ranging from 50 to 700 K). On the same data, the PropNet technique was applied, which could predict T_g values with s = 8.8 K and R^2 = 0.984.[858] Koehler and Hopfinger[855] combined molecular modeling with a GAP model to utilize 3D-molecular information in estimating the T_g of 30 structurally diverse linear polymers. T_g was considered as a function of the conformational entropy and the mass moments of the polymer calculated for the repeating unit, taking into account intermolecular interactions.[856] A highly significant QSPR (R^2 = 0.91, s = 15.6 K) was developed for 35 polymers using MLR analysis involving backbone and side chain entropies, backbone mass moments, and the intermolecular energies of the O^- and H^+ probes. Wiff et al.[854] found a semiempirical method for predicting the T_g of 178 linear polymers, 12 random copolymers, and selected cross-linked networks, from their chemical structure. For new moieties not included in the database, a scaling technique of similar moiety contributions was proven successful.

Joyce et al.[859] used ANNs with error back-propagation (BPANN) to build models for T_g prediction based on the monomer structures of 360 polymers. A series of indicator variable descriptors were calculated based on the SMILES representation of the monomers. The model predicted the T_g values for a testing set of polymers with an rms error of 35 K.

Gao and Harmon[864] correlated T_g for poly(p-alkyl styrenes), polyolefins, poly(alkyl methacrylates), and poly(alkyl acrylates) with bond radii-based structural parameters based on the repeat unit. The model provides prediction of the T_g for both the linear and highly branched polymers and also allows the extraction of the contribution of hydrogen bonding in polymethacrylates and polyacrylates. The model resulted in predictions with $R^2 > 0.97$ and $s < 7.5$ °C for all the studied polymer classes.

Waegell and co-workers[860,861] described an EVM (energy, volume, and mass) QSPR model based on molecular mechanics and molecular dynamics calculations for linear and branched aliphatic acrylate and methacrylate polymers with bulky ester substituents. In this approach intra- and interchain interactions were taken into account directly by calculating an energy density function related to the cylindrical volume of a 20 monomer unit polymer segment. The EVM method for 16 linear and branched alkyl acrylate and methacrylate polymers gave s = 12 K and R^2 = 0.96.[799] The EVM approach was also applied successfully to a set of polystyrenes.[863] The model correctly quantified the effects of the substituent position on the phenyl ring that have a high impact on the T_g. With these works the authors have argued also for the superiority of class-specific models in the form of the so-called “designer” models over global models.

Tan and Rode[865] found that the quantum chemical methods utilizing PM3 and especially AM1 parametrizations led to superior models compared to those obtained by the Gasteiger–Hückel method. Partial charges of some important atoms in the monomer together with the degree of substitution and chain length of the hydrocarbon group of ester or amide functions of the monomer were used as descriptors.

Katritzky et al.[866] introduced a four-parameter model with R^2 = 0.928 for 21 medium molecular weight polymers and copolymers based on their repeat units. The descriptors selected indicated the importance of intermolecular electrostatic interactions between the polymer chains, followed by the degree of branching and H-bonding capabilities. On a larger data set, CODESSA produced a five-parameter correlation (R^2 = 0.946, SE = 0.33 K mol/g)[867] relating glass transition temperatures (T_g/M) for a diverse set of 88 linear uncross-linked homopolymers including polyethylenes, polyacrylates, polymethylacrylates, polystyrenes, polyethers, and polyoxides. The descriptors, calculated based on a trimeric repeating unit, related to the shape/bulkiness of the repeat units (as reflected by the moment of inertia and the Kier shape index) and intermolecular electrostatic interactions

Table 15. Summary of QSPR Modeling of Glass Transition Temperatures, T_g[a]

no.	compound	N (n_{valid})	method[b]	model descriptors, n_d	R^2	s (K, °C)	R^2_{valid}	s_{valid} (K,°C)	ref
1	polymers		GAP	a weighted sum of scalar quantities of functional groups					Van Krevelen[853]
2	linear and copolymers	190		semiempirical					Wiff et al.[854]
3	diverse linear polymers	30	MLR	conformational entropy and mass moments, based on repeat unit	0.91	19			Hopfinger et al.[855]
4	diverse linear polymers	35	GAP, MLR	5 (backbone and side chain entropies, backbone mass moments, intermolecular energies of the O^- and H^+ probes)	0.91	16			Koehler and Hopfinger[856]
5	diverse polymers	320	MLR	14 (solubility param, weighted sum of 13 structural params)	0.95	25			Bicerano[802]
6	diverse polymers (50...700 K, incl tactic and cross-linked)	320	CNN	topological indices (repeat unit structure)					Sumpter and Noid[857]
7	diverse polymers (50...700 K, incl tactic and cross-linked)	320	CNN	(repeat unit structure)	0.98	8			Sumpter and Noid[858]
8	diverse linear homopolymers	360 (89)	BPANN	indicator variables from SMILES (based on monomers)				35	Joyce et al.[859]
9	acrylates and methacrylates with ester substituents	50	molecular mechanics	3 (energy of a polymer segment conformation, its volume, repeat unit molar mass)	0.83				Waegell et al.[860]
10	acrylates and methacrylates with bulky rigid substituents	23 (24)	EVM (MLR)	3 (same as above)	0.90	20	0.91	27	Waegell et al.[861]
11	acrylates and methacrylates	16 (18)	EVM (MLR)	3 (same as above)	0.96	12		13	Waegell et al.[862]
12	polystyrenes	10 (19)	EVM (MLR)	3 (same as above)	0.97	5.6			Waegell et al.[863]
13	poly(*p*-alkyl styrenes), polyolefins	10	structural parameter method	bond radii-based structural parameters for C, O, and H, based on repeat unit			0.99	6.1	Gao and Harmon[864]
	poly(alkyl methacrylates) and	10					0.99	7.1	
	poly(alkyl acrylates)	10					0.98	7.2	
		9					0.98	2.9	
14	poly(acrylic acid), poly(methacrylic acid), polyacrylamide and their derivatives			AM1; PM3 (based on monomers)					Tan and Rode[865]
16	medium molecular weight homo- and copolymers	21	MLR	4 (CODESSA descriptors based on repeat units, AM1)	0.93	3.05			Katritzky et al.[866]
17	diverse linear homopolymers (polyethylenes, polyacrylates, polymethylacrylates, polystyrenes, polyethers, and polyoxides)	88	MLR	5 (CODESSA descriptors based on monomers, AM1)	0.83	33			Katritzky et al.[867]
					(0.95[c])	(0.33[c])		(0.94[c])	
18	diverse linear homopolymers	88	MLR	10 topological indices	(0.89[c])	(0.44[c])			Garcia-Domenech and de Julian-Ortiz[868]
19	diverse linear homopolymers	88	MLR	5 ($\sum MV_{(ter)}(R_{ter})$, L_F, ΔX_{SB}, $\sum$PEI, Q±)	0.91	21			Cao and Lin[869]
20	diverse linear homopolymers	84	RBF NN	5 ($\sum MV_{(ter)}(R_{ter})$, L_F, ΔX_{SB}, $\sum$PEI, Q±)	0.93				Afantitis et al.[870]
21	polystyrene monomers	96 (11)	stepwise MLR	rigidness of side chain R_{SC}, stiffness of main chains S_{MC}, density of H-bond D_{HB}, molecular polarizability effect M_{PE}	0.92	15	0.89		Yu et al. (Gao)[871]
22	random styrenic and ANMA copolymers	32 (16)	stepwise MLR	3 quantum chemical (DFT) based on repeat units	0.983	6.92	0.978		Yu et al.[872]
23	polyvinyls, polyethylenes, and polymethacrylates	22 (38)	stepwise MLR	2 quantum chemical (DFT) based on repeat units	0.908	26.7	0.906		Yu et al.[873]

Table 15. Continued

no.	compound	N (n_{valid})	method[b]	model descriptors, n_d	R^2	s (K, °C)	R^2_{valid}	s_{valid} (K,°C)	ref
24	polystyrenes, polyacrylates, and polymethacrylates		BPANN	4 (rigidness, chain mobility, polarizability, and net charge on most negative atom)	0.912	20.5 (rms)	0.912	20.2 (rms)	Yu et al.[874]
25	polyamides	63 (15)	MLR	B3LYP/6-31G(d) level descriptors on repeat units	0.82	22	0.79	23	Gao et al.[875]
			BPANN		0.86	15	0.81	16	
26	diverse polymers (188...475 K)	148 (17)	SA, GA, 10 CNNs	10 (topological, electronic, and geometric descriptors based on monomers)	0.98	10 (rms)	0.92	22 (rms)	Mattioni and Jurs[876]
27	diverse polymers (188...673 K)	226 (25)	SA, CNN	11 (topological descriptors based on repeat units)	0.96	21 (rms)	0.96	22 (rms)	Mattioni and Jurs[876]
28	(meth)acrylic polymers (197...501 K)	80 (15)	RecNN	variable-size labeled structures	0.98	10	0.92	13	Duce et al.[877]
29	(meth)acrylic polymers (162...501 K)	127 (27)	RecNN	directed positional acyclic graphs (DPAGs)	0.997	3.6	0.90	19	Duce et al.[878]
30	poly(meth)acrylates	217 (54)	RecNN	graphical representation as labeled trees	0.97	11	0.87	19	Duce et al.[879]
31	polyphosphates and polyphosphonates	10	MLR	2 (molecular mechanics and AM1 based on dimers)	0.88	4.2			Funar-Timofei et al.[880]
32	small molecules, including 24 OLED materials (73...455 K)	81 (22)	GA, MLR	7 (topological, spatial, electrostatic, thermodynamic, and structural)	0.99	8.8	0.98	14	Kim et al.[881]
33	diverse OLED materials (311...468 K)	73 (15)	MLR	6 (CODESSA descriptors, AM1)	0.93	10.7		18 (AAE)	Yin et al.[882]
34	OLED materials (311...468 K)	80	MLR	5 (topological indices)	0.93	10.5			Xu and Chen[883]
35	amine-cured epoxy copolymers	13	MLR	4 (CODESSA descriptors based on repeat units, AM1)	0.998				Morrill et al.[884]
36	polyacrylates and polymethacrylates	22	MLR	3 (topological)	0.98	8.3			Dai et al.[885]
37	diverse homopolymers	251 (20)	ANN			8		7	Sun et al.[886]
38	diverse homopolymers (152...550 K)	241 (15)	fuzzy set theory	12 structural groups, based on repeat unit	0.95	19	0.90	30	Sun et al.[887]
39	diverse homopolymers (183...450 K)	235 (10)	fuzzy set theory	95 structural groups and their interactions, based on repeat unit	0.98	8	0.98	9	Sun et al.[888]
40	monosaccharides	6	MLR	3 (moment of inertia/atoms, LUMO energy, max. partial charge for an O atom)	0.99	1			Dyekjaer and Jonsdottir[889]
41	polymethacrylates	35	BPANN	4 (3 quantum chemical (DFT), 1 length of side chain minus 1.356, based on repeat units)	0.98				Liu et al.[890]
			stepwise MLR		0.92	15.9			
42	linear homopolymers	261	stepwise MLR	37 signature molecular descriptors	0.93	28 (MAE)			Brown et al.[891]
43	polyacrylamides	20	stepwise MLR	2 (thermal energy and total energy using DFT, based on repeat units)	0.92	21.7			Liu et al.[892]
44	L-tyrosine derived homo, co-, and terpolymers: polycarbonates and polyarylates	100	linear correlation	1 mass-per-flexible-bond (*M*/*f*) of a polymer repeat unit	0.905	6.4 (AAE)			Schut et al.[893]
45	fluorine-containing polybenzoxazoles	52 (17)	BPANN	3 (chain mobility and rigidness, and a group indicator, based on repeat units)	0.96	2.35 (rms)	0.956	2.30(rms)	Ning[894]
			MLR		0.88	10.5	0.91		

[a] N, number of all compounds; R^2, squared correlation coefficient; s, standard deviation. [b] EVM, energy, volume, and mass; RBF, radial basis function; GAP, group additive property; SA, simulated annealing; GA, genetic algorithm; OLED, organic light-emitting-diodes; GIM, group interaction modeling. [c] T_g/M (K mol/g); AAE, average absolute error; MAE, mean absolute error.

(accounted for by the most negative atomic charge, HASA-2/TFSA and FPSA-3). The model yielded a standard error of 32.9 K for the predicted T_g values. Cao and Lin[869] tested the same set of 88 polymers using their own designed descriptors expressing chain stiffness and intermolecular forces for polymers having polar groups to derive a QSPR with $R^2 = 0.906$ and $s = 20.86$ K. Afantitis[870] treated the same polymers and descriptors, using a radial basis function

(RBF) ANN to obtain an improved correlation coefficient (see Table 15).

Sun et al. conducted a series of works for modeling T_g values of diverse sets of over 230 homopolymers based on ANN[886] and fuzzy set theory.[887,888] Better results were obtained by the fuzzy theory with consideration of the interactions between the structural groups expressed as the entropy of the fuzziness ($R^2 = 0.98$, $s = 8$ K)[888] compared to the model without them (R^2=0.83, $s = 25$ K).[887]

Garcia-Domenech and de Julian-Ortiz[868] used graph theoretical indices based on monomers to predict T_g of a group of addition polymers by a model with 10 variables for T_g/M ($R^2 = 0.893$, SEE = 0.439). The average error (AAE) in the prediction of T_g was 12.7%.

Mattioni and Jurs[876] developed various CNNs with 10 and 11 descriptors, to predict T_g values for two diverse sets of polymers based on the monomers and the repeat units, respectively. The descriptors were selected using simulated annealing (SA) and genetic algorithms (GA). The test sets rms errors of the two models were above 21 K.

Kim et al.[881] used MLR to predict T_g for 103 molecules including OLED materials. GA was used to select the model descriptors (one topological, one thermodynamic, one spatial, one structural, and three electrostatic). The model was developed using a randomly chosen training set of 81 and a prediction set of 22 compounds. R^2 for the training set was 0.989, and the AAE was 8.8 K. For the prediction set, R^2 was 0.976, and AAE was 13.9 K, respectively.

Yin et al.[882] proposed a six-parameter correlation with $R^2 = 0.927$ and AAE of 8.5 K for a diverse set of 73 OLED materials. For a test set of 15 OLED materials, an AAE of 17.9 K was obtained. The descriptors involved reflect the effect of chain stiffness, electrostatic interactions, as well as vibrational motions on T_g.

Dai et al.[885] presented a side chain pulled-along model of polymers to correlate T_g to the chemical structure of 13 polyacrylates and 9 polymethacrylates. Three-parameter models, which include the number of carbon branches of side chains (B), the group charge of side chains (C), and the topological length of side chains (L), were obtained with $R^2 = 0.989$ ($s = 3.8$ K) and $R^2 = 0.993$ ($s = 4.8$ K), respectively. The number of carbon branches of side chains was used as the descriptor for the stereoeffect. For a common three-descriptor model covering both the 13 polyacrylates and the 9 polymethacrylates, the R^2 was 0.980 ($s = 8.3$ K).

Properties of monosaccharides were modeled by Dyekjaer and Jonsdottir[889] based on molecular descriptors obtained from molecular mechanics and quantum chemical calculations. Saccharides exhibit a large degree of conformational flexibility; therefore, a methodology for selecting the energetically most favorable conformers was developed. The QSPR for the T_g of 6 monosaccharides including 17 conformations contained 3 descriptors calculated at the B3LYP/6-31++g level and extracted with the CODESSA software, $R^2 = 0.99$, $s = 0.96$ °C.

Morrill et al.[884] developed a designer QSPR ($R^2 = 0.998$), based upon molecular properties calculated using the AM1 semiempirical quantum mechanical method, to predict T_g of amine-cured epoxy resins based on the diglycidyl ether of bisphenol A. By applying an ad hoc treatment based on the elementary probability theory to the QSPR analysis, a method was developed for computing bulk polymer T_g for stoichiometric and nonstoichiometric monomeric formulations. For the validation of the model predictions, a model polymer was synthesized.

Xu and Chen[883] performed a QSPR study between topological indices and the T_g of a diverse set of 80 OLED materials. A five-parameter correlation with $R^2 = 0.930$ and an AAE of 7.7 K was obtained through stepwise MLR analysis with $R^2_{CV} = 0.916$. The results were comparable to those of Yin et al.[882]

Multilinear QSPR models for series of polyphosphates and polyphosphonates were reported by Funar-Timofei et al.[880] using molecular mechanics and AM1 calculated descriptors for polymer dimers. R^2 of 0.88 was obtained for 10 samples with the Sterimol B1 parameter and torsion angle C1, associated with packing preferences and the polymer backbone flexibility, respectively.

A method for solving the inverse QSPR problem which facilitates the design of novel polymers with targeted properties has recently been presented by Brown et al.[891] The signature molecular descriptor, consisting of a column of molecular fragments and a column of the occurrences of these fragments, was used in both the forward and inverse QSPR approaches on T_g, heat capacity, and density. The forward stepwise MLR method was employed to develop QSPRs for T_g of 261 linear homopolymers. Using 37 descriptors, a model with $R^2 = 0.93$ and $R^2_{CV} = 0.81$ (with mean absolute error of 27.97) was reported. The QSPRs obtained were used for the inverse task to design a poly(N-methyl hexamethylene sebacamide) with a desired T_g, which was not part of the training set.

Duce et al.[877,878] predicted polymer properties from structured molecular representations using recursive neural networks (RecNN). For this purpose, a hierarchical set of labeled vertexes connected by edges that belong to subclasses of graphs, such as rooted trees, constituted the input. This representation allows variable-size structures with specified tacticity as input and bypasses the limitations associated with vectorial representations of data. The data-features are implicitly generated starting from the structured molecular representation and according to the specific task. This method was applied to calculate the T_g of 90 (meth)acrylic polymers containing alkyl, thiaalkyl, and fluoroalkyl ester groups, polyacrylamides, and α-substituted polyacrylics with different stereoregularity. The mean error of a preliminary hold-out cross-validation test was in the same range as the literature data spread (30 K) for stereoregular polymers. The same approach was applied to the prediction of the T_g of 277 poly(meth)acrylates.[879] As an improvement, the molecular representation through hierarchical structures was extended by two methods, named *group* and *cycle breaking*, in order to render cyclic structures. Standard unique molecular description systems, i.e. Unique SMILES and InChI, were exploited.

Gao et al.[895] developed QSPRs for T_g of 78 polyamides using both MLR and error back-propagation ANNs. All descriptors were calculated from molecular structures optimized at the B3LYP/6-31G(d) level. The MLR model had $R^2 = 0.823$ and $s = 22.47$, $R^2_{test} = 0.792$ and $s_{test} = 23.24$. The ANN model was better: $R^2 = 0.859$ and $s = 14.90$, $R^2_{test} = 0.810$ and $s_{test} = 16.44$. The value of T_g was affected mainly by the molecular energy and polarity.

Yu et al.[871] used a set of new molecular descriptors, the rigidness of side chain R_{SC}, the stiffness of main chains S_{MC}, the density of hydrogen bonds D_{HB}, and the molecular polarizability effect M_{PE}, obtained directly from polystyrenes monomer structures, to predict the values of 96 polystyrenes and generate a QSPR model by stepwise MLR analysis, with $R^2 = 0.920$ and $s = 15.20$ K. In a following work by the same authors,[872] a QSPR model, based on three quantum chemical descriptors (α, q^+, and C_v) obtained from the monomers using the density functional theory (DFT), was developed to predict T_g of random copolymers, such as poly(styrene-*co*-acrylamide) (SAAM), poly(styrene-*co*-acrylic acid) (SAA), poly(styrene-*co*-acrylonitrile) (SAN), poly(styrene-*co*-butyl acrylate) (SBA), poly(styrene-*co*-methyl acrylate) (SMA), poly(styrene-*co*-ethyl acrylate) (SEA), and poly(acrylonitrile-*co*-methyl acrylate) (ANMA). The QSPR models having R^2 of 0.982 and $s = 6.924$ K had good predictive power. These descriptors can reflect the relative rigidity of the side groups and the chain backbone, the essential parameters governing the nature of glass formation in polymers. By carrying out DFT calculations for 60 polyvinyls, polyethylenes, and polymethacrylates repeating units at the B3LYP/6-31G(d) level, two quantum chemical descriptors, the molecular traceless quadrupole moment Θ and the molecular average hexadecapole moment Φ, were used to predict the T_g.[873] A physically meaningful QSPR model having R^2 of 0.908 for the training set and 0.952 for the test set was generated using stepwise MLR analysis. Compared with the existing QSPR models, the proposed model with only two multipole moment descriptors was the most simple.

Liu et al.[890] optimized the structural units of 35 polymethacrylates and calculated their quantum chemical descriptors *via* the DFT method at the 6-31G(d) level. The model obtained by BPANN performed better than the one found by MLR. Four descriptors were selected by the stepwise regression: the thermal energy, E_{thermal}, polarizability, α, atomic net charge of C^6, $Q^6{}_C$, and, finally, the length of the side chain minus 1.356, $|L - 1.356|$, where $L = 1.356$ nm is the length of the side chain at which T_g is lowest. The latter descriptor was the most significant according to the *t* test, and its correlation with T_g was $R^2 = 0.89$.

Liu et al.[892] have used DFT calculations to derive descriptors of the repeat units for QSPR modeling of T_g. A model with two quantum chemical descriptors was selected with $R^2 = 0.92$, $s = 21.7$ K based on 20 polyacrylamides. The model descriptors were easily relatable to the relevant structural features of polymers; for example, the thermal energy, E_{thermal}, is larger for longer side chains, decreasing the T_g, and the more negative the total energy, E_{HF}, the more it enhances intermolecular forces and the stiffness of the chains, thus increasing the T_g.

Schut et al.[893] have employed a semiempirical method based on the mass-per-flexible-bond (*M*/*f*) principle to explain the large variation in T_g values in a library of 132 L-tyrosine derived homo-, co-, and terpolymers. Polymer class-specific behavior was observed in T_g vs *M*/*f* plots and explained in terms of different densities, steric hindrances, and intermolecular interactions of chemically distinct polymers. The method was found to be useful in the prediction of polymer T_g: AAE ranged from 6.4 to 3.7 K for the three polymer classes that yielded straight lines on the plot. The proposed method can also be used for structure prediction of polymers to match a target T_g value, by keeping the thermal behavior of a terpolymer constant while independently choosing its chemistry.

Ning[894] has successfully correlated T_g of fluorine-containing polymers ($n = 35$, $R^2 = 0.96$, rms = 2.35 K) using a [3-1-1] BPANN with three descriptors: the number of atoms in the flexible group of the main chain, n_A, the total number of $-CF_2CF_2O-$ groups in the repeating unit, *m*, and the number of certain $-CF_2-$ groups, n_{CF2}. The authors experienced a turning point in T_g near $n_{\mathrm{CF2}} = 3$. Therefore, the final descriptor was defined as $N_{\mathrm{CF2}} = |\, n_{\mathrm{CF2}} - 3|$, and it can express rigidness of the polymer chain. For the prediction set of 17 polymers, rms = 2.3 K.

The extensive research reviewed in this paragraph has produced useful knowledge toward the discovery of practically applicable polymeric (bio)materials. QSPRs have a great potential in this field as a tool for presynthesis T_g estimation for effective polymer selection.

6.11. Contact Angles for Pharmaceutical Solids

The surface energies of solids are of great technological importance for interfacial phenomena, such as adhesion, adsorption, and wettability of compounds. Unlike those of liquids, the surface energies of solids cannot be determined directly. Indirect measurements of the surface tension of solids by contact angle (CA) methods were reviewed recently.[896] Consequently, the prediction of the contact angle from structure is of considerable value. In pharmaceutical practice, the contact angle is used for determining and understanding the performance of a pharmaceutical solid. This includes the form design, selecting a suitable binder for a drug to ensure its bioavailability, estimating the suspension stability between drugs and excipients, and choosing a suitable coating material. QSARs could thus prove useful for predicting the wettability of pharmaceutical powders at an early stage during development of the formulation.

Sheridan et al.[897] presented equations to predict surface properties of 16 pharmaceutical powders of three structural types including homologous series of alkyl *p*-hydroxybenzoates and imidazoles, and HMG-CoA reductase inhibitors: all are examples of comparatively high molecular weight drugs. The descriptors were calculated for structures optimized at the MNDO semiempirical level of the molecular orbital theory. A relationship with $R^2 = 0.808$ between the water contact angle and a set of superdelocalizability indices for functional groups involved in hydrogen bonding, $\sum \mathrm{SI}_{r(\mathrm{HB}+)}$, was reported. Two outliers were identified: methyl *p*-hydroxybenzoate and L-679,336: both had crystal unit cell structures different from the other members in their respective series, which may lead to an unequal distribution of hydrogen bonding groups on the crystal surface. The predicted results may therefore depend on the processing of the powder surface, which would alter the orientation of the surface molecules and the net surface energy.

A more general model was developed by Suihko et al.[898] for 25 structurally heterogeneous pharmaceutical materials. Molecular descriptors calculated included those derived from 3D molecular interaction fields containing attractive and repulsive forces between a chemical probe and a target molecule. Water (hydrophilic), hydrophobic, and carboxyl oxygen were used as probes. The experimental water contact angle was modeled using the PLS method employing fractional factorial design for descriptor selection. Two models provided promising predictions for use in cutting the cost of dosage form design. The model based on computed conformations contained 52 descriptors with $R^2 = 0.57$, $R_{\mathrm{cv}}{}^2 = 0.42$, and the model based on optimized conformations from the crystal structure database contained 32 descriptors,

$R^2 = 0.80$, $R_{cv}^2 = 0.66$. Both models yielded contact angles with standard deviation of errors of prediction of 15 degrees for an external test set of 7 pharmaceuticals. The first model was considered superior in terms of practicality due to the fact that it was accomplished utilizing only computerized techniques.

QSPR models for air−water contact angles have also been used as a measure of hydrophobicity of the material surface for the design of polymers. For the design of a diverse and focused library of synthetic biodegradable polymers, important properties such as the glass transition temperature, T_g, and the CA were selected for modeling by Reynolds.[899] QSPR equations derived using GA led to the selection of two molecular topology descriptors of the repeat unit. A subset of 17 out of 112 polymers was selected for the training set using a stochastic diversity method, SimSearch-SCA. The obtained models were tested, with the remaining 95 polymers giving R^2 for the CA between the computed and experimental values of 0.92. The QSPR models were further used to build focused libraries with specific values of T_g and CA. The focused libraries successfully identified polymers falling within specified ranges of T_g and CA.

7. Summary and Future Prospects

Quantitative structure−property relationship (QSPR) techniques have become indispensable in many aspects of the molecular interpretation of physical, chemical, biological, and technological properties. Today it would be inconceivable for any commercial, governmental, or academic group to research these fields without the help of sophisticated calculations. This paper reviews the applicability and power of the QSPR approaches for the prediction of diverse properties of chemical compounds and materials. This is due to substantial progress in the development of new, more adequate molecular descriptors and methods of derivation of multiple linear and nonlinear relationships. The QSPRs are empirical equations for formal interpolation or extrapolation of missing data. In many cases, they also give insight into the physical interactions and processes determining the properties of substances. Moreover, the ability to use exclusively theoretical molecular descriptors has provided the means to predict properties of molecular structures that are difficult to determine experimentally or even of those compounds not yet synthesized.

8. Acknowledgments

We thank Ms. Iva Stoyanova-Slavova for her help in various ways and the Kenan Trust for financial support.

9. References

(1) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125.
(2) Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, 1940.
(3) Taft, R. W. *J. Am. Chem. Soc.* **1952**, *74*, 2729.
(4) Taft, R. W. *J. Am. Chem. Soc.* **1952**, *74*, 3120.
(5) Taft, R. W. *J. Am. Chem. Soc.* **1953**, *75*, 4231.
(6) Taft, R. W., Jr. *J. Am. Chem. Soc.* **1953**, *75*, 4538.
(7) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(8) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 2636.
(9) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419.
(10) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. *J. Pharm. Sci.* **1975**, *64*, 1971.
(11) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(12) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.
(13) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(14) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic: New York, 1976.
(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(16) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vols. *1 and 2*.
(17) King, R. B., Ed. *Chemical Applications of Topology and Graph Theory*; Elsevier: Amesterdam, The Netherlands, 1983.
(18) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structure*; Wiley-Interscience: New York, 1983.
(19) Rouvary, D. H. *Sci. Am.* **1986**, *254*, 40.
(20) Seybold, P. G.; May, M.; Bagal, U. A. *J. Chem. Educ.* **1987**, *64*, 575.
(21) Randić, M.; Guo, Xiaofeng; Oxley, T.; Krishnapriyan, H. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 709.
(22) Katritzky, A. R.; Gordeeva, E. V. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835.
(23) Estrada, E.; Rodriguez, L. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037.
(24) Schultz, H. P.; Schultz, T. P. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 107.
(25) Pogliani, L. *Chem. Rev.* **2000**, *100*, 3827.
(26) Randić, M.; Balaban, A. T.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 593.
(27) Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.
(28) Roy, K. *Mol. Diversity* **2004**, *8*, 321.
(29) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.
(30) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
(31) Randić, M. *J. Mol. Graph. Model.* **2001**, *20*, 19.
(32) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley & Sons: New York, 1986.
(33) Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 645.
(34) Stupper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Functions*; Wiley-Interscience: New York, 1979.
(35) Basak, S. C. *Med. Sci. Res.* **1987**, *15*, 605.
(36) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.
(37) Katritzky, A. R.; Fara, D. C. *Energy Fuels* **2005**, *19*, 922.
(38) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.
(39) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679.
(40) Katritzky, A. R.; Fara, D. C.; Yang, H.; Tämm, K.; Tamm, T.; Karelson, M. *Chem. Rev.* **2004**, *104*, 175.
(41) Katritzky, A. R.; Fara, D. C.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. *Curr. Top. Med. Chem.* **2002**, *2*, 1333.
(42) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551.
(43) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. *EXCLI J.* **2009**, *8*, 74.
(44) Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11.
(45) Duchowicz, P. R.; Castro, E. A. *Int. J. Mol. Sci.* **2009**, *10*, 2558.
(46) Jónsdóttir, S.; Ó, Jørgensen, F. S.; Brunak, S. *Bioinformatics* **2005**, *21*, 2145.
(47) Sutter, J. M.; Jurs, P. C. *Data Handling Sci. Technol.* **1995**, *15*, 111.
(48) Novič, M.; Vračko, M. *Data Handling Sci. Technol.* **2003**, *23*, 231.
(49) Baumann, K. *Trends Anal. Chem.* **1999**, *18*, 36.
(50) Randić, M. Similarity Methods of Interest in Chemistry. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, S. I., Ed.; Gordon and Breach Publishers: 1996; pp 1−100.
(51) Randić, M. Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; John Wiley & Sons: New York, 1990; pp 77−145.
(52) Eriksson, L.; Johansson, E.; Muller, M.; Wold, S. *J. Chemometrics* **2000**, *14*, 599.
(53) Tropsha, A.; Gramatica, P.; Gombar, V. *QSAR Comb. Sci.* **2003**, *22*, 69.
(54) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Anal. Chim. Acta* **2004**, *515*, 199.
(55) Topliss, J.; Edwards, P. *J. Med. Chem.* **1979**, *22*, 1238.
(56) Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chem. Inf. Model.* **2009**, *49*, 1669.
(57) Dobson, C. M. *Nature* **2004**, *432*, 824.
(58) Tetko, I. V. *Drug Discovery Today* **2005**, *10*, 1497.
(59) Cramer, K. *Essentials of Computational Chemistry. Theory and Models*; John Wiley &Sons: New York, 2002.
(60) Mendenhall, W.; Scheaffer, R. *Mathematical Statistics with Applications*; Duxbury Press: 1973.

(61) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993.
(62) Gedeck, P.; Rohde, B.; Bartels, C. *J. Chem. Inf. Model.* **2006**, *46*, 1924.
(63) Lao, Y.; Leong, H. W. *Trends in Artificial Intelligence*; 7th Pacific Rim International Conference on Artificial Intelligence; Tokyo, Japan, August 2002; Proceedings, p 345.
(64) Fletcher, R. *Practical Methods of Optimization*; John Wiley & Sons: New York, 1980; Vol. 1.
(65) Gill, P. E.; Murray, W.; Wright, M. H. *Practical Optimization*; Academic Press, Inc.: New York, 1981.
(66) Saunders, M. *J. Am. Chem. Soc.* **1987**, *109*, 3150.
(67) Yang, Z. Z.; Wang, C. S. *J. Theor. Comput. Chem.* **2003**, *2*, 273.
(68) Arulmozhiraja, S.; Morita, M. *Chem. Res. Toxicol.* **2004**, *17*, 348.
(69) Liu, S. S.; Cui, S. H.; Yin, D. Q.; Shi, Y. Y.; Wang, L. S. *Chin. J. Chem.* **2003**, *21*, 1510.
(70) Chiu, T. L.; So, S. S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 147.
(71) Chin, T. L.; So, S. S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 154.
(72) Lin, Z. H.; Wu, Y. Z.; Quan, X. J.; Zhou, Y. G.; Ni, B.; Wan, Y. *Lett. Pept. Sci.* **2002**, *9*, 273.
(73) Agrawal, V. K.; Mishra, K.; Khadikar, P. V. *Oxid. Commun.* **2003**, *26*, 14.
(74) CODESSA PRO, University of Florida, www.codessa-pro.com.
(75) POLLY, University of Minnesota, Duluth.
(76) ADAPT, The Pennsylvania State University,www.research.chem.psu.edu/pcjgroup/adapt.html.
(77) OASIS, Laboratory of Mathematical Chemistry, Bulgaria, www.oasis-lmc.org.
(78) Dragon, TALETE, Italy, www.talete.mi.it/products/dragon_description.htm.
(79) Chem-X, Accelrys, www.accelrys.com.
(80) Tsar, Accelrys Software Inc.,www.accelrys.com/products/tsar.
(81) QSARModel, MolCode Ltd., http://www.molcode.com.
(82) Sotomatsu, T.; Murata, Y.; Fujita, T. *J. Comput. Chem.* **1989**, *10*, 94.
(83) Diercksen, G. H. F., Wilson, S., Eds. *Methods in Computational Molecular Physics*; Reidel: Dordrecht, Holland, 1983.
(84) *Perspectives in Quantum Chemistry*; Jortner, J., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, 1989.
(85) Krishnan, R.; Schlegel, H. B.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 4654.
(86) Kucharski, S. A.; Bartlett, R. J. *Adv. Quantum Chem.* **1986**, *18*, 281.
(87) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
(88) Gauss, J.; Cremer, D. *Adv. Quantum Chem.* **1992**, *23*, 205.
(89) Dewar, M. J. S. *Science* **1975**, *187*, 1037.
(90) Dobchev, D. A.; Karelson, M. *J. Mol. Model.* **2006**, *12*, 503.
(91) Systat, www.systat.com/Default.aspx.
(92) Statistica, www.statsoft.com.
(93) JMP, www.jmp.com/software/.
(94) Hocking, R. R. *Biometrics* **1976**, *32*, 1.
(95) Drapper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1981.
(96) Hawkins, D. M.; Yin, Y. A. *Comput. Stat. Data Anal.* **2002**, *40*, 253.
(97) Hawkins, D. M.; Basak, S. C.; Shi, X. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663.
(98) Hao, L.; Naiman, D. Q. *Quantile Regression*; SAGE Publications: Thousand Oaks, 2007.
(99) Eriksson, L.; Andersson, P.; Johansson, E.; Tysklind, M. *Mol. Diversity* **2006**, *10*, 169.
(100) Jolliffe, I. T. *Principal Component Analysis Series; Springer Series in Statistics*, 2nd ed.; Springer: New York, 2002; Vol. XXIX, p 487.
(101) Pearson, K. *Philos. Mag.* **1901**, *2*, 559.
(102) Eriksson, L.; Andersson, P.; Johansson, E.; Tysklind, M. *Mol. Diversity* **2006**, *10*, 187.
(103) Jurs, P. C.; Bakken, G. A.; McClelland, H. E. *Chem. Rev.* **2000**, *100*, 2649.
(104) Topliss, J. G.; Costello, R. J. *J. Med. Chem.* **1972**, *15*, 1066.
(105) Rücker, C.; Rücker, G.; Meringer, M. *J. Chem. Inf. Model.* **2007**, *47*, 2345.
(106) Katritzky, A. R.; Dobchev, D. A.; Slavov, S.; Karelson, M. *J. Chem. Inf. Model.* **2008**, *48*, 2207.
(107) Calabrese, E. J. *EMBO Rep.* **2004**, *5*, S37.
(108) Hopfield, J. J. *Proc. Natl. Acad. Sci.* **1984**, *81*, 3088.
(109) Burns, J. A.; Whitesides, G. *Chem. Rev.* **1993**, *93*, 2583.
(110) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an Introduction*; VCH-Verlag: Weinheim, 1993; pp 213−228.
(111) Baskin, I. I.; Ait, A. O.; Halberstam, N. M.; Palyulin, V. A.; Zefirov, N. S. *SAR QSAR Environ. Res.* **2002**, *13*, 35.
(112) Haykin, S. *Neural Networks. A Comprehensive Foundation*; Prentice-Hall: Upper Saddle River, NJ, 1999.
(113) Jurs, P. C. *Computer Applications in Chemistry: A University Course. Comput. Educ. Chem. (Symp.)* **1984**, 1.
(114) Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.
(115) Karelson, M.; Sild, S.; Maran, U. *Mol. Simulat.* **2000**, *24*, 229.
(116) Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 360.
(117) Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 6598.
(118) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
(119) Friedman, J. *Multivariate Adaptive Regression Splines*; Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University: Stanford, CA, 1988.
(120) Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
(121) So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4347.
(122) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Conner, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189.
(123) *Machine Learning and Data Mining in Pattern Recognition*; 4th International Conference, MLDM 2005, Leipzig, Germany, July 9−11, 2005; Proceedings By Petra Perner, Atsushi Imiya; Springer: 2005; pp 62−70.
(124) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, U.K., 2000.
(125) Ignizio, J. Introduction to Expert Systems; 1991.
(126) Giarratano, J. C.; Riley, G. Expert Systems, Principles and Programming; 2005.
(127) Jackson, P. Introduction to Expert Systems; 1998.
(128) Retention and selectivity in liquid chromatography: prediction, standardisation and phase comparisons. In *Journal of Chromatography Library*; Smith, R. M., Ed.; Elsevier: Amsterdam, New York, Vol. 57; 1995.
(129) Heller, S. R.; Bigwood, D. W.; May, W. E. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 627.
(130) CAMEO distribution system, http://www.cemcomco.com /CAMEO_Distribution1139.html.
(131) Gushurst, A. J.; Jorgensen, W. L. *J. Org. Chem.* **1986**, *51*, 3513.
(132) Carreira, L. A.; Hilal, S.; Karickhoff, S. W. *Estimation of Chemical Reactivity Parameters and Physical Properties of Organic Molecules Using SPARC*; Theoretical and Computational Chemistry, Quantitative Treatment of Solute/Solvent Interactions; Politzer, P., Murray, J. S., Eds.; Elsevier Publishers: 1994.
(133) Valko, K.; Szabo, G.; Rohricht, J.; Jemnitz, K.; Darvas, F. *J. Chromatogr.* **1989**, *485*, 349.
(134) Szepesi, G.; Valkó, K. *J. Chromatogr.* **1991**, *550*, 87.
(135) Mahalanobis, P. C. *Proc. Natl. Inst. Sci. Ind.* **1936**, *2*, 49.
(136) Hotelling, H. *Ann. Math. Statist.* **1931**, *2*, 360.
(137) Crown, H., Ed. *Comprehensive Drug Design*; Pergamon Press: New York, 1990; p 19.
(138) Katritzky, A. R.; Slavov, S. H.; Dobchev, D. A.; Karelson, M. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 371.
(139) Edgington, E. S. *Randomization Tests*; Marcel Dekker, Inc.: New York, 1980; pp 195−216.
(140) Lingren, F.; Hansen, B.; Karcher, W.; Sjostrom, M.; Eriksson, L. *J. Chemom.* **1996**, *10*, 521.
(141) Wehrens, R.; Putter, H.; Buydens, L. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 35.
(142) Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chem. Inf. Model.* **2009**, *49*, 1669.
(143) Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423.
(144) Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109.
(145) Fisher, C. H. *Chem. Eng.* **1989**, *96*, 157.
(146) Satyanarayana, K.; Kakati, M. C. *Fire Mater.* **1991**, *15*, 97.
(147) Rechsteiner, C. E. In *Handbook of Chemical Property Estimation Methods*; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; Chapter 12.
(148) Walker, J. *J. Chem. Soc.* **1894**, *65*, 193.
(149) Meissner, H. P. *Chem. Eng. Progr.* **1949**, *45*, 149.
(150) Horvath, A. L. *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds*; Elseiver: Amsterdam, 1992; Chapter 2.
(151) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987.
(152) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546.
(153) Copeman, T. W.; Mathias, P. M.; Klotz, H. C. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: New York, 1988; pp 351.
(154) Joback, K. G.; Reid, R. C. *Chem. Eng. Commun.* **1987**, *57*, 233.
(155) Stein, S. E.; Brown, R. L. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 581.
(156) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem.* **1996**, 10400.
(157) Katritzky, A. R.; Lobanov, V.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28.
(158) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(159) Needham, D. E.; Wei, I.-C.; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, *110*, 4186.

(160) Randić, M.; Hansen, P.; Jurs, P. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 60.

(161) Stanton, D. T.; Jurs, P. C.; Hicks, M. G. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.

(162) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517.

(163) Mihalić, Z.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28.

(164) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233.

(165) Balaban, A. T.; Kier, L. B.; Joshi, N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237.

(166) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306.

(167) Egolf, L. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616.

(168) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118.

(169) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947.

(170) Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 68.

(171) Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841.

(172) Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039.

(173) Hall, L. H.; Story, C. T. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004.

(174) Kuanar, M.; Mishra, R. K.; Mishra, B. K. *Indian J. Chem.* **1996**, *35A*, 1026.

(175) Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. *Int. J. Quantum Chem.* **1997**, *63*, 215.

(176) Trinajstić, N.; Nikolić, S.; Lučić, B.; Amić, D.; Mihalić, Z. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 631.

(177) Kuanar, M.; Mishra, B. K. *J. Serb. Chem. Soc.* **1997**, *62*, 289.

(178) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. *Tetrahedron* **1998**, *54*, 9129.

(179) Plavšić, D.; Trinajstić, N.; Amić, D.; Šoškić, M. *New J. Chem.* **1998**, *22*, 1075.

(180) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395.

(181) Bunz, A.; Braun, B.; Janowsky, R. *Ind. Eng. Chem. Res.* **1998**, *37*, 3043.

(182) Randić, M.; Basak, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261.

(183) Balaban, A. T.; Basak, S.; Mills, D. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 769.

(184) Balaban, A. T.; Mills, D.; Basak, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 758.

(185) Goll, E. S.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974.

(186) Randić, M.; Basak, S. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899.

(187) Randić, M. *New J. Chem.* **2000**, *24*, 165.

(188) Lučić, B.; Lukovits, I.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 527.

(189) Randić, M.; Basak, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 650.

(190) Randić, M.; Plavšić, D.; Lerš, N. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 657.

(191) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. *Internet Electron. J. Mol. Des.* **2002**, *1*, 252.

(192) Espinosa, G.; Yaffe, D.; Arenas, A.; Cohen, Y.; Giralt, F. *Ind. Eng. Chem. Res.* **2001**, *40*, 2757.

(193) Zhou, C.; Nie, C.; Li, S.; Li, Z. *J. Comput. Chem.* **2007**, *28*, 2413.

(194) Zhou, C.; Nie, C. *Chromatographia* **2007**, *66*, 545.

(195) Zhou, C.; Chu, X.; Nie, C. *J. Phys. Chem. B* **2007**, *111*, 10174.

(196) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker: New York, 1992.

(197) Meylan, W. M.; Howard, P. H.; Boethling, R. S. *Environ. Toxicol. Chem.* **1996**, *15*, 100.

(198) Ran, Y.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354.

(199) Nikmo, J.; Kukkonen, J.; Riikonen, K. *J. Hazard. Mater.* **2002**, *91*, 43.

(200) Benoit-Guyod, J. L.; Andre, C.; Taillandier, G.; Rochat, J.; Boucherle, A. *Ecotoxicol. Environ. Saf.* **1984**, *8*, 227.

(201) Devillers, J.; Chambon, P. *Bull. Environ. Contam. Toxicol.* **1986**, *37*, 599.

(202) Barratt, M. D. *Toxicol. Lett.* **1995**, *75*, 169.

(203) Dearden, J. C. *Sci. Total Environ.* **1991**, *109*, 59.

(204) Dearden, J. C. In *Advances in Quantitative Structure-Property Relationships*; Charton, M., Charton, B. I., Eds.; JAI Press Inc.: Stamford, 1999; Vol. 2; pp 127−175.

(205) Dearden, J. C. *Environ. Toxicol. Chem.* **2003**, *22*, 1696.

(206) Kitaigorodsky, A. I. In *Molecular Crystals and Molecules*; Loebel, E. M., Ed.; Academic Press: New York, 1973.

(207) Mackay, D.; Shiu, W. T.; Bobra, A.; Billington, J.; Chan, E.; Yeun, A.; Ng, C.; Szeto, F. *Volatilization of Organic Pollutants from Water. U. S. Environmental Agency Report PB 82-230939*; U. S. Environmental Agency: Athens, GA, 1982.

(208) Simamora, P.; Miller, A. H.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 437.

(209) Constantinou, L.; Gani, R. *AIChE J.* **1994**, *40*, 1697.

(210) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530.

(211) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. *Cryst. Growth Des.* **2001**, *1*, 261.

(212) Hanson, M. P.; Rouvary, D. H. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvary, D. H., Eds.; Elsevier Science: Amsterdam, 1987; Vol. 51, pp 201−208.

(213) Abramowitz, R.; Yalkowsky, S. H. *Pharm. Res.* **1990**, *7*, 942.

(214) Charton, M.; Charton, B. *J. Phys. Org. Chem.* **1994**, *7*, 196.

(215) Murugan, R.; Grendze, M. P.; Toomey, J. E., Jr.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. *CHEMTECH* **1994**, *24*, 17.

(216) Katritzky, A. R.; Lobanov, V. S.; Karelson, M.; Murugan, R.; Grendze, M. P.; Toomey, J. E. *Rev. Roum. Chim.* **1996**, *41*, 851.

(217) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913.

(218) Randić, M. *J. Mol. Graphics Modelling* **2001**, *20*, 19.

(219) Estrada, E. *J. Phys. Chem. A* **2002**, *106*, 9085.

(220) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 64.

(221) Randić, M. Graph Theoretical Approach to Structure-Activity Studies: Search for Optimal Antitumor Compounds. In *Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens and Oncogens*; Rein, R., Ed.; Alan R. Liss, Inc. Publ: 1985; pp 309318.

(222) Johnson-Restrepo, B.; Pacheco-Londono, L.; Olivero-Verbel, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1513.

(223) Burch, K. J.; Whitehead, E. G. *J. Chem. Eng. Data* **2004**, *49*, 858.

(224) Gramatica, P.; Navas, N.; Todeschini, R. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 53.

(225) Welton, T. *Chem. Rev.* **1999**, *99*, 2071.

(226) Wasserscheid, P.; Keim, W. *Angew. Chem.* **2000**, *39*, 3772.

(227) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71.

(228) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 225.

(229) Carrera, G.; Aires-de-Sousa, J. *Green Chem.* **2005**, *7*, 20.

(230) Bergstroem, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177.

(231) Modarresi, H.; Dearden, J. C.; Modarress, H. *J. Chem. Inf. Model.* **2006**, *46*, 930.

(232) (a) Gao, J.; Wang, X.; Yu, X.; Li, X.; Wang, H. *J. Mol. Model.* **2006**, *12*, 521. (b) Holder, A. J.; Yourtee, D. M.; White, D. A.; Glaros, A. G.; Smith, R. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 223.

(233) Boys, S. F. *Proc. R. Soc. (London) A* **1950**, *200*, 542.

(234) *Gaussian 03*; Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; Gaussian, Inc., Wallingford, CT, 2004.

(235) Monnery, W. D.; Svrcek, W. Y.; Mehrotra, A. K. *Can. J. Chem. Eng.* **1995**, *73*, 3.

(236) Suzuki, T.; Ohtaguchi, K.; Koide, K. *Comput. Chem. Eng.* **1996**, *20*, 161.

(237) Suzuki, T.; Ebert, R.; Schüürmann, G. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1122.

(238) Ivanciuc, O.; Ivanciuc, T.; Filip, P. A.; Cabrol-Bass, D. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 515.

(239) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lučić, B.; Trinajstić, N.; Suzuki, T.; Schüürmann, G. *J. Phys. Org. Chem.* **2000**, *13*, 80.

(240) Lučić, B.; Bašić, I.; Nadramija, D.; Miličević, A.; Trinajstić, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. *ARKIVOC* **2002**, *4*, 45.

(241) Katritzky, A. R.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 840.

(242) Katritzky, A. R.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171.

(243) Xu, J.; Chen, B.; Zhang, Q.; Guo, B. *Polymer* **2004**, *45*, 8651.

(244) Cocchi, M.; De Benedetti, P. G.; Seeber, R.; Tassi, L.; Ulrici, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1190.

(245) Cao, C.; Jiang, L.; Yuan, H. *Internet Electron. J. Mol. Des.* **2003**, *2*, 621.

(246) Koziol, J. *Internet Electron. J. Mol. Des.* **2003**, *2*, 315.

(247) Ha, Z.; Ring, Z.; Liu, S. *Energy Fuels* **2005**, *19*, 152.

(248) (a) Xu, J.; Liang, H.; Chen, B.; Xu, W.; Shen, X.; Liu, H. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 152. (b) Holder, A. J.; Ye, L.; Eick, J. D.; Chappelow, C. C. *QSAR Comb. Sci.* **2006**, *25*, 342. (c) Holder, A. J.; Ye, L.; Eick, J. D.; Chappelow, C. C. *QSAR Comb. Sci.* **2006**, *25*, 905.

(249) Liu, Z.-Y.; Chen, Z.-C. *Chem. Eng. J. Biochem. Eng. J.* **1995**, *59*, 127.

(250) Gakh, A. A.; Gakh, E. G.; Sumpter, B. G.; Noid, D. W. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832.

(251) Zhang, R.; Liu, S.; Liu, M.; Hu, Z. *Comput. Chem.* **1997**, *21*, 335.

(252) Karelson, M.; Perkson, A. *Comput. Chem.* **1999**, *23*, 49.

(253) Toropov, A. A.; Toropova, A. P. *THEOCHEM* **2003**, *637*, 1.

(254) Morgan, H. L. *J. Chem. Soc.* **1965**, *5*, 197.

(255) Razinger, M. *Theor. Chim. Acta* **1982**, *61*, 581.

(256) Rucker, C.; Rucker, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 534.

(257) Toropov, A. A.; Toropova, A. P.; Nesterova, A. I.; Nabiev, O. M. *Chem. Phys. Lett.* **2004**, *384*, 357.

(258) Randić, M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105.

(259) Cao, C.; Gao, S. *Internet Electron. J. Mol. Des.* **2005**, *4*, 671.

(260) Hasted, J. *Aqueous Dielectrics*; Chapman and Hall: London, 1973.

(261) Tomasi, J.; Mennucci, B.; Capelli, C. In *Handbook of Solvents*; Wypich, G., Ed.; ChemTec Publishing: Toronto, 2001; Chapter 8, pp 419−504.

(262) Schweitzer, R. C.; Morris, J. B. *Anal. Chim. Acta* **1999**, *384*, 285.

(263) Schweitzer, R. C.; Morris, J. B. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1253.

(264) Pauling, L.; Pressman, D. *The Nature of the Chemical Bond*; Cornell University Press: Berlin, 1960; p 607.

(265) Batsanov, S. S. *Strukturnaya Refraktometriya [Structural Refractometry]*; Vysshaya Shkola: Moscow, 1976; p 302.

(266) Batsanov, S. S. *Eksperimentalnye Osnovy Strukurnoi Khimii [Experimental Foundations of Structural Chemistry]*; Izd-vo standartov: Moscow, 1986; Chapters 3.2 and 3.3.

(267) Biobyte Corporation, 201 West 4th St., Suite 204, Claremont, CA 91711.

(268) Vogel, A. I. *J. Chem. Soc.* **1948**, 1833.

(269) Vogel, A. I.; Cresswell, W. T.; Jeffery, G. J.; Leicester, J. *Chem. Ind.* **1950**, *358*.

(270) Jeffery, G. H.; Parker, R.; Vogel, A. I. *J. Chem. Soc.* **1961**, 570.

(271) Randić, M.; Pompe, M. *SAR QSAR Environ. Res.* **1999**, *10*, 451.

(272) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533.

(273) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8543.

(274) Stout, J. M.; Dykstra, C. E. *J. Am. Chem. Soc.* **1995**, *117*, 5127.

(275) Kagawa, H.; Ichimura, A.; Kamka, N. A.; Mori, K. *THEOCHEM* **2001**, *546*, 127.

(276) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952.

(277) Jensen, L.; Astrand, P.-O.; Sylvester-Hvid, K. O.; Mikkelsen, K. V. *J. Phys. Chem. A* **2000**, *104*, 1563.

(278) Nagle, J. K. *J. Am. Chem. Soc.* **1990**, *112*, 4741.

(279) Hati, S.; Datta, D. *J. Phys. Chem.* **1995**, *99*, 10742.

(280) Pal, S.; Chandra, A. K. *J. Phys. Chem.* **1995**, *99*, 13865.

(281) Fricke, B. B. *J. Chem. Phys.* **1986**, *84*, 862.

(282) Cao, C.; Yuan, H. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 667.

(283) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512.

(284) Ghanty, T. K.; Ghosh, S. K. *J. Phys. Chem.* **1993**, *97*, 4951.

(285) Hati, S.; Datta, D. *J. Phys. Chem.* **1994**, *98*, 10451.

(286) Meyers, F.; Marder, S. R.; Pierce, B. M.; Bredas, J. L. *J. Am. Chem. Soc.* **1994**, *116*, 10703.

(287) Staikova, M.; Wania, F.; Donaldson, D. J. *Atmos. Environ.* **2004**, *38*, 213.

(288) Bradley, M.; Waller, C. L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1301.

(289) Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1154.

(290) Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. *Russ. Chem. Bull.* **2003**, *52*, 1061.

(291) Pogliani, L. *New J. Chem.* **2003**, *27*, 919.

(292) Hansch, C.; Steinmetz, W. E.; Leo, A. J.; Mekapati, S. B.; Kurup, A.; Hoekman, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 120.

(293) Agin, D.; Hersh, L.; Holtzman, D. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 952.

(294) Hansch, C.; Kurup, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1647.

(295) Verma, R. P.; Kurup, A.; Hansch, C. *Bioorg. Med. Chem.* **2005**, *13*, 237.

(296) Verma, R. P.; Hansch, C. *Bioorg. Med. Chem.* **2005**, *13*, 2355.

(297) Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M. *J. Mol. Model.* **2007**, *13*, 951.

(298) Martin, D.; Slid, S.; Maran, U.; Karelson, M. *J. Phys. Chem. C* **2008**, *112*, 4785.

(299) Shiu, W. Y.; Doucette, W.; Gobas, F. A. P. C.; Andren, A.; Mackay, D. *Environ. Sci. Technol.* **1988**, *22*, 651.

(300) Reinhard, M.; Drefahl, A. *Handbook for Estimating Physicochemical Properties of Organic Compounds*; Wiley & Sons: New York, 1999.

(301) Mackay, D.; Bobra, A.; Chan, D. W.; Shiu, W.; Ying, *Environ. Sci. Technol.* **1982**, *16*, 645.

(302) Banerjee, S.; Howard, P. H.; Lande, S. S. *Chemosphere* **1990**, *21*, 1173.

(303) Staikova, M.; Messih, P.; Lei, Y. D.; Wania, F.; Donaldson, D. J. *J. Chem. Eng. Data* **2005**, *50*, 438.

(304) Yuan, W.; Hansen, A. C.; Zhang, Q. *Fuel* **2005**, *84*, 943.

(305) Asher, W. E.; Pankow, J. F. *Atmos. Environ.* **2006**, *40*, 3588.

(306) Acree, W. E., Jr.; Chickos, J. S. *Fluid Phase Equilib.* **2006**, *243*, 198.

(307) Kühne, R.; Ebert, R.-U.; Schüürmann, G. *Chemosphere* **1997**, *34*, 671.

(308) Godavarthy, S. S.; Robinson, R. L., Jr.; Gasem, K. A. M. *Fluid Phase Equilib.* **2006**, *246*, 39.

(309) Basak, S. C.; Gute, B. D.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651.

(310) Basak, S. C.; Mills, D. *ARKIVOC* **2005**, *10*, 308.

(311) Liang, C.; Gallagher, D. A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321.

(312) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720.

(313) Goll, E. S.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1081.

(314) McClelland, H. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967.

(315) Cash, G. G. *Chemosphere* **1999**, *39*, 2583.

(316) Chalk, A. J.; Beck, B.; Clark, T. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1053.

(317) Yaffe, D.; Cohen, Y. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 463.

(318) Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 692.

(319) Katritzky, A. R.; Slavov, S. H.; Dobchev, D. A.; Karelson, M. *Comput. Chem. Eng.* **2007**, *31*, 1123.

(320) Girifalco, L. A.; Good, R. J. *J. Phys. Chem.* **1957**, *61*, 904.

(321) Masterton, W. L.; Slowinski, E. J. *Chemical Principles*, 4th ed.; W. B. Saunders: Philadelphia, PA, 1973; p 207.

(322) Le Grand, D. G.; Gaines, G. L., Jr. *J. Colloid Interface Sci.* **1975**, *51*, 338.

(323) Maguna, F. P.; Ninez, M. B.; Okulik, N. B.; Castro, E. A. *Russ. J. Gen. Chem.* **2003**, *73*, 1792.

(324) Thakur, A. *ARKIVOC* **2005**, *XIV*, 49.

(325) Gutman, I.; Popović, L.; Mishra, B. K.; Kuanar, M.; Estrada, E.; Guevara, N. *J. Serb. Chem. Soc.* **1997**, *62*, 1025.

(326) Liu, S.; Cai, S.; Cao, C.; Li, Z. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1337.

(327) Wiener, H. *J. Phys. Colloid Chem.* **1948**, *52*, 1082.

(328) Kavun, S. M.; Chalykh, A. E.; Palyulin, V. A. *Colloid J. (Translation of Kolloidnyi Zhurnal)* **1995**, *57*, 767.

(329) Knotts, T. A.; Wilding, V.; Oscarson, J. L.; Rowley, R. L. *J. Chem. Eng. Data* **2001**, *46*, 1007.

(330) Kauffman, G. W.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 408.

(331) Wang, Z. W.; Li, G. Z.; Mu, J. H.; Zhang, X. Y.; Lou, A. J. *Chin. Chem. Lett.* **2002**, *13*, 363.

(332) Grigoras, S. *J. Comput. Chem.* **1990**, *11*, 493.

(333) Turner, B. E.; Costello, C. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 639.

(334) Katritzky, A. R.; Mu, L.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 293.

(335) Bonchev, D. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; CRC Press: 1999; pp 361−401.

(336) Bonchev, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 582.

(337) Gutman, I.; Tomović, Z.; Mishra, B. K.; Kuanar, M. *Indian J. Chem.* **2001**, *40A*, 4.

(338) Lučić, B.; Miličević, A.; Nikolić, S.; Trinajstić, N. *Croat. Chem. Acta* **2002**, *75*, 847.

(339) Yao, X.; Wang, Y.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217.

(340) Duchowicz, P.; Castro, E. A. *Russ. J. Gen. Chem.* **2002**, *72*, 1867.

(341) Krenkel, G.; Castro, E. A. *J. Theor. Comput. Chem.* **2003**, *2*, 33.

(342) Ponce, Y. M. *Molecules* **2003**, *8*, 687.

(343) Shacham, M.; Brauner, N.; Cholakov, G. S.; Stateva, R. P. *AIChE J.* **2004**, *50*, 2481.

(344) Shamsipur, M.; Ghavami, R.; Hemmateenejad, B.; Sharghi, H. *QSAR Comb. Sci.* **2004**, *23*, 734.

(345) Ni, C. H.; Zeng, X. Y.; Huang, H. *Chin. Chem. Lett.* **2005**, *16*, 709.

(346) Charton, M. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 197.
(347) Godavarthy, S. S.; Robinson, R. L., Jr.; Gasem, K. A. M. *Fluid Phase Equilib.* **2008**, *264*, 122.
(348) Bae, H.; Lee, S.; Teja, A. *Fluid Phase Equilib.* **1991**, *66*, 225.
(349) Sola, D.; Ferri, A.; Banchero, M.; Manna, L.; Sicardi, S. *Fluid Phase Equilib.* **2008**, *263*, 33.
(350) Yuan, H.; Cao, C. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 501.
(351) Koutek, B.; Hoskovec, M.; Streinz, L.; Vrkocova, P.; Ruzicka, K. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1351.
(352) Emel'yanenko, V. N.; Roganov, G. N. *Zh. Fiz. Khim.* **2001**, *75*, 204.
(353) Makitra, R. G.; Polyuzhin, I. P. *Russ. J. Gen. Chem.* **2005**, *75*, 739.
(354) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. *QSAR Comb. Sci.* **2003**, *22*, 565.
(355) Puri, S.; Chickos, J. S.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 299.
(356) (a) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959. (b) Code, J. E.; Holder, A. J.; Eick, J. D. *QSAR Comb. Sci.* **2008**, *27*, 841.
(357) Garbalena, M.; Herndon, W. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 37.
(358) Kuanar, M.; Kuanar, S. K.; Mishra, B. K.; Gutman, I. *Indian J. Chem.* **1999**, *38*, 525.
(359) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 631.
(360) Golovanov, I. B.; Tsygankova, I. G. *Russ. J. Gen. Chem.* **2001**, *71*, 839.
(361) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. *Internet Electron. J. Mol. Des.* **2002**, *1*, 467.
(362) Ren, B. *Comput. Chem.* **2002**, *26*, 357.
(363) Ivanciuc, O. *Rev. Roum. Chim.* **2003**, *47*, 577.
(364) Dyekjaer, J. D.; Jonsdottir, S. O. *Ind. Eng. Chem. Res.* **2003**, *42*, 4241.
(365) Marino, D. J. G.; Peruzzo, P. J.; Krenkel, G.; Castro, E. A. *Chem. Phys. Lett.* **2003**, *369*, 325.
(366) Tsygankova, I. G. *Russ. J. Phys. Chem.* **2005**, *79*, S14.
(367) Stull, D. R. *The Chemical Thermodynamics of Organic Compounds*; J. Wiley: New York, 1969.
(368) Wagman, D. D.; Evans, W. H.; Parker, V. B.; Schumm, R. H.; Halow, I.; Bailey, S. M.; Churney, K. L.; Nuttall, R. L. *J. Phys. Chem. Ref. Data* **1982**, *11*, 1 Suppl. 2.
(369) Thinh, T. P.; Trong, T. K. *Can. J. Chem. Eng.* **1976**, *54*, 344.
(370) Lias, S. G.; Liebman, J. F.; Levin, R. D. *J. Phys. Chem. Ref. Data* **1984**, *13*, 695.
(371) Habibollahzadeh, D.; Grice, M. E.; Concha, M. C.; Murray, J. S.; Politzer, P. *J. Comput. Chem.* **1995**, *16*, 654.
(372) Ibrahim, M. B.; Schleyer, P. v. R. *J. Comput. Chem.* **1985**, *6*, 157.
(373) Yala, Z. *THEOCHEM* **1990**, *207*, 217.
(374) Castro, E. A. *THEOCHEM* **1994**, *304*, 93.
(375) Castro, E. A. *THEOCHEM* **1995**, *339*, 239.
(376) Dewar, M. J. S.; Storch, D. M. *J. Am. Chem. Soc.* **1985**, *107*, 3898.
(377) Castro, E. A. *J. Chem. Soc. Pakistan* **1995**, *17*, 156.
(378) Vericat, C.; Castro, E. A. *Commun. Math. Comp. Chem. MATCH* **1996**, *34*, 167.
(379) Vericat, C.; Castro, E. A. *Egypt. J. Chem.* **1998**, *41*, 109.
(380) Hu, L. H.; Wang, X. J.; Wong, L. H.; Chen, G. H. *J. Chem. Phys.* **2003**, *119*, 11501.
(381) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864.
(382) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133.
(383) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622.
(384) Duan, X.-M.; Li, Z.-H.; Song, G.-L.; Wang, W.-N.; Chen, G.-H.; Fan, K.-N. *Chem. Phys. Lett.* **2005**, *410*, 125.
(385) Thanikaivelan, P.; Subramanian, V.; Rao, J. R.; Nair, B. U. *Chem. Phys. Lett.* **2000**, *323*, 59.
(386) Mercader, A.; Castro, E. A.; Toropov, A. A. *Int. J. Mol. Sci.* **2001**, *2*, 121.
(387) Cash, G. G. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 815.
(388) Gallegos, A.; Girones, X. *J. Chem. Inf. Model.* **2005**, *45*, 321.
(389) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. *ACH-Model. Chem.* **2000**, *137*, 57.
(390) Sukhachev, D. V.; Pivina, T. S.; Volk, F. S. *Propellants, Explosives, Pyrotechnics* **1994**, *19*, 159.
(391) Roy, K.; Saha, A. *J. Mol. Model.* **2003**, *9*, 259.
(392) Golovanov, I. B.; Zhenodarova, S. M.; Smolyaninova, O. A. *Russ. J. Gen. Chem.* **2003**, *73*, 519.
(393) Vilkov, L. V.; Pentin, Y. A. *Physical Methods of Investigation in Chemistry, Structural Methods and Optical Spectroscopy*; Vysshaya Shkola: Moscow, 1987; p 367.
(394) Stull, D. R.; Westrum, E. F.; Sinke, G. C. *The Chemical Thermodynamics of Organic Compounds*; Wiley: New York, 1969.
(395) Coutsikos, P.; Voutsas, E.; Magoulas, K. M.; Tassios, D. P. *Fluid Phase Equilib.* **2003**, *207*, 263.
(396) Chickos, J. S.; Acree, W. E., Jr.; Liebman, J. F. *J. Phys. Chem. Ref. Data* **1999**, *28*, 1535.
(397) Pankratov, A. N. *Afinidad* **1999**, *482*, 257.
(398) Duchowicz, P.; Castro, E. A.; Toropov, A. A. *Comput. Chem.* **2002**, *26*, 327.
(399) Golovanov, I. B.; Zhenodarova, S. M. *Russ. J. Gen. Chem.* **2004**, *74*, 828.
(400) Dannenfelser, R.-M.; Yalkowsky, S. H. *J. Pharm. Sci.* **1999**, *88*, 722.
(401) Johnson, J. L. H.; Yalkowsky, S. H. *Ind. Eng. Chem. Res.* **2005**, *44*, 7559.
(402) Drakenberg, T.; Daffiqvist, K. J.; Forsen, S. *J. Phys. Chem.* **1972**, *76*, 2178.
(403) Feigel, M.; Strassner, T. *THEOCHEM* **1993**, *283*, 33.
(404) Wiberg, K. B.; Rablen, P. R.; Rush, D. J.; Keith, T. A. *J. Am. Chem. Soc.* **1995**, *117*, 4261.
(405) Del Valle, J. C.; De Paz, J. L. G. *THEOCHEM* **1991**, *86*, 481.
(406) Ross, B. D.; True, N. S. *J. Am. Chem. Soc.* **1984**, *106*, 1.
(407) Taha, A. N.; Neugebauer-Crawford, S. M.; True, N. S. *J. Phys. Chem. A* **1998**, *102*, 1425.
(408) Leis, L.; Klika, K. D.; Pihlaja, M.; Karelson, M. *Tetrahedron* **1999**, *55*, 5227.
(409) Wiberg, K. B.; Laidig, K. E. *J. Am. Chem. Soc.* **1987**, *109*, 5935.
(410) Wiberg, K. B.; Breneman, C. M. *J. Am. Chem. Soc.* **1992**, *114*, 831.
(411) Bader, R. F. W.; Cheeseman, J. R.; Laidig, K. E.; Wiberg, K. B.; Breneman, C. *J. Am. Chem. Soc.* **1990**, *112*, 6530.
(412) Leis, J.; Karelson, M. *Comput. Chem.* **2001**, *25*, 171.
(413) Briggs, G. *J. Agric. Food Chem.* **1981**, *29*, 1050.
(414) Pollack, G. *Science* **1991**, *251*, 1323.
(415) Yalkowsky, S. H.; Valvani, S. C. *J. Pharm. Sci.* **1980**, *69*, 912.
(416) Yalkowsky, S. H.; Valvani, S. C.; Roseman, T. J. *J. Pharm. Sci.* **1983**, *72*, 866.
(417) Jain, N.; Yalkowsky, S. H. *J. Pharm. Sci.* **2001**, *90*, 234.
(418) Ran, Y.; Jain, N.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208.
(419) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. *Chem. Pharm. Bull.* **1986**, *34*, 4663.
(420) Suzuki, T. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149.
(421) Klopman, G.; Zhu, H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439.
(422) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. *Chemosphere* **1995**, *30*, 2061.
(423) Ruelle, P.; Kesselring, U. W. *Chemosphere* **1997**, *34*, 275.
(424) Lombardo, F.; Gifford, E.; Shalaeva, M. Y. *Mini-Rev. Med. Chem.* **2003**, *3*, 861.
(425) Eros, D.; Keri, G.; Kovesdi, I.; Szantai-Kis, C.; Meszaros, G.; Orfi, L. *Mini-Rev. Med. Chem.* **2004**, *4*, 167.
(426) Nirmalakhandan, N. N.; Speece, R. E. *Environ. Sci. Technol.* **1988**, *22*, 328.
(427) Nirmalakhandan, N. N.; Speece, R. E. *Environ. Sci. Technol.* **1989**, *23*, 708.
(428) Sutter, J. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100.
(429) Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489.
(430) Huibers, P. D. T.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283.
(431) Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450.
(432) Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773.
(433) Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155.
(434) Yaffe, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177.
(435) Bruneau, P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605.
(436) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488.
(437) Liu, R.; So, S.-S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633.
(438) Delgado, E. J. *Fluid Phase Equilib.* **2002**, *199*, 101.
(439) Gao, H.; Shanmugasundaram, V.; Lee, P. *Pharm. Res.* **2002**, *19*, 497.
(440) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. *J. Comput. Chem.* **2002**, *23*, 275.
(441) Engkvist, O.; Wrede, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247.
(442) Cheng, A.; Merz, K. M. *J. Med. Chem.* **2003**, *46*, 3572.
(443) Schaper, K.-J.; Kunz, B.; Raevsky, O. A. *QSAR Comb. Sci.* **2003**, *22*, 943.
(444) Nohair, M.; Zakarya, D. *J. Mol. Mod.* **2003**, *9*, 365.
(445) Yan, A.; Gasteiger, J. *QSAR Comb. Sci.* **2003**, *22*, 821.
(446) Yan, A.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429.
(447) Duchowicz, P. R.; Castro, E. A.; Toropov, A. A.; Nesterov, I. V.; Nabiev, O. M. *Mol. Diversity* **2004**, *8*, 325.
(448) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. *Chem. Biodiversity* **2004**, *1*, 1829.
(449) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 170.
(450) Sacan, M. T.; Oezkul, M.; Erdem, S. S. *SAR QSAR Environ. Res.* **2005**, *16*, 443.

(451) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794.
(452) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806.
(453) Katritzky, A. R.; Tulp, I.; Fara, D.; Lauria, A.; Maran, U.; Acree, W. *J. Chem. Inf. Model.* **2006**, *45*, 913.
(454) Martin, D.; Maran, U.; Sild, S.; Karelson, M. *J. Phys. Chem. B* **2007**, *111*, 9853.
(455) Kühne, R.; Ebert, R.-U.; Schüürmann, G. *J. Chem. Inf. Model.* **2006**, *46*, 636.
(456) Yamashita, F.; Itoh, T.; Hara, H.; Hashida, M. *J. Chem. Inf. Model.* **2006**, *46*, 1054.
(457) Lu, G. N.; Dang, Z.; Tao, X. Q.; Yang, C.; Yi, X. Y. *QSAR Comb. Sci.* **2008**, *27*, 618.
(458) Zhou, D. S.; Alelyunas, Y.; Liu, R. F. *J. Chem. Inf. Model.* **2008**, *48*, 981.
(459) Toropov, A. A.; Rasulev, B. F.; Leszczynska, D.; Leszczynski, J. *Chem. Phys. Lett.* **2008**, *457*, 332.
(460) Hemmateenejad, B.; Shamsipur, M.; Miri, R.; Elyasi, M.; Foroghinia, F.; Sharghi, H. *Anal. Chim. Acta* **2008**, *610*, 25.
(461) Duchowicz, P. R.; Talevi, A.; Bruno-Blanch, L. E.; Castro, E. A. *Bioorg. Med. Chem.* **2008**, *16*, 7944.
(462) Kim, J.; Jung, D. H.; Rhee, H.; Choi, S. H.; Sung, M. J.; Choi, W. S. *Korean J. Chem. Eng.* **2008**, *25*, 865.
(463) Huuskonen, J.; Livingstone, D. J.; Manallack, D. T. *SAR QSAR Environ. Res.* **2008**, *19*, 191.
(464) Du-Cuny, L.; Huwyler, J.; Wiese, M.; Kansy, M. *Eur. J. Med. Chem.* **2008**, *43*, 501.
(465) Mackay, D.; Shiu, W. Y. *J. Phys. Chem. Ref. Data* **1981**, *10*, 1175.
(466) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777.
(467) Staudinger, J.; Roberts, P. V. *Crit. Rev. Environ. Sci. Technol.* **1996**, *26*, 205.
(468) Hine, H.; Mookerjee, P. K. *J. Org. Chem.* **1975**, *40*, 292.
(469) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563.
(470) Russell, C. J.; Dixon, S. L.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 1350.
(471) *ACD/Absolv*; Advanced Chemistry Development, Inc.: Toronto, ON, Canada, www.acdlabs.com.
(472) Katritzky, A. R.; Mu, L.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162.
(473) Katritzky, A. R.; Tatham, D. B.; Maran, U. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 358.
(474) Pierotti, G. J.; Deal, C. H.; Derr, E. L. *Ind. Eng. Chem.* **1959**, *51*, 95.
(475) Tsonopoulos, C.; Prausnitz, J. M. *Ind. Eng. Chem. Fundam.* **1971**, *10*, 593.
(476) Mackay, D.; Shiu, W. Y. *J. Chem. Eng. Data* **1977**, *22*, 399.
(477) Medir, M.; Giralt, F. *AIChE J.* **1982**, *28*, 341.
(478) Tochigi, K.; Tiegs, D.; Gmehling, J.; Kojima, K. *J. Chem. Eng. Jpn.* **1990**, *23*, 453.
(479) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. *Ind. Eng. Chem. Res.* **1991**, *30*, 2352.
(480) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. *AIChE J.* **1975**, *21*, 1086.
(481) Thomas, E. R.; Eckert, C. A. *Ind. Eng. Chem. Process. Des. Dev.* **1984**, *23*, 194.
(482) Sherman, S. R.; Trampe, D. B.; Bush, D. M.; Schiller, M.; Eckert, C. A.; Dallas, A. J.; Li, J.; Carr, P. W. *Ind. Eng. Chem. Res.* **1996**, *35*, 1044.
(483) Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 200.
(484) Rani, Y. K.; Dutt, N. V. K. *Chem. Eng. Commun.* **2002**, *189*, 372.
(485) He, J. T.; Zhong, C. L. *Fluid Phase Equilib.* **2003**, *205*, 303.
(486) Estrada, E.; Diaz, G. A.; Delgado, E. J. *J. Comput. Aided Mol. Des.* **2006**, *20*, 539.
(487) Giralt, F.; Espinosa, G.; Arenas, A.; Ferre-Gine, J.; Amat, L.; Girones, X.; Carbo-Dorca, R.; Cohen, Y. *AIChE J.* **2004**, *50*, 1315.
(488) Xu, H. Y.; Min, J. Q. *Chin. J. Struct. Chem.* **2008**, *27*, 491.
(489) Eike, D. M.; Brennecke, J. F.; Maginn, E. J. *Ind. Eng. Chem. Res.* **2004**, *43*, 1039.
(490) Tämm, K.; Burk, P. *J. Mol. Model.* **2006**, *12*, 417.
(491) Wang, J.; Sun, W.; Li, C.; Wang, Z. *Fluid Phase Equilib.* **2008**, *264*, 235.
(492) Katritzky, A. R.; Kuanar, M.; Stoyanova-Slavova, I. B.; Slavov, S. H.; Dobchev, D. A.; Karelson, M.; Acree, W. E. *J. Chem. Eng. Data* **2008**, *53*, 1085.
(493) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616.
(494) Rekker, R. *The Hydrophobic Fragment Constant*; Elsevier: New York, 1977.
(495) Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.* **1984**, *19*, 71.
(496) van de Waterbeemd, H.; Mannhold, R. *QSAR* **1996**, *15*, 410.
(497) Mannhold, R.; Dross, K. *Quant. Struct.-Act. Relat.* **1996**, *15*, 403.
(498) Mannhold, R.; Rekker, R. F. *Perspect. Drug Discovery Des.* **2000**, *18*, 1.
(499) Mannhold, R.; Van de Waterbeemd, H. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337.
(500) (a) Leo, A. In *Handbook of Property Estimation Methods for Chemicals*; Mackay, D., Boethling, R. S., Eds.; CRC Press: 2000; pp 89−114. (b) Holder, A. J.; Ye, L.; Yourtee, D. M.; Agarwal, A.; Eick, J. D.; Chappelow, C. C. *Dent. Mater.* **2005**, *21*, 591.
(501) Livingstone, D. J. *Curr. Top. Med. Chem.* **2003**, *3*, 1171.
(502) Klopman, G.; Zhu, H. *Mini-Rev. Med. Chem.* **2005**, *5*, 127.
(503) Leo, A. J. *Chem. Rev.* **1993**, *93*, 1281.
(504) Leo, A. J.; Hoekman, D. *Perspect. Drug Discovery Des.* **2000**, *18*, 19.
(505) Klopman, G.; Li, J.; Wang, S.; Dimayuga, M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752.
(506) Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868.
(507) Viswanadhan, V. N. *Perspect. Drug Discovery Des.* **2000**, *19*, 85.
(508) Mannhold, R.; Rekker, R. F.; Dross, K.; Bijloo, G.; De Vries, G. *QSAR* **1998**, *17*, 517.
(509) Meylan, W. M. *Perspect. Drug Discovery Des.* **2000**, *18*, 67.
(510) Petrauskas, A. A.; Kolovanov, E. A. *Perspect. Drug Discovery Des.* **2000**, *19*, 99.
(511) Wang, R.; Fu, Y.; Lai, L. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615.
(512) Wang, R.; Gao, Y.; Lai, L. *Perspect. Drug Discovery Des.* **2000**, *19*, 47.
(513) Rogers, K. S.; Cammarata, A. *Biochim. Biophys. Acta* **1969**, *193*, 22.
(514) Klopman, G.; Iroff, L. D. *J. Comput. Chem.* **1981**, *2*, 157.
(515) Bodor, N.; Huang, M. J. *J. Pharm. Sci.* **1992**, *81*, 272.
(516) Bodor, N.; Buchwald, P. *J. Phys. Chem. B* **1997**, *101*, 3404.
(517) Buchwald, P. *Perspect. Drug Discovery Des.* **2000**, *19*, 19.
(518) Gombar, V. K.; Enslein, K. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127.
(519) Gombar, V. K. *SAR QSAR Environ. Res.* **1999**, *10*, 371.
(520) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. *Chem. Pharm. Bull.* **1992**, *40*, 127.
(521) Devillers, J.; Domine, D.; Karcher, W. *Polycyclic Aromat. Compd.* **1996**, *11*, 211.
(522) Devillers, J.; Domine, D.; Guillon, C. *Eur. J. Med. Chem.* **1998**, *33*, 659.
(523) Raevsky, O. A. *SAR QSAR Environ. Res.* **2001**, *12*, 367.
(524) Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaja, O. E. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540.
(525) Tetko, I. V.; Tanchuk, V. Yu.; Villa, A. E. P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407.
(526) Tetko, I. V.; Tanchuk, V. Yu. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136.
(527) Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83.
(528) Kellogg, G. E.; Abraham, D. J. *Analysis* **1999**, *27*, 19.
(529) Kellogg, G. E.; Abraham, D. J. *Eur. J. Med. Chem.* **2000**, *35*, 651.
(530) Zhu, H.; Sedykh, A.; Chakravarti, S. K.; Klopman, G. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 3.
(531) Sedykh, A. Y.; Klopman, G. *J. Chem. Inf. Model.* **2006**, *46*, 1598.
(532) Eros, D.; Kovesdi, I.; Orfi, L.; Takacs-Novak, K.; Acsady, G.; Keril, G. *Curr. Med. Chem.* **2002**, *9*, 1819.
(533) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565.
(534) Molnar, L.; Keseru, G. M.; Papp, A.; Gulyas, Z.; Darvas, F. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 851.
(535) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. *J. Med. Chem.* **2000**, *43*, 2922.
(536) Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1998**, *63*, 4305.
(537) No, K. T.; Kim, S. G.; Cho, K.-H.; Scheraga, H. A. *Biophys. Chem.* **1999**, *78*, 127.
(538) In, Y.; Chai, H. H.; No, K. T. *J. Chem. Inf. Model.* **2005**, *45*, 254.
(539) Machatha, S. G.; Yalkowsky, S. H. *Int. J. Pharm.* **2005**, *294*, 185.
(540) Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *Bioorg. Med. Chem.* **2006**, *14*, 1021.
(541) Basak, S. C.; Mills, D. *ARKIVOC* **2005**, *2*, 60.
(542) Shamsipur, M.; Ghavami, R.; Hemmateenejad, B.; Sharghi, H. *Internet Electron. J. Mol. Des.* **2005**, *4*, 882.
(543) Al-Fahemi, J. H.; Cooper, D. L.; Allan, N. L. *THEOCHEM* **2005**, *727*, 57.
(544) Lamarche, O.; Platts, J. A.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 848.
(545) Oliferenko, A. A.; Oliferenko, P. V.; Huddleston, J. G.; Rogers, R. D.; Palyulin, V. A.; Zefirov, N. S.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1042.
(546) Gao, S.; Cao, C. *Int. J. Mol. Sci.* **2008**, *9*, 962.

(547) Zou, J.-W.; Zhao, W.-N.; Shang, Z.-C.; Huang, M.-L.; Guo, M.; Yu, Q.-S. *J. Phys. Chem. A* **2002**, *106*, 11550.
(548) Wegner, J. K.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077.
(549) Peruzzo, P. J.; Marino, D. J. G.; Castro, E. A.; Toropov, A. A. *Internet Electron. J. Mol. Des.* **2003**, *2*, 334.
(550) Leahy, D. E.; Taylor, P. J.; Wait, A. R. *Quant. Struct.-Act. Relat.* **1989**, *8*, 17.
(551) Leahy, D. E.; Morris, J. J.; Taylor, P. J.; Wait, A. R. *J. Chem. Soc., Perkin Trans. 2* **1992**, 723.
(552) Khadikar, P. V.; Mandloi, D.; Bajaj, A. V.; Joshi, S. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 419.
(553) Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 934.
(554) Albertsons, P. A. *Nature* **1958**, *182*, 709.
(555) Albertsons, P. A. *Partition of Cell Particles and Macromolecules*; Wiley & Sons: New York, 1986.
(556) Rogers, R. D.; Eiteman, M. A. *Aqueous Biphasic Separations: Biomolecules to Metal Ions*; Plenum Press: New York, 1995.
(557) Zaslavsky, B. *Aqueous Two Phase Partitioning: Physical Chemistry and Bioanalytical Applications*; Marcel Dekker: New York, 1995.
(558) Gulyaeva, N.; Zaslavsky, A.; Lechner, P.; Chait, A.; Zaslavsky, B. *J. Chromatogr., B* **2000**, *743*, 187.
(559) Gulyaeva, N.; Zaslavsky, A.; Lechner, P.; Chait, A.; Zaslavsky, B. *J. Pharm. Sci.* **2001**, *90*, 1366.
(560) Gulyaeva, N.; Zaslavsky, A.; Chait, A.; Zaslavsky, B. *J. Pept. Res.* **2002**, *59*, 277.
(561) Eiteman, M. A.; Gainer, J. L. *Biotechnol. Prog.* **1990**, *6*, 479.
(562) Eiteman, M. A.; Gainer, J. L. *Bioseparation* **1991**, *2*, 31.
(563) Gulyaeva, N.; Zaslavsky, A.; Chait, A.; Zaslavsky, B. *J. Pept. Res.* **2003**, *61*, 129.
(564) Rogers, R. D.; Willauer, H. D.; Griffin, S. T.; Huddleston, J. G. *J. Chromatogr., B* **1998**, *711*, 255.
(565) Willauer, H. D.; Huddleston, J. G.; Griffin, S. T.; Rogers, R. D. *Sep. Sci. Technol.* **1999**, *34*, 1069.
(566) Huddleston, J. G.; Ingenito, C. C.; Rogers, R. D. *Sep. Sci. Technol.* **1999**, *34*, 1091.
(567) Willauer, H. D.; Huddleston, J. G.; Rogers, R. D. *Ind. Eng. Chem. Res.* **2002**, *41*, 1892.
(568) Willauer, H. D.; Huddleston, J. G.; Rogers, R. D. *Ind. Eng. Chem. Res.* **2002**, *41*, 2591.
(569) Katritzky, A. R.; Tämm, K.; Kuanar, M.; Fara, D. C.; Oliferenko, A.; Oliferenko, P.; Huddleston, J. G.; Rogers, R. D. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 136.
(570) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernäs, H.; Karlén, A. *J. Med. Chem.* **1998**, *41*, 4939.
(571) Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. *Drug Metab. Dispos.* **2004**, *32*, 132.
(572) Colmenarejo, G. *Med. Res. Rev.* **2003**, *23*, 275.
(573) Ishigami, M.; Honda, T.; Takasaki, W.; Ikeda, T.; Komai, T.; Ito, K.; Sugiyama, Y. *Drug Metab. Dispos.* **2001**, *29*, 282.
(574) Poulin, P.; Theil, F. P. *J. Pharm. Sci.* **2002**, *91*, 129.
(575) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. *J. Med. Chem.* **2004**, *47*, 1242.
(576) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(577) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F. *J. Med. Chem.* **2001**, *44*, 2490.
(578) Xing, L.; Glen, R. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796.
(579) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. *J. Pharm. Sci.* **1997**, *86*, 865.
(580) Tetko, I. V.; Bruneau, P. *J. Pharm. Sci.* **2004**, *93*, 3103.
(581) Tetko, I. V.; Poda, G. *J. Med. Chem.* **2004**, *47*, 5601.
(582) Meyer, K. H.; Hopff, H. *Z. Physiol. Chem.* **1923**, *126*, 281.
(583) Meyer, K. H.; Hemmi, H. *Biochem. Z.* **1935**, *277*, 39.
(584) Meulenberg, C. J. W.; Vijverberg, H. P. M. *Toxicol. Appl. Pharmacol.* **2000**, *165*, 206.
(585) Meulenberg, C. J. W.; Wijnker, A. G.; Vijverberg, H. P. M. J. *Toxicol. Environ. Health, Part A* **2003**, *66*, 1985.
(586) Abraham, M. H.; Weathersby, P. K. *J. Pharm. Sci.* **1994**, *83*, 1450.
(587) Abraham, M. H.; Fuchs, R. *J. Chem. Soc., Perkin Trans. 2* **1988**, 523.
(588) Klopman, G.; Ding, C.; Macina, O. T. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 569.
(589) Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E. *Bioorg. Med. Chem.* **2004**, *12*, 4735.
(590) Abraham, M. H.; Ibrahim, A. *J. Chem. Inf. Model.* **2006**, *46*, 1735.
(591) Grob, R. L. *Modern Practice of Gas Chromatography*; John Wiley & Sons: New York, 1985.
(592) Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships*; John Wiley & Sons: New York, 1987.
(593) Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chem.* **1985**, *57*, 2770.
(594) Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chem.* **1986**, *58*, 1210.
(595) Hasan, M.; Jurs, P. C. *Anal. Chem.* **1988**, *60*, 978.
(596) Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chem.* **1988**, *60*, 2249.
(597) Hasan, M. N.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2318.
(598) Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. *Anal. Chem.* **1991**, *63*, 2021.
(599) Georgakopoulos, C. G.; Tsika, O. G.; Kiburis, J. C.; Jurs, P. C. *Anal. Chem.* **1991**, *63*, 2025.
(600) Woloszyn, T. F.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 3059.
(601) Woloszyn, T. F.; Jurs, P. C. *Anal. Chem.* **1993**, *65*, 582.
(602) Duvenbeck, C.; Zinn, P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 211.
(603) Duvenbeck, C.; Zinn, P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 220.
(604) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. *Anal. Chem.* **1994**, *66*, 1799.
(605) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. *Anal. Chem.* **1997**, *342*, 113.
(606) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610.
(607) Katritzky, A. R.; Chen, K.; Maran, U.; Carlson, D. A. *Anal. Chem.* **2000**, *72*, 101.
(608) Ren, B. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 29.
(609) Junkes, B.; Amboni, R.; Yunes, R.; Heinzen, V. *Internet Electron. J. Mol. Des* **2003**, *2* (33), S1–S12.
(610) Alves de Lima Ribeiro, F.; Ferreira, M. *THEOCHEM* **2003**, *663*, 109.
(611) Hodjmohammadi, M. R.; Ebrahimi, P.; Pourmorad, F. *QSAR Comb. Sci.* **2004**, *23*, 295.
(612) Hu, Q.; Liang, Y.; Peng, X.; Yin, H.; Fang, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 437.
(613) Garkani-Nejad, Z.; Karlovits, M.; Demuth, W.; Stimpfl, T.; Vycudilik, W.; Jalali-Heravi, M.; Varmuza, K. *J. Chromatogr., A* **2004**, *1028*, 287.
(614) Hu, R.; Liu, H.; Zhang, R.; Xue, C.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. *Talanta* **2005**, *68*, 31.
(615) Zarei, K.; Atabati, M. *J. Anal. Chem.* **2005**, *60*, 732.
(616) Liu, F.; Liang, Y.; Cao, C. *Internet Electron. J. Mol. Des.* **2006**, *5*, 102.
(617) Nakajima, N.; Lay, C.; Du, H.; Ring, Z. *Energy Fuels* **2006**, *20*, 1111.
(618) Lu, C.; Guo, W.; Hu, X.; Wang, Y.; Yin, C. *Chem. Phys. Lett.* **2006**, *417*, 11.
(619) Stein, S. E.; Babushok, V. I.; Brown, R. L.; Linstrom, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 975.
(620) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.
(621) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1989**, *61*, 1328.
(622) Whalen-Pedersen, E. K.; Jurs, P. C. *Anal. Chem.* **1981**, *53*, 2184.
(623) Buydens, L.; Massart, D. L.; Geerlings, P. *Anal. Chem.* **1983**, *55*, 738.
(624) Donovan, W. H.; Famini, G. R. *J. Chem. Soc., Perkin Trans. 2.* **1996**, 83.
(625) Burchmann, A.; Zinn, P.; Haffer, C. M. *Anal. Chim. Acta* **1993**, *283*, 869.
(626) Pompe, M.; Razinger, M.; Novič, M.; Veber, M. *Anal. Chim. Acta* **1997**, *348*, 215.
(627) Pompe, M.; Novič, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59.
(628) Zarei, K.; Atabati, M. *J. Anal. Chem.* **2005**, *60*, 732.
(629) Schomburg, G. *Gas Chromatography—A Practical Course*; VCH: Weinheim, 1990.
(630) Sternberg, J. C.; Gallaway, W. S.; Jones, D. T. L. In *Gas Chromatography*; Berner, N., Callen, J. E., Weiss, M. D., Eds.; Academic Press: New York, 1962; Chapter 18.
(631) Ackman, R. G. *J. Gas Chromatogr.* **1964**, 173.
(632) Ackman, R. G.; Sipos, J. C. *J. Chromatogr.* **1964**, *16*, 298.
(633) Dietz, W. A. *J. Gas Chromatogr.* **1967**, 68.
(634) Scanlon, J. T.; Willis, D. E. *J. Chromatogr. Sci.* **1985**, *23*, 333.
(635) Musumarra, G.; Pisano, D.; Katritzky, A. R.; Lapucha, A. R.; Luxem, F. J.; Murugan, R.; Siskin, M.; Brons, G. *Tetrahedron Comput. Methodol.* **1989**, *2*, 17.
(636) Huang, Y.; Ou, Q.; Yu, W. *Anal. Chem.* **1990**, *62*, 2063.
(637) Morvai, M.; Palyka, I.; Molnar-Perl, I. *J. Chromatogr. Sci.* **1992**, *30*, 448.
(638) Jalali-Heravi, M.; Fatemi, M. H. *J. Gas Chromatogr., A* **1998**, *825*, 161.
(639) Jalali-Heravi, M.; Fatemi, M. H. *J. Chromatogr.* **2000**, *897*, 227.
(640) Saradhi, U. V. R. V.; Suryanarayana, M. V. S.; Gupta, A. K.; Semwal, R. P.; Jayaram, B. *J. Chromatogr., A* **2001**, *911*, 63.
(641) Jalali-Heravi, M.; Garkani-Nejad, Z. *J. Chromatogr., A* **2002**, *950*, 183.
(642) Jalali-Heravi, M.; Noroozian, E.; Mousavi, M. *J. Chromatogr.* **2004**, *1023*, 247.
(643) Prabhakar, Y. S. *Internet Electron. J. Mol. Des.* **2004**, *3*, 150.
(644) Katritzky, A. R.; Dobchev, D. A.; Karelson, M. *Z. Naturforsch.* **2006**, *61b*, 373.

(645) Sukhachev, D. V.; Pivina, T. S.; Zhokhova, N. I.; Zefirov, N. S.; Zeman, S. *Russ. Chem. Bull.* **1995**, *44*, 1585.

(646) Hiob, R.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1062.

(647) Hiob, R.; Karelson, M. *Comput. Chem.* **2002**, *26*, 237.

(648) Griffin, G. W.; Dzidic, I.; Carroll, D. I.; Stillwell, R. N.; Horning, E. C. *Anal. Chem.* **1973**, *45*, 1204.

(649) Wessel, M. D.; Jurs, P. C. *Anal. Chem.* **1994**, *66*, 2480.

(650) Wessel, M. D.; Sutter, J. M.; Jurs, P. C. *Anal. Chem.* **1996**, *68*, 4237.

(651) Agbonkonkon, N.; Tolley, H. D.; Asplund, M. C.; Lee, E. D.; Lee, M. L. *Anal. Chem.* **2004**, *76*, 5223.

(652) Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. *Talanta* **2007**, *71*, 258.

(653) Kuhnt, G.;, Muntau, H., Eds. *EUROSOILS—Identification, Collection, Treatment, Characterization*; European Commission Special Publication No. 1.94.60;, Ispra: 1994; p 154.

(654) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. *Chemosphere* **1997**, *34*, 2525.

(655) Winget, P.; Cramer, C. J.; Thrular, D. G. *Environ. Sci. Technol.* **2000**, *34*, 4733.

(656) Meylan, W. M.; Howard, P. H.; Boethling, R. S. *Environ. Sci. Technol.* **1992**, *26*, 1560.

(657) Liao, Y.-Y.; Wang, Z.-T.; Chen, J.-W.; Han, S.-K.; Wang, L.-S.; Lu, G.-Y.; Zhao, T.-N. *Bull. Environ. Contam. Toxicol.* **1996**, *56*, 711.

(658) Hong, H.; Wang, L.; Han, S.; Zhang, Z.; Zou, G. *Chemosphere* **1997**, *34*, 827.

(659) Baker, J. R.; Mihelcic, J. R.; Luehrs, D. C.; Hickey, J. P. *Water Environ. Res.* **1997**, *69*, 136.

(660) Baker, J. R.; Mihelcic, J. R.; Shea, E. *Chemosphere* **2000**, *41*, 813.

(661) Baker, J. R.; Mihelcic, J. R.; Sabljić, A. *Chemosphere* **2001**, *41*, 213.

(662) Müller, M. *Chemosphere* **1997**, *35*, 365.

(663) Szabo, G.; Guczi, J.; Kördel, W.; Zsolnay, A.; Major, V.; Keresztes, P. *Chemosphere* **1999**, *39*, 431.

(664) Hansen, B. G.; Paya-Perez, A. B.; Rahman, M.; Larsen, B. R. *Chemosphere* **1999**, *39*, 2209.

(665) Dai, J.; Sun, C.; Han, S.; Wang, L. *Bull. Environ. Contam. Toxicol.* **1999**, *62*, 530.

(666) Dai, J.; Xu, M.; Wang, L. *Bull. Environ. Contam. Toxicol.* **2000**, *65*, 190.

(667) Gramatica, P.; Corradi, M.; Consonni, V. *Chemosphere* **2000**, *41*, 763.

(668) Tao, S.; Lu, X. X.; Cao, J.; Dawson, R. A. *Water Environ. Res.* **2001**, *73*, 307.

(669) Klamt, A.; Eckert, F.; Diedenhofen, M. *Environ. Toxicol. Chem.* **2002**, *21*, 2562.

(670) Wu, C.-D.; Wei, D.-B.; Liu, X.-H.; Wang, L.-S. *Bull. Environ. Contam. Toxicol.* **2001**, *66*, 777.

(671) Wu, C. D.; Wei, D. B.; Hu, G. P.; Wang, L. S. *Bull. Environ. Contam. Toxicol.* **2003**, *70*, 513.

(672) Huuskonen, J. *Environ. Toxicol. Chem.* **2003**, *22*, 816.

(673) Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1457.

(674) Delgado, E. J.; Alderete, J. B.; Jana, G. A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1928.

(675) Wei, D. B.; Wu, C. D.; Wang, L. S.; Hu, H. Y. *SAR QSAR Environ. Res.* **2003**, *14*, 191.

(676) Liu, G. S.; Yu, J. G. *Water Res.* **2005**, *39*, 2048.

(677) Kahn, I.; Fara, D.; Karelson, M.; Maran, U.; Andersson, P. L. *J. Chem. Inf. Model.* **2005**, *45*, 94.

(678) Lu, C. H.; Wang, Y.; Yin, C. S.; Guo, W. M.; Hu, X. F. *Chemosphere* **2006**, *63*, 1384.

(679) Gonzalez, M. P.; Helguera, A. M.; Collado, I. G. *Mol. Diversity* **2006**, *10*, 109.

(680) Ivanciuc, T.; Ivanciuc, O.; Klein, D. J. *Int. J. Mol. Sci.* **2006**, *7*, 358.

(681) Gramatica, P.; Giani, E.; Papa, E. *J. Mol. Graph. Model.* **2007**, *25*, 755.

(682) Duchowicz, P. R.; Perez Gonzalez, M.; Morales Helguera, A.; Dias Soeiro Cordeiro, M. N.; Castro, E. A. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 197.

(683) Lu, G.; Yang, C.; Tao, X.; Yi, X.; Dang, Z. *J. Theor. Comput. Chem.* **2008**, *7*, 67.

(684) Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 684.

(685) Katritzky, A. R.; Fara, D. C.; Wang, H.; Tämm, K.; Karelson, M. *Chem. Rev.* **2004**, *104*, 175.

(686) Katritzky, A. R.; Fara, D. C.; Kuanar, M.; Hur, E.; Karelson, M. *J. Phys. Chem.* **2005**, *109*, 10323.

(687) Katritzky, A. R.; Mu, L.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 756.

(688) Hoffman, H.; Ulbright, W. In *Thermodynamic Data for Biochemistry and Biotechnology*; Hinz, H. J., Ed.; Springer-Verlag: Berlin, 1986; p 297.

(689) Rosen, M. J. *Surfactants and Interfacial Phenomena*; Wiley: New York, 1987.

(690) Elimelech, M.; Gregory, J.; Jis, X.; Williams, R. A. *Particle Deposition and Aggregation, Measurement, Modeling, and Simulation*; Butterworth: Oxford, 1995.

(691) Hiemenz, P. C. *Principles of Colloid and Surface Chemistry*; Marcel Dekker: New York, 1977.

(692) Klevens, H. B. *J. Am. Oil Chem. Soc.* **1953**, *30*, 74.

(693) Schick, M. J. *Nonionic Surfactants*; Marcel Dekker: New York, 1967.

(694) Rosen, M. J. *J. Colloid Interface Sci.* **1976**, *56*, 320.

(695) Becher, P. *J. Dispersion Sci. Technol.* **1984**, *5*, 81.

(696) Ravey, J. C.; Gherbi, A.; Stebe, M. *Prog. Colloid Polym. Sci.* **1988**, *76*, 234.

(697) Chen, C.-C. *AIChE J.* **1996**, *42*, 3231.

(698) Puvvada, S.; Blankschtein, D. *J. Chem. Phys.* **1990**, *92*, 3710.

(699) Couper, A. In *Surfactants*; Tadros, Th. F., Ed.; Academic Press: London, 1984; p 39.

(700) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M. *Langmuir* **1996**, *12*, 1462.

(701) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M. *J. Colloid Interface Sci.* **1997**, *187*, 113.

(702) Kuanar, M.; Kuanar, S. K.; Mishra, B. K. *Indian J. Chem.* **1999**, *38A*, 113.

(703) Roberts, D. W. *Langmuir* **2002**, *18*, 345.

(704) Wang, Z.-W.; Li, G.-Z.; Zhang, X.-Y. *Huaxue Xuebao* **2002**, *60*, 1548.

(705) Wang, Z.-W.; Huang, D.-Y.; Gong, S.; Li, G. Z. *Chin. J. Chem.* **2003**, *21*, 1573.

(706) Li, X.; Zhang, G.; Dong, J.; Zhou, X.; Yan, X.; Luo, M. *THEOCHEM* **2004**, *710*, 119.

(707) Qiu, M.-H.; Cao, C.-Z.; Zhao, L.-G.; Luo, J. *Yingyong Huaxue* **2004**, *21*, 276.

(708) Ahmed, E.; Gad, M. *J. Dispersion Sci. Technol.* **2007**, *28*, 231.

(709) Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M. *J. Chem. Inf. Model.* **2007**, *47*, 782.

(710) Huibers, P. D. T.; Shah, D. O.; Katritzky, A. R. *J. Colloid Interface Sci.* **1997**, *193*, 132.

(711) Connors, K. A. *Chem. Rev.* **1997**, *97*, 1325.

(712) Szejtli, J. *Chem. Rev.* **1998**, *98*, 1743.

(713) Hedges, A. R. *Chem. Rev.* **1998**, *98*, 2035.

(714) Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529.

(715) Solov'ev, V. P.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847.

(716) Fitch, W. L.; McGregor, M.; Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 830.

(717) Yanagita, M.; Kanda, S.; Tokita, S. *Mol. Cryst. Liq. Cryst. Sci. Technol. Sect. A* **1999**, *327*, 53.

(718) Türker, L. *THEOCHEM* **2002**, *588*, 133.

(719) Horiguchi, E.; Shirai, K.; Matsuoka, M.; Matsui, M. *Dyes Pigments* **2002**, *53*, 45.

(720) Al-Hazmy, S. M.; Kassab, K. N.; El-Daly, S. A.; Ebeid, E. Z. M. *Spectrochim. Acta A* **2000**, *56*, 1773.

(721) Machado, A. E. H.; Miranda, J. A.; Guilardi, S.; Nicodem, D. E.; Severino, D. *Spectrochim. Acta A* **2003**, *59*, 345.

(722) de Melo, S.; Fernandes, P. F. *THEOCHEM* **2001**, *565*, 69.

(723) Maud, J. M. *Synth. Met.* **1999**, *101*, 575.

(724) Breza, M.; Lukeš, V.; Vrábel, I. *THEOCHEM* **2001**, *572*, 151.

(725) Lukeš, V.; Breza, M.; Laurinc, V. *THEOCHEM* **2002**, *582*, 213.

(726) Lukeš, V.; Breza, M.; Végh, D.; Hrdlovič, P.; Krajčovič, J.; Laurinc, V. *Synth. Met.* **2002**, *129*, 85.

(727) Ridley, J.; Zerner, M. C. *Theor. Chim. Acta* **1973**, *32*, 111.

(728) Pople, J. A.; Beveridge, D. L.; Dobosh, P. A. *J. Chem. Phys.* **1967**, *47*, 2026.

(729) Pople, J. A.; Beveridge, D. *Approximate Molecular Orbital Theory*; McGraw-Hill: 1970.

(730) Gross, E.; Dobson, J.; Petersilka, M. *Top. Curr. Chem.* **1996**, *181*, 81.

(731) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(732) Williams, G. M.; Carr, R. A. E.; Congreve, M. S.; Kay, C.; McKeown, S. C.; Murray, P. J.; Scicinski, J. J.; Watson, S. P. *Angew. Chem., Int. Ed. Engl.* **2000**, *39*, 3293.

(733) Molnar, S. P.; King, J. W. *Int. J. Quantum Chem.* **1997**, *65*, 1047.

(734) Tämm, K.; Fara, D. C.; Katritzky, A. R.; Burk, P.; Karelson, M. *J. Phys. Chem. A* **2004**, *108*, 4812.

(735) Jover, J.; Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1727.

(736) Toropova, A. P.; Toropov, A. A.; Ishankhodzhaeva, M. M.; Parpiev, N. A. *Zh. Neorg. Khim.* **2000**, *45*, 1169.

(737) Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem. (Transl. Koord. Khim.)* **2000**, *26*, 792.

(738) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107.

(739) Bonchev, D.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1992**, *9*, 75.

(740) Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem. (Transl. Koord. Khim.)* **2001**, *27*, 574.

(741) Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem. (Transl. Koord. Khim.)* **2002**, *28*, 877.

(742) Toropov, A. A.; Toropova, A. P.; Nesterova, A. I.; Nabiev, O. M. *Russ. J. Coord. Chem. (Transl. Koord. Khim.)* **2004**, *30*, 611.

(743) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. *J. Chem. Inf. Model.* **2006**, *46*, 808.

(744) Mendenhall, W.; Scheafter, R. *Mathematical Statistics with Applications*; Duxburry Press: 1973.

(745) Solov'ev, V. P.; Kireeva, N. V.; Tsivadze, A. Yu.; Varnek, A. *J. Struct. Chem.* **2006**, *47*, 298.

(746) *ISIDA Project*, http://infochim.u-strasbg.fr/recherche/isida/index.php, 2004.

(747) Ghasemi, J.; Saaidpour, S. *J. Inclusion Phenom. Macrocycl. Chem.* **2008**, *60*, 339.

(748) Svetlitski, R.; Lomaka, A.; Karelson, M. *Sep. Sci. Technol.* **2006**, *41*, 197.

(749) Varnek, A.; Fourches, D.; Solov'ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1365.

(750) Catalan, J.; Diaz, C.; Garcia-Blanco, F. *J. Org. Chem.* **2000**, *65*, 3409.

(751) Katritzky, A. R.; Perumal, S.; Petrukhin, R. *J. Org. Chem.* **2001**, *66*, 4036.

(752) Sabljić, A.; Peijnenburg, W. *Pure Appl. Chem.* **2001**, *73*, 1331.

(753) Güsten, H.; Medven, Z.; Sekušak, S.; Sabljić, A. *SAR QSAR Environ. Res.* **1995**, *4*, 197.

(754) Güsten, H. *Chemosphere* **1999**, *38*, 1361.

(755) Meylan, W. M.; Howard, P. H. *Environ. Toxicol. Chem.* **2003**, *22*, 1724.

(756) *AOPWIN*, Version 1.90; Environmental Protection Agency: U.S.A., 2000.

(757) Atkinson, R. A. *Int. J. Chem. Kinet.* **1987**, *19*, 799.

(758) Klamt, A. *Chemosphere* **1993**, *26*, 1273.

(759) Klamt, A. *Chemosphere* **1996**, *32*, 717.

(760) Neeb, P. *J. Atmos. Chem.* **2000**, *35*, 295.

(761) Gramatica, P.; Consonni, V.; Todeschini, R. *Chemosphere* **1999**, *38*, 1371.

(762) Gramatica, P.; Pilutti, P.; Papa, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794.

(763) Bakken, G. A.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1064.

(764) Medven, Z.; Güsten, H.; Sabljić, A. *J. Chemom.* **1996**, *10*, 135.

(765) Pompe, M.; Veber, M.; Randić, M.; Balaban, A. T. *Molecules* **2004**, *9*, 1160.

(766) Pompe, M.; Randić, M. *J. Chem. Inf. Model.* **2006**, *46*, 2.

(767) Öberg, T. A. *Atmos. Environ.* **2005**, *39*, 2189.

(768) Kumar, M.; Thurow, K.; Stoll, N.; Stoll, R. *Eur. J. Med. Chem.* **2007**, *42*, 675.

(769) Bakken, G. A.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 508.

(770) Heberger, K.; Borosy, A. P. *J. Chemom.* **1999**, *13*, 473.

(771) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa Prediction for Organic Acids and Bases*; Chapman and Hall Ltd.: London, 1981.

(772) CompuDrug NA, Inc., pKalc version 3.1, 1996.

(773) ACD Inc., ACD/pKa Version 1.0, 1997.

(774) Santili-Kakoulidou, A. T.; Panderi, I.; Sizmadia, F. C.; Darvas, F. *J. Pharm. Sci.* **1997**, *86*, 1173.

(775) da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **1999**, *103*, 11194.

(776) Citra, M. J. *Chemosphere* **1999**, *38*, 191.

(777) Oberoi, H.; Allewell, N. M. *Biophys. J.* **1993**, *65*, 48.

(778) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415.

(779) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458.

(780) Warwicker, J. *Protein Sci.* **1999**, *8*, 418.

(781) Grüber, C.; Buss, V. *Chemosphere* **1989**, *19*, 1595.

(782) Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J. S.; Politzer, P. *J. Org. Chem.* **2001**, *66*, 6919.

(783) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457.

(784) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. *Quant. Struct.-Act. Relat.* **2002**, *21*, 473.

(785) Xing, L.; Glen, R. C.; Clark, R. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870.

(786) Polanski, J.; Gieleciak, R.; Bak, A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184.

(787) Cherkasov, A.; Sprous, D. G.; Chen, R. *J. Phys. Chem. A* **2003**, *107*, 9695.

(788) Soriano, E.; Cerdan, S.; Ballesteros, P. *THEOCHEM* **2004**, *684*, 121.

(789) Luan, F.; Ma, W.; Zhang, H.; Zhang, X.; Liu, M.; Hu, Z.; Fan, B. *Pharm. Res.* **2005**, *22*, 1454.

(790) Chaudry, U. A.; Popelier, P. L. A. *J. Org. Chem.* **2004**, *69*, 233.

(791) Popelier, P. L. A.; Smith, P. J. *Eur. J. Med. Chem.* **2006**, *41*, 862.

(792) Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. *J. Mol. Model.* **2006**, *12*, 338.

(793) Pompe, M.; Randić, M. *Acta Chim. Slov.* **2007**, *54*, 605.

(794) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. *J. Chem. Inf. Model.* **2007**, *47*, 2172.

(795) Ghasemi, J.; Saaidpour, S.; Brown, S. D. *THEOCHEM* **2007**, *805*, 27.

(796) Lee, P. H.; Ayyampalayam, S. N.; Carreira, L. A.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. *Mol. Pharm.* **2007**, *4*, 498.

(797) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348.

(798) Jover, J.; Bosque, R.; Sales, J. *QSAR Combin. Sci.* **2007**, *26*, 385.

(799) Jover, J.; Bosque, R.; Sales, J. *QSAR Combin. Sci.* **2008**, *27*, 563.

(800) Giri, S.; Roy, D. R.; Van Damme, S.; Bultinck, P.; Subramanian, V.; Chattaraj, P. K. *QSAR Combin. Sci.* **2008**, *27*, 208.

(801) Shan, X.; Qin, W.; Zhou, Z.; Dai, Y. *J. Chem. Eng. Data* **2008**, *53*, 331.

(802) Bicerano, J. *Prediction of Polymer Properties*, 2nd ed.; Marcel Dekker Inc.: New York, 1996.

(803) Xu, J.; Guo, B.; Chen, B.; Zhang, Q. J. *J. Mol. Model.* **2005**, *12*, 65.

(804) Ignatz-Hoover, F.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 295.

(805) Vidal, M.; Rogers, W. J.; Holste, J. C.; Mannan, M. S. *Process Saf. Prog.* **2004**, *23*, 47.

(806) Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. *Russ. Chem. Bull.* **2003**, *52*, 1885.

(807) Gramatica, P.; Navas, N.; Todeschini, R. *TrAC Trends Anal. Chem.* **1999**, *18*, 461.

(808) Gramatica, P.; Battaini, F.; Papa, E. *Fresenius Environ. Bull. Spec. Iss. SI* **2004**, *13*, 1258.

(809) Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 491.

(810) Suzuki, T.; Ohtaguchi, K.; Koide, K. *J. Chem. Eng. Jpn.* **1991**, *24*, 258.

(811) Katritzky, A. R.; Petrukhin, R.; Jain, R.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521.

(812) Catoire, L.; Naudet, V. *J. Phys. Chem. Ref. Data* **2004**, *33*, 1083.

(813) Stefanis, E.; Constantinou, L.; Panayiotou, C. *Ind. Eng. Chem. Res.* **2004**, *43*, 6253.

(814) Vidal, M.; Rogers, W. J.; Mannan, M. S. *Process Saf. Environ. Protect.* **2006**, *84*, 1.

(815) Catoire, L.; Paulmier, S.; Naudet, V. *Process Saf. Prog.* **2006**, *25*, 33.

(816) Catoire, L.; Paulmier, S.; Naudet, V. *J. Phys. Chem. Ref. Data* **2006**, *35*, 9.

(817) Pan, Y.; Jiang, J.; Wang, Z. *J. Hazard. Mater.* **2007**, *147*, 424.

(818) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Dobchev, D. A.; Karelson, M. *J. Mol. Graph. Model.* **2007**, *26*, 529.

(819) Egolf, L. M.; Jurs, P. C. *Ind. Eng. Chem. Res.* **1992**, *31*, 1798.

(820) Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538.

(821) Tetteh, J.; Metcalfe, E.; Howells, S. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 177.

(822) Tetteh, J.; Howells, S.; Metcalfe, E.; Suzuki, T. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 17.

(823) Yoshida, H.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1115.

(824) Kim, Y. S.; Lee, S. K.; Kim, J. H.; Kim, J. S.; No, K. T. *J. Chem. Soc., Perkin Trans. 2* **2002**, *12*, 2087.

(825) Albahri, T. A.; George, R. S. *Ind. Eng. Chem. Res.* **2003**, *42*, 5708.

(826) Pan, Y.; Jiang, J. C.; Wang, R.; Cao, H. Y.; Zhao, J. B. *J. Hazard. Mater.* **2008**, *2*−*3*, 510.

(827) Pan, Y.; Jiang, J. C.; Wang, R.; Cao, H. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 169.

(828) Suzuki, T. *Fire Mater.* **1994**, *18*, 81.

(829) Pan, Y.; Jiang, J. C.; Wang, R.; Cao, H.; Cui, Y. *J. Hazard. Mater.* **2009**, *164*, 1242.

(830) Myers, M. E. *Anal. Chem.* **1975**, *47*, 2301.

(831) Meusinger, R.; Moros, R. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 67.

(832) Meusinger, R.; Moros, R. *Fuel* **2001**, *80*, 613.

(833) Estrada, E.; Gutierrez, Y. *MATCH Commun. Math. Comput. Chem.* **2001**, *44*, 155.

(834) Nikolić, S.; Plavšić, D.; Trinajstić, N. *MATCH Commun. Math. Comput. Chem.* **2001**, *44*, 361.

(835) Hosoya, H. *Croat. Chem. Acta* **2002**, *75*, 433.

(836) Albahri, T. A. M.; Riazim, M. R.; Alqattan, A. A. *Energy Fuels* **2003**, *17*, 689.

(837) Albahri, T. A. *Ind. Eng. Chem. Res.* **2003**, *42*, 657.

(838) Podlipnik, C.; Šolmajer, T.; Koller, J. *MATCH Commun. Math. Comput. Chem.* **2004**, *52*, 55.

(839) Ghosh, P.; Hickey, K. J.; Jaffe, S. B. *Ind. Eng. Chem. Res.* **2006**, *45*, 337.

(840) Pasadakis, N.; Gaganis, V.; Foteinopoulos, C. *Fuel Process. Technol.* **2006**, *87*, 505.
(841) Smolenskii, E. A.; Ryzhov, A. N.; Bavykin, V. M.; Myshenkova, T. N.; Lapidus, A. L. *Russ. Chem. Bull.* **2007**, *56*, 1681.
(842) Ladommatos, N.; Goacher, J. *Fuel* **1995**, *74*, 1083.
(843) Yang, H.; Fairbridge, C.; Ring, Z. *Petrol. Sci. Technol.* **2001**, *19*, 573.
(844) Yang, H.; Ring, Z.; Briker, Y.; McLean, N.; Friesen, W.; Fairbridge, C. *Fuel* **2002**, *81*, 65.
(845) Kapur, G. S.; Ecker, A.; Meusinger, R. *Energy Fuels* **2001**, *15*, 943.
(846) Basu, B.; Kapur, G. S.; Sarpal, A. S.; Meusinger, R. *Energy Fuels* **2003**, *17*, 1570.
(847) Ghosh, P.; Jaffe, S. B. *Ind. Eng. Chem. Res.* **2006**, *45*, 346.
(848) Santana, R. C.; Do, P. T.; Santikunaporn, M.; Alvarez, W. E.; Taylor, J. D.; Sughrue, E. L.; Resasco, D. E. *Fuel* **2006**, *85*, 643.
(849) Smolenskii, E. A.; Bavykin, V. M.; Ryzhov, A. N.; Slovokhotova, O. L.; Chuvaeva, I. V.; Lapidus, A. L. *Russ. Chem. Bull.* **2008**, *57*, 461.
(850) Smolenskii, E. A.; Vlasova, G. V.; Platunov, D. Y.; Ryzhov, A. N. *Russ. Chem. Bull.* **2006**, *55*, 1508.
(851) Morita, E. *Rubber Chem. Technol.* **1984**, *57*, 744.
(852) Ignatz-Hoover, F.; Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Rubber Chem. Technol.* **1999**, *72*, 318.
(853) van Krevelen, D. W. *Properties of Polymers Their Estimation and Correlation with Chemical Structure*, 2nd ed.; Elsevier: New York, 1976.
(854) Wiff, D. R.; Altieri, M. S.; Goldfarb, I. J. *J. Polym. Sci., Part B: Polym. Phys.* **1985**, *23*, 1165.
(855) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. *J. Polym. Sci., Part B: Polym. Phys.* **1988**, *26*, 2007.
(856) Koehler, M. G.; Hopfinger, A. J. *Polymer* **1989**, *30*, 116.
(857) Sumpter, B.; Noid, D. *Macromol. Theory Simul.* **1994**, *3*, 363.
(858) Ulmer, C. W.; Smith, D. A.; Sumpter, B. G.; Noid, D. I. *Comput. Theor. Polym. Sci.* **1998**, *8*, 311.
(859) Joyce, S. J.; Osguthorpe, D. J.; Padgett, J. A.; Price, G. J. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 2491.
(860) Camelio, P.; Lazzeri, V.; Waegell, B. *Polym. Prepr.: Am. Chem. Soc., Div. Polym. Chem.* **1995**, *36*, 661.
(861) Cypcar, C. C.; Camelio, P.; Lazzeri, V.; Mathias, L. J.; Waegell, B. *Macromolecules* **1996**, *29*, 8954.
(862) Camelio, P.; Cypcar, C. C.; Lazzeri, V.; Waegell, B. *J. Polym. Sci., Part A: Polym. Chem.* **1997**, *35*, 2579.
(863) Camelio, P.; Lazzeri, V.; Waegell, B.; Cypcar, C.; Mathias, L. J. *Macromolecules* **1998**, *31*, 2305.
(864) Gao, H.; Harmon, J. P. *J. Appl. Polym. Sci.* **1997**, *64*, 507.
(865) Tan, T. T. M.; Rode, B. M. *Macromol. Theory Simul.* **1996**, *5*, 467.
(866) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879.
(867) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300.
(868) Garcia-Domenech, R.; de Julian-Ortiz, J. V. *J. Phys. Chem. B* **2002**, *106*, 1501.
(869) Cao, C. Z.; Lin, Y. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 643.
(870) Afantitis, A.; Melagraki, G.; Makridima, K.; Alexandridis, A.; Sarimveis, H.; Iglessi-Markopoulou, O. *THEOCHEM* **2005**, *716*, 193.
(871) Yu, X.; Wang, X.; Li, X.; Gao, J.; Wang, L. *Macromol. Theory Simul.* **2006**, *15*, 94.
(872) Yu, X.; Wang, X.; Wang, H.; Liu, A.; Zhang, C. *THEOCHEM* **2006**, *766*, 113.
(873) Yu, X.; Yi, B.; Wang, X.; Xie, Z. *Chem. Phys.* **2007**, *332*, 115.
(874) Yu, X. L.; Yi, B.; Wang, X. Y. *J. Theor. Comput. Chem.* **2008**, *7*, 953.
(875) Gao, J.; Wang, X.; Li, X.; Yu, X.; Wang, H. *J. Mol. Model.* **2006**, *12*, 513.
(876) Mattioni, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232.
(877) Duce, C.; Micheli, A.; Starita, A.; Tine, M. R.; Solaro, R. *Macromol. Rapid Commun.* **2006**, *27*, 711.
(878) Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *Macromol. Symp.* **2006**, *234*, 13.
(879) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *Polymer* **2007**, *48*, 7121.
(880) Funar-Timofei, S.; Kurunczi, L.; Iliescu, S. *Polym. Bull.* **2005**, *54*, 443.
(881) Kim, Y. S.; Kim, J. H.; Kim, J. S.; No, K. T. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 75.
(882) Yin, S. W.; Shuai, Z.; Wang, Y. L. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 970.
(883) Xu, J.; Chen, B. *J. Mol. Model.* **2005**, *12*, 24.
(884) Morrill, J. A.; Jensen, R. E.; Madison, P. H.; Chabalowski, C. F. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 912.
(885) Dai, J. F.; Liu, S. L.; Chen, Y.; Cao, C. Z. *Acta Polym. Sin.* **2003**, *3*, 343.
(886) Sun, H.; Tang, Y. W.; Wu, G. S. *Macromol. Res.* **2002**, *10*, 13.
(887) Sun, H.; Tang, Y. W.; Wu, G. S.; Zhang, F. S. *J. Polym. Sci., Part B: Polym. Phys.* **2002**, *40*, 454.
(888) Sun, H.; Tang, Y. W.; Zhang, F. S.; Wu, G. S.; Chan, S. K. *J. Polym. Sci., Part B: Polym. Phys.* **2002**, *40*, 2164.
(889) Dyekjaer, J. D.; Jonsdottir, S. O. *Carbohydr. Res.* **2004**, *339*, 269.
(890) Liu, W. Q.; Yi, P. G.; Tang, Z. L. *QSAR Comb. Sci.* **2006**, *25*, 936.
(891) Brown, W. M.; Martin, S.; Rintoul, M. D.; Faulon, J. L. *J. Chem. Inf. Model.* **2006**, *46*, 826.
(892) Liu, A.; Wang, X.; Wang, L.; Wang, H.; Wang, H. L. *Eur. Polym. J.* **2007**, *43*, 989.
(893) Schut, J.; Bolikal, D.; Khan, I. J.; Pesnell, A.; Rege, A.; Rojas, R.; Sheihet, L.; Murthy, N. S.; Kohn, J. *Polymer* **2007**, *48*, 6115.
(894) Ning, L. W. *J. Mater. Sci.* **2009**, *44*, 3156.
(895) Gao, J.; Wang, X.; Li, X.; Yu, X.; Wang, H. *J. Mol. Model.* **2006**, *12*, 513.
(896) Karbowiak, T.; Debeaufort, F.; Voilley, A. *Crit. Rev. Food Sci. Nutr.* **2006**, *46*, 391.
(897) Sheridan, P. L.; Buckton, G.; Storey, D. E. *Int. J. Pharm.* **1995**, *125*, 141.
(898) Suihko, E.; Forbes, R. T.; Korhonen, O.; Ketolainen, J.; Paronen, P.; Gynther, J.; Poso, A. *J. Pharm. Sci.* **2005**, *94*, 745.
(899) Reynolds, C. H. *J. Comb. Chem.* **1999**, *1*, 297.

CR900238D