Physical, Chemical, and Technological Property Correlation with Chemical Structure: The Potential of QSPR*

Alan R. Katritzky[a], Dimitar A. Dobchev[a, b], and Mati Karelson[b, c]

[a] Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida, 32611, USA
[b] Department of Chemistry, University of Tartu, 2 Jakobi St., Tartu 51014, Estonia
[c] Department of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia

Reprint requests to A. R. Katritzky. E-mail: katritzky@chem.ufl.edu

Z. Naturforsch. **61b,** 373 – 384 (2006); received January 4, 2006

Correlations of simple and complex physical, and chemical, biological and technological properties with chemical structure are reviewed. When an adequate training set of structures and experimentally determined property values are available, the equations produced enable the prediction of these properties of molecules as yet synthesized or indeed as yet unknown. Frequently they also offer considerable insights into the manner in which the structure controls the property. Many further applications of this methodology can be anticipated.

Key words: QSPR, QSAR, CODESSA PRO, Multilinear Regression, Molecular Descriptors

Introduction

"Quantitative Structure Property Relationships" (QSPR) and "Quantitative Structure – Activity Relationships" (QSAR) relate a property or activity of interest, defined quantitatively by a numerical measure, to characteristic "descriptors" derived theoretically from the chemical structures of the compounds. In the last 35 years QSAR methodology has expanded exponentially out of analytical chemistry; it is now indispensible in the pharmaceutical chemistry and in drug design [1 – 7].

QSPR has also become a well-established and proven technique to correlate diverse physicochemical properties of compounds, ranging from simple to complex, with molecular structure, through a variety of "descriptors" (as discussed below) [8, 9] of the chemical structures. QSPR has received important contributions from the groups of Abraham [10], Balaban [11], Dearden [12], Hilal [13], Jurs [14], Kier and Hall [15], Politzer [16], Randić [17], Trinajstic [18] and many others including ourselves Katritzky and Karelson [19]. The basic strategy is to find reliable quantitative relationships, which can then be used for the prediction of that same property for other structures not yet measured or not yet prepared.

The molecular descriptors utilized for QSAR and QSPR equations are numerical parameters that are defined quantitatively from a chemical structure alone. Conventionally, molecular descriptors are classified in five main classes: (i) *constitutional*, describing the atomic composition of the compound, (ii) *topological*, which describe the way in which the atoms in the compounds are mutually bonded, (iii) *geometric*, (iv) *electrostatic* relating respectively to geometry and charge distribution, and (v) a very large number of *quantum chemical* descriptors obtained by quantum mechanics from the structure.

Most QSAR/QSPR treatments utilize a program to calculate descriptors and then try to select a small number of descriptors in a purely empirical fashion to form an equation. This is derived from a so-called "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. Very importantly, this helps researchers to predict and prepare structures with optimum properties. The software is also of great assistance for chemical and physical interpretation.

In the past fifteen years, our groups at the University of Florida, and at Tartu/Tallinn Estonia, have developed multipurpose statistical analysis soft-

* Presented in part at the 7th Conference on Iminium Salts (ImSaT-7), Bartholomä/Ostalbkreis, September 6 – 8, 2005.

0932–0776 / 06 / 0400–0373 $ 06.00 © 2006 Verlag der Zeitschrift für Naturforschung, Tübingen · http://znaturforsch.com

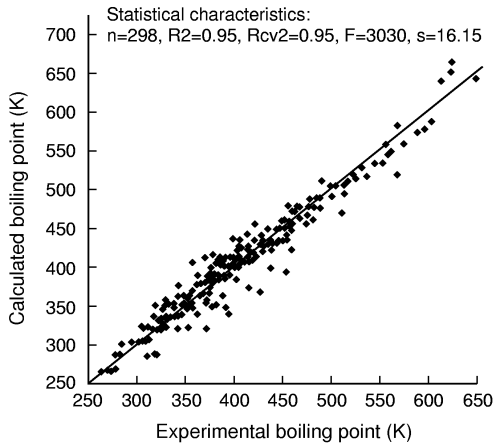

Fig. 1. Plot of experimental *vs.* calculated boiling points by the 2-parameter model. Descriptors: 1) cubic root of the gravitational index; 2) hydrogen donor charge surface area.

ware in the form of the CODESSA (COmprehensive Descriptors for Structure and Statistical Analysis) program, recently updated as the CODESSA PRO program [20].

For a satisfactory treatment, it is essential that good quality input data is utilized in the form of a set of structures and quantitative measurements of the property, measured under the similar conditions with satisfactory reproducibility and accuracy. The preparation of the input data utilizes a molecular editor or direct import of the structures from a chemical database. The 3D-geometries are generated and optimized using molecular mechanics and semiempirical quantum-chemical methods such as MM+, AM1 as in MOPAC [21] *etc*.

The next stage is the generation of descriptors. By default, the CODESSA PRO program enables 507 basic descriptors to be calculated for each structure. These descriptors have diverse molecular and atomic variations and hence, the total number of descriptors can reach many thousands. Definitions of these descriptors together with the original references are freely available on the CODESSA PRO homepage [20]. A search for the best set of molecular descriptors is based on various algorithms such as (i) the Heuristic method or (ii) the Best Multilinear Regression (BMLR). The selected parameters are then combined with the measured property values in a statistical analysis in an attempt to extract an equation which utilizes a small number of descriptors (usually not more than four or five) to correlate the measured values of the quantity satisfactorily. The higher the number of compounds employed in the training set, the higher is the acceptable number of descriptors. The descriptors involved in any proposed model should not be highly intercorrelated.

The efficiency of QSPR models for prediction is estimated using (i) *internal validation*, and (ii) *cross-validation* (Leave One Out) methods, correlation both for the full set and each training set. In these methods, all available data are used for both fitting and assessing. A new "ABC" method for cross validation, developed by our group, is based on a general "Leave-Group-Out" technique: here, the parent data set is divided into three parts denoted A, B, C; a QSPR equation is derived from each pair of these subsets and used to predict the third remaining set.

Illustrative applications of QSPR to simple physical properties

Boiling points

Boiling point was one of the first properties for which we derived QSPRs [22]. A training set of the boiling points at atmospheric pressure of 298 diverse compounds was fitted by a two-parameter equation (Fig. 1). The dataset includes saturated and unsaturated hydrocarbons, halogenated compounds, and hydroxyl, cyano, amino, ester, ether, carbonyl and carboxyl functionalities. The two descriptor straight-line equation has a high R^2 of 0.954 and is robust as shown by the statistically significant squared cross-correlated correlation coefficient of 0.953. Importantly, the two parameters selected by the descriptor forward selection procedure, the cubic root of the gravitation index and the hydrogen donor charged surface area, are physically well understandable. As its name suggests, the gravitation index describes the distribution of the mass of a molecule about its center of gravity and is connected 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. Therefore, our 2-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 attractive force between molecules is hydrogen bonding.

Melting points

The correlation and prediction of melting points is a far more difficult task. The melting point is defined

Table 1. Nine-parameter model of melting points. The squared correlation coefficient (R^2) and the Fischer criterion (F) relate to the models involving the descriptor on the given line and all descriptors above it.

Descriptor name	R^2	F
HDSA2 [Zefirov PC]	0.3829	273.63
Average valency oh atom H	0.5546	273.95
Total molecular surface area	0.7477	433.67
Average structural information content (1 order)	0.7504	329.17
Average information content (2 order)	0.8061	363.41
Maximum interaction of a C-H bond	0.8155	321.26
Average nuclephilic reactivity index	0.8256	294.27
BETA polarizability	0.8315	267.66
Symmetry number	0.8373	247.62

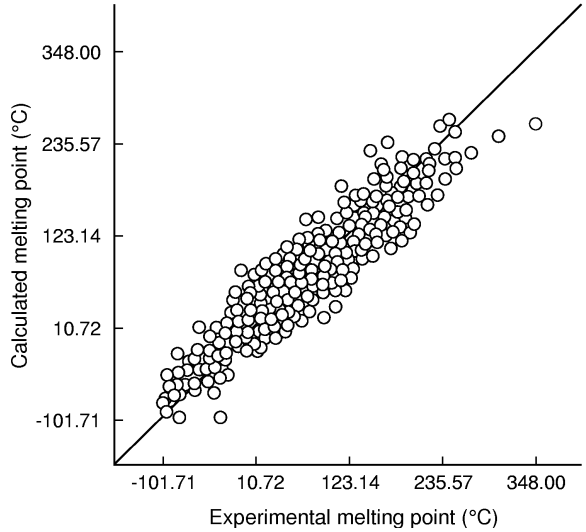

Fig. 2. Plot of experimental *vs.* calculated melting points by the 9-parameter model. $R^2 = 0.8373$, $F = 247.62$, $s = 30.19$, $n = 443$.

as the temperature of the transition between the solid and liquid phases. 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 much affected by the purity of a compound and experimental error. In such a situation, it is necessary to restrict the range of structures. A correlation equation, shown in Table 1 and Fig. 2, including nine descriptors $(R^2 = 0.8373)$ was developed for a large set of 443 melting points of substituted benzenes [23].

The melting point is a challenging property for physico-chemists to correlate, but also rewarding because as recently discussed [24], it can be used as a tool for prediction or modeling other properties. For more accurate correlation we need to be able to predict the crystal habit (or habits) in which a compound would crystallize and then estimate more exactly the interactive forces of attraction and repulsion existing in the crystal. The crystal lattice into which each molecule fits differs not only for each compound but also for each polymorph. The melting point of a crystal is governed by the hydrogen bonding ability of the molecules, 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. The correlation of the melting points of some ionic liquids has already received attention [25].

Refractive index

The QSPR models for refractive indices are much more tractable than those for melting points. A five-parameter correlation was developed for the refractive indices of 125 diverse liquid organic compounds $(R^2 = 0.945)$ [26]. The most important descriptor is the HOMO-LUMO energy gap, the energy difference between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). Both the refractive index and the HOMO-LUMO energy gap are related to the polarizability of a molecule. A small HOMO-LUMO energy difference usually means that the molecule is relatively easy to polarize.

Descriptors for polymeric molecules can not be calculated using the same techniques as for small molecules. However, for linear polymers above a certain chain length, it is possible to use the repeating unit to calculate appropriate descriptors; this was done for 95 amorphous homopolymers [27] of known refractive indices to afford a good correlation $(R^2 = 0.940)$. The majority of the polymers examined were homochain polymers (only carbon atoms in the main chain) or polyoxides, but several polyamides and polycarbonates were also included. The HOMO-LUMO energy gap is again the most important descriptor.

Viscosity

Viscosity is decisive for the transfer or movement of bulk quantities of liquids *e. g.* in the petroleum industries and in chemical engineering in general. A five-descriptor quantitative structure–property relationship (QSPR) model developed by us [28] for the liquid viscosity of 361 organic compounds containing C, H, N, O, S and/or halogens had a good squared correlation

coefficient (R^2) of 0.854 and a standard error (*s*) of 0.22 log units.

Density of organic liquids

The normal density (*i.e.* the density at 1 atm and 208 °C) is a main physicochemical property for the characterization and identification of a compound. Densities are used to predict or estimate other physical properties, such as critical pressure, viscosity, thermal conductivity, diffusion coefficients, and surface tension.

Successful one to four parameter correlation equations were developed for the densities of various subsets of organic compounds containing various heteroatoms [29], with standard errors ranging from 0.027 for hydrocarbons to 0.085 g/cm^3 for halogenated compounds. A general two-parameter correlation model was also developed for the prediction of the density of any organic compound containing C, H, O, N, S, and/or halogen atoms. This correlation model covers a large diversity of organic structures with a standard prediction error of 0.046 g/cm^3.

Dielectric constants

The static dielectric constant, also called the relative permittivity ε, a well defined molecular bulk property, is measured as the ratio of the capacitance of a condenser with the material as dielectric to its capacitance with vacuum as dielectric. Experimental data for many organic and inorganic compounds are available. We have developed multilinear regression and neural network QSPR models for the satisfactory prediction of both the dielectric constant (ε) and the Kirkwood function $(\varepsilon-1)/(2\varepsilon+1)$ of organic liquids [30]. The QSPR models were developed from a training set of 155 diverse compounds using theoretical molecular descriptors that encode the electronic properties of the molecule and the intermolecular interactions between molecules. The average prediction errors of the best models are 27% for the dielectric constant and 4.1% for the Kirkwood function.

Complex Physical Properties

GC retention indices

We correlated the retention times [31] of 152 structures incorporating a wide cross section of classes of organic compounds in a six-descriptor equation with $R^2 = 0.955$ and $R^2_{\mathrm{cv}} = 0.881$.

A specific application of GC retention times is the analysis of insect pheromones. Insects produce a great variety of methyl-branched alkanes as pheromones [32], but the structural variation is usually quite limited; most have a straight-chain backbone of 21 – 37 carbons, although it may extend to 51 carbons. Methyl branches appear at restricted locations on these backbones. Many insects produce monomethyl alkanes with the methyl branch located on carbon 2, 3, 7, 9, 11, 13, or 15. The next most commonly found series consists of dimethyl alkanes, in which the methyl branches are separated by a chain of 3, 7, 9, or 11 methylene (-CH_2-) groups; in these dimethyl derivatives the methyl branches are seldom separated by an even-number of carbons. The same pattern appears for the trimethyl alkanes, where three methyl branches separated by chains of three -CH_2- groups are again the most common. In tetramethyl alkanes, those with the four methyl branches each separated by three -CH_2- groups are the only types observed so far. The principal method used for the identification of these alkane-pheromones is gas chromatography (GC) and GC-mass spectrometry (GC-MS) [32].

A general QSPR model (squared correlation coefficient of 0.9585 and a standard error of 5.8) including mainly topological descriptors was obtained for 178 data points by our group for the GC retention indexes of methylalkanes produced by insects [33]. The error of the model was similar to the experimental error. The model was supported by (i) leave-one-out cross validation and (ii) division into three sets and prediction of each set from the other two. As a further test of the utility of the model, retention indices were successfully predicted for an external set of 30 methyl-branched hydrocarbons not involved in the deduction of the correction equation from the main data set. The average error was 4.6 overall and 4.3 for the 165 compounds remaining after leaving out small monomethyl alkanes. General trends of the structural variation of compounds in any given range of retention index were established by the analysis of the molecular descriptors appearing in the best QSPR model [34]. Topological descriptors were found to have high coding capabilities for the GC retention index and were selected to represent the chemical structures effectively.

Rat blood partition coefficient

The absorption, distribution and elimination (in animals and humans) of volatile organic compounds are

important in pharmacokinetics. Partition coefficients between air, water, blood and other liquids are important to explain these phenomena.

The partition coefficient, $PC_{A/B}$, for a given organic compound is defined as the ratio of concentrations achieved at equilibrium between the two different media as expressed mathematically in eq. (1), where A can be blood, saline, oil etc., and B is air.

$$PC_{A/B} = \frac{\textit{concentration in media A}}{\textit{concentration in media B}} \qquad (1)$$

QSPR treatment [35] of a data set of 100 diverse organic compounds related the logarithmic function of rat blood : air, saline : air and olive oil : air partition coefficients (denoted by $\log K(b:a)$, $\log K(s:a)$, and $\log K(o:a)$, respectively) to theoretical molecular and fragment descriptors had resulted in models with R^2 of 0.881, 0.926, and 0.922, respectively. The verification of the predictive power of these models on a test set of 33 organic chemicals not included in the training set gave satisfactory squared correlation coefficients: 0.791 for rat blood : air, 0.794 for saline : air and 0.846 for olive oil : air.

Cyclodextrin complexation energies

Applications of computational chemistry (including QSAR/QSPR) to the study of complexation with cyclodextrins (CD) have been well reviewed by Lipkowitz [36]. Our QSAR investigation jointly with the group of Varnek and Suzuki of the CD binding energies [37], used both multilinear CODESSA PRO and TRIAL fragment approaches. CODESSA-PRO modeled binding energies for 1 : 1 complexation systems between 218 organic guest molecules and α-cyclodextrin, with a seven-parameter equation with R^2 of 0.796 and R^2_{cv} of 0.779. Fragment-based TRAIL calculations (involving 79 fragment parameters) gave a better fit with R^2 of 0.943 and R^2_{cv} of 0.848 for 195 data points in the database. The study indicated that charge-related and topological descriptors connected to the branching of the molecules were the most important for the complexation binding energies.

Uranyl extractants

Again in collaboration with Professor Varnek, we developed a computer-aided design of new phosphoryl-containing podands, which efficiently extract the uranyl cation from water to an organic solvent [38]. This study was devoted to computer-aided design of new extractants of the uranyl cation involving three main steps: (i) a QSPR study, (ii) generation and screening of a virtual combinatorial library, and (iii) synthesis of several predicted compounds and their experimental extraction studies. First, QSPR modeling was performed of the distribution coefficient $(\log D)$ of structure of the cation extracted by phosphoryl-containing podands from water to 1,2-dichloroethane. Two different approaches were used for modeling purposes: one based on classical structural and physicochemical descriptors (implemented in the CODESSA PRO program) and another one based on fragment descriptors (implemented in the TRAIL program). Three statistically significant models obtained with TRAIL involved as descriptors either sequences of atoms and bonds or atoms with their close environment (augmented atoms). The best models of CODESSA PRO included its own molecular descriptors as well as fragment descriptors obtained with TRAIL. At the second step, a virtual combinatorial library of 2024 podands was generated with the CombiLib program, followed by the assessment of $\log D$ values using developed QSPR models. At the third step, eight of these hypothetical compounds were synthesized and tested experimentally. Comparison with experiment showed that the QSPR models developed successfully predicted $\log D$ values for 7 out of 8 compounds [39].

Biphasic partitioning

Aqueous biphasic systems (ABS) are formed by mixing two (or more) water-soluble polymers or adding a salt to an aqueous solution of a polymer above a certain critical concentrations or temperature. ABS are noteworthy because each of their two nonmiscible phases possesses different solvent properties although each is over 80% water on a molal basis. Due to their highly aqueous and hence mild nature, which is consonant with the maintenance of macromolecular structure, ABS have been employed for the separation of biological macromolecules for over 40 years [40].

ABS media are nonvolatile, nontoxic, and nonflammable and have recently found applications in many fields of science and technology, representing unique alternatives to traditional solvent based biphasic systems for the separation of metal ion species [41], small organic molecules [42], and lignins from cellulose in the paper and pulping process [43].

We have investigated [44] the partitioning of 29 small organic probes in a PEG-2000/$(NH_4)_2SO_4$

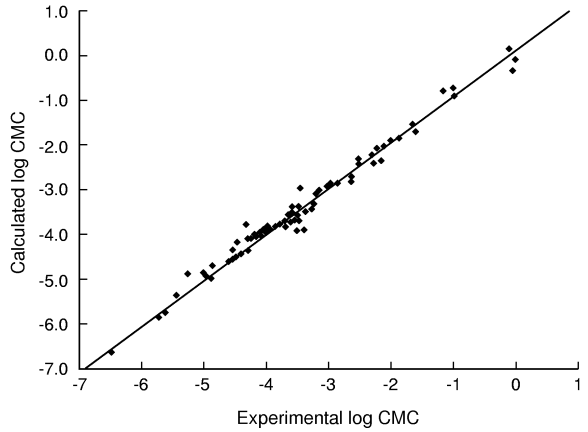

Fig. 3. Correlation of critical micelle concentration of non-ionic surfactants. Descriptors: 1) Kier and Hall index (0); 2) Average Information Content; 3) Relative number of N and D atoms. Comments: * The two fragments topologically express the bulk and branchiness of the tail. * The relative number of nitrogen and oxygen atoms represents the size of the hydrophilic fragment.

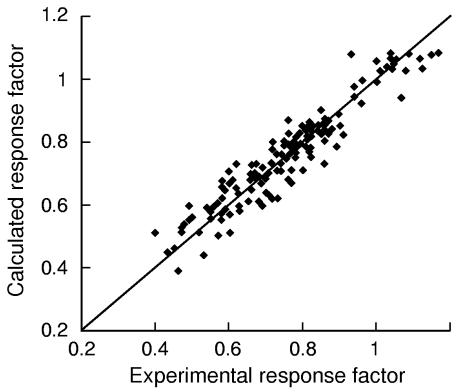

Fig. 4. Calculated *vs.* experimental values of response factors. Statistical characteristics: $n = 152$, $R^2 = 0.892$, $R^2_{cv} = 0.881$, $s = 0.054$. Descriptors: 1. relative weight of atoms; 2. total molecular center electron repulsion; 3. relative number of atom; 4. minimum total bond order of C atom; 5. minimum valency of a H; 6. total hybridization component molecular.

biphasic system by QSPR. A three-descriptor equation R^2 of 0.97 for the partition coefficient $(\log D)$ was obtained. Using the same descriptors derived solely from the chemical structure of the compounds, a three parameter model was also obtained for $\log P$ (octanol/water, $R^2 = 0.89$); predicted $\log P$ values were used as an external descriptor for modeling $\log D$.

Critical micelle concentrations

A hydrophobic tail and a hydrophilic head characterize surfactants. We have studied critical micelle concentrations of 77 non-ionic surfactants with a diversity of tails including straight chain, branched chain, aromatic and fluorinated [45]. The heads included polyethylene oxide, glycols, sugars and others to find a good correlation with just three descriptors (Fig. 3). The three empirically found fragment descriptors are physically very meaningful. The two fragment descriptors for the molecular tail are both topological, describing its length and branching. The third descriptor relates to the size of the hydrophilic head. A similar result was obtained with anionic surfactants [46].

GC response factors

We have correlated response factors for a gas flame ionization detector [31] as shown in Fig. 4. Here, we have introduced as descriptors the concept of "effective carbon atoms", from flame ionization detector theory. This has shown [31] that certain carbon atoms in a structure are more effective in producing pyrolysis products that conduct electricity better than others.

UV spectral intensities

In high throughput screening split-pool libraries, sub nanomole amounts of compound are synthesized, their structures are confirmed by LC/MS, and the LC/UV signal used to assess purity. It is difficult to assess these parameters quantitatively as weighing, NMR, ELSD (Evaporative Light Scattering Detector), CLND (Combustion-based Chemiluminiscent Nitrogen Detector) have insufficient sensitivity.

We have attempted [47] to predict response in typical HPLC UV detectors directly from structure. For a diverse set of 460 compounds, use of the sum of ZINDO [48] oscillator strengths in the integration range as an additional descriptor, produced a robust five-descriptor model with $R^2 = 0.857$.

Solubility

The phenomenon of solubility is of both fundamental importance and high practical interest. For 406 structurally diverse organic compounds we modeled environmentally important air: water partition coefficients, [49] (see Fig. 5) to obtain a five-descriptor equation with $R^2 = 0.939$ and $R^2_{cv} = 0.936$. The data set includes saturated and unsaturated hydrocarbons, halo-

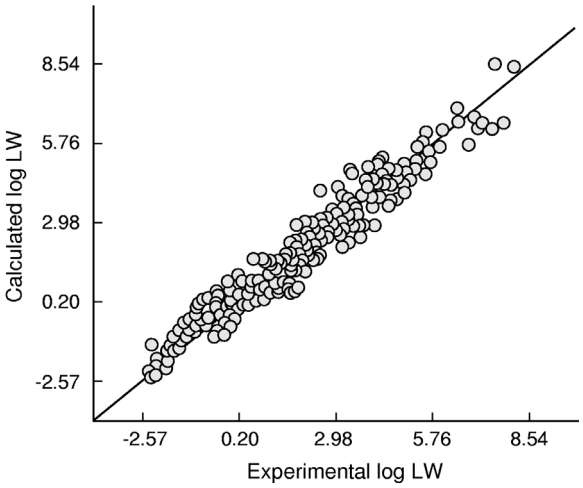

Fig. 5. Correlation of air : water partition coefficients for 406 organic compounds. $R^2 = 0.939$, $R^2_{cv} = 0.936$, $F = 1232$ (five descriptors).

Descriptor name	R^2	R^2_{cv}
HDCA(2)	0.522	0.518
nO + 2nN	0.684	0.68
$E_{HOMO} - E_{LUMO}$	0.879	0.876
PCWTE	0.916	0.913
N_{rings}	0.941	0.936

genated compounds, and compounds containing hydroxyl, cyano, amino, nitro, thio, ester, ether, carbonyl, and carboxyl functional groups plus furan, pyran, pyridine, and pyrazine rings.

Analogous work has been carried out for other systems, for example, the solubility of gases and vapors in ethanol [50].

More recently, we have been involved in a general treatment of solubility. A carefully chosen overall data set included 154 solvents and 397 solutes. In the results [51–53], the number of solutes used for each solvent varied from 226 for hexadecane to 2 for several solvents. We have reported our results [51] on the modeling of 69 of these solvents, which ranged from 14 to 226 solutes, with the average number of solutes per solvent being 48. The remaining 72 solvents have less than 14 available experimental points. We also analyzed and reported the modeling of 80 selected solutes [52] by choosing only those that had reliable solubility data for at least 15 solvents. The correlations were examined by holding the solute constant and varying the solvent and vice versa. Good quality statistical correlations were obtained which should ultimately enable us to construct a large matrix and fill in the missing points (work in progress). A principal component analysis should give considerable insight into the whole phenomenon of solubility, is being carried out.

Soil sorption coefficients

The soil sorption potential of chemicals is an important parameter in environmental risk assessment procedures to estimate the persistence and mobility of chemicals. This refers to assessments of the bioavailability of chemicals for both soil- and water-living organisms. Soil sorption is most often expressed as a coefficient defined as the concentration of the chemical in soil divided by the concentration in the aqueous phase.

Diverse chemical descriptors were explored for use in QSPR models aimed to screen the soil sorption potential of 351 organic compounds [54]. These compounds were divided into 11 groups. Five general models were developed with R^2 ranging from 0.34–0.99. Another recent investigation [55] of the soil sorption by us resulted in general and class-specific QSPR models for the soil sorption, $\log K_{OC}$, of 344 organic pollutants $(0 < \log K_{OC} < 4.94)$ using a large variety of theoretical molecular descriptors based only on molecular structure. Two general models were obtained. The first, two-parameter model was derived for a structurally representative set of 68 chemicals had ($R^2 = 0.76$ and $s = 0.44$). The second, four-parameter model was based on data for 344 compounds ($R^2 = 0.76$, $s = 0.41$). The first model was validated using the data for the remaining 276 pollutants ($R^2 = 0.70$, $s = 0.45$). An additional validation of both models was performed using an independent set of 48 pollutants. Both models predict the $\log K_{OC}$ at the level of experimental precision, while the theoretical molecular descriptors appearing in the QSPR models give further insight into the mechanisms of soil sorption. The analysis of the distribution of the residuals of the $\log K_{OC}$ values calculated by both general models indicated the need and possible advantages of modeling soil sorption for smaller data sets related to individual classes of chemicals.

Chemical Properties

Lithium-cation basicities

The reactivity of metal ions toward ligands is usually quite straightforward: in general, they form adducts or clusters, which can be considered as ions "solvated" by one or several ligands. The basicity of

ligands is very important for this process of cluster formation. We investigated gas-phase lithium-cation basicity for 205 diverse compounds [56] to obtain a six-parameter general QSPR model with good statistical characteristics: $R^2 = 0.80$, $R^2_{cv} = 0.79$, $s^2 = 8.78$. In addition, the theoretical descriptors, such as minimum net atomic charge, highest occupied molecular orbital energy, total point-charge component of the molecular dipole, etc., logically explain the reaction equilibrium and electrostatic interaction between the lithium cation and a base.

Decarboxylation rates

Reaction rates depend greatly on the nature of the solvents employed. We have used QSPR to investigate the decarboxylation rates of 6-nitrobenzisoxazole-3-carboxylic acid [57] in 24 solvents. The results of multilinear correlation with theoretical molecular descriptors demonstrated that CODESSA can produce a good QSPR model even for a relatively small number of data points.

Chain transfer constants

Kinetic chain-transfer constants play an important role in polymer chemistry since an understanding of chain transfer clarifies the microkinetic processes in polymerization reactions. QSPR treatments of the respective reaction parameters may have great potential from both practical and theoretical standpoints [58]. The Quantitative Structure-Reactivity Relationships (QSRR) developed by us for kinetic chain-transfer constants for 90 agents in styrene polymerization at 60 °C produced three- and five-parameter correlations with R^2 of 0.725 and 0.818, respectively. The descriptors involved in the correlations were consistent with the proposed mechanism of the chain-transfer reactions. In other words, the descriptors in the models explained the mechanism of reaction at different stages. Descriptors such as LUMO, HDCA-1, and Kier and Hall topological indices are important in this QSRR model.

Flash points

The flash point, the temperature at which the mixture of a vapor and air spontaneously ignites, is of obvious importance in many connections. We have derived QSPRs for flash points [59] using the available experimental data for 271 various organic compounds ($R^2 = 0.924$). Flash points correlate well with boiling points, the first parameter involved in the model is the boiling point predicted by our previously derived equation [60]. This enables the model to be used to predict flash points of compounds for which no measured boiling point is available. The other two parameters involved in the regression equation developed for 271 diverse organic compounds were the relative negative charge and the hydrogen acceptor surface area in the molecule.

Gas–phase homolysis

Simple bond fission eq. (2) is one of the simplest elementary chemical reactions that plays an important role in solving fundamental kinetic problems in chemistry.

$$\mathrm{R}i\mathrm{R}j \longrightarrow \mathrm{R}i^{*} + \mathrm{R}j^{*} \qquad (2)$$

Knowledge of the kinetic parameters of reactions of type (3) also provides benefits for the design of industrially and environmentally important processes. In particular, the modeling of the pyrolysis and combustion processes requires the knowledge of reliable values of kinetic parameters of many elementary gas-phase homolytic reactions. A successful 5-parameter QSPR model was derived by us for the rate constants of the gas phase pyrolysis of C-X chemical bonds [61].

An extension of this work improved the chemical picture behind the homolysis phenomenon [62]. The kinetic parameters of the gas-phase homolysis for 58 different C-CH_3 bonds were treated using the CODESSA program. The resulting six-parameter models were developed for prediction of $\log k$ (1047 K) and the parameters of the Arrhenius equation, $\log A$ and E.

Rotational activation energies for amides

A novel approach to predict the gas-phase rotational activation energies of amides was presented by our group using the CODESSA program [63] for a QSPR treatment gave a three-parameter equation with $R^2 = 0.982$ for the free energies of activation for the amide bond rotation for a set of 24 N,N-dialkylamides. The descriptors that appeared in this model were explained from the chemical point of view by taking into account the nature of the compounds.

Biological properties

Toxicity of aqueous pollutants

Quantitative structure-toxicity relationships were developed by our group for the prediction of aque-

ous toxicities for *Poecilia reticulata* (guppy) using the CODESSA treatment [64]. Experimental LC_{50} (and $\log P$) values from the literature for 293 compounds were divided into four groups (according to the functional group), for each of which a QSAR model was obtained. A two-parameter correlation with $R^2 = 0.96$ was found for class 1 toxins. The five-parameter correlations were derived with $R^2 = 0.92$ for class 2 toxins and with $R^2 = 0.85$ for class 3 toxins, respectively. The correlations for class 4 toxins had $R^2 = 0.85$. Again, all the descriptors utilized were calculated solely from the structure of the molecules, which made it possible to predict the LC_{50} values for unavailable or unknown toxins. Our results [64] were generally consistent with the results obtained by others [65].

Nitrobenzene toxicities

Nitroaromatics are hazardous chemicals that display several manifestations of toxicity, including skin sensitization, immunotoxicity, germ cell degeneration, inhibition of liver enzymes and also a conjectured carcinogenicity.

We have developed a five-parameter QSAR correlation [$R^2 = 0.723$, $R^2_{cv} = 0.676$, in terms of $\log(IGC_{50})^{-1}$] based on CODESSA-PRO methodology for the aquatic toxicity of 97 substituted nitrobenzenes to the ciliate *Tetrahymena pyriformis*. The results support previous conclusions that hydrophobicity and electrophilic reactivity control nitrobenzene toxicity [66]. Correction of the data for the ionization of acidic species (picric and nitrobenzoic acids) improved the results to $R^2 = 0.813$, $R^2_{cv} = 0.787$. Consideration of the results for a total set of 97 compounds suggested two mechanisms of toxic action. A subset containing 43 compounds favorably disposed to reversible reduction of nitro group with respect to the single occupied molecular orbital energy, ESOMO correlated well with just four theoretically derived descriptors: $R^2 = 0.915$, $R^2_{cv} = 0.890$. Another set of 49 substances predisposed to aromatic nucleophilic substitution modeled well ($R^2 = 0.915$, $R^2_{cv} = 0.888$) with five descriptors.

Oxazolidinone antibacterials

The increase during the last decade of bacterial resistance to antibiotics poses a serious concern for medicine. Oxazolidinones, a new class of synthetic antibacterials with activity against gram-positive pathogenic and anaerobic bacteria, bind selectively to the 50S ribosomal subunit and inhibit bacterial translation at the initiation phase of protein synthesis.

Few QSAR/QSPR investigations have been published in this area of biochemistry. We have established that the minimum inhibitory concentrations (MIC) required to inhibit the growth of *S. aureus* for 60 3-aryloxazolidin-2-one antibacterials [67], successfully correlate with theoretical molecular and fragment descriptors. The use of CODESSA PRO [20] descriptors led to a significant seven-parameter model with $R^2 = 0.820$ and $R_{cv} = 0.758$. Our results demonstrate that in characterizing a complex biological property (as antibacterial activity) by multilinear QSPR equation, the descriptors related to the whole molecule can provide superior model versus that obtained using just fragmental descriptors. However, the best model utilized both fragmental and molecular descriptors.

PDGF receptor activities

Insights into the biology of tumor angiogenesis have led to the identification of various molecules that promote tumor development. Of particular interest are such factors as the platelet-derived growth factor (PDGF), which plays a major role as a regulator of cell growth. We have investigated the QSAR models describing the activity of 1-phenylbenzimidazoles as promising selective inhibitors of PDGF [68]. Two approaches were applied to 123 reported activity of $\log (IC_{50})^{-1}$ for the 1-phenylbenzimidazoles; (i) linear (multilinear regression) and (ii) nonlinear (artificial neural network). The results obtained in this work indicate that the regression and ANN models exhibit significant prediction capabilities. The linear model was developed mainly for the purpose of structure-activity interpretation, whereas the ANN model was primarily developed for predictive ability and classification.

Human milk to plasma concentration ratios

The milk to plasma concentration ratio (M/P ratio) of a drug estimates an infant's exposure to drugs through breast milk. The M/P ratio is an attempt to quantify the equilibrium concentration between breast milk and blood. It is defined as the ratio of the drug concentration in the breast milk (C_{BM}) and its concentration in the maternal plasma (C_{MP}) eq. (3).

$$M/P = C_{BM}/C_{MP}. \tag{3}$$

Recently, a satisfactory model has been developed using CODESSA PRO for the correlation and predic-

tion of M/P ratios for diverse pharmaceuticals [69]. The experimentally derived M/P ratio values for 100, widely used pharmaceuticals gave a seven-parameter QSAR model with $R^2 = 0.791$.

Activity of receptor antagonists

For rationalization of biological responses (binding affinity, selectivity, and efficacy) of ligands for G-protein coupled receptors, the large molecular diversity of the receptors and their subtypes, is crucial.

In our study [70], theoretical descriptors derived by means of the program CODESSA and *ad hoc* defined size and shape descriptors have been employed for deciphering, quantitatively the molecular features responsible for affinity and selectivity in a series of potent N^4-substituted arylpiperazines antagonists acting at postsynaptic 5-HT1A receptor. We obtained 10 QSAR models or the binding affinity with correlation coefficients between 0.70 and 0.83. The theoretical descriptors involved in the selected QSAR models can be classified as: (a) ad hoc size and shape descriptors defined with respect to a super-molecule of high affinity ligands, and (b) descriptors derived on a single structure, *i. e.* molecular orbital indexes and charged partial surface area descriptors. The QSAR models were also developed for the wide series of structurally diverse α_1-adrenergic receptor antagonists [71].

Genotoxicities

The carcinogenicity and mutagenicity of small organic molecules are important chronic toxicity manifestations. They are closely related: some 90% of carcinogenic compounds are known or potential mutagens. While the experimental assessment of carcinogenicity is complex and time consuming, several tests allow easy detection of mutagenicity. A quantitative structure-activity relationship with $R^2 = 0.834$, was derived by our group for a set of 95 heteroaromatic and aromatic amines to correlate and predict their mutagenic activity [72]. It consists of six descriptors calculated from the molecular structures with quantum chemical methods. The descriptors in the model reveal the importance in mutagenic interactions of heteroaromatic amines of hydrogen bonding, of effects induced by the solvent, and of the size of compound. The model also suggests that the amino group is critical for the reactive site. Later the investigation was continued applying the nonlinear approaches of Chebyshev polynomial expansion and neural networks [73]. The dependence of molecular descriptors in these models on environmental effects and molecular conformations were analyzed and models were found superior to the linear QSAR treatment.

Technological Properties

Rubber vulcanization rates

Many heterocyclic disulfides, sulfenamides, or sulfenimides are accelerators used for the vulcanization of rubber [74]. In the vulcanization process, there should be a delay before the onset of cross-linking; after this delay the vulcanization should proceed rapidly and irreversibly. Compounds are tested as accelerators for this process by constructing a "rubber rheometer curve" [74] in a machine that measures the change in the torque ("stiffness" of the rubber undergoing vulcanization). It is also important that the torque does not immediately increase but for a certain period change is delayed enough to formulate the tire or other article but then proceeds rapidly and irreversibly to a maximum hardness. We have carried out the QSPR treatment for 23 compounds that have been measured for their potential as accelerators. Together with colleagues at the Flexsys Company [74], we investigated the possibility of using CODESSA to correlate the structure of accelerators with (i) the time to scorch, *ts2*, and (ii) the maximum rate of vulcanization, *mxr*. Modeling was done on both the parent molecular accelerator (12 sulfenamides, 11 sulfenimides) and also on a zinc complex of the accelerator with thiolate fragments. The correlation coefficients of the QSPR models range from 0.925 to 0.967 and show that the developed equations are statistically satisfactory and useful.

Concluding Remarks

Quantitative structure-activity/property relationship (QSAR/QSPR) techniques have become indispensable in all aspects of research into the molecular interpretation of physical, chemical, biological and technological properties. Today it would be inconceivable for any commercial, governmental, or academic group to research in these fields without the help of sophisticated calculations. The results reviewed in this paper witness the applicability and power of the QSAR and QSPR approaches for the prediction of very diverse properties of chemical compounds and materials. This has become possible 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 (QSARs)

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 the properties for molecular structures that are difficult to measure experimentally or even for those not yet synthesized.

Acknowledgements

The Estonian Science Foundation grant No. 4548 and the W. R. Kenan, Jr. Trust are acknowledged for the partial support of this work. We thank Dr. D. Fara for his help.

[1] Y. C. Martin, Perspect. Drug Discov. **12**, 3 (1998).
[2] U. Norinder, Perspect. Drug Discov. **12**, 25 (1998).
[3] D. J. Maddalena, Expert Opin. Ther. Pat. **8**, 249 (1998).
[4] H. Kubinyi, Drug Discov. Today **2**, 538 (1997).
[5] C. Hansch, T. Fujita, in C. Hansch and T. Fujita (eds): Classical and Three-Dimensional QSAR in Agrochemistry, p. 1, ACS, Washington (1995).
[6] C. Hansch, A. Leo, Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, ACS, Washington (1995).
[7] A. R. Katritzky, D. C. Fara, R. O. Petrukhin, D. B. Tatham, U. Maran, A. Lomaka, M. Karelson, Curr. Top. Med. Chem. **2**, 1333 (2002).
[8] M. Karelson, Molecular Descriptors in QSAR/QSPR, John Wiley & Sons, New York (2000).
[9] R. Todeschini, V. Consonni, Handbook of Molecular Desciptors, Wiley-VCH, Weinheim (2000).
[10] M. H. Abraham, in P. Politzer, J. S. Murray (eds): Quantitative Treatments of Solute/Solvent Interactions, p. 83, Elsevier, Amsterdam (1994).
[11] A. T. Balaban, J. Chem. Inf. Comput. Sci. **37**, 645 (1997).
[12] J. C. Dearden, Trends QSAR Mol. Modell. **92**, 163 (1993).
[13] S. H. Hilal, L. A. Carreira, S. W. Karickhoff, in P. Politzer, J. S. Murray (eds): Quantitative Treatments of Solute/Solvent Interactions, p. 291, Elsevier, Amsterdam (1994).
[14] A. J. Stuper, W. E. Brugger, P. C. Jurs, Computer-assisted Studies of Chemical Structure and Biological Function, While, New York (1979).
[15] L. B. Kier, L. H. Hall, Molecular Connectivity in Structure-Activity Analysis, John Wiley & Sons, New York (1986).
[16] J. S. Murray, P. Politzer, in J. S. Murray, P. Politzer (eds): Quantitative Treatments of Solute/Solvent Interactions, p. 243, Elsevier, Amsterdam (1994).
[17] M. Randić, M. Razinger, in A. T. Balaban (ed.): From Chemical Topology to Three-Dimensional Geometry, p. 159, Plenum Press, New York (1996).
[18] B. Lucic, N. Trinajstic, J. Chem. Inf. Comput. Sci. **39**, 121 (1999).
[19] A. R. Katritzky, V. S. Lobanov, M. Karelson, Chem. Rev. **96**, 1027 (1996).
[20] www.codessa-pro.com
[21] J. J. P. Stewart, MOPAC 7.0, QCPE # 455, http://qcpe.chem.indiana.edu/
[22] A. R. Katritzky, L. Mu, V. S. Lobanov, M. Karelson, J. Phys. Chem. **100**, 10400 (1996).
[23] A. R. Katritzky, U. Maran, M. Karelson. V. S. Lobanov, J. Chem. Inf. Comput. Sci. **37**, 913 (1997).
[24] A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, U. Maran, M. Karelson, Crystal Growth Design. **1**, 261 (2001).
[25] A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, M. Karelson, A. E. Visser, R. D. Rogers, J. Chem. Inf. Comput. Sci. **42**, 225 (2002).
[26] A. R. Katritzky, S. Sild, M. Karelson, J. Chem. Inf. Comput. Sci. **38**, 840 (1998).
[27] A. R. Katritzky, S. Sild, M. Karelson, J. Chem. Inf. Comput. Sci. **38**, 1171 (1998).
[28] A. R. Katritzky, K. Chen, Y. Wang, M. Karelson, B. Lucic, N. Trinajstic, T. Suzuki, G. Schüürmann, J. Phys. Org. Chem. **13**, 80 (2000).
[29] M. Karelson, A. Perkson, Comp. & Chem. **23**, 49 (1999).
[30] S. Sild, M. Karelson, J. Comp. Inf. Chem. Sci. **42**, 360 (2002).
[31] A. R. Katritzky, E. S. Ignatchenko, R. A. Barcock, V. S. Lobanov, M. Karelson, Anal. Chem. **66**, 1799 (1994).
[32] D. A. Carlson, U. R. Bernier, B. D. Sutton, J. Chem. Ecol. **24**, 1845 (1998).
[33] A. R. Katritzky, K. Chen, U. Maran, D. Carlson, Anal. Chem. **72**, 101 (2000).
[34] A. R. Katritzky, K. Chen, U. Maran, D. Carlson, Anal. Chem. **72**, 101 (2000).
[35] A. R. Katritzky, M. Kuanar, D. C. Fara, M. Karelson, W. E. Acree (Jr.), Bioorg. Med. Chem. **12**, 4735 (2004).
[36] K. B. Lipkowitz, Chem. Rev. **98**, 1829 (1998).
[37] A. R. Katritzky, D. C. Fara, H. Yang, M. Karelson, T. Suzuki, V. P. Solov'ev, A. Varnek, J. Chem. Inf. Comput. Sci. **44**, 529 (2004).
[38] A. Varnek, D. Fourches, V. P. Solov'ev, V. E. Baulin, A. N. Turanov, V. K. Karandashev, D. Fara, A. R. Katritzky, J. Chem. Inf. Comput. Sci. **44**, 1365 (2004).
[39] A. Varnek, D. Fourches, V. P. Solov'ev, V. E. Baulin, A. N. Turanov, V. K. Karandashev, D. Fara, A. R. Katritzky, J. Chem. Inf. Comput. Sci. **44**, 1365 (2004).
[40] P.-A. Albertsson, Partition of Proteins in Liquid

Polymer-Polymer Two-Phase Systems, Nature **182**, 709 (1958).
[41] R. D. Rogers, A. H. Bond, C. B. Bauer, Sep. Sci. Technol. **28**, 139 (1993).
[42] R. D. Rogers, H. D. Willauer, S. T. Griffin, J. G. Huddleston, J. Chromatogr. B. **711**, 255 (1998).
[43] M. Li, H. D. Willauer, J. G. Huddleston, R. D. Rogers, Sep. Sci. Technol. **36**, 835 (2001).
[44] A. R. Katritzky, K. Tamm, M. Kuanar, D. C. Fara, A. Oliferenko, P. Oliferenko, J. G. Huddleston, R. D. Rogers, J. Chem. Inf. Comput. Sci. **44**, 136 (2004).
[45] P. D. T. Huibers, V. S. Lobanov, A. R. Katritzky, D. O. Shah, M. Karelson, Langmuir **12**, 1462 (1996).
[46] P. D. T. Huibers, V. S. Lobanov, A. R. Katritzky, D. O. Shah, M. Karelson, J. Colloid Interface Sci. **187**, 113 (1997).
[47] L. F. William, M. McGregor, A. R. Katritzky, A. Lomaka, R. Petrukhin, M. Karelson, J. Chem. Inf. Comput. Sci. **42**, 830 (2002).
[48] ZINDO, version 99.1, M. C. Zerner, J. E. Ridley, A. D. Bacon, W. D. Edwards, J. D. Head, J. McKelvey, J. C. Culberson, P. Knappe, M. G. Cory, B. Weiner, J. D. Baker, W. A. Parkinson, D. Kannis, J. Yu, N. Roesch, M. Kotzian, T. Tamm, M. Karelson, X. Zheng, G. Pearl, A. Broo, K. Albert, J. M. Cullen, C. J. Cramer, D. G. Truhlar, J. Li, G. D. Hawkins, D. A. Liotard, QTP, Gainesville, July-Sept. (1999).
[49] A. R. Katritzky, L. Mu, M. Karelson, J. Chem. Inf. Comput. Sci. **36**, 1162 (1996).
[50] A. R. Katritzky, D. B. Tatham, U. Maran, J. Chem. Inf. Comput. Sci. **41**, 358 (2001).
[51] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, R. Petrukhin, D. B. Tatham, U. Maran, A. Lomaka, W. E. Acree (Jr.), J. Chem. Inf. Comput. Sci. **43**, 1794 (2003).
[52] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, R. Petrukhin, D. Tatham, U. Maran, A. Lomaka, W. E. Acree (Jr.), J. Chem. Inf. Comput. Sci. **43**, 1806 (2003).
[53] A. R. Katritzky, I. Tulp, D. C. Fara, A. Lauria, U. Maran, A. E. Acree (Jr.), J. Chem. Inf. Mod. **45**, 913 (2005).
[54] P. L. Anderson, U. Maran, D. Fara, M. Karelson, J. L. M. Hermens, J. Chem. Inf. Comput. Sci. **42**, 1450 (2002).
[55] I. Kahn, D. Fara, M. Karelson, U. Maran, P. L. Andersson, J. Chem. Inf. Mod. **45**, 94 (2005).
[56] K. Tamm, D. Fara, A. R. Katritzky, P. Burk, M. Karelson, J. Phys. Chem. **108**, 4812 (2004).
[57] A. R. Katritzky, S. Perumal, R. Petrukhin, J. Org. Chem. **66**, 4036 (2001).
[58] F. Ignatz-Hoover, R. Petrukhin, M. Karelson, A. R. Katritzky, J. Chem. Inf. Comput. Sci. **41**, 295 (2001).
[59] A. R. Katritzky, R. Petrukhin, M. Karelson, J. Chem. Inf. Comput. Sci. **41**, 1521 (2001).
[60] A. R. Katritzky, V. S. Lobanov, M. Karelson, J. Chem. Inf. Comput. Sci. **38**, 28 (1998).
[61] R. Hiob, M. Karelson, J. Chem. Inf. Comput. Sci. **40**, 1062 (2000).
[62] R. Hiob, M. Karelson, Comp. & Chem. **26**, 237 (2002).
[63] J. Leis, M. Karelson, Comp. & Chem. **25**,171 (2001).
[64] A. R. Katritzky, D. Tatham, U. Maran, J. Chem. Inf. Comput. Chem. **41**, 1162 (2001).
[65] J. L. M. Hermens, in O. Hutzinger (ed.): The Handbook of Environmental Chemistry, Vol. 2, Part E, p. 111, Springer-Verlag, Berlin (1989).
[66] A. R. Katritzky, P. Oliferenko, A. Oliferenko, A. Lomaka, M. Karelson, J. Phys. Org. Chem. **16**, 811 (2003).
[67] A. R. Katritzky, D. Fara, M. Karelson, Bioorg. Med. Chem. **12**, 3027 (2004).
[68] A. R. Katritzky, D. Dobchev, D. Fara, M. Karelson, Bioorg. Med. Chem. **13**, 6598 (2005).
[69] A. R. Katritzky, D. Dobchev, D. C. Fara, M. Karelson, Bioorg. Med. Chem. **13**, 1623 (2005).
[70] M. C. Menziani, P. G. De Benedetti, M. Karelson, Bioorg. Med. Chem. **6**, 535 (1998).
[71] M. C. Menziani, P. G. De Benedetti, M. Karelson, Bioorg. Med. Chem. **7**, 2437 (1999).
[72] U. Maran, M. Karelson, A. R. Katritzky, QSAR **18**, 3 (1999).
[73] M. Karelson, S. Sild, U. Maran, Mol. Simulat. **24**, 229 (2000).
[74] F. Ignatz-Hoover, A. R. Katritzky, V. S. Lobanov, M. Karelson, Rubber Chem. Technol. **72**, 318 (1999).