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

* Introduction and scope: - the manifestations of molecular structure fall into one of two major classes: (i) the influence of a specific portion of the molecule, (ii) the influence of the whole molecule.
* The effects of structural variation in a molecule are distinct in 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.
* But many quantitative structure activity relationships (QSAR) are strongly linked to specific regions of molecules, and thus into class (i).
* Overview of QSPR Approaches: - the beginning of QSPR dates back to more than a century. In 1884 Mills developed a QSPR for predicting the melting points and boiling points of homologous series.
* 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, and between narcosis and chain length.
* The first theoretical whole molecule descriptor, the Wiener index and Platt number, were proposed in 1947 to model the boiling points of hydrocarbons.
* QSPR or quantitative structure related analysis of physicochemical properties before 1970 had major applications only in analytical chemistry.
* 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 event unknown.
* QSPR became more attractive for chemists with development of new software tools, but there are still tremendous opportunities for developments in the application of purely structure-based molecular description in QSAR models and in the application of quantitative property-activity relationship (QPAR) through the use of physicochemical properties predicted with QSPR.
* The QSPR approach has been applied in many different areas, like (i) properties of single molecules; (ii) interactions between different molecular species; (iii) surfactant properties; (iv) complex properties and properties of polymers.
* Multi-linear (MLR) methods are now accompanied by computational neural networks (NN) that have been utilized to described non-linear relationships between structure and property.
* Available Software: - commercially available statistical software packages such as SAS, SPSS family, STATISTICA, UNISTAT, STATAGRAPHICS etc. 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; (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.
* The problems may be solved on the level of standards for binary compatibility supported by many statistical packages.
* 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 that can be used for QSPR tasks; Pharamacopeia, 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.
* CambridgeSoft has developed the ChemSAR COM plug-in for their Chem3D molecular 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.
* Many QSPR/QSAR packages were developed in the academic environment. ADAPT (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.
* CODESSA PRO (COmprehensive Descriptors for Structural and Statistical Analysis PROfessional) was developed in Center for Heterocyclic Chemical, University of Florida, U.S.A., and the Institute of Chemistry 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 calculated 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 models, and to predict the properties for compounds previously unknown or unavailable, and to test the model extensively.
* A design tool with a focus on analyzing QSARs, Tsar, distributed by Accelrys, has been widely used throughout drug discovery, from initial compound selection for primary screening to reagent selection to reagent selection and creation of focused libraries for lead optimization.
* Multidimensional QSPR/QSAR: - two most relevant QSPR/QSAR opinion is: (i) multidimensional QSPR/QSAR models (2D-, 3D-, and recently 4D- QSPR/QSAR), and (ii) the multivariate statistical analysis of specific variable of the different phenomenon.
* 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.
* 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.
* 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, and (iii) the so-called interaction pharamacophore.
* 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.
* 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.
* 3D-QSPR formalism has been applied by Burke et al. to an analog series of pyridobenzodiazepinone inhibitors of muscarinic 2 and 3 receptors.
* 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.
* 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 PO3H2 group hydrogen bonding to the surface, and (iii) a nonpolar region of space favorable to inhibition potency.
* 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 descriptors are the main activity correlates.
* The advantage of using multivariate statistical analysis to provide insight into how these variables interrelate quantitatively has been confirmed by many studies.
* PCA has been used frequently in QSAR studies, to extract uncorrelated and useful information from independent variable.
* 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, (iv) to classify diverse sets of toxic compounds into subsets by MOA.
* The application of PCA for data reduction has provided insight into (a) the concept of solvent polarity scales, and should polarity scales, 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.
* The researchers 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 extracted three PCs explaining 75% of the variance.
* 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.
* Several groups have estimated retention indices using various descriptors: topological, charged partial surface area, and quantum-chemical descriptors for a large variety of compounds; substituted pyrazines, polycyclic aromatics, stimulants and narcotics, and anabolic steroids.
* Different measures of toxicity have been used depending on species, concentration, mode of action, and duration.
* Various measures of toxicity for different endpoints can be analyzed independently in terms of simple QSAR models combined with pre-selection of descriptors.
* There is an obvious interrelation between multidimensional QSPR/QSAR models and multivariate statistical analysis of variables.
* Solubility: - it is very important in several areas related to medicinal chemistry and also to the properties/activities, correct estimates of solubility are required for understanding the environment fate of possible pollutants and how easily compounds enter into the environment and thereafter into the living organisms.
* Solubility is also crucial in determining the bioavailability and thus the effectiveness and bio-degradation of pharmaceuticals.
* 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 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.
* 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.
* Various descriptors selection methods have been developed, including stepwise forward selection (SFS) procedures, generic algorithms (GA) and simulated annealing (SA) routines in conjunction with MLR and NN.
* Yalkowsky and Banerjee classified the different approaches for the prediction of solubility into three categories: (i) correlations (ii) correlations based on group contributions; (iii) correlations with parameters calculated solely from the molecular structure.
* Into the first category, one can also add correlations using descriptions based on empirical measurements.
* 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.
* But the 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. Therefore, it is difficult to elucidate how molecular structure affects the observed property.
* In the second category, group contribution methods have also gained much attention in prediction of solubility. This approach provides less understanding of the physical nature of the relationship between the molecular structure and solubility process itself.
* Recently the group contribution methods were evaluated for their ability to predict water solubility.
* The third category comprises 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).
* These descriptors explicitly involve structural properties of the compounds, and more importantly, they can be calculated for any structure.
* Recent rapid enhancements in computers and semi-empirical quantum chemical programs have encouraged the application of various quantum chemical descriptors in QSPR analysis.
* A rapidly growing number and variety of descriptors makes crucial the selection of the descriptors for the final solubility’s models. It leads to the application of techniques for efficient descriptors selection.
* The researchers 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.
* Based on their model the researchers concluded that the solubility of gases and vapors depends on 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.
* In other/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.
* The structural-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.
* The structural-based whole molecule descriptors in QSPR models for both solvents led to conclusion that descriptors cover solute-solvent interactions like polarizability, dipole-dipole interactions, hydrogen bonding, and lipophilicity.
* 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.
* 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.
* Bioavailability: - the definition of bioavailability depends on the field of study. The pharmacological bioavailability is the most intensively studied in the QSAR/QSPR literature.
* It estimates the relative fraction of the orally administered does that is absorbed into the systemic circulation when compared to the data measured for a solution, suspension, or intravenous dosage form.
* The pharmacological bioavailability reflects not only the characteristics of a chemical and its environmental specification, but also the behavior and physiology of the organism.
* The bioavailability studies also provide useful pharmacokinetic information related to the distribution, elimination, the effects of nutrients on absorption of the drug, and dose proportionality.
* 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 may be in *vitro* assays.
* 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 *in vitro* nature of the screening techniques provides no bias towards properties with favorable oral activity.
* It takes more time 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 permission measurements.
* Lipinski et al. discuss that poor solubility and permeability as causes of low bioavailability.
* The first obvious choice of such properties is molecular weight 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).
* Hydrogen donor ability can be expressed in terms of the solvatochromic parameter of a donor group. Various researchers have compiled experimental values of the parameter.
* However, it has found that a simple sum of the number of NH and OH bonds can also perform well.
* The above considerations, together with the analysis of a compound library with favorable physiochemical properties, led to the formulations of the “rule of 5”, so called because the cutoff values for the respective parameters were close to 5 or 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; (ii) Molecular weight is over 500; (iii) log P is over 5; (iv) More than 10 H-bond acceptors.
* 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.
* 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.
* 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 and Pfizer.
* The lipophilicity is unchanged in Merck drug candidates, it is increased in Pfizer candidates, because the most reliable method to increase *in virto* potency is with an appropriately positioned lipophilic functionality.
* The H-bond acceptor trend, unchanged in Pfizer candidates, is increased in Merck probably because of the strong focus on peptide-mimetic like structures in rational drug design the typically interact through hydrogen bonding.
* Merck-like rational drug design leads to poorer permeability while Pfizer-like high throughput screening leads to poorer solubility.
* Several QSPR models estimate membrane permeability, as an example, corneal permeability data have been analyzed for quantitative relationships with physicochemical properties.
* Good parabolic correlations were established between lipophilicity, as expressed by the octanol-water partition coefficient, log P, and the permeability in individual analyses of compound classes such as adrenoceptor blocks and steroids.
* 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, intermolecular interactions, hydrogen-bonding capacity, and conformational stability of molecules can determine permeability.
* 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.
* 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 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 Atkinsons’ group contribution method.
* 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.
* It fails for the 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.
* QSPR/QSAR models utilizing different kinds of molecular descriptors have been developed for reaction rate constants with OH and also with NO3 radicals, which are the most important reactive species in the troposphere at night.
* 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 R2 = 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 NO3 radicals, models were obtained separately for 58 aliphatic compounds and 16 aromatic compounds.
* It shows that tropospheric degradation models based on whole molecule descriptors and that the performance of such model is comparable with the higher dimensional parameter Atkinson’s model.
* Atmospheric half-life is a common criterion used to study persistence in the environment and tendency to undergo long-range transport.
* Various structure-based biodegradation estimation methods have been compared in a recent review of Raymond et al. Biodegradation of organic chemicals in natural systems can be classified as primary, ultimate, or acceptable.
* 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.
* Heterologous models able to predict biodegradability for a diverse set of chemical structure are scarce.
* Modeling biodegradation is complicated by a multitude of factors including temperature, population of microorganisms, accessibility of metabolic cofactors, cellular transport properties etc.
* The models are used to predict the probability of biodegradation ranging from 0 and 1 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 *R2* =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.
* 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.
* QSAR on Toxicity: - the rapid development of QSAR analysis for the prediction of toxicity was initiated by Hansch and Fujita, they demonstrate that the relationships exist between biological activities and the hydrophobic, electronic and steric properties of compounds.
* When dealing with toxic specific interactions, Hansch type QSAR models often gave moderate prediction of the toxicity of compounds. It is particularly the case for the carcinogenicity and mutagenicity.
* 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.
* The most common classification is based on researcher’s experimental knowledge into the modes of action (MOA). This classification was introduced with the concept of “baseline toxicity” by Könemann and coworkers while studying relationship between toxicity and the octanol-water partition coefficient for inert narcotic pollutants.
* 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.
* Recent Hermens derived rule-based system, his rules rely on the presence or absence of certain structural or substructural features, it 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.
* 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.
* The hydrophobic interaction generally expressed by the octanol-water partition coefficient (log P) has been a major determinant of the toxic behavior of compounds.
* The researchers used only structure-based whole molecule descriptors to correlate the acute toxicity of 293 compounds toward *Poecilia reticulate*.
* A greater advantage of whole molecule descriptors was apparent for the unselectively and selectivity reacting toxic compounds.
* Importantly, replacements were found for commonly used log P for those subclasses, 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.
* The researchers 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.
* The researchers are 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.
* General Conclusions: - there is no doubt that QSAR/QSPR approaches will gain significantly in popularity in the coming years.
* 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.

Alan R. Katritzky, Dan C. Fara, Ruslan O. Petrukhin, Douglas B. Tatham, Uko Maran, Andre Lomaka and Mati Karelson