AUTHOR'S REVIEW

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2016**, *89*(4), 419–429
Published online: January 29, 2017
DOI: 10.5562/cca3029

HKD 1926

Use of Graph Invariants in Quantitative Structure-Activity Relationship Studies

Subhash C. Basak

Natural Resources Research Institute & Department of Chemistry and Biochemistry, University of Minnesota Duluth,
5013 Miller Trunk Highway, Duluth, Minnesota 55811, USA
Author's e-mail address: sbasak@nrri.umn.edu

RECEIVED: November 5, 2016 * REVISED: December 9, 2016 * ACCEPTED: December 12, 2016

THIS PAPER IS DEDICATED TO PROF. NENAD TRINAJSTIĆ ON THE OCCASION OF HIS 80TH BIRTHDAY

Abstract: This chapter reviews results of research carried out by Basak and collaborators during the past four decades or so in the development of novel mathematical chemodescriptors and their applications in quantitative structure-activity relationship (QSAR) studies related to the prediction of toxicities and bioactivities of chemicals. For chemodescriptors based QSAR studies, we have used graph theoretical, three dimensional (3-D), and quantum chemical indices. The graph theoretic chemodescriptors fall into two major categories: (a) Numerical invariants defined on simple molecular graphs representing only the adjacency and distance relationship of atoms and bonds; such invariants are called topostructural (TS) indices; (b) Topological indices derived from weighted molecular graphs, called topochemical (TC) indices. Collectively, the TS and TC descriptors are known as topological indices (TIs). The set of independent variables used for modeling also includes a group of three-dimensional (3-D) molecular descriptors. Semi-empirical and various levels of *ab initio* quantum chemical indices have also been used for hierarchical QSAR (HiQSAR) modeling. Results indicate that in many cases of property / activity / toxicity analyzed by us, a TS + TC combination explains most of the variance in the data.

Keywords: molecular structure, model object, theoretical model, graph invariant, quantitative structure-activity relationship (QSAR), topological indices (TIs), three dimensional (3-D) or geometrical descriptors, quantum chemical descriptors, principal component analysis (PCA), leave-one-out (LOO) cross-validation, k-fold cross-validation, external validation, rank-deficient, two-deep cross validation, naïve q^2, true q^2, dibenzofurans, aryl hydrocarbon (Ah) receptor, Interrelated two way clustering (ITC), congenericity principle, diversity begets diversity principle, applicability domain (AD), mutagenicity, anticancer activity.

INTRODUCTION

IN order to describe an aspect of holistic reality we have to ignore certain factors such that the remainder separates into facts. Inevitably, such a description is true only within the adopted partition of the world, that is, within the chosen context.

Hans Primas
In: *Chemistry, Quantum Mechanics and Reductionism*

A recent trend in structural chemistry, new drug discovery, and environmental toxicology is the use of computed molecular structure descriptors in predicting their properties and bioactivities [1–7] In particular, during the past half century or so we have witnessed an upsurge of interest in the use of numerical invariants derived from molecular graphs for quantitative structure-activity / property relationship (QSAR / QSPR) studies.[1–8] The structure of an assembled entity, *e.g.,* a molecule, is the pattern of relationship among its parts. The various concepts of molecular structure, *e.g.* classical valence bond representations, various chemical graph theoretic representations, ball and spoke model of a molecule, representation of molecules by minimum energy conformations, representation of chemical species by Hamiltonian operators, are model objects[9,10] derived through different abstractions from the same chemical reality. In each instance, the equivalence class (concept or model of molecular structure) is generated by selecting certain aspects while ignoring some unique properties of those objects.

The modeling process consists of selecting certain aspects of molecular structure while ignoring others. As noted by Albert Einstein[11] in his remarks on the philosopher

Bertrand Russell's theory of knowledge: "The more, however, we tum to the most primitive concepts of everyday life, the more difficult it becomes amidst the mass of inveterate habits to recognize the concept as an independent creation of thinking. It was thus that the fateful conception – fateful, that is to say, for an understanding of the here-existing conditions – could arise, according to which the concepts originate from experience by way of "abstraction," *i.e.,* through omission of a part of its content."

As pointed out by Basak *et al*[10,12] regarding the philosophy of modeling of molecular structure: "Any concept of molecular structure is a hypothetical sketch of the organization of molecules. Such a *model object* is a general theory and remains empirically untestable. A model object has to be grafted onto a specific theory to generate a *theoretical model*. A theoretical model of an object can be empirically tested. For example, when it was suggested by Sylvester[13] in 1878 that the structural formula of a molecule is a special kind of graph, it was an innovative general theory without any predictive potential. When the idea of combinatorics was applied on chemical graphs (model objects), it could be predicted that "there should be exactly two isomers of butane (C_4H_{10})" because "there are exactly two tree graphs with four verüces" when one considers only the non-hydrogen atoms present in C_4H_{10}. This is a theoretical model of limited predictive potential. Although it predicts the existence of chemical species, given a set of molecules, *e.g.* isomers of hexane (C_6H_{14}), the model is incapable of predicting any property. This is because of the fact that any empirical property P maps a set of chemical structures into the set ʀ of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, we need a nonempirical (structural) ordering scheme which closely resembles the empirical ordering of structures as determined by P. This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariant(s)."

MATHEMATICAL CHEMO-DESCRIPTORS: TOPOLOGICAL INDICES, 3-D DESCRIPTORS AND QUANTUM CHEMICAL INDICES

As discussed in section 1 above, optimal characterization of structure has remained elusive. Different groups of researchers have used different methods for the representation and quantification of molecular structure. In our quantitative structure-activity relationship (QSAR) and quantitative molecular similarity analysis (QMSA) research, we have used mainly three classes of descriptors for the quantification of structure, *viz*., (a) graph invariants defined on molecular graphs, also known as topological indices, (b) three dimensional (3-D) or geometrical descriptors, and (c) quantum chemical descriptors.

The author of this chapter (Basak) and his coworkers have been involved since the early 1970s in the development of novel numerical graph invariants or topological indices (TIs)[1,2] as well as biodescriptors derived from DNA / RNA sequences[14,15] and proteomics maps.[6,7] It may be mentioned here that graph theoretical numerical indices of molecules were called "topological indices" by Hosoya[16] for the first time in a paper published in 1971.

Our team at the University of Minnesota Duluth routinely uses the software MolconnZ,[17] POLLY,[18] Triplet[19] APProbe,[20] MOPAC,[21] and Gaussian[22] for the calculation of molecular descriptors for QSAR / QSPR studies. A typical list of descriptors used by us is shown in Table 1 (Supplementary material).

Basak *et al.*[2] have divided the topological indices (TIs) into two major groups: topostructural (TS) indices and topochemical (TC) indices. TS indices are calculated from skeletal graph models of molecules which do not distinguish among different types of atoms in a molecule or the various types of chemical bonds, *e.g*., single bond, double bond, triplet bond, *etc*. Thus, TS indices quantify information regarding the connectivity, adjacency, and distances between vertices ignoring their distinct chemical nature. TC indices, on the other hand, are sensitive to both the pattern of connectedness of the vertices (atoms), as well as their chemical / bonding characteristics. Therefore, the TC indices are more complex and chemically informative as compared to the TS descriptors.

The geometrical or 3-D parameters quantify the volume, size, and shape of molecules from various models. The three-dimensional Wiener index calculated on the hydrogen-suppressed and hydrogen-filled graphs are also quantifiers of molecular shape and size. With respect to calculation of quantum chemical descriptors, we have used both the AM1 semi-empirical method as well as *ab initio* calculations based on the STO-3G, 6-31G(d), 6-311G, 6-311G(d), and aug-cc-pVTZ basis sets.

THE QSAR PARADIGM

Many physiological, pathological, toxicological, and biomedicinal processes are determined by interactions of small molecules such as endogenous ligands, drugs, xenobiotics, and substrates as well as inhibitors of enzymes related to metabolic pathways with their appropriate biological targets. The maintenance of the integrity and continuity of such key ligand-biotarget interactions is critical for the smooth functioning of biological systems ranging from the single celled organism to the complex ecosystems. A large number of drugs are small molecules that interact with specialized

enzymes / receptors in appropriate physiological compartments and thereby produce effect(s) that bring a pathologically perturbed biological system back to a healthy state.[1–4] Biological properties of molecules, beneficial or deleterious, can be looked upon as the result of ligand-biotarget interactions and can be expressed by the relationship:

$$BR = f(S,B) \quad (1)$$

where *BR* represents the normal biological or pathological / toxicological response produced by the ligand (drug or toxicant) in the target biological system, and *B* represents the relevant biochemical part of the target system which is perturbed by the ligand to produce the measurable effect. It is believed that a major determinant of *BR* is the nature or structure (*S*) of the ligand. The structure becomes the sole determinant of the variation of the measured BR from one chemical to another when the biological system, *B*, remains practically the same during the course of the experiment and there is alternation only in the structure of the ligands. [Eq. 1] under such a condition approximates to:

$$BR = f(S) \quad (2)$$

A lot of research conducted in drug discovery, toxicology, environmental sciences, and biochemistry follows the paradigm expressed in [Eq. 2] and researchers attempt to decipher the effects as well as the modes, and mechanism(s) of action of molecules on some selected biotargets, which are assumed not to change significantly during the course of the experiment.

When we embark on the characterization of *BR* based on chemical structure alone following [Eq. 2], we really attempt to understand which characteristics of the chemical structure are recognized by the biomolecular target. What are the factors involved in recognition: Molecular size, shape, chirality, stereo-electronic nature or charge? Which ones are more important and which have a marginal impact on *BR*? This is often accomplished by the development of molecular descriptors, referred to by us as chemodescriptors, which quantify various aspects of molecular structure such as shape, size, symmetry, chirality, stereo-electronic nature, *etc*. using various mathematical techniques.

HIERARCHICAL CLASSIFICATION OF DESCRIPTORS

The combination of topological, geometrical, and quantum chemical chemodescriptors, and biodescriptors derived from proteomics, genomics, and DNA / RNA sequence characterization, gives a hierarchy of descriptors that begins with the simplest graph invariants and ends with the biodescriptors, which require the use of massive amounts of expensive and time-intensive laboratory test data (Big Data) (Figure 1). It should be clearly stated here that descriptors in the higher levels of the hierarchy are not necessarily superior to those placed at lower levels. The scheme simply shows a gradation based on the need for computational and laboratory resources.

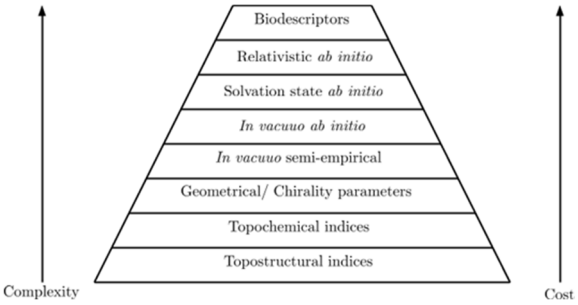

Figure 1. Hierarchical classification of chemodescriptors and biodescriptors used in QSAR.

Table 1 (supplementary material) provides a list of TS, TC, 3-D, and quantum chemical chemodescriptors used by Basak and coworkers[2,3] over the years in their QSAR and quantitative molecular similarity analysis (QMSA) studies.

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) STUDIES USING CHEMODESCRIPTORS

Current industrialized societies routinely use a large number of natural and anthropogenic chemicals in the form of drugs, solvents, synthetic intermediates, cosmetics, herbicides, pesticides, *etc*. to maintain the lifestyle. But in many cases a large fraction of these chemicals do not have the experimental data necessary for the prediction of their beneficial and deleterious effects.[23] Table 2 gives a partial list of properties, both physicochemical and biochemical / pharmacological / toxicological, needed for the effective screening of chemicals for new drug discovery and protection of human as well as ecological health. Because determination of such properties for so many chemicals in the laboratory is prohibitively costly, one solution of this quagmire has been the use of QSARs and molecular similarity based analogs to obtain acceptable estimated values of properties.

Statistical Methods for QSAR Model Development And Validation

In God we trust. All others must bring data.
– W. Edwards Deming

Table 2. List of properties needed for screening of chemicals.

Physicochemical	Pharmacological / Toxicological
Molar volume	Macromolecular level
Boiling point	: Receptor binding (K_d)
Melting point	: Michaelis constant (K_m)
Vapor pressure	: Inhibitor constant (K_i)
Water solubility	: DNA alkylation
Dissociation constant (pK_a)	: Unscheduled DNA synthesis
Partition coefficient	Cell level
: Octanol-water (log *P*)	: Salmonella mutagenicity
: Air-water	: Mammalian cell transformation
: Sediment-water	Organism level (acute)
Reactivity (electrophilicity)	LD_{50} (mouse, rat)
	LC_{50} (fathead minnow)
	Organism level (chronic)
	: Bioconcentraton factor
	: Carcinogenicity
	: Reproductive toxicity
	: Delayed neurotoxicity
	: Biodegradation profile

In the early 1970s, when this author (Basak) started carrying out research on the development and use of calculated chemodescriptors in QSAR, only a few such descriptors were available. But now, with the availability of various software[17–22] the landscape of availability and calculation of molecular descriptors is very different. The four major pillars[2] of a useful QSAR system development are:

(a) Availability of high quality experimental data (veracity of dependent variable)
(b) Data on sufficient number of compounds (reasonably good sample size)
(c) Availability of relevant descriptors (independent variables of QSAR) which quantify aspects of molecular structure relevant to the activity / toxicity of interest
(d) Use of appropriate methods for model building and validation.

The various pathways for the development of structure-activity relationship (SAR) and property-activity relationship (PAR) models either from calculated molecular descriptors or from experimentally determined as well as calculated properties as independent variables may be expressed by the scheme provided in Figure 2.

The use of computed molecular descriptors and experimental property data in PAR / SAR / QSAR may be illuminated through a formal exposition of the *structure-property similarity principle*—the central paradigm of the field of SAR.[24] Figure 2 depicts the determination of an experimental property, *e.g.*, measurement of octanol-water partition coefficient of a chemical in the laboratory, as a function α: C→R which maps the set C of compounds into the real line R. A non-empirical QSAR may be looked upon as a composition of a description function β_1: C→D mapping each chemical structure of C into a space of non-empirical structural descriptors (D) and a prediction function β_2: D→R which maps the descriptors into the real line. When $[\alpha(C) - \beta_2 \circ \beta_1 (C)]$ is within the range of experimental errors, we say that we have a good QSAR model. On the other hand, PAR is the composition of θ_1: C→M which maps the set C into the molecular property space M and θ_2: M→R mapping those molecular properties into the real line R. Property-activity relationship seeks to predict one property (usually a complex physicochemical property) or bioactivity of a molecule in terms of other (usually simpler or easily determined experimentally) properties.

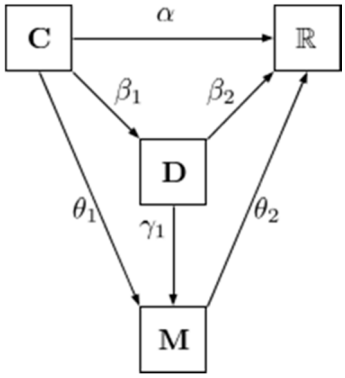

Figure 2. Composition functions of various mappings for structure-activity relationship (SAR) and property-activity relationship (PAR).

In the process of formulating a scientifically interpretable and technically sound QSAR model, one needs to keep in mind some important issues. First and foremost, we have to check whether a specific method is the best technique in modeling a specific QSAR scenario. In a regression set up, for example, when the number of independent variables or descriptors (*p*) is much larger than the number of data points (dependent variable, *n*) *i.e.* $p >> n$, the estimate of the coefficient vector is non-unique. This is also the case when predictors in the study are highly correlated with one another to the extent that the 'design matrix' is rank-deficient. Both of these factors are relevant to QSARs. In many contemporary QSAR studies, the number of initial set of predictors typically is in the range of hundreds or thousands, whereas more often than not, mostly to manage experimental cost, the experimenter can collect only a much smaller number (tens or hundreds) of samples. This effectively makes the problem high-dimensional and

rank-deficient ($p >> n$) in nature. Also, when a large number of descriptors on a set of chemicals are used to develop QSARs, one should expect that some predictors within a single class, *e.g.*, TC descriptors, or even predictors belonging to apparently different classes could be highly correlated with one another. Such situations can be tackled either by attempting to pick important variables through model selection or 'sparsity'-type approaches, e.g. forward selection, least absolute shrinkage and selection operator (LASSO),[25] adaptive LASSO,[26] or finding a lower-dimensional transformation that preserves most of the information present in the original set of descriptors, *e.g.,* principal component analysis (PCA) or envelope methods.[27]

One also needs to check the ability of QSAR models in providing competent predictions on 'similar' datasets *via* validation on out-of-sample test sets.[28–32] For a relatively small sample, *i.e*., a small collection of compounds, this is done by following a *leave-one-out (LOO) cross-validation* method. For data sets with a large number of compounds, a more computationally economical way is to do a *k-fold cross-validation*: split the data set randomly into *k* (previously decided) equal subsets, take each subset in turn as test set and use the remaining compounds as training sets and use the model to obtain predictions. Comparing cross-validation with the somewhat prevalent method in QSAR research of *external validation*, *i.e*. choosing a single train-test split of compounds, it should be pointed out that in external validation the splits of data sets are carried out only once using the experimenters' *a priori* knowledge or some subjectively chosen *ad hoc* criterion. But in cross-validation the splits are chosen randomly, thus providing a more unbiased estimate of the general nature of the QSAR model. Furthermore, Hawkins *et al*[28] showed theoretically that compared to external validation, the LOO method of cross-validation is a better estimator of the actual predictive ability of a statistical model for small datasets, while for large sample sizes both perform equally well. To quote Hawkins *et al*,[28] "*The bottom line is that in the typical QSAR setting where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly*." Specific drawbacks of holding out only one test set in the external validation method include: 1) Structural features of the held out chemicals may not be included in the modeling process, resulting in a loss of information, 2) Predictions are made on only a subset of the available compounds, whereas the LOO method predicts the activity value for all compounds, 3) There is no scientific tool that can guarantee similarity between chemicals in the training and test sets, and 4) Personal bias can easily be present in the selection of the external test set.

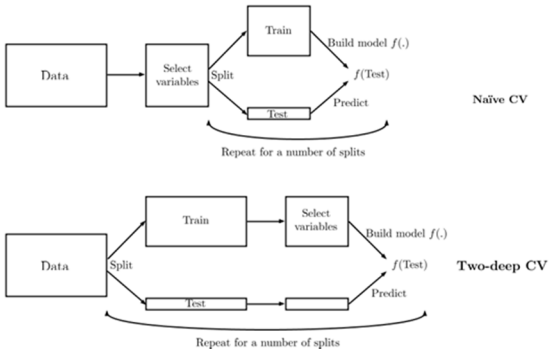

Figure 3. Difference between naïve and two-deep cross validation (CV) schemes (Reprinted with permission from Bentham Science Publishers).

In the rank-deficient QSAR development scenarios, special care should be taken in combining conventional modeling with the additional step of variable selection or dimension reduction. An intuitive, but frequently misunderstood and wrong, procedure would be to perform the first stage of pre-processing first, selecting important variables or determining the optimal transformation, and then using the transformed data / selected variables to build the predictive QSAR models and obtain predictions for each train-test split. The reason why this is not appropriate is that the data is split only after the variable selection / dimension reduction step is already completed. Essentially this method ends up using information from the holdout compound/ split subset to predict activity of those very samples. This *naïve cross-validation* procedure causes synthetic inflation of the cross-validated q^2, hence compromises the predictive ability of the model[29–32] (Figure 3). A two-step approach (referred in Figure 3 as '*Two-deep CV*') helps avoid this tricky situation. Instead of doing the pre-model building step first and then taking multiple splits for out-of-sample prediction, for each split of the data the initial steps are performed only using the training set of compounds every time. Because calculations on two different splits are not dependent on each other, for large data sets the increased computational demand arising out of the repeated variable selection can be handled using substantial computer resources, *e.g.*, parallel processing. It should be emphasized that the naïve cross-validation (naïve CV) method gives *naïve* or wrong q^2 values whereas the two-deep cross-validation (two deep CV) approach gives us the correct or *"true"* q^2.

The quality of the model, in terms of its predictive ability, is evaluated based on the associated q^2 value, which is defined as:

$$q^2 = 1 - (\text{PRESS} / \text{SSTotal}) \quad (3)$$

where PRESS is the prediction sum of squares and SSTotal is the total sum of squares. Unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, thus providing a reliable measure of model quality.

Some Examples of Hierarchical QSAR Using Calculated Chemodescriptors

BINDING AFFINITY OF DIBENZOFURANS FOR ARYL HYDROCARBON (AH) RECEPTOR

Dibenzofurans constitute an important class of environmental contaminants that are produced as undesirable by-products in natural and industrial processes. The toxic effects of these compounds are thought to be mediated through binding to the aryl hydrocarbon (*Ah*) receptor. We developed HiQSAR models based on a set of 32 dibenzofurans with *Ah* receptor binding affinity values obtained from the literature. Descriptor classes used to develop the models included the TS, TC, 3D, and the STO-3G class of *ab initio* QC descriptors. For this data set, the initial number of descriptors calculated were as follows: TS = 114; TC = 248; 3D = 7, and QC = 6. There were 188 descriptors retained for QSAR formulation after removing those that were: a) Constant for all compounds, b) perfectly correlated with another descriptor, and c) Triplet indices with missing values. Statistical metrics for the ridge regression (RR), partial least square (PLS), and principal component regression (PCR) models are provided in Table 3. We saw that the RR models were superior to those developed using either PLS or PCR. Examining the RR metrics, it is evident that the TC and the TS + TC descriptors provide high quality predictive models, with R^2_{cv} values of 0.820 and 0.852, respectively. The addition of the 3-D and STO-3G QC descriptors does not result in significant improvement in model quality. When either of the 3-D or QC classes is used alone, the results are quite poor. This indicates that the topological indices are capable of adequately quantifying those structural features which are relevant to the binding of dibenzofurans to the *Ah* receptor. Comparison of the experimentally determined binding affinity values and those predicted using the TS + TC descriptors and RR model is available in Table 4 (Supplementary material). The details of this QSAR analysis has been published.[33]

Table 3. Summary statistics for predictive *Ah* receptor binding affinity models.

Independent Variables	$R^2_{c.v.}$			PRESS		
	RR	PCR	PLS	RR	PCR	PLS
TS	0.731	0.690	0.701	16.9	19.4	18.7
TS+TC	0.852	0.683	0.836	9.27	19.9	10.3
TS+TC+3D	0.852	0.683	0.837	9.27	19.9	10.2
TS+TC+3D + STO-3G	0.862	0.595	0.862	8.62	25.4	8.67
TS	0.731	0.690	0.701	16.9	19.4	18.7
TC	0.820	0.694	0.749	11.3	19.1	15.7
3D	0.508	0.523	0.419	30.8	29.9	36.4
STO-3G	0.544	0.458	0.501	28.6	33.9	31.3

HiQSAR MODELING OF A DIVERSE SET OF 508 CHEMICAL MUTAGENS

TS, TC, 3D, and QC descriptors for 508 structurally diverse chemicals were calculated and QSARs were developed hierarchically using the four types of descriptors. For details of calculations and model building, see ref. [31], [32], and [34]. The method Interrelated two way clustering, ITC,[34] which falls under the unsupervised class of approaches,[35] was used for variable selection. Table 5 gives results of ridge regression (RR) alone as well as those where RR was used after descriptors were selected by ITC. For both RR only and ITC + RR models the TS + TC combination gave the best QSARs for predicting mutagenicity of the 508 diverse chemicals. The addition of 3-D and QC descriptors to the set of independent variables made minimum or no improvement in the quality of the models. Of the 508 chemicals, 256 were mutagens and 252 were non-mutagens based on Ames' Salmonella / microsome mutagenicity assay. Regarding the number of indices in the various classes for this data set, the make-up was as follows: TS (103); TS + TC (298); TS+TC+3D+QC (307).[31]

Table 5. HiQSAR model (RR and ITC+RR) for a diverse set of 508 chemical mutagens / non-mutagens. All four means the model used TS + TC + 3D + QC descriptors.

Predictor Type	Predictor Number	Correct classification / %	Sensitivity	Specificity
Model Type: RR				
TS	103	53.14	52.34	53.97
TS + TC	298	76.97	83.98	69.84
All four	307	77.17	84.38	69.84
Model Type: ITC				
TS	103	66.34	73.83	58.73
TS + TC	298	73.23	77.34	69.05
TS + TC + 3D	301	74.80	77.34	72.22
All four	307	72.05	76.17	67.86

QSAR OF ANTICANCER ACTIVITY OF PHENYLINDOLES

QSARs were developed for a set of 89 phenylindole derivatives using TS, TC indices and atom pairs,[36] a specific class of substructures. Table 6 below summarizes the results of this modeling effort.[37] In total, 691 APs and a set of 369 topological indices were calculated for this data set.

Recent review of results of HiQSARs carried out by Basak and coworkers[2,38–40] using topostructural, topochemical, 3-D, and quantum chemical indices for diverse properties, *e.g.*, acute toxicity of benzene derivatives, dermal penetration of polycyclic aromatic hydrocarbons (PAHs), mutagenicity of a congeneric set of amines (heteroaromatic and aromatic) and others indicate that in most of the above mentioned cases TS + TC combination of indices give reasonable predictive models. The addition of 3-D and quantum chemical indices after the use of TS and TC descriptors did very little improvement in model quality.

How can we explain the above-mentioned trend in HiQSAR?

One plausible explanation is that for the recognition of a receptor, *e.g.*, the specific recognition of dibenzofuran by the Ah receptor, discussed above, the dibenzofuran derivatives probably need some specific geometrical and stereoelectronic arrangements or a specific pharmacophore. But once this minimal requirement of the recognition process is present in the molecule, alterations in bioactivities from one molecule to another in the same structural class are governed by more general structural features which are quantified reasonably well by the TS and TC indices derived from the conventional bonding topology of molecules and features like sigma bond, π bond, lone pair of electrons, hydrogen bond donor acidity, hydrogen bond acceptor basicity, *etc*. More studies with different groups of molecules with diverse bioactivities are needed to validate or falsity this hypothesis in line with the falsifiability principle of Sir Karl Popper,[41] a basic scientific paradigm in the philosophy of science which defines the inherent testability of any scientific hypothesis.

TWO QSAR PARADIGMS

Congenericity Principle versus Diversity Begets Diversity Principle- Analyzed Using Computed Mathematical Chemodescriptors of Homogeneous and Diverse Sets of Chemical Mutagens

The well-known and age old paradigm of quantitative structure-activity relationship (QSAR) is the *congenericity principle* which states that similar structures usually have similar properties. But these days a lot of large and structurally diverse data sets of chemicals with a common experimental property (dependent variable) are being available. Starting with the same classes of descriptors we extracted the two subsets of statistically most significant predictors for the formulation of QSARs for two different sets of molecules: A homogeneous set of 95 amine mutagens and a diverse set of 508 structurally diverse mutagens / non-mutagens. The predictors included calculated TS, TC, geometrical, and QC indices. Whereas for the homogeneous amines, a small group of only 7 descriptors were found to be significant in model building, for the 508 diverse set 42 descriptors were found to be statistically significant.[42] This preliminary and empirical study supports the *diversity begets diversity* principle of QSAR formulated for the first time by Basak.[2]

Table 6. Ridge Regression Results with TI, AP, and TI + AP for 89 phenylindoles.

Descriptor class	q^2	PRESS
TI	0.678	13.72
AP	0.703	12.66
TI + AP	0.730	11.48

DIFFERENTIAL QSAR

A Computational Approach to Understand the Molecular Basis of Drug Resistance

The development of drug resistance, the emergence of multiple drug-resistant (MDR) organisms in particular, is well documented in the medical field today.[2] This problem has been identified for tuberculosis,[43] Hepatitis B,[44] Influenza A viruses,[45] different types of cancer cells,[46] to mention just a few cases. Curt *et al.*[47] discussed multiple mechanisms, *e.g.*; reduced drug accumulation and / or retention, conformational changes in and / or overproduction of the biochemical target, and reduced activation and / or increased catabolism of drugs may be involved in the emergence of resistance. When resistance develops, the original effectiveness of the drug is gradually compromised. The target-ligand interaction thus altered may be analyzed using multiple methods including: (a) Characterizing the altered sensitivity of the target using QSARs based on computed descriptors of the ligands, and (b) Analyzing altered structures of the biological target and their impact on drug-target interaction patterns resulting in the observed phenomenon of resistance.

In probably the first study of its kind, using a set of 58 cycloguanil derivatives tested against the dihydrofolate reductase (DHFR) enzyme from sensitive and mutant

strains of *Plasmodium falciparum*, Basak and Mills[48] developed QSARs for two types of DHFRs, one wild and the other resistant, using the same set of calculated mathematical descriptors.

When key amino acid residues in the DHFR enzyme sequences are altered, the parasite develops resistance to the antimalarial drug cycloguanil. For example, *P. falciparum* strains having a pair of point mutations from Ala-16 to Val-16 and from Ser-108 to Thr-108 are substantially resistant to cycloguanil as compared to the sensitive strains.[49]

A comparison of the first twenty most influential molecular descriptors from the two QSARs, based on sensitive and resistant DHFRs, showed that only two of the twenty descriptors were common between the two QSAR models. Such differential QSARs (DiffQSARs) using a high dimensional chemodescriptor space shed light on the manner in which ligand (cycloguanil) – target (DHFR) interaction was modified as a result of the mutation underlying resistance. Subsequently, Basak and Mills[50] extended this approach to five (one wild and four mutants) varieties of DHFRs. Such mathematical chemodescriptor based QSARs can help in the *in silico* screening of libraries in the design of drugs active against resistant organisms like *P. falciparum* and others mentioned above once sufficient test data for model building are available.

APPLICABILITY DOMAIN OF QSAR MODELS

A very important issue in the development of a QSAR model is that of defining the applicability domain (AD) of the model. This is necessary for any valid implementable QSAR model according to OECD principles.[51] There are a few methods of defining the AD of statistical models which can be roughly divided into two classes: (a) AD methods that define the active predictor space through some method like bounding box, PCA or convex hulls; and (b) Distance-based methods which compute the similarity / dissimilarity of a new compound to the set of compounds which have been used in formulating the training model. To obtain predictions for any incoming sample set using the model, the first group of methods are used to ensure that the compounds are within the so-called 'active subspace': which essentially means that we are actually performing interpolation only, not extrapolation.[52,53] For the distance-based approach, a pre-defined statistic is calculated to quantify the degree of nearness of the test compounds to the training set and based on whether that statistic is above or below a certain cutoff value, predictions for those compounds are considered acceptable or not.[52,54]

PRACTICAL APPLICATIONS OF QSAR

Knowledge is of no value unless you put it into practice.
– Anton Chekhov

Practical applications of good quality QSARs, particularly those based on easily calculated molecular descriptors, can be very useful tools in pharmaceutical drug design, pollution prevention, and specialty chemical design.

In pharmaceutical drug design, the journey of identified lead molecules in the drug discovery pipeline is a long and risky one. Average cost of developing a new drug (including the cost of failures) during 2000s to early 2010s was US $2.6 billion.[55] One important contributing factor to this astronomical cost is that the drug developer has to produce and test a large number of derivatives of the lead structure for their beneficial and toxic side effects before one marketable drug is identified. QSAR can play a very important role in drug design providing a cheaper and fast alternative to the medium throughput *in vitro* and low throughput *in vivo* testing of chemicals, the latter two methods generally being used more frequently in the later stages of the discovery cascade. It has been noted that currently no drug is developed without going through the prior evaluation by QSAR methods.[56]

In Figure 4, a general scheme is provided for the use of QSAR in drug discovery. Beginning with a "lead compound," modern combinatorial chemistry techniques can produce millions, even billions, of derivatives. Such real or hypothetical chemicals must be evaluated in real time to prioritize them for synthesis and experimental testing. QSARs based on easily calculated descriptors can help us in accomplishing this task.[2]

The era of "Big Data" has arrived in the realm of drug discovery. QSAR based on easily calculated molecular

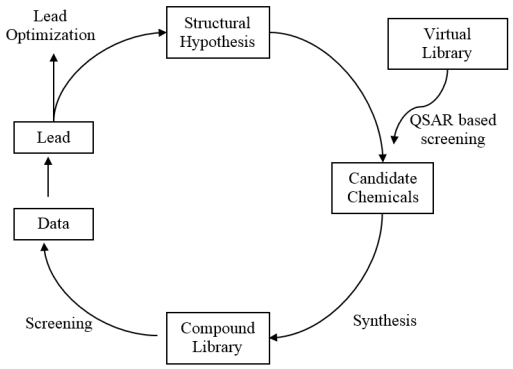

Figure 4. A suggested generic scheme for the use of QSARs in drug discovery protocols.

descriptors, the TS and TC graph invariants in particular, may find applications both in new drug discovery and risk assessment of environmental pollutants.[2] For a concise description of trends in this realm, please see Basak *et al.*[57]

Here we give some examples where TIs have been used for practical drug design. Galvez and coworkers[58,59] used topological approaches to design drugs for cancer chemotherapy and Alzheimer's disease. In the 1980s the software POLLY and quantitative molecular similarity analysis (QMSA) method derived from principal components (PCs) calculated from 90 POLLY indices were installed at the Upjohn Company. They used this method – called the Basak method[60,61] – for the discovery of numerous lead structures for drugs. Another landmark study was by Grassy *et al*[62] on the rational design of immunosuppressive peptides without any information about their receptors or biochemical mechanisms of action. The authors used a variety of topological and shape descriptors in combination with an analysis of molecular dynamics trajectories to down select subsets of potential drug candidates. The lead compounds were peptides, derived from the heavy chain of the human leukocyte antigen (HLA) class I, that has the capability of modulating immune responses under *in vitro* and *in vivo* conditions. The molecule predicted to be most potent by the descriptor based approach was found to have an immunosuppressive activity approximately 100 times higher than the starting lead compound.

DISCUSSION

Everything should be made as simple as possible, but not simpler.
– Albert Einstein

This chapter has presented a brief review of our research in the use of mathematical chemodescriptors in the prediction of bioactivity / toxicity of chemicals.

In the chemodescriptor realm, our major objective has been to investigate the utility of graph theoretical parameters, also known as topological indices, in QSAR / QSPR studies. At present a large number of descriptors can be calculated for chemicals using the currently available software. If the number of data points (dependent variables) for QSAR model building is much smaller than the number of descriptors, *i.e.*, the situation is rank deficient, one needs to be cautious. We have discussed variable selection methods including ITC[34] which, to our knowledge, has probably been for the first time imported to QSAR from the genomics area in our research. In the calculation of q^2 in the rank deficient case, one must follow the two-deep cross-validation procedure; otherwise the calculated q^2 will reflect overfitting.[28–31] In HiQSAR modeling, we found that of the four types of calculated molecular descriptors, *viz.*, TS, TC, 3-D, and QC indices, in most cases a TS + TC combination gave good quality models; the addition of 3-D or QC descriptors after the utilization of TS and TC combination did not improve the model quality significantly. This is a good news in view of the fact that we already reached the age of big data[57] and easily calculated indices like TS and TC descriptors, if they give good models in many areas, could find wide applications in the *in silico* evaluation of chemicals. The congenericity principle has been a major theme of QSAR whereby there has been a tendency in developing QSARs of congeneric or structurally related sets of chemicals. When the same property, *viz.*, mutagenicity, of congeneric versus diverse sets were used to develop QSAR models, the congeneric set of 95 amines needed much lower number of significant descriptors as compared to the diverse set of 508 molecules. This gives support to the diversity begets diversity principle formulated by us.[2]

In the post-genomic era, the omics technologies are generating a lot of data on the effects of chemicals, *e.g.*, drugs and xenobiotics, on the genetic system, *viz.*, transcription, translation, and post-translational modification, of cells and tissues. We have been involved in the development of biodescriptors from DNA / RNA sequences and two-dimensional gel electrophoresis (2DE) data as well as mass spectrometry derived proteomics data from cells / tissue exposed to drugs and toxicants including nanosubstances. Results of our research in this area show that the biodescriptors developed from proteomics data are capable of characterizing the pharmacological / toxicological profiles of chemicals.[2,6,7,65] Some preliminary studies have been carried out by us on the use of the combined set of chemodescriptors and biodescriptors in predicting bioactivity of chemicals. Further research is needed to test the relative effectiveness of these two classes of descriptors, chemodescriptors versus biodescriptors, in predictive pharmacology and toxicology.

At this juncture, after reviewing results of QSAR studies using computed chemodescriptors and biodescriptors, we may ask: *Quo Vadimus*? We found that calculated chemodescriptors are capable of predicting and characterizing bioactivity and toxicity as well as toxic modes of action[66] of chemicals. Research using different types of biodescriptors also showed that such descriptors derived from proteomics maps have reasonable power of discriminating among structurally closely related toxicants. Can we, at this stage, opt for either of the two classes of descriptors, *viz.*, chemodescriptor or biodescriptors alone? The answer is no, as is evident from our models developed for predictive toxicology using both chemodescriptors and biodescriptors. Therefore, in the foreseeable future we will need integrated approaches combining chemodescriptors and biodescriptors in order to obtain the best results (Figure 5).

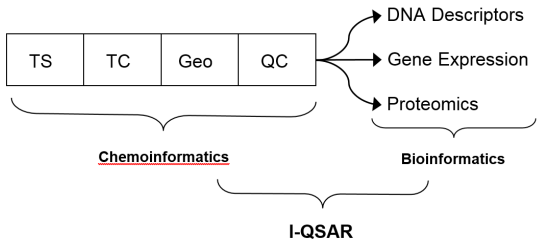

Figure 5. Integrated QSAR, combining chemodescriptors and biodescriptors.

As discussed by this author[67] in a recent book on Advances in Mathematical Chemistry and applications: "Mathematical chemistry or more accurately discrete mathematical chemistry had a tremendous growth spurt in the second half of the twentieth century and the same trend is continuing now. This growth was fueled primarily by two major factors: 1) Novel applications of discrete mathematical concepts to chemical and biological systems, and 2) Availability of high speed computers and associated software whereby hypothesis driven as well as discovery oriented research on large data sets could be carried out in a timely manner. This led to the development of not only a plethora of new concepts, but also various useful applications to such important areas as drug discovery, protection of human as well as ecological health, bioinformatics, and chemoinformatics. Following the completion of the Human Genome Project in 2003, discrete mathematical methods were applied to the "omics" data to develop descriptors relevant to bioinformatics, toxicoinformatics, and computational biology."

The results of various types of QSAR / QSPR research using computed chemodescriptors and biodescriptors derived through applications discrete mathematics on chemical and biological systems give us hope that an exciting future is in waiting ahead of us.[1 2,6,7,67]

Acknowledgment. I am especially thankful to Gregory D. Grunwald for providing excellent technical support and collaboration to my research team at the University of Minnesota Duluth for the past three decades. I am thankful to Kanika Basak, Douglas Hawkins, Jessica Kraker, Brian Gute, Subhabrata Majumdar, Denise Mills, Ashesh Nandy, Frank Witzmann, Kevin Geiss, Krishnan Balasubramanian, Ramanathan Natarajan, Gerald J. Niemi, Alexandru T. Balaban, late Alan Katritzky, Milan Randić, Sonja Nikolic, Marjan Vracko, Xiaofeng Guo, Terry Neumann, Qianhong Zhu and Marissa Harle, for their collaboration in my research.

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: http://dx.doi.org/10.5562/cca3029.

REFERENCES

[1] S. Nikolic, N. Trinajstić, D. Amic, D. Beslo, S. C. Basak in *QSAR / QSPR Studies by Molecular Descriptors* (Ed. M. Diudea), Nova Science Publishers, Huntington, New York, USA, **2001**, pp. 63–81.
[2] S. C. Basak, *Curr. Comput. Aided Drug Des*. **2013**, *9*, 449.
[3] S. C. Basak, *J Eng Sci Manage. Educ*. **2014**, *7*, 178.
[4] L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotopological State,* Academic Press, San Diego, CA, **1999**.
[5] J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Amsterdam, **1999**.
[6] S. C. Basak, *Drug Dev. Res.* **2010**, *72*, 1.
[7] D. M. Hawkins, S. C. Basak, J. J. Kraker, K. T. Geiss, F. A. Witzmann, *J. Chem. Inf. Model*. **2006**, *46*, 9.
[8] N. Trinajstić, *Chemical Graph Theory*, *2nd Ed*.; CRC Press: Boca Raton, FL, **1992**.
[9] M. Bunge, *Method, Model and Matter*, Reidel, Dordrecht-Boston, **1973**.
[10] S. C. Basak, G. J. Niemi, G. D. Veith, *J. Math. Chem*. **1991**, *7*, 243.
[11] A. Einstein in *Ideas and Opinions by Albert Einstein*, (Ed. Carl Seelig), Crown Publishers, Inc., New York, **1954**, pp. 18–24,
[12] S. C. Basak, *HYLE* **2013**, *19*, 3.
[13] J. J. Sylvester, *Am. J. Math*. **1878**, *1*, 105.
[14] A. Nandy, M. Harle, S. C Basak, *Arkivoc* **2006**, *9*, 211.
[15] M. Randić, M. Vracko, A. Nandy, S. C. Basak, *J Chem. Inf. Comput Sci.* **2000**, *40*, 1235.
[16] H. Hosoya, *Bull. Chem. Soc. Jpn*. **1971**, *44*, 2332.
[17] MolconnZ, Version 4.05, Hall Ass. Consult. Quincy, MA, **2003**.
[18] S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY v. 2.3, Copyright of the University of Minnesota, USA, **1988**.
[19] P. A. Filip, T. S. Balaban, A. T. Balaban, *J Math Chem*. **1987**, *1*, 61.
[20] S. C. Basak, G. D. Grunwald, APProbe, Copyright of the University of Minnesota, USA, **1990**.
[21] J. J. P. Stewart, MOPAC Version 6.00, QCPE #455, Frank J Seiler Research Laboratory, US Air Force Academy, CO, **1990**.
[22] M. J. Frisch *et al.*, Gaussian 98 (Revision A.11.2), Pittsburgh, PA, Gaussian Inc, **1998**.
[23] C. M. Auer, J. V. Nabholz, K. P. Baetcke, *Environ. Health Perspect.* **1990**, *87*, 183.
[24] M. Johnson, S. C. Basak, G. A. Maggiora, *Mathl. Comput. Modelling* **1988**, *11*, 630.
[25] R. Tibshirani, *J Royal Stat Soc. Ser. B*. **1996**, *58*, 267.
[26] H. Zou, *J Amer. Stat Assoc.* **2006**, *101*, 1418.
[27] R. D. Cook, B. Li Chiaromonte, *F. Stat Sinica* **2010**, *20*, 927.

[29] D. M. Hawkins, S. C. Basak, D. Mills, *J Chem. Inf. Comput. Sci.* **2003**, *3*, 579.
[30] D. M. Hawkins, S. C. Basak, D. Mills, *Environ Toxicol Pharmacol.* **2004**, *16*, 37.
[31] S. C. Basak, D. Mills, D. M. Hawkins, J. J. Kraker in *Computation in Modern Science and Engineering* (Eds. T. E. Simos and G. Maroulis), American Institute of Physics, Melville, New York, **2007**; pp. 548–551.
[32] S. C. Basak, S. Majumdar in *Advances in Mathematical Chemistry and Applications, Vol. 1* (Eds. S. C. Basak, G. Restrepo, J. L. Villaveces), Bentham eBooks, Elsevier & Bentham Science Publishers, Amsterdam-Boston-Heidelberg, **2016**, pp. 251–328
[33] S. C. Basak, S. Majumdar, "Hierarchical quantitative structure-activity relationships (HiQSARs) for the prediction of physicochemical and toxicological properties of chemicals using computed molecular descriptors", Mol2Net Conference, **2015**. The full text can be found under: http://sciforum.net/conference/MOL2NET-1/paper.
[34] S. C. Basak, D. Mills, M. M. Mumtaz, K. Balasubramanian, *Ind. J. Chem.* **2003**, *42A,* 1385.
[35] C. Tang, L. Zhang, A. Zhang, M. Ramanathan in *Proceedings of BIBE* (Eds. R. Bilof, L. Palagi), 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, IEEE Computer Society: Los Alamitos, CA, **2001**, pp. 41–48.
[36] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, Z. Yakhini, *J. Comput. Biol*. **2000**, *7*, 559.
[37] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
[38] S. C. Basak, Q. Zhu, D. Mills, *Curr. Comput. Aided Drug Des.* **2011**, *7*, 98.
[39] B. D. Gute, S. C. Basak, *SAR QSAR Environ. Res*. **1997**, *7*, 117.
[40] B. D. Gute, G. D. Grunwald, S. C. Basak, *SAR QSAR Environ. Res.* **1999**, *10*, 1.
[41] S. C. Basak, D. Mills, A. T. Balaban, B. D. Gute, *J. Chem. Inf. Comput Sci.* **2001**, *41*, 671.
[42] K. Popper, *The Logic of Scientific Discovery,* Taylor & Francis e-Library, **2005**.
[43] S. C. Basak, S. Majumdar, *Curr. Comput.-Aided Drug Des.* **2016**, *12*, 1.
[44] L. Nguyen, C. Thompson, *Trends Microbiol.* **2006**, *14*, 304.
[45] F. Zoulim, *Liver Int*. **2011**, *31*, 111.
[46] R. A. Bright, D. K. Shay, B. Shu, N. J. Cox, A. I. Klimov, *J. Am. Med. Assoc*. **2006**, *295*, 891.
[47] R. N., Ganapathi, M. K. Ganapathi, *Front. Pharmacol*. **2013**, *4*, 1.
[48] G. A. Curt, N. J. Clendeninn, B. Chabner, *Cancer Treat. Rep.* **1984**, *68*, 87.
[49] S. C. Basak, D. Mills, *SAR QSAR in Environ. Res*. **2010**, *21*, 215.
[50] D. S. Peterson, W. K. Milhoust, T. E. Welleams, *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 3018.
[51] S. C. Basak, D. Mills, D. M. Hawkins, *Chem. Biodiversity* **2011**, *8*, 440.
[52] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, *Molecules* **2012**, *17*, 4791.
[53] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *Altern. Lab. Anim.* **2005**, *33*, 445.
[54] F. P. Preparata, M. I. Shamos in *Computational Geometry: An Introduction* (Eds. F. P, Preparata, M. I. Shamos), Springer-Verlag, New York, USA, **1991**; pp. 95–148.
[55] A. P. Worth, A. Bassan, A. Gallegos, T. I. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, M. Vracko, *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance,* ECB Report EUR 21866 EN, European Commission, Joint Research Centre, Ispra, Italy, **2005**, p. 95
[56] "Pharmaceutical Research and Manufacturers of America. Biopharmaceutical Research Industry Profile", **2014**, The full text can be found under: http://www.phrma.org/sites/default/files/pdf/2014_PhRMA_PROFILE.pdf.
[57] O. A. Santos-Filho, A. J. Hopfinger, A. Cherkasov, R. B. de Alencastro, *Med Chem (Shariqah)* **2009**, *5*, 359.
[58] S. C. Basak, A. K. Bhattacharjee, M. Vracko, *Curr Comp Aided Drug Des.* **2015**, *11*, 197.
[59] R. Zanni, M. Galvez-Llompart, C. Morell, N. Rodríguez-Henche, I. Diaz-Laviada, M. Carmen Recio-Iglesias, R. Garcia-Domenech, J. Galvez, *Plos One* **2015**, *10*, e0124244.
[60] J. Wang, D. Land, K. Ono, J. Galvez, W. Zhao, P. Vempati, J. W. Steele, A. Cheng, M. Yamada, S. Levine, P. Mazzola, G. M. Pasinetti, *Plos One* **2014**, *9*, e92750.
[61] M. Lajiness in *Computational Chemical Graph Theory* (Ed. D. H. Rouvray), Nova, New York, **1990**, pp. 299–316.
[62] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal, *Discrete Appl. Math*. **1988**, *19*, 17.
[63] G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h, R. Buelow, *Nat. Biotechnol*. **1998**, *16*, 748.
[64] R. Hoffmann, V. L. Minkin, B. K. Carpenter, *HYLE* **1997**, *3*, 3.
[65] S. Majumdar, S. C. Basak, *Curr. Comp. Aided Drug Des.* **2016**, *12*, 294.
[66] S. C. Basak, M. Vracko, F. A. Witzmann, *Curr. Comp. Aided Drug Des.* **2016**, *12*, 259.
[67] S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, S. P. Bradbury, *Environ. Toxicol. Chem*. **1998**, *17*, 1056.
[68] S. C. Basak in *Advances in Mathematical Chemistry and Applications, Vol 1* (Eds. S. C. Basak, G. Restrepo, J. L. Villaveces), Bentham eBooks, Elsevier & Bentham Science Publishers, **2016**. pp. 3–23.