QSPR and QSAR Models Derived with CODESSA Multipurpose Statistical Analysis Software

Mati Karelson and Uko Maran

Department of Chemistry, University of Tartu
2 Jakobi Str.
Tartu 51014, Estonia
mati@chem.ut.ee and uko@chem.ut.ee

Yilin Wang and Alan R. Katritzky

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida
P.O. Box 117200
Gainesville, FL 32611-7200, U.S.A.
katritzky@chem.ufl.edu

Abstract

An overview on the development of QSPR/QSAR equations using various descriptor mining techniques and multilinear regression analysis in the framework of program CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) is given. The description of the methodologies applied in CODESSA is followed by the presentation of the QSAR and QSPR models derived for eighteen molecular activities and properties. The properties cover single molecular species, interactions between different molecular species, properties of surfactants, complex properties and properties of polymers.

Historical Introduction

The fast progress in modern computer technology has created an entirely new environment for the efficient use of the theoretical constructions of natural science in many areas of applied research. The theoretical approach has proven to be especially beneficial in chemistry and allied sciences, where the experimental study and synthetic development of new compounds and materials can frequently be time consuming, expensive or even hazardous. Contemporary quantum theory of molecular matter and the corresponding *ab initio* computational methods can, in principle, predict the properties of isolated small molecules with an accuracy comparable to the experimental precision. However, the majority of industrially and environmentally important chemical processes, and all biochemical transformations in living cells take place in heterogeneous condensed media. The extreme complexity of such systems usually prohibits use of *ab initio* theory and thus the relationship between the chemical and physical properties and the molecular structure in these systems is often poorly described and understood.

The direct development of empirical equations that are commonly referred to as the quantitative structure-activity/property relationships (QSAR/QSPR) has been an attractive alternative approach to predict molecular properties in complex systems. Notably, the QSAR methodology has been extremely productive in pharmaceutical chemistry and in computer-assisted drug design. Thousands of potential new therapeutic agents have been first developed on a computer screen before the attempted implementation of selected examples in a synthetic laboratory. In analytical chemistry, QSPR equations are commonly used to predict spectroscopic, chromatographic and other analytical properties of compounds. In recent years, the QSPR approach has been rapidly expanding to diverse areas of industrial and environmental chemistry.

In most contemporary applications, *empirical* molecular descriptors that rely on some experimental data have been used in the development of QSAR/QSPR equations. Such descriptors, ranging from the original Hammett substituent σ-constants to the highly popular partition coefficients between water and octanol (*logP*) are, strictly speaking, restricted to those compounds for which the necessary experimental data are available. Another shortcoming of experimental descriptors evolves from the fact that many of them reflect a complicated combination of different physical interactions and thus their appearance in a QSAR/QSPR equation may be difficult to interpret. An alternative approach is to use molecular descriptors which can be derived using only the information encoded in the chemical structure of the compound. Importantly, such *theoretical* descriptors can be developed for compounds that are experimentally unexplored, unavailable, or even unknown.

The objective of this review is to provide a compilation of the utility of theoretical molecular descriptors in a variety of topics of chemistry, technology and related areas of research. In this review we have, for reason of space, restricted ourself mainly to work carried out using the CODESSA software, developed by our groups on both the MS Windows (Katritzky, Lobanov, and Karelson 1994a; Katritzky, Lobanov, and Karelson 1995) and the Unix platforms (Semichem 1995). QSPR treatments have been developed by many other groups. Pioneer work was and is being done in the groups of (in alphabetical order) Balaban (Balaban 1997), Bodor (Bodor, Harget, and Huang 1991), Benfenati (Benfenati and Gini 1997), Clementi (Pastor, Cruciani, and Clementi 1997), Hilal (Hilal, Carreira, and Karichoff 1994), Hopfinger (Hopfinger, Koehler, and Rogers 1995), Jurs (Jurs, Chou, and Yuan 1978), Kier and Hall (Kier and Hall 1986), Randic (Randic, Jerman-Blazic, and Trinajstic 1990), Trinajstic (Randic and Trinajstic 1993), and the references quoted are but illustrative. In our groups, we have tried to obtain as general a relationship as possible, utilizing data sets of wide structural diversity and we have tried to address problems of technological as well as of academic interest.

Methodology

Geometry Optimization

The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compound. Accordingly, the property of interest and corresponding structures need to be prepared in a format acceptable for the computer. In practice there are numerous ways to prepare the data and each researcher can work out an individual approach. For the users of CODESSA, the key points of the data preparation are determined by the available computer-readable formats of the structure of compounds. The CODESSA software accepts various standard structure formats as input: MDL .mol file; Hyperchem .hin file; SYBYL .mol file; MOPAC/AMPAC regular .out file. In most cases, the use of the correct 3D molecular structure of the compounds is vitally important to predict correctly the molecular properties or the biological activity. Therefore, the geometry of the molecules needs to be optimized to obtain the correct shape and conformation of the molecule. A variety of molecular modeling programs are available that employ different molecular mechanics algorithms for the geometry optimization. The following steps have been frequently used for the preparation of data and generation of 3D structures: (i) input of the molecular geometry using a graphical interface or by downloading from the corresponding database; (ii) preliminary geometry optimization using molecular mechanics; (iii) refinement of the 3D molecular structure and calculation of electronic properties of compounds using (semiempirical) quantum mechanical methods. The next step, the calculation of molecular descriptors from these data, comprises the kernel of the CODESSA software.

Descriptor Generation

The CODESSA software package includes a tool for effective descriptor generation based on the information given by the input file for the structure. All these descriptors are derived only from the structure and calculated electronic properties of the molecules. The number of descriptors calculated depends on the constitution of the molecule and the selections made by the user. In most cases, more than 400 molecular descriptors can be calculated for a single molecule in the first instance. By combining the available standard descriptors using a special tool within the CODESSA program, this number can be substantially increased. An option for the development of new descriptors also provides the possibility to calculate fragment descriptors.

The molecular descriptors available in CODESSA are subdivided into various subsets according to the molecular features they reflect. Constitutional, topological, geometric, electrostatic, quantum-chemical, thermodynamic and solvation descriptors can thus be distinguished. However, such classification is somewhat arbitrary, because some descriptors are sensitive to several molecular features. The origin of various descriptors has been extensively described elsewhere (Katritzky, Lobanov, and Karelson 1994b; Murugan et al. 1994; Katritzky et al. 1996a; Katritzky et al. 1996b; Katritzky et al. 1997a) and we therefore limit ourselves here to a short classification of theoretical molecular descriptors.

Constitutional descriptors depend only on the chemical composition of the molecule and describe very simple dependencies such as the additivity of molecular properties from constant fragment contributions. *Topological* descriptors represent one of the most widely used class of molecular descriptors that are derived from the two-dimensional structural formula of the molecule. These descriptors are sensitive to molecular connectivity and reflect the branching of the molecule. *Electrostatic* descriptors reflect the structural charge distribution in the molecules. In many cases they are also related to the molecular topology and composition. Partial charge distributions can be calculated using various empirical schemes that are based on the electronegativities of atoms (Katritzky, Lobanov, and Karelson 1994a) or Mulliken charges obtained from the quantum mechanical calculations. *Geometrical* descriptors reflect the three-dimensional structure and shape of the molecule. A large number of molecular and local quantities characterizing the reactivity, shape and binding of a molecule as well as its molecular fragments and substituents can be defined as *quantum chemical descriptors* (Karelson, Lobanov, and Katritzky 1996). Quantum mechanical calculations have become routine even for rather large molecular systems and therefore the information related to the structure and

electronic distribution can be easily and efficiently used in deriving new descriptors and explaining the properties of molecules. *Thermodynamic descriptors*, which include the heat of formation, entropy and heat capacity of the compound, can be derived using the MOPAC/AMPAC software. *Solvatational* descriptors are also based on the quantum mechanical calculations and can be obtained using SCRF2.2 program of self-consistent reaction field model implemented in the MOPAC package (Karelson et al. 1989). A recent trend is the development of descriptors that reflect the 3-dimensional properties of molecules (Cramer, Famini, and Lowrey 1993; Hopfinger, Burke, and Dunn 1994; Cho, Garsia, and Bier 1996; Goodford 1996; Balaban 1997) which should be more appropriate to describe intermolecular interactions under real conditions.

Statistical Methods

The CODESSA software involves two menus that are needed for the development of QSAR/QSPR equations on the large molecular descriptor basis. The first of them features the preliminary analysis of data whereas the second incorporates various multivariate regression analysis techniques (Katritzky, Lobanov, and Karelson 1994a; Katritzky, Lobanov, and Karelson 1995).

The *preliminary analysis tool* involves one-dimensional, two-dimensional and multivariate analysis of property(s) and descriptor(s). Statistical characteristics of data such as the mean values, dispersions, standard deviations and variation coefficients can be calculated automatically both for the descriptors and for the properties. Also, the requirement of the normal distribution of data is checked according to various criteria. Two-dimensional analysis makes it possible to analyze the intercorrelation of descriptors. This procedure is mandatory to avoid the chance correlations due to the collinearity of the descriptors. The multivariate statistical analysis methods represented in CODESSA involve principal component analysis (PCA), nonlinear iterative partial last squares (NIPALS) and target transformation PCA techniques.

The *regression analysis tool* involves various techniques based on the (multi-)linear regression analysis to find the best QSPR/QSAR representation of a property studied. Simple linear and multi-linear regression methods can be used to develop relationships between the property and specific descriptors. Several strategies are encoded in CODESSA that can be used to develop the QSPR/QSAR equations with the maximum predictive and descriptive power. The strategies of the *heuristic* and the *best multi-linear regression* approaches are usually those chosen first (Katritzky, Lobanov, and Karelson 1994a; Katritzky, Lobanov, and Karelson 1995; Katritzky et al. 1996c; Katritzky et al. 1996b; Katritzky, Mu, and Karelson 1997b). These strategies are both based on the stepwise forward selection of scales that proceed from the statistical significance and collinearity control of the descriptors selected into the correlation equations. One can also use principal component regression analysis, NIPALS regression analysis, or non-linear regression analysis to develop the model with the best predictive and descriptive power. However, it should be emphasized that the development of the best QSAR/QSPR model for each particular property/activity often involves a combination of different approaches. Powerful strategies, already utilized by other groups in QSPR analysis, include methods that rely on neural networks (Bodor, Harget, and Huang 1991; Gasteiger and Zupan 1993; Egolf and Jurs 1993), simulated annealing (Sutter, Dixon, and Jurs 1995) and various data- and knowledge-mining techniques.

Available Programs

Many commercially available statistical software packages (STAT-GRAPHICS, MATLAB, LINPACK, etc.) (Meloun, Militky, and Forina 1992) include the standard multi-linear least squares technique and can in principle be used to develop QSAR/QSPR correlations. However, their extensive use in the QSAR/QSPR development is often inconvenient because of the need (i) to calculate and format the molecular descriptors separately using different software , (ii) to select manually individual descriptor into correlations, which is impractical for a large number (several hundreds) of virtual descriptor scales.

A number of software packages have been developed specifically for structure-activity/property relationship studies. These packages include, as a rule, modules for structure input and for the calculation of empirical and also non-empirical descriptors. In most cases, various techniques for the statistical data treatment are also incorporated into the package. For instance, the ADAPT (Automated Data Analysis and Pattern Recognition Toolkit) program of Jurs (Jurs, Chou, and Yuan 1987; Stuper, Brugger, and Jurs 1979) includes several methods to select the best subset of descriptors and the mapping of these descriptors onto the known biological activity or physical property using regression analysis or computational neural networks. ADAPT has also a large set of modules to generate structure-based descriptors classified as topological, geometrical, electronic and physicochemical. The TSAR software is a fully integrated QSAR package distributed by the Oxford Molecular Group (OMG). Using TSAR the molecular structures, properties and associated data can be conveniently treated in one straightforward chemical spreadsheet. It also combines the data visualization with the complex statistical analysis and works as a front-end to several software packages distributed by OMG. The QSAR+ is a module for the Cerius^2 program distributed by the Molecular Simulations Inc. QSAR+ allows the calculation of various electronic, conformational, shape and thermodynamic descriptors. It also offers linear regression analysis, stepwise and multiple linear regression analysis, principal component analysis and principal component regression, and partial least squares techniques for developing QSAR models. The resulted models can be validated with cross-validated

regression coefficients (R^2_{cv}), bootstrap R^2, and the Fisher significance test. The module Descriptor+ extends the range of QSAR analysis in Cerius^2 by supplying a wide range of generic descriptors. The SPARC (Performs Automated Reasoning in Chemistry) (Hilal, Carreira, and Karichoff 1994) software is a specific package for the prediction of physical properties and chemical reactivity parameters of organic compounds from the molecular structure data. It involves statistical methods related to the conventional linear free energy relationship (LFER) and structure activity relationships.

Some of the QSAR/QSPR programs are designed to handle specific data or compounds. A good example is TOPKAT (developed by Health Design, Inc. and distributed by OMG) which computes and automatically validates an assessment of the toxic and environmental effects of chemicals based on developed QSPR/QSAR models.

Results

Properties of Single Molecule Species

Boiling Point. The boiling point of a compound is predetermined by the intermolecular interactions in the liquid and by the difference in the molecular internal partition function in the gas phase and in the liquid at the boiling temperature. Therefore, it is expected to be related directly to the chemical structure of the molecule and indeed numerous methods have been developed for estimating the normal boiling point of a compound from its structure.

Our first study considered the boiling points of pyridines and piperidines (Murugan 1994). A data set of 84 compounds was used to generate a QSPR model (R^2 = 0.898) which involved six descriptors based only on molecular structure (Murugan 1994). A subsequent related study was limited to the boiling point of substituted pyridines (Katritzky et al. 1996c). A set of 64 non-associated (incapable of hydrogen bonding) pyridines resulted in a good two-parameter correlation (R^2 = 0.927) for the boiling points. The descriptors showed the importance of the effects related to the molecular mass (expressed by the gravitation index) and intermolecular dipole-dipole interactions in the liquid media. The full set of pyridines (85 compounds) included also those derivatives, which form hydrogen bond(s). A six-parameter correlation model was derived (R^2 = 0.948) with four additional descriptors which describe the hydrogen bond accepting/donating capability of compounds at the intra- or intermolecular level (Katritzky et al. 1996c).

These successful studies of small groups of compounds encouraged us to work with data sets which comprised a large structural and functional variability. Such a QSPR treatment of normal boiling points was carried out for a set of 298 structurally variable organic compounds (Katritzky et al. 1996b). A highly significant two-parameter correlation (R^2 = 0.9544, s = 16.2K) was obtained that involved theoretical descriptors with clear physical meaning. The first descriptor (gravitation index) is connected with the bulk cohesiveness, dispersion and cavity-formation effects in liquids. The second descriptor, the area-weighted surface charge of the hydrogen bonding donor atom(s), is connected with the hydrogen bonding ability of the molecule. A more refined QSPR model (with R^2 = 0.9732 and s = 12.4K) included, in addition, the most negative atomic partial charge and the number of the chlorine atoms in the molecule. The four parameter equation offered an average predicted error of 2.3% for a standard set of compounds with an average experimental error of 2.1%. The QSPR equations developed allowed remarkably accurate predictions of the normal boiling points for a number of simple inorganic compounds, including water.

In follow-up work, the data set of 298 compounds was extended to provide a still more diverse and general data set of 584 organic compounds containing C, H, N, O, S, F, Cl, Br and I atoms, compiled and divided into subsets by molecular functionalities (Katritzky, Lobanov, and Karelson 1998a). Additional descriptors were sought for each subset which together with the gravitation index and the charged surface area of hydrogen donor atoms, would model the boiling points. A final global eight-parameter correlation model had R^2 = 0.965 and a standard error of 15.5K that is close to the estimated experimental error. The model appears to be general for a wide variety of organic compounds and expands and refines the conclusions of previous correlation models of boiling point.

Melting Point. Another important physical property of pure compounds is the melting point. The melting point is a fundamental physical property specifying the transition temperature when the solid and liquid phases can coexist. Besides its direct utility as an indicator to whether a compound is solid or liquid under normal conditions, melting points have numerous applications in biochemical and environmental sciences due to their relationship with solubilities. Because of the complex interactions involved, the melting temperature is expected to be a difficult property to describe by a uniform QSPR model for compound sets with large structural variability. Additionally, many compounds crystallize in more than one polymorphic form, with different melting points. Therefore, our studies have been limited to distinct groups of compounds.

The melting points of 141 pyridines and piperidines were used to develop a QSPR model for these heterocycles. Six descriptors gave a reasonably good correlation of melting points with R^2 = 0.831 and cross-validated R^2_{cv} = 0.816 (Murugan et al. 1994). Later (Katritzky et al. 1996c), the data set was limited to pyridines only and updated with additional data points. The melting points of pyridine and 140 substituted pyridines yielded in a six-parameter

correlation with R^2 = 0.857, R^2_{cv} = 0.843 and standard deviation s = 36.1K. The most important descriptor reflects the importance of the hydrogen bonding ability of the compound. The other descriptors can be related to intermolecular interactions in condensed media, crystal lattice packing and the fact that solid insulators with a smaller energy gap between the valence band and the unoccupied band are more resistant to disordering (melting).

Another set comprised included 443 mono- and disubstituted benzenes. A correlation equation including nine descriptors (R^2 = 0.8373, s = 30.19K) was obtained for the whole set (Katritzky et al. 1997a). Three other six-parameter equations described the ortho-, meta-, and para-substituted compounds subsets. The importance of hydrogen bonding descriptors was again reflected in these QSPR models. Notably, the same hydrogen bonding descriptor (the area-weighted surface charge of the hydrogen bonding donor atom(s)) was also important in the prediction of the boiling points (Katritzky et al. 1996b). Apart from the hydrogen bonding ability of the molecules, the melting point is governed by the molecular packing in crystals (effects from molecular shape, size and symmetry), and other intermolecular interactions such as charge-transfer and dipole-dipole interactions in the solid phase.

Critical Temperature. The critical temperature is one of the important properties revealing the intermolecular interactions between molecules in the liquid state. The development of QSPR models for critical temperatures using CODESSA methodology has been successful. One- and three-parameter QSPR models were developed for sets of 76 hydrocarbons and of 165 structurally diverse molecules, respectively (Katritzky, Mu, and Karelson 1998b). The one parameter model utilizing the cube root of the gravitation index allows the prediction of critical temperatures for hydrocarbons with an average error of 13.9K (with R^2 = 0.9526, R^2_{cv} = 0.9472), while the three parameter prediction of critical temperatures for diverse molecules has an average error of 16.8K (with R^2 = 0.955, R^2_{cv} = 0.9547). The models confirmed that molecular size-dependent bulk effects (dispersion and cavity-formation) in the liquid state can be represented by functions of the gravitation index, whereas the hydrogen-bonding self-association interactions can be represented by the area weighted surface charge or hydrogen bonding donor atoms. However, the donor hydrogen structural features alone do not account for the differences among various hydrogen-bonding acceptors, and this inadequacy is more serious for the critical temperature than for the boiling point. The supplementary descriptors needed to account for the differences of hydrogen-bonding acceptors and branching effects in isomers differ for the two properties.

Flash Point. Preliminary correlations of flash points have given moderate results. A modest correlation (R^2 = 0.758) was obtained for the flash points of 126 pyridines (Murugan et al. 1994). Flash point appears to be a difficult physical property to predict unless some provision has been made to separate compounds into similar functional groups.

A reduced data set of 121 pyridines with the exclusion of experimentally questionable data was used to develop a six parameter equation for the flash points (Katritzky et al. 1996c) with R^2 = 0.837 (R^2_{cv} = 0.832, s = 16.7K). The descriptors employed in this equation indicate the importance of molecular bulk and hydrogen bonding effects in determining the flash point (Katritzky et al. 1996c).

Vapor Pressure. Vapor pressure determines the volatility of a chemical. It governs the exchange rate of a chemical across an air-water interface through Henry's Law Constant. Accurate vapor pressures of chemicals of low-volatility are often not available due to analytical difficulties. In such cases, the vapor pressure may be predicted using either the Clapeyron-Clausius equation and known values of the enthalpy of vaporization and the respective compressibility factor, or by a group-contribution method. Alternatively, the quantitative structure-property relationship approach is highly promising for the estimation of vapor pressures from descriptors derived solely from the molecular structure by fitting into experimental data. The method is more general and is particularly suitable for the prediction of the vapor pressure of new chemical products.

We applied regression analysis tools in CODESSA to develop a QSPR model for the vapor pressure. The best linear five-parameter correlation model (R^2 = 0.949, R^2_{cv} = 0.947, s = 0.331) applied to a set of 411 compounds (Katritzky et al. 1998c). The model indicates that vapor pressure is governed by structure factors similar to those already found for the boiling point. The gravitation index over all bonded atoms reflects the effective mass distribution in the molecule and effectively describes the molecular dispersion forces in the bulk liquid media. The hydrogen-bonding donor charged surface area also represents the forces of intermolecular attraction, particularly the hydrogen bonding ability of the compound. Three additional descriptors compensate for an inadequate description of the intermolecular interactions occurring in molecules containing fluorine, chlorine or nitrogen atoms. The cross-validated correlation coefficient shows the regression equation is of high stability and that the standard error approaches the experimental error of 0.32 log units.

Refractive Index. The refractive index (n) is one of the most important optical properties and is frequently employed to characterize organic compounds. The refractive index is defined as the ratio of the velocity of light in vacuum to the velocity of light in the substance of interest. It has been used as an indicator of the purity of organic compounds, but the relationship of refractive index to other optical, electrical and magnetic properties has more significance. The refractive index is connected to polarizability, critical temperature, surface tension, density, and boiling point. Refractive index is also widely used in

material science to evaluate the applicability of materials for various purposes. Prior to our work, no general QSPR relationship relating refractive index of organic compounds with the chemical structure had been proposed.

A five-parameter correlation equation (R^2 = 0.945, R^2_{cv} = 0.937, s = 0.0155) was obtained for a diverse set of 125 organic compounds (Katritzky, Sild, and Karelson 1998d). The descriptors reveal several interaction mechanisms important for the refractive index. Specifically, they include the polarizability and the polarity of the molecule, the charge distribution in the molecule, hydrogen bonding interactions in the medium, and molecular size dependent effects in the molecule. The calculated cross-validated correlation coefficient confirms the stability of the final QSPR model. The predicted values have an the average error of 0.8% when compared with the experimental values, therefore this QSPR relationship can be used for the prediction of refractive indices with a high degree of confidence.

Density. The normal density (i.e. the density at 1 atm and 20°C) is one of the major physicochemical properties used to characterize and identify a compound. Besides being an indicator for the physical state (condensed phase or gas) of a compound, the density also provides an indication of its utility in certain industrial applications. In addition, densities can be used to predict or estimate other physical properties such as critical pressures.

A general QSPR treatment of 303 structures (containing C, H, N, O, S, F, Cl, Br and I) incorporating a wide cross section of classes of liquid organic compounds provided a good two-parameter correlation for densities (R^2 = 0.9749, s = 0.0458 for density ρ^{20}) (Karelson and Perkson 1999). The main descriptor involved in this correlation represents the intrinsic density of the compound calculated as the ratio of the molecular mass and the molecular volume (represented by the overlapping van der Waals' atomic spheres model) of the molecule. The second term is defined as the average electrostatic interaction per atom in the molecule, a term that is formally analogous to the Madelung energy in ionic crystals. Correlations were also developed for individual classes of organic liquids.

Interactions Between Different Molecular Species

Octanol-water Partition Coefficient. A six-parameter CODESSA correlation model constructed for octanol-water partition coefficient of 71 pyridines showed R^2 = 0.943, R^2_{cv} = 0.929, s = 0.19 (Katritzky et al. 1996c). The descriptors indicate the importance of the constitution and topology of the compounds. The electrostatic and structural features of the N atom were reflected by four descriptors connected with the hydrogen bond acceptor ability of pyridines in water and in octanol.

Aqueous Solubility of Liquids and Solids. The aqueous solubilities (S_w) of organic compounds are very important in many research areas, such as pharmaceutical or environmental science. A confident prediction of the aqueous solubility of a compound could greatly assist drug design by avoiding the synthesis of unsuitable compounds. The many different predictive methods available fall into the following types: (1) Group contribution methods derived from measured aqueous solubilities; (2) Correlations with experimentally determined physicochemical properties such as boiling point, molecular surface area, molar volume, chromatographic retention time and others; (3) Correlations with descriptors calculated only from molecular structure.

The aqueous solubilities of a set of 96 hydrocarbons and 126 halogenated hydrocarbons excluding compounds capable of forming hydrogen bonds were correlated by a three term equation using descriptors calculated solely from molecular structure, with a correlation coefficient of 0.980 and a standard error of 0.386 log units, compared to an estimated average experimental error of 0.24 log units (Huibers and Katritzky 1998). This allows the estimation of aqueous solubilities of hydrocarbons and halogenated hydrocarbons (including PCBs). The key descriptor is the molecular volume, modified by topological and constitutional terms to account for features that increase the solubility of the molecules.

To develop a general QSPR model for calculating the aqueous solubilities of diverse organic compounds, the data set was enlarged to 411 compounds (Katritzky et al. 1998c) and a six-parameter correlation model (R^2 = 0.879, R^2_{cv} = 0.874, s = 0.573) was derived. Solute-solvent interactions are major determining factors for the aqueous solubilities of compounds and accordingly the descriptors involved in the model are related to the polarizability of the molecule, cavity-size effects (dispersion and cavity formation), shape of the molecule and specific solute-solvent interactions. The standard error of the model is within the estimated experimental error of 0.58 log units.

Aqueous Solubility of Gases and Vapors (Water-Air Partition Coefficients). The partitioning of non-electrolytes between air and water or aqueous solutions is of significant chemical and thermodynamical interest as well as of great practical importance. The partitioning of organic gases and vapors into water (L_w) has been studied using CODESSA on two sets of compounds (Katritzky, Mu, and Karelson 1996d). The first correlation equation (R^2 = 0.977, R^2_{cv} = 0.975, s = 0.20K) gives an excellent prediction for 95 alkanes, cycloalkanes, alkylarenes, and alkynes with two descriptors which reflect the effective mass distribution and the degree of branching of the hydrocarbon molecule, and adequately represent the effective dispersion and cavity formation effects for the solvation of nonpolar solutes in water. An enlarged set of organic compounds (406) with far greater structural variability gives a good correlation equation (R^2 = 0.941,

R^2_{cv} = 0.939, s = 0.53) involving five descriptors. These descriptors, which are completely different to those for the set of 95 nonpolar solutes, account for the dispersion energy of polar solutes in solution, the electrostatic part of the solute-solvent interaction and hydrogen-bonding interactions in liquids.

Vapor pressure (VP) and water solubility (S_w) are fundamental physical parameters and they can be used to derive many other properties. The relationship between water-air partition coefficient (L_w), water solubility (S_w) and vapor pressure (VP) is important because water solubility and vapor pressure can be determined more easily than water-air partition coefficients. Using the direct relationship L_w=24.45S_w/VP (Katritzky et al. 1998c) we could predict L_w by VP and S_w through experimental data and/or through the appropriate QSPR models for VP and S_w. The QSPR models also help the understanding of the different structural factors which determine VP, S_w and L_w.

Values for vapor pressure and aqueous solubility were predicted by the models described above for the diverse set of 411 compounds (Katritzky et al. 1998c). They were then used to predicted water-air partition coefficients according to the derived formula. The result was compared with experimental data. The mean standard error of this prediction is 0.63 log units, which is close to the standard error of L_w predicted using the equation derived directly from the experimental values of L_w (Katritzky, Mu, and Karelson 1996d). We conclude that hence this procedure is a valid approach to calculate L_w by using QSPR predicted values of VP and Sw. It is apparent that the QSPR models of S_w and VP (Katritzky et al. 1998c) have similar leading structural determining factors in comparison with the direct QSPR equation of L_w (Katritzky, Mu, and Karelson 1996d).

Solvent Polarity Scales. The use of solvents is fundamental to the practice of chemistry, and the choice of an appropriate solvent can be anything but trivial. To assist chemists in their understanding of solvent properties and in the choice of solvent, many solvent polarity scales have been developed. These scales are based on diverse physico-chemical phenomena including reaction rates, solvatochromic effects, reaction enthalpies, etc. Frequently the actual mechanism of the solvent influence on a physical or chemical process is unclear. The same is often true about the individual polarity scales.

A three-parameter QSPR equation with R^2 = 0.936 (R^2_{cv} = 0.900) was developed for the unified nonspecific solvent polarity scale (S') on the basis of theoretical molecular descriptors (Katritzky, Mu, and Karelson 1997b). It correlates S' for 25 structurally diverse solvents within a 5% average absolute error. The correlation equation includes the following three orthogonal theoretical molecular descriptors: (i) the average structural information content (order 0); (ii) the weighted partial negative surface area; and (iii) the hydrogen-bonding acceptor surface area. These descriptors provide insight into nonspecific solvation at the molecular level. They reflect adequately the solvent-solute interactions in the internal cavity of the solvents. Predictions using this three-parameter model are used to extend available S' values to a total of 67 solvents. The same solvent polarity scale has been also studied using CODESSA to enable the prediction of the S' values from quantum-mechanical calculations (Mu, Drago, and Richardson 1998).

In a more comprehensive study, the most important solvent polarity scales were collected and QSPR models developed for each of them. Altogether 45 different solvent polarity scales and 350 solvents were analyzed. The QSPR models for each of the scales were constructed using only theoretical descriptors. From these, 27 of the 45 models give $R^2 > 0.90$ and only two had $R^2 < 0.82$ (Katritzky et al. 1999a). This study allowed a unified PCA treatment of solvent polarity where the missing values in the polarity scales are calculated from correlation models derived with CODESSA (Katritzky, Tamm, and Karelson 1999b). A set of 40 scales and 40 solvents showed that three main principle components accounted for a total of 74% of the variance. Moreover for 29% of the scales these three components described ≥ 88% of the variance. The PCA loadings showed clear clustering of the scale is a 3 dimensional space in a chemical rational manner. Similarly the PCA scores classified the solvent intelligently (Katritzky, Tamm, and Karelson 1999b).

CODESSA has been also used to examine the dimensionality of intermolecular interactions in liquids and solutions (Stavrev, Tamm, and Zerner 1996; Karelson 1997a; Karelson 1997b).

GC Retention Time and Response Factor. A good six-parameter QSPR model was obtained for the retention indices of 50 polyalkylated pyridines (R^2 = 0.971, R^2_{cv} = 0.966, s = 017.8) (Katritzky et al. 1996c). The descriptors involved in the equation reflect the relative position and size of alkyl groups connected to the pyridine ring. They also show the importance of intermolecular interactions between solute and stationary phase, upon which gas chromatographic retention depends.

A general QSPR treatment on 152 individual structures incorporating a wide cross-section of classes of organic compounds provided good six-parameter correlations for gas chromatographic retention times (R^2 = 0.959, R^2_{cv} = 0.955, s = 0.515 for t_R) and for Dietz flame-ionization response factors (R^2 = 0.892, R^2_{cv} = 0.881, s = 0.0543 for RF_{Dietz}) (Katritzky et al. 1994a). In the case of t_R, the most important descriptors were α-polarizability and the minimum valency at an H atom, describing the dispersional and hydrogen-bonding interaction between the compound studied and the gas chromatographic medium, respectively. In the case of RF, the most important descriptors were found to be the relative weight of the "effective" carbon atoms and the total molecular one-center one-electron repulsion energy in the molecule. The possibility to predict value is of particular significance for the response factors, which are independent of GC column parameters. These results are recently reevaluated using improved procedures

in CODESSA and new methods for the efficient variable selection for multilinear regression analysis (Lucic et al. 1999).

Surfactant properties

Critical Micelle Concentration. The strategies implemented in CODESSA have been successful in developing QSAR models for complex surfactant properties such as critical micelle concentrations. We found that for these studies fragment descriptors, based on the two clearly differentiated parts of a surfactant, were applicable.

The first QSPR study was performed on the critical micelle concentrations (cmc) of nonionic surfactants (Huibers et al. 1996). A general three-parameter structure-property relation was developed for a diverse set of 77 nonionic surfactants (R^2 = 0.9849, R^2_{cv} = 0.9823, s = 0.1697) employing topological descriptors calculated separately for the hydrophobic and hydrophilic fragments of the surfactant molecule. The three descriptors represent contributions from the topology of the hydrophobic group, and the size of the hydrophilic group. The cmc of nonionic surfactants in aqueous solution is primarily determined by the hydrophobic part of the molecule. The logarithm of the cmc decreases with an increase in the size of the hydrophobic fragment and increases with an increase in the relative size of the hydrophilic fragment. Hydrophobicity is affected by the branching of the hydrophobic fragment and by the presence of heteroatoms.

Relationships between the molecular structure and the cmc of anionic surfactants were investigated in a second study (Huibers et al. 1997a). The measured cmc for 119 anionic structures were considered, representing sodium alkyl sulfates and sodium sulfonates with a wide variety of hydrophobic tails. The best multiple linear regression model involved three descriptors and had a correlation coefficient of R^2 = 0.940 (s = 0.2173). A still better correlation (R^2 = 0.986) was obtained using three descriptors for a subset of 63 structures, with variation only in the hydrophobic domain.

Cloud Point. 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 loses sufficient water solubility and a cloud dispersion results.

A general empirical relationship (R^2 = 0.937) has been developed for estimating the cloud point of pure nonionic surfactants of the alkyl ethoxylate class (Huibers, Shah, and Katritzky, 1997b). For a set of 62 structures, composed of linear alkyl, branched alkyl, cyclic alkyl, and alkylphenyl ethoxylates, cloud points can be estimated to an accuracy of ± 6.3ºC (3.7ºC median error) using the logarithm of the number of ethylene oxide residues and three topological descriptors that account for hydrophobic domain variation.

Complex Properties and Properties of Polymers

Polymer Glass Transition Temperature. The QSPR description of polymer properties represents an interesting challenge since many theoretical molecular descriptors for the high molecular weight compounds are difficult to calculate, or cannot be calculated directly. The glass transition temperature, T_g, also known as the glass temperature or the glass-rubber transition temperature, is one of the most important properties of amorphous polymers. In the vicinity of T_g, a polymer experiences a sudden increase in the rate of molecular motions and as a result undergoes a series of conformational transformations.

Using the CODESSA software, an optimum four-parameter QSAR model (R^2 = 0.928, R^2_{cv} = 0.890) was derived for glass transition temperatures for a homogenous set of 22 homo- and co-polymers (Katritzky et al. 1996a). Removing an obvious outlier from the data set improved the correlation to R^2 = 0.983. The descriptors in the correlation equations reveal that the glass transition temperatures of the polymers studied are strongly influenced by the difference between the positive and negative partial surface areas normalized by the number of atoms (DPSA). As expected the polymers with large DPSA values have stronger intermolecular electrostatic interactions and therefore display higher glass transition temperatures. The next most important descriptor is the topological Randic index computed for the repeating unit and then extrapolated through multiplication by log N (N - number of fragments), which reflects the branching level of a molecule. According to the QSPR model developed, a higher degree of branching in the repeating fragment structure elevates the glass transition temperature. The third parameter, the number of OH groups, is of the expected significance because it accounts for the presence of hydrogen bonds in the polymer matrix.

A five-parameter QSPR correlation (R^2 = 0.946 and the standard error 0.33 K g mol-1) of molar glass transition temperatures (T_g/M) for a diverse set of 88 high molecular weight polymers was developed as an extension of the earlier work to more diverse structures (Katritzky et al. 1998e). The polymers were modeled with three repeating units for each polymer and the descriptors were calculated only for the middle unit of the trimeric structure. In this way the influence from adjacent repeating units was also taken into account. The descriptors in the model relate to the rotational flexibility of the molecules at the T_g, the branching of the polymer molecules, hydrogen bonding interactions, and electrostatic interactions between the polymer molecules. This approach is applicable, in principle, to all linear polymers of regular structure, and encourages the further application of QSPR methods to other types of polymers such as copolymers, crosslinked polymers and biopolymers.

Polymer Refractive Index. For polymers the refractive index (n) is a fundamental optical property directly related

to other optical, electrical, and magnetic properties. Therefore a satisfactory quantitative structure-property relationship (QSPR) that would allow quantitative prediction of the refractive index of both known and of as yet unsynthesized polymers would clearly be of significant utility.

A general QSPR model (R^2 = 0.940, R^2_{cv} = 0.934, s = 0.018) was developed for the prediction of the refractive index for a diverse set of 95 amorphous homopolymers (Katritzky, Sild, and Karelson 1998f). The five descriptors, involved in the model, are calculated from the structure of the repeating unit of the polymer. The QSAR model was derived with an intercept fixed to a value of one, i.e. to the refractive index of a vacuum. The correlation model shows that the polarizability (described by the HOMO - LUMO energy gap) has an important influence on the refractivity index of polymers just as for low molecular weight compounds (see above) (Katritzky, Sild, and Karelson 1998f). Compounds of lower stability (described by the heat of formation) possess higher refractive indices. Other descriptors show the importance of charge distribution and the hybridization of carbon atoms in the repeating unit of the polymer. The average prediction error of the model is 0.9%, and the highest prediction error is 3.2%.

Rubber Vulcanization Acceleration. In spite of the fact that the vulcanization of rubber has been studied for many years, its precise mechanism has remained unclear. CODESSA QSPR treatment has assisted the understanding of some key features of this process. The regression analysis was carried out to correlate various parameters, including ts2 (the onset of cure) and mxr (the maximum rate of vulcanization), with molecular descriptors (Ignatz-Hoover et al. 1999). Correlations were performed on four data sets and two classes of accelerator molecules. The first class comprised disulfides and the other represented a combination of sulfenamides and sulfenimides. Parent molecules of the accelerators and also their zinc complexes with thiolate fragments were both modeled for each class to give a total of four data sets, all of which gave good correlations.

Biological Properties. The QSAR approach is widespread in the prediction of the biological activity of compounds. The CODESSA software has been used to study the mutagenic toxicity. A QSAR model with R^2 = 0.834 was derived for a set of 95 heteroaromatic and aromatic amines to correlate and predict their mutagenic activity measured by the *Ames test* (Maran, Katritzky, and Karelson 1999). It consists of six quantum-chemical descriptors, which indicate the importance in the mutagenic activity of heteroaromatic amines of hydrogen bonding, of effects induced by the solvent, and of the size of compound.

The theoretical descriptors and statistical methods implemented in CODESSA have been used to develop interpretative and predictive QSAR models for structurally diverse 5-$HT1_A$ receptor antagonists (Menziani, De Benedetti, and Karelson 1998). Altogether ten correlation models were analysed. The descriptors involved in these models show the importance of electrostatic interaction between the protonated amine function and a primary nucleophilic site of the receptor necessary for recognition, as expressed by the molecular orbital indexes localized on the $N\text{-}H^+$ group. Short-range attractive and repulsive intermolecular interactions which modulate the binding affinities are described by MO indexes computed on the whole molecules (polar and dispersive forces), by CPSA descriptors computed on the whole molecules or on the bicyclic fragments (polar forces) and by *ad hoc* defined size and shape descriptors (dispersive and steric forces).

Theoretical molecular descriptors in QSAR models also elucidated the role of the main pharmacophoric components and developed a model for the interaction of the 5-TH_3 ligands related to quipazine with their receptor (Cappelli et al. 1998). The essential nature of the arylpiperazine interaction mode toward the receptor can be summarized as follows: (i) a charge assisted hydrogen bond, (ii) a hydrogen-bonding interaction, and (iii) an aromatic specific interaction.

Problems and Future Potential

The development of QSAR/QSPR models on large theoretical descriptor spaces represents a powerful tool not only for the experimentally meaningful prediction of the chemical, physical and biological properties of compounds, but also for the deeper understanding of the detailed mechanisms of interactions in complex systems that predetermine these properties. Two directions seem to be especially promising for the further development of this approach. First, it is essential to derive new theoretical descriptors that correspond to clearly defined physical interactions in complex molecular systems. In particular, descriptors that account correctly for the properties of mixtures, blends and other multi-component systems would be of immense practical applicability in many areas of chemical technology and engineering (Hopfinger, Koehler, and Rogers 1995). It is also important to develop molecular descriptors that could properly account for environmental conditions such as the temperature, pressure and solvent.

The second direction in the further development of the QSAR/QSPR approach is undoubtedly connected with the extensive use of modern computer intelligence methods in the development of quantitative relationships between the molecular structure and properties. The methods that rely on neural networks (Bodor, Harget, and Huang 1991; Gasteiger and Zupan 1993; Egolf and Jurs 1993), simulated annealing (Sutter, Dixon, and Jurs 1995) and various data- and knowledge-mining techniques promise to be much more efficient in developing QSAR/QSPRs in large and very large molecular descriptor spaces. Notably, such approaches should help overcome problems often encountered in regression analysis due to the collinearity of scales or heterosedasticity of the data.

References

Balaban, A.T. 1997. From Chemical Topology to 3D Geometry. *J. Chem. Inf. Comput. Sci.* 37:645-650.

Benfenati, E. and Gini, G. 1997. Computational Predictive Programs (Expert Systems) in Toxicology. *Toxicology* 119:213-225.

Bodor, N.; Harget, A.; and Huang, M. 1991. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* 113:9480-9483.

Cappelli, A.; Anzini, M.; Vomero, S.; Mennuni, L.; Makovec, F.; Doucet, E.; Hamon, M.; Bruni, G.; Romeo, M. R.; Menziani, M. C.; De Benedetti, P.G.; and Langer, T. 1998. Novel Potent and Selective Central 5-HT_3 Receptor Ligands Provided with Different Intrinsic Efficacy. 1. Mapping the Central 5-HT_3 Receptor Binding Site by Arylpiperazine Derivatives. *J. Med. Chem.* 41: 728-741.

Cho, S.J.; Garsia, M.L.S.; Bier, J.; and Tropsha, A. 1996. Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors, *J. Med. Chem.* 39:5064-5071.

Cramer, C. J.; Famini, G. R.; and Lowrey, A. H. 1993. Use of Calculated Quantum Chemical Properties as Surrogates for Solvatochromic Parameters in Structure-Activity Relationships. *Acc. Chem. Res.* 26:599-605.

Egolf, L. M. and Jurs, P. C. 1993. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* 33:616-625.

Goodford, P. 1996. Multivariate characterization of molecules for QSAR analysis, *J. Chemometrics.* 10:107-117.

Gasteiger, J. and Zupan, J. 1993. Neural Netwoeks in Chemistry. *Angew. Chem. Int. Ed. Engl.* 32:503-527.

Hilal, S. H.; Carreira, L. A.; and Karichoff, S. W. 1994. Estimation of Chemical Reactivity Parameters and Physical Properties of Organic Molecules Using SPARC. In *Quantitative Treatment of Solute/Solvent Interactions, Theoretical and Computational Chemistry*, 291-353. Amsterdam: Elsevier Science B. V.

Hopfinger, A. J.; Burke, B. J.; and Dunn, W. J. 1994. A Generalized Formalism of 3-Dimensional Quantitative Structure-Property Relationship Analysis for Flexible Molecules Using Tensor Representation, *J. Med. Chem.* 37:3768-3774.

Hopfinger, A. J.; Koehler, M. G.; Rogers, D. 1995. Molecular Modelling of Polymers. 14. Quantitative Structure-Property Relationship Analyses of Multicomponent Systems Containing Polymers, *Macromolecular Symposia*. 98:1087-1100.

Huibers, P. D. T.; Lovanov, V. S.; Katritzky, A. R.; Shah, D. O.; and Karelson, M. 1996. Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants. *Langimuir* 12:1462-1470.

Huibers, P. D. T.; Lobanov V. S.; Katritzky, A. R.; Shah, D. O.; and Karelson, M. 1997a. Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 2. Anionic Surfactants. *J.Colloid Interface Sci.* 187:113-120.

Huibers, P. D. T.; Shah, D. O., and Katritzky, A. R. 1997b. Predicting Surfactant Cloud Point from Molecular Structure. *J. Colloid Interface Sci.* 193:132-136.

Huibers, P. D. T.; and Katritzky, A. R. 1998. Correlation of Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* 38:283-292.

Ignatz-Hoover, F.; Katritzky, A. R.; Lobanov, V. S.; and Karelson, M. 1999a. Applications of Semiempirical Molecular Orbital Calculations and Quantitive Structure Property Relations in the Study of Accelerated Sulfur Vulcanization. *Rubber Chem.and Technol.* Forthcoming.

Jurs, P. C.; Chou, J. T.; and Yuan, M. 1978b. Studies of chemical structure-biological activity relations using pattern recognition. In Computer-Assisted Drug Design, ACS Symposium Series 112, 103-129. Washington, DC: American Chemical Society.

Karelson, M.; and Perkson, A. 1999. The Correlation and Prediction of Densities of Organic Liquids, *Computers & Chemistry* Forthcoming.

Karelson, M. 1997a. Molecular Properties and Spectra in Solution. In *Problem Solving in Computational Molecular Science: Molecules in Different Environments*, 353-387. Dordrecht: Kluwer Academic Publ.

Karelson, M. 1997b Quantum Chemical Treatment of Molecules in Condensed Disordered Media. *Adv. Quant. Chem.* 28:142-159.

Karelson, M.; Lobanov, V. S.; and Katritzky, A. R. 1996. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* 96:1027-1043.

Karelson, M; Tamm, T.; Katritzky, A. R.; Cato, S. J.; and Zerner, M. C. 1989. Application of Self-Consistent Reaction Field Method in Semiempirical Quantum-

Chemical Calculations. *Tetrahedron Comp. Methodol.* 2:295-304.

Katritzky, A. R.; Lobanov, V. S.; and Karelson, M. 1994a. CODESSA: Reference Manual, version 2.0. Gainesville, FL.

Katritzky, A. R.; Ignatchenko, E. S.; Barcock, A. R.; Lobanov, V. S.; and Karelson, M. 1994b. Predicton of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure-Property Realationship Treatment. *Anal. Chem.* 66:1799-1807.

Katritzky, A. R.; Lobanov, V. S.; and Karelson, M. 1995. CODESSA: Training Manual. Gainesville, FL.

Katritzky, A. R.; Rachwal, P.; Law, K.W.; Karelson, M.; and Lobanov, V. S. 1996a. Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure-Property Relationship Treatment. *J. Chem. Inf. Comput. Sci.* 36:879-884.

Katritzky, A. R.; Mu, L.; Lobanov, V. S.; and Karelson, M. 1996b. Correlation of Boiling Points with Molecular Structure. 1. A Tarining Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* 100:10400-10407.

Katritzky, A. R.; Lobanov, V.; Karelson, M.; Murugan, R.; Grendze, M. P.; and Toomey, J. E. 1996c. Comprehensive Descriptors for Structural and Statistical Analysis. 1. Correlations Between Structure and Physical Properties of Substituted Pyridines. *Rev. Roum. Chim.* 41:851-867.

Katritzky, A. R.; Mu, L.; and Karelson, M. 1996d. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci.* 36:1162-1168.

Katritzky, A. R.; Maran, U.; Karelson, M.; and Lobanov, V. S. 1997a. Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci.* 37:913-919.

Katritzky, A. R.; Mu, L.; and Karelson, M. 1997b. QSPR Treatment of the Unified Nonspecific Solvent Polarity Scale. *J. Chem. Inf. Comput. Sci.* 37:756-761.

Katritzky, A. R.; Lobanov, V. S.; and Karelson, M. 1998a. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure-Property Relationship. *J. Chem. Inf. Comput. Sci.* 38:28-41.

Katritzky, A. R.; Mu, L.; and Karelson, M. 1998b. Relationship of Critical Temperatures to Calculated Molecular Properties. *J. Chem. Inf. Comput. Sci.* 38:293-299.

Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; and Karelson, M. 1998c. QSPR Studies on Vapor Pressure, Aqueous Solubility and the Prediction of Water-Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* 38:720-725.

Katritzky, A. R.; Sild, S.; and Karelson, M. 1998d. A General QSPR Treatment of the Refractive Index of Organic Compounds. *J. Chem. Inf. Comput. Sci.* 38:840-844.

Katritzky, A. R.; Sild, S.; Lobanov, V. S.; and Karelson, M. 1998e. Quantitative Structure-Property (QSAR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* 38:300-304.

Katritzky, A. R.; Sild, S.; and Karelson, M. 1998f. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J. Chem. Inf. Comput. Sci.* 38:1171-1176.

Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; and Karelson, M. 1999a. QSPR Treatment of Solvent Polarity Scale. *J. Chem. Inf. Comput. Sci.* Forthcoming.

Katritzky, A. R.; Tamm, T.; Wang, Y.; and Karelson, M. 1999b. A Unified Treatment of Solvent Polarity *J. Chem. Inf. Comput. Sci.* Forthcoming.

Kier, L. B. and Hall, L. H. 1986. Molecular Connectivity in Structure-Activity Analysis. New York: Wiley.

Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; and Katritzky, A. R. 1999. A New Efficient approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* Forthcoming

Maran, U.; Katritzky, A. R.; and Karelson, M. 1999. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.* Forthcoming.

Meloun, M.; Militky, M.; and Forina, M. 1992 *Chemometrics in Analytical Chemistry*. New York: Ellis Horwood.

Menziani, M. C.; De Benedetti, P. G.; and Karelson, M. 1998. Theoretical Descriptors in Quantitative Structure-Affinity and Selectivity Relationship Study of Potent N4-Substituted Arylpiperazine 5-$HT1_A$ Receptor Antagonists. *Bioorg. Med. Chem.* 6:535-550.

Mu, L.; Drago, R. S.; and Richardson, D. E. 1998. A Model Based QSPR Analysis of the Unified non-specific solvent polarity scale. *J. Chem. Soc., Perkin Trans. 2* 159-167.

Murugan, R.; Grendze, M. P.; Toomey, J. E.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; and Rachwal, P. 1994. Predicting Physical Properties from Molecular Structure. *CHEMTECH* 24:17-23.

Randic, M.; Jerman-Blazic, B.; and Trinajstic, N. 1990. Development of 3-Dimentional Molecular Descriptors. *Comput. Chem*. 14:237-246.

Pastor, M.; Cruciani, G.; and Clementi, S. 1997. Smart Region Definition: A New Way to Improve the Predictive Ability and Interpretability of Three - Dimensional Quantitative Structure-Relationships. *J. Med. Chem*. 40:1455-1464.

Randic, M. and Trinajstic, M. 1993. Comparative Structure-Property Studies: the Connectivity Basis. *J. Mol. Struc*. 284:209-215.

Semichem, 7128 Summit, Shawnee, KS 66216. CODESSA, 1995.

Stavrev, K. K.; Tamm, T.; and Zerner, M. C. 1996. Comparison of Theoretical Models of Solvation *Int. J. Quant. Chem.* 30S:1585-1594.

Stuper, A. J.; Brugger, W. E.; and Jurs, P. C. 1979. Computer-assisted Studies of Chemical Structure and Biological Function., New York: While.

Sutter, J. M.; Dixon, S. L.; and Jurs, P. C. 1995. Automated Descriptor Selection for Quantitative Structure-Activity Relationship Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* 35:77-84.