




Applications of Multidimensional Space of Mathematical Molecular Descriptors in Large-Scale Bioactivity and Toxicity Prediction- Applications to Prediction of Mutagenicity and Blood-Brain Barrier Entry of Chemicals

 Subhash C. Basak,^{1,*}  Subhabrata Majumdar,²  Claudiu Lungu³

¹ Department of Chemistry and Biochemistry, University of Minnesota, Duluth, 1802 Stanford Avenue, Duluth MN 55811, USA

² Center for Interdisciplinary Research and Education, Kolkata, India

³ Department of Surgery, Country Emergency Hospital Braila, 810249 Brăila, Romania

* Corresponding author's e-mail address: sbasak@d.umn.edu

RECEIVED: April 12, 2021 * REVISED: June 10, 2021 * ACCEPTED: June 10, 2021

— THIS PAPER IS DEDICATED TO PROF. MILAN RANDIĆ ON THE OCCASION OF HIS 90TH BIRTHDAY, AND TO THE MEMORY OF PROF. MIRCEA DIUDEA —

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, Chemistry, Quantum Mechanics and Reductionism^[1]

Unless you try to do something beyond what you have already mastered, you will never grow.

Ralph Waldo Emerson

Abstract: In this chapter, we review our QSAR research in the prediction of toxicities, bioactivities and properties of chemicals using computed mathematical descriptors. Robust statistical methods have been used to develop high quality predictive quantitative structure-activity relationship (QSAR) models for the prediction of mutagenicity and BBB (blood-brain barrier) entry of two large and diverse sets chemicals.

Keywords: Quantitative Structure-Activity Relationship (QSAR).

INTRODUCTION

A large part of biological and toxicological processes is guided by the interaction of small molecules with their appropriate biological targets. For example, many drugs are small molecules that interact with specialized enzymes/receptors in appropriate compartments and thereby produce effect(s) that bring a pathologically perturbed biological system back to a healthy state.^[2,3] In toxicology and ecotoxicology, chemicals generated by

natural and anthropogenic processes enter the physiological or environmental milieu and precipitate undesirable effects. Such biological properties of molecules, beneficial or deleterious, can be looked upon as the result of ligand-biotarget interactions and may be expressed by the relationship:^[3,4]

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

where BR represents the normal biological or pathological/toxicological response produced by the biological system,

and B represents the relevant biochemical part of the target system which is perturbed by 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 BR from chemical to chemical when the biological system, B , is practically the same and there is alternation only in the structure of the ligand. Under such conditions Eq.(1) approximates to:

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

MATHEMATICAL CHARACTERIZATION OF STRUCTURE

*Computers are incredibly fast, accurate, and stupid.
Human beings are incredibly slow, inaccurate, and brilliant.
Together they are powerful beyond imagination."*

Albert Einstein

Molecular structure can be represented and quantified by various methods, e.g., topological, three-dimensional (3D) or geometrical and quantum chemical approaches, to name just a few. In this article, the majority of descriptors used to the formulation quantitative structure-activity relationship (QSAR) models are based on topological or graph theoretic formalism. We also used some 3D and quantum chemical descriptors which will be described below.

Graph Theoretical Formalism

A graph, G , is defined as an ordered pair consisting of two sets V and R , $G = [V(G), R]$, where $V(G)$ represents a finite nonempty set of points, and R is a binary relation defined on the set $V(G)$. The elements of V are called vertices and the elements of R , also symbolized by $E(G)$ or E , are called edges. When representing molecular structures as graphs, V represents the set of atoms and E represents the set of bonds present in the molecule. The set E is not limited only to covalent bonds, and may symbolize any type of bonds, viz., covalent, ionic, or hydrogen bonds. Basak *et al.*^[5] emphasized that weighted pseudographs constitute a versatile class of models for the representation of a wide range of chemical species. In depicting a molecule by a connected graph $G = [V(G), E(G)]$, $V(G)$ may contain either all atoms present in the empirical formula or only non-hydrogen atoms. Hydrogen-filled graphs are preferable to hydrogen-suppressed graphs when hydrogen atoms are involved in critical steric or electronic interactions intramolecularly or intermolecularly or when hydrogen atoms have different physicochemical properties due to differences in their bonding neighborhoods. Most stable chemical species can be represented by simple graphs or multigraphs. The structural formula, labeled hydrogen-

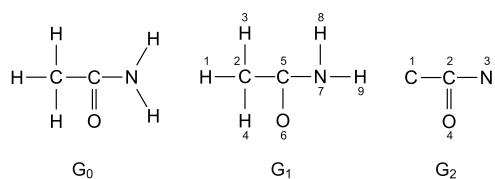


Figure 1. Structural formula (G_0), labeled hydrogen-filled graph (G_1), and labeled hydrogen-suppressed graph (G_2) of acetamide.

filled and the labeled hydrogen-suppressed graphs for acetamide are shown in Figure 1.

Topological Indices (Set # 1)

Graphs can be characterized by graph invariants. Numerical invariants of graphs are called topological indices.^[6] Many topological indices can be conveniently derived from various matrices including the adjacency matrix $A(G)$ and the distance matrix $D(G)$ of a chemical graph G . These matrices are usually constructed from labeled graphs of hydrogen-suppressed molecular skeletons^[6–8] (see Supplementary material for details). Dr. Harry Wiener^[9] was the first to put forward the idea of a structural index (topological index) for the estimation of properties of molecules from their structure. Other indices derived from the adjacency matrix include the Hosoya index $Z^{[10]}$ the zero order connectivity index, ${}^0\chi$,^[11] Randić's connectivity index, ${}^1\chi$,^[12] and the generalized connectivity index ${}^h\chi$ and its variants derived by Kier, Murray, Randić, and Hall.^[11] As a further extension, electrotopological state indices are calculated using the method of Kier and Hall.^[13] This class of indices combines the electronic nature and the topological neighborhood of each skeletal atom in the molecule.

In many cases, *invariants* have been used to answer a specific structural question or to "quantify" a hitherto "qualitative" concept.^[14] For example, in developing the branching index, ${}^1\chi$, Randić^[12] asked the question: *Which of a given collection of molecules is the most branched?* ${}^1\chi$ puts the molecules in a numerical scale and answers this question. In developing the various information theoretic indices, Basak and coworkers^[14–17] asked the question: *Which of a collection of chemicals is the most complex?* The answer to this question came from the different information theoretic indices, IC_r , SIC_r , and CIC_r , which provide quantitative scales for molecular complexity. Of course, complexity of a molecule is not uniquely defined; it varies with the order, r , of neighborhood of atoms considered (*vide infra*), and the specific equivalence relation used to decompose the set of atoms into disjoint subsets.

Information-theoretic topological indices are calculated by the application of information theory to chemical graphs. It is to be noted that information content of a graph G is not uniquely defined. It depends on the way

the set A is derived from the molecular graph G as well as on the equivalence relation which partitions A into disjoint subsets A_i . For example, when A constitutes the vertex set of a chemical graph G , two methods of partitioning have been widely used: (a) chromatic-number coloring of G , where two vertices of the same color are considered equivalent, and (b) determination of the orbits of the automorphism group of G where after vertices belonging to the same orbit are considered equivalent.

Based on an equivalent relation the information content of a molecular graph can be computed by Shannon's relation:^[18]

$$IC = -\sum_{i=1}^h p_i \log_2 p_i \quad (3)$$

IC_r ($r = 0, 1, 2, \dots, \rho$; ρ is the radius of the graph G) can be calculated for different order of neighborhood of the vertices of G .^[16]

Basak, *et al.*^[15] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as in Eq. (4):

$$SIC_r = IC_r / \log_2 n \quad (4)$$

where IC_r is calculated from Eq. (3) and n is the total number of vertices of the graph.

Another information-theoretic invariant, complementary information content (CIC_r) is defined as in Eq. (4):^[17]

$$CIC_r = \log_2 n - IC_r \quad (4)$$

CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r . Bonchev^[19] has pointed out that, in many cases, equivalent vertices in the neighborhood symmetry formalism belong to the same orbits of the automorphism group of the graph. A review of the information-theoretic indices, IC_r , SIC_r , and CIC_r , and their application in QSPR/QSAR/QSTR and QMSA studies is available by Basak.^[3]

The information-theoretic index on graph distance I_b^W is calculated from the distance matrix $D(G)$ of a chemical graph G as follows:^[20]

$$I_b^W = W \log_2 W - \sum g_h \cdot h \log_2 h \quad (5)$$

The mean information index, $\overline{I_b^W}$, is found by dividing the information index, by W .

The **triplet indices**, developed by Balaban and coworkers^[21] result from a matrix, a main diagonal column vector, and a free term column vector, converting the matrix into a system of linear equations whose solutions are the local vertex invariants. These local vertex invariants are then used in the following operations in order to obtain

the triplet descriptors:

1. Summation, $\sum x_i$;
2. Summation of squares, $\sum x_i^2$;
3. Summation of square roots, $\sum x_i^{1/2}$;
4. Sum of inverse square root of cross-product over edges ij , $\sum_{ij} (x_i x_j)^{-1/2}$;
5. Product, $N(\sum x_i)^{1/N}$.

Basak *et al.*^[3] have divided the topological indices (TIs) into two major groups: topostructural indices (TSIs) and topochemical indices (TCIs). TSIs are topostructural indices which 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, TSIs quantify information regarding the connectivity, adjacency and distances between vertices, ignoring their distinct chemical nature. Topochemical indices (TCIs), 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 TCIs are more complex than the TSIs.

Computation. In summary, a general approach in developing chemodescriptors and biodescriptors is as follows:

- a) Define a structural model,
- b) Associate a graph or matrix to the structural model,
- c) Calculate invariants for use as chemo- or biodescriptors.

The generic procedure of developing such descriptors and their application in QSAR is summarized in Figure 2.

Software used by Basak and coworkers group for the calculation of molecular descriptors includes *POLLY*,^[22] *MolConnZ-Z*,^[23] and *Triplet*.^[24] Supplementary Table S1 gives a more detailed background of the mathematical basis and the list of descriptors most often used by Basak *et al.*

For both the mutagenicity data sets the following 3D and quantum chemical indices were also calculated: V_w (Van der Waals' volume), 3DW (3D Wiener number based on the hydrogen-suppressed geometric distance matrix), 3DWH (3D Wiener number based on the hydrogen-filled geometric distance matrix), E_{HOMO} (Energy of the highest occupied molecular orbital), $E_{\text{HOMO}-1}$ (Energy of the second

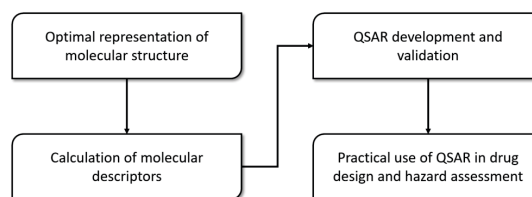


Figure 2. Schematic of the use of topological indices in QSAR.

highest occupied molecular), E_{LUMO} (Energy of the lowest unoccupied molecular orbital), $E_{\text{LUMO}+1}$ (Energy of the second lowest unoccupied molecular orbital, Heat of formation, and Dipole moment. Please see Table S1 in the supplementary material for details.

Topological Indices: Cluj Descriptors Based on Graph-Theoretic Properties

Apart from the adjacency matrix, a number of other matrices may be constructed from the topological structure of atoms and bonds inside a molecule. The Cluj set of chemical descriptors, utilized by Diudea and co-workers in several studies,^[25–32] are based on a number of such matrices. Below we provide brief descriptions of such matrices, details of which can be found in the supplementary material.

In cycle-containing graphs, when the shortest path is replaced by the longest path between two vertices, the maximum path matrix, or the *Detour matrix*, can be constructed:

$$[\Delta]_{ij} = \begin{cases} \delta_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The 3D distance matrices (D3D), in a full analogy with the construction of distance matrices, D , from the topological distances: the entries in 3D are the actual 3D distances between the vertices of a graph which was geometrically optimized (i.e., by a MMX calculus). The Cluj matrices CJD_u and CJ3D_u have been recently proposed by Diudea,^[31] both of which are square but non-symmetric matrices. Variants of these include the Cluj Fragmental Distance matrix and Cluj Fragmental Detour matrix, which are useful when descriptors are meant to represent real (connected) chemical fragments. These graph-theoretical Cluj matrices are calculable as "basic matrices" by TOPOCLUJ.^[33]

Given a property of a specific atom (vertex of the molecular graph), the layer matrix (**LM**) and shell matrix (**SM**) summarize the properties of atoms situated at or within a certain distance k from an atom. Thus, layer/shell matrices are specific to the molecular property, and can be constructed given some vertex property p_i , or a square *info matrix* **M** supplying local/vertex properties as *row sum RS*, *column sum CS* or *diagonal entries* given by the *Walk matrix* (see below). The novel shell matrix provides a partitioning of the entries in a square matrix according to the vertex (distance) partitions in the graph. It means a true decomposition of the property collected in the square matrix in contributions brought by vertices pertaining to shells located at distance k around each vertex. However, the property depends on the vertex-pair relationship. Both layer and shell matrices can be constructed from symmetric

and non-symmetric matrices, including the adjacency matrix and the above Cluj matrices.

Layer matrices are used to derive two topological indices: (i) indices of *centrality* $C(LM)$ and (ii) indices of *centrocomplexity* $X(LM)$. Indices of centrality $C(LM)$ look for the center of a graph and are defined as

$$C(LM)_i = \left[\sum_{k=1}^{ecc} ([\text{LM}]_{ik}^{2k})^{1/(ecc)^2} \right]^{-1} \quad C(LM) = w \sum_i C(LM)_i$$

where ecc is the maximal distance in G (i.e., $\max_{i,j} d(i,j)$) and w is a weighting factor.

Indices of centrocomplexity express the location vs. a vertex of high complexity (e.g., a vertex of high degree or a heteroatom), and are defined as:

$$X(LM)_i = \sum_{k=0}^{ecc} [\text{LM}]_{ik} 10^{-k} \quad X(LM) = w \sum_i X(LM)_i$$

In Supplementary Table S2, we list the Cluj descriptors calculated using the above matrices that are used in the case studies.

In the QSAR of the two properties reported in this paper we also used descriptors calculated by the Software Schrodinger. Please see the Supplementary material for details.

QSAR USING CALCULATED INDICES

The most fundamental and lasting objective of synthesis is not production of new compounds, but production of properties.

Norris Award Lecture, 1968

George S. Hammond

Background of QSARs

Use of molecular descriptors and experimental properties in QSAR may be clearly understood through a formal exposition of the *structure-property similarity principle*—the central paradigm of SAR.^[3,34] Figure 3 represents an empirical property as a function $\alpha: C \rightarrow R$ which maps the set C of compounds into the real line R . A non-empirical SAR may be looked upon as a composition of a description function $\beta_1: C \rightarrow D$ mapping each chemical structure of C into a space of non-empirical structural descriptors (D) and a prediction function $\beta_2: D \rightarrow R$ which maps the descriptors into the real line. When $[\alpha(C) - \beta_2 \beta_1(C)]$ is within the range of experimental errors, we say that we have a good non-empirical predictive model. On the other hand, the property-activity relationship (PAR) is the composition of $\theta_1: C \rightarrow M$ which maps the set C into the molecular property space M and $\theta_2: M \rightarrow R$ mapping those molecular properties into the real line R . PAR seeks to predict one property

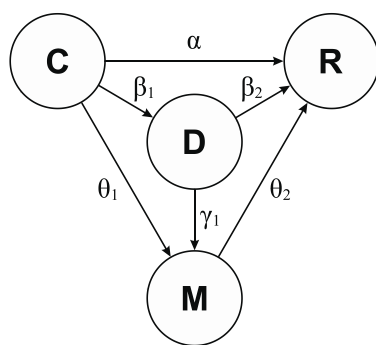


Figure 3. Composition functions for structure-activity relationship (SAR) and property-activity relationship (PAR).

(usually a complex property) of a molecule in terms of another (usually simpler) property.^[35] The latter group of properties may consist either of a number of experimentally determined quantities (e.g. melting point, boiling point, vapor pressure, partition coefficient) or substituent constants or solvatochromic parameters (e.g. steric, electronic, hydrophobic, charge transfer substituent constants, hydrogen bond donor acidity, hydrogen bond acceptor basicity).

PAR using a calculated property (e.g. calculated partition coefficient, $\log P$, octanol-water) may be looked upon as a mapping of θ_2 or γ_1 or β_1 : $C \rightarrow R$, which is a composition of β_1 , γ_1 : $D \rightarrow M$ mapping the descriptor space into the molecular property space (e.g. calculation of $\log P$ from fragments using the additivity rule), and θ_2 , as described above.

Both in drug design and predictive toxicology, SAR can be used to manage a combinatorial explosion. In drug design, one can synthesize many derivatives from a "lead" structure. It is not unusual that one must test 200,000 or more chemicals to discover a molecule that is marketable. The TSCA Inventory contains approximately 86,000 chemicals.^[36] **These substances and their possible metabolites together may result in many thousands of chemical structures.** We need to know many properties and activities (or endpoints) of these chemicals to perform a reasonable risk assessment. Table 1 provides a partial list of endpoints necessary for pharmacological/ toxicological evaluation of chemicals. Although many of the properties listed in Table 4 can be determined experimentally, the combination of these properties and the number of candidate chemicals is a combinatoric explosion! Cost and time limitations will not allow us to test a large fraction of existing chemicals in a rigorous way. Therefore, there is a need to develop procedures which can rapidly screen chemicals for their toxicological properties and allow us to focus scarce resources on chemicals with the greatest potential risk.

Table 1. Important SAR endpoints.

Physicochemical	Pharmacological / Toxicological
Molar volume	Macromolecule level
Boiling point	: Receptor binding (KD)
Melting point	: Michaelis constant (Km)
Vapor pressure	: Inhibitor constant (Ki)
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 (electrophile)	: Algae
	: Invertebrates
	: Fish
	: Birds
	: Mammals
	Organism level (chronic)
	: Bioconcentration
	: Carcinogenicity
	: Reproductive toxicity
	: Delayed neurotoxicity
	: Biodegradation
	Ecosystem level
	: ??

In the sections that follows, we describe the utility of our hierarchical QSAR (HiQSAR) approach^[3] in the prediction of bioactivity/toxicity of chemicals at the levels of enzymes, receptors, cells, and whole animal, as well as properties related to drug action and environmental pollutants from calculated descriptors.

The approach in HiQSAR is to include the more complex and resource intensive descriptors only if they result in significant improvement in the quality of the predictive model. We begin by building a model using only the TS descriptors, followed by the creation of additional models based on the successive inclusion of the hierarchically ranked descriptor classes. In comparing the resulting models, the contribution of each descriptor class is elucidated. In addition, the hierarchical approach enables us to determine whether the higher-level descriptors are necessary in predicting the property or activity under consideration. In situations where these complex descriptors are not useful, we can avoid spending the time required for their calculation. A general scheme for the use of easily calculated molecular descriptors is shown in Figure 4.

Methods for Model Development

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Ronald Fisher

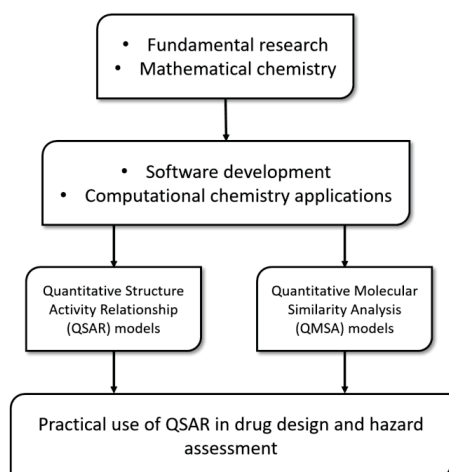


Figure 4. Development and use of topological indices in QSAR and QMSA models.

In much of our earlier work, we developed ordinary least squares (OLS) regression models using the REG procedure of the SAS statistical program. The single goal of OLS regression is to minimize the sample response prediction error, seeking linear functions of the predictors that explain as much variation in each response as possible. However, OLS should never be used in situations where the number of independent variables (descriptors) available to the regression procedure is large with respect to the number of observations (chemicals), which is often the case in the field of computational chemistry. Using OLS under these circumstances can result in chance correlations wherein the predictive quality of the resulting model is vastly overestimated. One approach in circumventing this problem is to reduce the number of independent variables prior to the regression process. We have used the SAS VARCLUS procedure to accomplish this. With this procedure, a set of n descriptors is reduced by dividing them into disjoint clusters which are essentially unidimensional. From each cluster, we select the descriptor that is most correlated with the cluster, as well as any which are poorly correlated with the cluster ($R < 0.70$). This reduced set of descriptors, then, is used with the REG procedure to produce predictive models.

Studies have shown, however, that subsetting of available descriptors followed by OLS regression is inferior to using alternative regression methodologies that retain all available descriptors and deal with rank deficiency in another way.^[37–40] Ridge regression (RR), principal components regression (PCR), and partial least squares (PLS) regression all accomplish this. We have used these three methods, comparatively, in much of our recent work and have found that RR generally outperforms both PLS and PCR.

RR, like PCR, transforms the descriptors to their principal components (PCs) and uses the PCs as descriptors. However, unlike PCR, RR retains all the PCs, and ‘shrinks’ them differentially according to their eigenvalues. As with PCR and RR, PLS also involves new axes in predictor space, however, they are based on the dependent variable as well as the independent.

Validation. For the sake of brevity, we do not report the highly parameterized RR, PCR, and PLS models. Rather, we have reported summary statistics for the models, including the cross-validated R^2 and the prediction sum of squares (PRESS). The cross-validated R^2 is calculated using the leave-one-out approach wherein each compound is removed, in turn, from the data set and the regression is fitted based on the remaining $n-1$ compounds. The cross-validated R^2 mimics the results of applying the final regression to a future compound; large values can be interpreted unequivocally and without regard to the number of compounds or descriptors as indicating that the model will accurately predict the activity of a compound of the same chemical type as those used to calibrate the regression. Although some QSAR proponents routinely recommend partitioning the available data into training and test sets, where the model is developed based on the training set compounds and the activity of the test compounds is then predicted by the model, this is unnecessary and wasteful when one is working with small data sets, and the leave-one-out cross-validation approach should be used. The cross-validated R^2 is defined by:

$$R_{cv}^2 = 1 - \frac{PRESS}{SSTotal} \quad (13)$$

where $SSTotal$ is the total sum of squares. Unlike the conventional R^2 , the cross-validated R^2 may be negative if the model is very poor. It should be stressed that the conventional R^2 is unreliable in assessing modeling predictability when rank deficiency is an issue. In fact, the R^2 value will increase upon the addition of any descriptor, even the irrelevant. In contrast, the cross-validated R^2 will decrease upon the addition of irrelevant descriptors, providing a reliable measure of model quality.

Another statistical metric often utilized in conjunction with our ridge regression studies is the absolute value of t , where t represents the model coefficient divided by its standard error. Those descriptors with large $|t|$ values are known to be important in the model under consideration. Therefore, we use this metric in hopes of gaining some mechanistic insight. It should be noted, however, the no conclusions can be drawn with respect to descriptors associate with small $|t|$ values.

Pre-processing. Routinely, prior to model development, any descriptor with a constant value for all compounds is

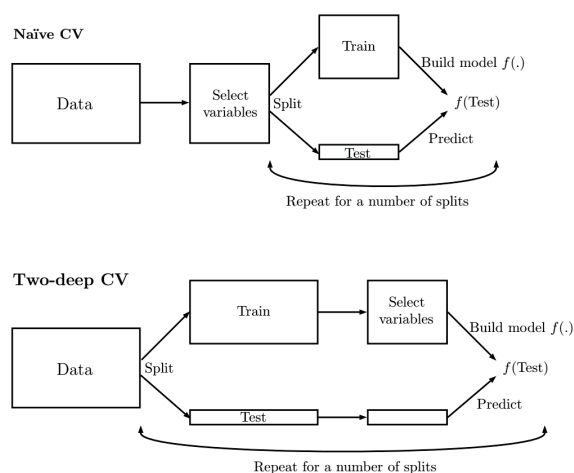


Figure 5. Schematic of naïve vs. two-deep cross-validation.

omitted. In addition, only one descriptor of any perfectly correlated pair (i.e., $r = 1.0$), as identified by the CORR procedure of the SAS statistical package, is retained. Because the variable scales differ from one another by many orders of magnitude, they are typically scaled by the natural logarithm prior to modeling. Some methodological papers^[40] on the proper use of descriptors provide good guidance in the proper use of indices in the formulation of robust QSAR models.

Two-deep validation. A number of statistical and machine learning (ML) methods used frequently in large-scale data analytic applications involve the selection of one or more tuning parameter(s). For example, results from PCR or PLS depends on the selection of a number of top descriptor projections, and RR and sparse methods like LASSO, SCAD^[41] depends on the value of shrinkage parameter. While an intuitive method to select the tuning parameters may be to use performance on the test set, this is wrong and overestimates predictive performance. Since for the evaluation phase the information from test set is being used, the final cross-validation metric values do not, in principle, remain out-of-sample. Instead of this naïve procedure, a two-layer process should be undertaken. After a train-test split, tuning parameters or any iterations in the training phase should be performed only using the training data, and the resulting model should be used to evaluate the test data.^[37,42] Figure 5 details and contrasts this two-step process with the naïve, single-step process.

QSARs for Chemical Mutagens

We report here formulation of QSARs on sets of mutagens:

- A congeneric set of 95 aromatic and heteroaromatic amines originally collated by the group of Corwin Hansch^[43] and
- A structurally diverse data set of 260 mutagens and 260 non-mutagens collated by Basak et al.^[44] from the CRC

Handbook of Identified Carcinogens and Non-carcinogens and consisted of those compounds that had a positive or negative response to the Ames mutagenicity test.^[45] The distribution of chemical classes of the compounds (non-exclusive) in Table 2 summarizes the vast diversity in this dataset and the complexity of the underlying prediction problem.

In one of our past papers with Prof. Diudea,^[46] we explored the use of structural descriptors in toxicity/mutagenicity prediction, the points mentioned above. The two datasets analyzed come from two different representative situations:

- Mutagenic activity of 95 congeneric amines, measured as the log number of revertants per nmol when a sample compound is applied to *S. typhimurium* cultures.
- Binary mutagen/non-mutagen status, determined by the Ames mutagenicity test, of 508 diverse chemical compounds belonging to broad array of structural classes.

For the compounds in each sample we calculated two sets of descriptors, one each to the Basak and Cluj set described earlier. We applied a number of statistical and ML techniques to evaluate the comparative performance of these descriptor sets separately and combined, as well as obtain some possible mechanistic interpretations behind the observed chemical activity.

Table 2. Chemical classes of samples in the 508 compound diverse dataset

Chemical class	Number of compounds
Aliphatic alkanes, alkenes, alkynes	124
Monocyclic compounds	260
Monocyclic carbocycles	186
Monocyclic heterocycles	74
Polycyclic compounds	192
Polycyclic carbocycles	119
Polycyclic heterocycles	73
Nitro compounds	47
Nitroso compounds	30
Alkyl halides	55
Alcohols, thiols	93
Ethers, sulfides	38
Ketones, ketenes, imines, quinones	39
Carboxylic acids, peroxy acids	34
Esters, lactones	34
Amides, imides, lactams	36
Carbamates, ureas, thioureas, guanidines	41
Amines, hydroxylamines	143
Hydrazines, hydrazides, hydrazones, traizines	55
Oxygenated sulfur and phosphorus	53
Epoxides, peroxides, aziridines	25

As we see in the results in Table 3, contrary to the common and intuitive view that novel or more descriptors will help in developing better QSAR models do not always hold. This is the case for the smaller set of compounds, but not for the diverse, larger set. As further observation, both the sparsity based methods — LASSO, SCAD — perform poorly on both datasets. This result indicates the potential presence of high collinearity and lower-dimensional subspaces in the predictor space than individual predictors. These subspaces also encode significant information about the corresponding response variables, as the good performances of PLS indicate.

Table 3. Average and standard deviations (in brackets) of performance measures over 100 random splits for different methods applied on the 508 compound heterogeneous dataset. PCR = RF = Random Forest, GBM = Gradient Boosting Machine

508 compounds data (Area Under Curve)			
Method	Descriptor set used		
	Combined	Basak set	Cluj set
PCR	0.59 (0.055)	0.78 (0.038)	0.58 (0.057)
PLS	0.86 (0.035)	0.58 (0.057)	0.79 (0.038)
Lasso	0.72 (0.048)	0.75 (0.045)	0.63 (0.06)
SCAD	0.57 (0.061)	0.58 (0.059)	0.62 (0.063)
RF	0.81 (0.036)	0.80 (0.042)	0.79 (0.040)
GBM	0.80 (0.04)	0.82 (0.04)	0.75 (0.042)
95 amines data (Mean Squared Prediction Error)			
Method	Descriptor set used		
	Combined	Basak set	Cluj set
PCR	29.1 (13.79)	57.1 (93.83)	76.0 (24.72)
PLS	18.9 (6.03)	19.9 (7.46)	75.7 (24.69)
Lasso	26.9 (9.05)	28.7 (8.83)	72.8 (18.0)
SCAD	25.8 (8.96)	31.8 (21.44)	74.9 (18.32)
RF	17.3 (6.50)	19.0 (6.59)	84.6 (21.74)
GBM	14.8 (5.84)	18.0 (6.30)	74.8 (17.43)

Table 4. Top 5 PCs and their loadings (in brackets) for each dataset. Brackets in the headings indicate the percentage of variance explained by each PC

95 amines data			
PC1 (15 %)	PC2 (14.1 %)	PC3 (12 %)	PC4 (9.3 %)
E_tor (0.96)	vsurf_DW23 (0.95)	density (-0.93)	GCUT_SlogP_0 (0.96)
vsurf_DW23 (-0.25)	E_tor (0.23)	GCUT_SlogP_0 (-0.19)	density (-0.19)
GCUT_SlogP_0 (0.06)	GCUT_SlogP_0 (0.15)	vsurf_ID2 (0.12)	vsurf_DW23 (-0.14)
vsurf_ID3 (-0.04)	density (-0.07)	vsurf_ID3 (0.12)	E_tor (-0.09)
vsurf_ID4 (-0.04)	vsurf_ID2 (-0.06)	vsurf_ID4 (0.11)	vsurf_CP (0.05)
508 compounds data			
PC1 (12.5 %)	PC2 (10.3 %)	PC3 (7.1 %)	PC4 (6.9 %)
E_ele (-0.42)	E_ele (-0.42)	E_ele (-0.35)	E_vdw (-0.73)
vsurf_EWmin1 (-0.43)	vsurf_EWmin1 (-0.27)	vsurf_EWmin1 (-0.33)	E_nb (-0.48)
vsurf_EWmin2 (-0.4)	vsurf_EWmin2 (-0.25)	E_vdw (-0.31)	E_ele (0.34)
vsurf_EWmin3 (-0.3)	vsurf_DW13 (-0.21)	vsurf_EWmin2 (-0.31)	vsurf_EWmin1 (0.19)
vsurf_DW13 (-0.17)	vsurf_EWmin3 (-0.18)	E_nb (-0.27)	vsurf_EWmin2 (0.18)

Looking at the descriptors that have highest loading in Principal Component Analysis (PCA) of the combined feature matrix reveal some interesting observations (Table 4). Contrary to the poor performance of PCR, the top 4 PCs explain high proportions of overall variability (37 % for 95 amines, 50 % for 508 compounds). The top loadings are dominated by the Diudea set of graph connectivity descriptors. Specifically, a few descriptors with high loadings across multiple PCs include E_ele, E_tor, vsurf_DW23, density, and GCUT_SlogP_0. Looking at their mechanistic roles, E_ele is a measure of the stored potential energy in the 3D structure of the chemical compound, and the others are absorption–distribution–metabolism–excretion (ADME) features. However, the comparative poor predictive performance of the PCR model indicates that the effect of these features on the specific mutagenic activity being modeled is limited. This analysis underlines the importance of using such deeper statistical reasoning while interpreting outputs of QSAR models.

QSAR of Blood-Brain Entry of Chemicals

Continuing our collaboration with Prof. Diudea,^[37] in a subsequent paper we focused on the important problem of developing QSAR models for measuring Blood-Brain Barrier (BBB) permeability of chemicals, the collection of BBB data on a diverse set of 415 chemicals taken from Li et al.^[47] The binary dependent variable was whether a compound is BBB-permeable or not. A schematic representation of the BBB with its luminal and abluminal sides is shown in Figure 6 below. Building up on our work on obtaining parsimonious representation of chemical spaces, in this work we attempted to see if it is possible to use state-of-the-art ML methods in tandem with a subset of descriptors can provide to develop well-performing models.

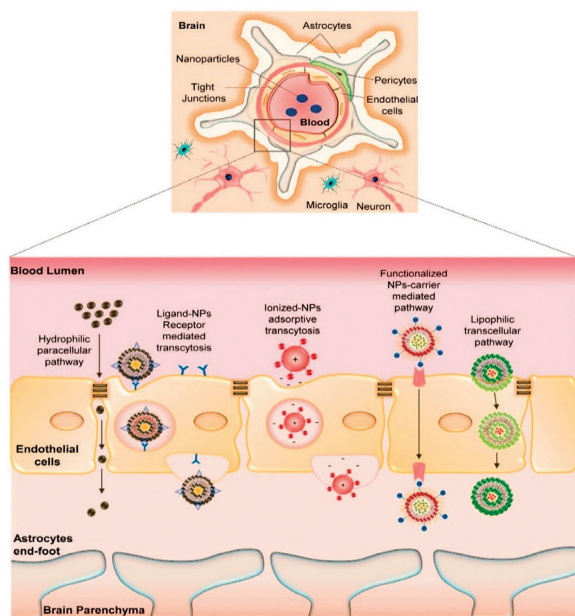


Figure 6. A schematic representation of the essential features of the blood-brain barrier (BBB) showing the two luminal and abluminal membranes that separate the blood from the brain. Shared from Ref. [48] Under Creative Commons Attribution License 4.0.

The results in Figure 7 plot four metrics — AUC, lift in percent of positive samples captured in top 20 % highest predicted probabilities, sensitivity, and specificity. We discovered that the combined and Basak set perform well in terms of parsimony. A random forest model composed of only top 5 % important descriptors perform better than the full model in both these cases. However this does not hold for the Cluj set. Taking a further look at the important descriptors selected by both methods (Table 5), we see that the top descriptors for the Cluj and combined set are very similar. Given the similar prediction performances of all the descriptor sets (see Table 2 in Ref. [37]), this suggests a high degree of shared information among the Basak and Cluj set of descriptors.

Analysis of mechanistic interpretation of the top descriptors bring up some interesting points. For example, most top descriptors from the Basak set relate to structural heterogeneity within atomic neighborhoods (IC indices), and presence of multiple bonds and/or heteroatoms (triplet descriptors). Influential indices from the Cluj set are related to topological distances and connectivity. A number of descriptors from both sets—ANZ4, AZN4, ANZ5, and DN2N3, ALOGP3—relate to activity properties relevant to BBB, such as polarity and hydrophobicity. Previous studies corroborate these properties as predictive of BBB permeability.^[47,49,50] These outcomes are also in line with the findings of the original analysis of Li et al.^[47]

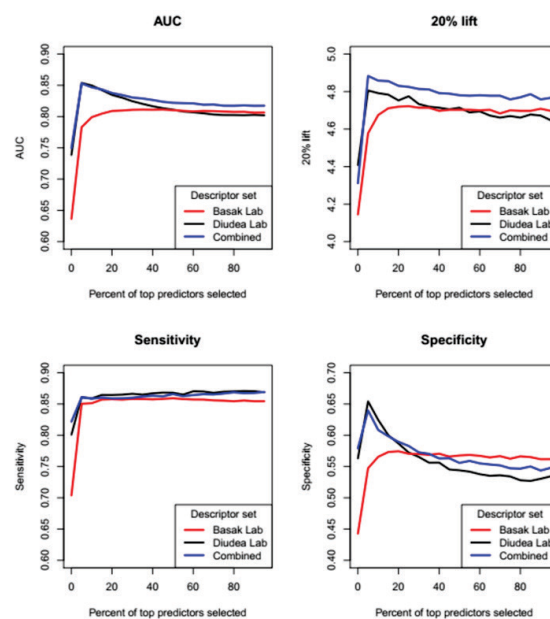


Figure 7. BBB activity prediction performance comparisons for all-descriptor vs. partial models. First a random forest model is fit with all descriptors, from which descriptor importance behind the activity are obtained. Then a second set of models are trained with top x % important descriptors only, $x = 5, 10, \dots, 95$.

GENERAL DISCUSSION

All generalizations are dangerous, even this one.

Alexandre Dumas

Prediction is very difficult, especially if it's about the future.

Niels Bohr

In this paper, we reviewed robust QSAR models derived by Basak and Diudea groups in the prediction of important biological properties like mutagenicity and BBB permeability of chemicals.

Mutagenicity testing is important for human health risk assessment,^[51] pharmaceutical drug design^[52] and ecological risk assessment of chemicals.^[53] Therefore, assessment of mutagenicity of many chemicals, both in new drug discovery protocols and risk estimation of environmental pollutants, is done routinely by the regulatory agencies all over the world. Topological indices can be calculated fast for any chemical structure, real or hypothetical. High quality QSAR models of mutagenicity, particularly those on the structurally diverse set of 508 mutagens, may find practical application in drug design and environmental protection.

The blood-brain barrier is a unique biological barrier critical for the protection of the brain from the entry of undesirable chemicals present in the blood. BBB is

Table 5. Top descriptors for BBB permeability prediction

Basak lab			
Variable	Importance	Variable	Importance
IC1	2.23	IC4	1.32
IC2	1.90	SIC3	1.27
ANZ4	1.58	AZN4	1.22
SIC1	1.56	ANZ5	1.21
DN2N3	1.54	SIC2	1.20
Diudea lab (Cluj)			
Variable	Importance	Variable	Importance
PSA	2.51	PEOE1	1.83
Sum.of.topological.distances.between.O..O	2.35	E.state	1.66
E.state.topological.parameter	2.20	Superpendentic	1.49
ALOGP3	2.06	Topological.charge.index.of.order.5	1.31
Sum.of.topological.distances.between.N..O	1.99	PEOE12	1.28
Combined			
Variable	Importance	Variable	Importance
Sum.of.topological.distances.between.O..O	2.30	E.state	1.31
E.state.topological.parameter	2.22	Sum.of.topological.distances.between.N..O	1.21
PSA	1.56	Molecular.electrotopological.variation	1.12
Superpendentic	1.44	PEOE1	0.99
ALOGP3	1.37	PEOE12	0.87

implicated in the design of psychoactive drugs,^[54,55] assessment of potential neurotoxicity of industrial chemicals and pollutants.^[56]

As shown in Figure 6, the tight junctions (TJs) of endothelial cells control the entry and efflux of substances in and out of the BBB. The passage of molecules across BBB is based on their physicochemical properties like lipophilicity, ionization, polarity, etc.^[54,55] Our QSAR studies on the BBB entry of a diverse set of chemicals (Table 5) shows that easily calculated topological descriptors are capable of quantifying aspects of molecular structure which are relevant for the estimation of BBB transport of chemicals.

It is interesting to note that for both properties, mutagenicity, and BBB entry, the two groups of molecular descriptors used by the Basak group and the Diudea team gave QSAR models of similar quality. When QSARs are developed for congeneric sets of structures, a few simple descriptors may suffice, But for diverse chemical sets one needs a diversity of descriptors for the formulation of good QSAR models.^[57] In the case of both the two mutagenicity data sets and the BBB set augmentation of the number of descriptors by combining the Basak and Diudea set of descriptors did not improve the model quality. It is tempting to speculate that for the two biological endpoints the two sets of descriptors were probably qualifying very similar aspects of molecular structure needed for QSAR formulation. **That is why the quality of the QSARs reached a plateau.** This does not discourage the formulation of novel indices, but to be recognized as sufficiently novel such new descriptors must be able to quantify aspects of molecular structure not characterized by the already existing indices.

Acknowledgment. The authors would like to dedicate this paper to Professors Milan Randić and Mircea Diudea for their outstanding contributions to the field of mathematical chemistry and its applications over many decades of their pioneering research on diverse topics of the field.

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: <https://doi.org/10.5562/cca3772>.

PDF files with attached documents are best viewed with Adobe Acrobat Reader which is free and can be downloaded from [Adobe's web site](https://www.adobe.com/acrobat/).

REFERENCES

- [1] H. Primas, *Chemistry, Quantum Mechanics and Reductionism*, Springer-Verlag, Berlin, **1981**.
<https://doi.org/10.1007/978-3-662-11314-1>
- [2] *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 8th ed. (eds.: A. G. Gilman, T. W. Rall, Alan S. Nies, Palmer Taylor) Pergamon Press, New York, **1990**.
- [3] S. C. Basak, *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 449–462.
<https://doi.org/10.2174/15734099113096660041>
- [4] C. Hansch, A. Leo, *Exploring QSAR: Volume 1: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society; 1st ed., **1995**.
- [5] S. C. Basak, V. R. Magnuson, G. J. Niemi and R. R. Regal, *Discrete Appl. Math.* **1988**, *19*, 17–44.
[https://doi.org/10.1016/0166-218X\(88\)90004-2](https://doi.org/10.1016/0166-218X(88)90004-2)

- [6] N. Trinajstić, *Chemical Graph Theory*, Vols. I. & II., CRC Press, Boca Raton, Florida **1983**.
- [7] S. C. Basak, *Information-theoretic indices of neighborhood complexity and their applications*, in: (Eds. J. Devillers, A. T. Balaban) *Topological Indices and Related Descriptors in QSAR and QSPR*, 563, Gordon and Breach Science Publishers, The Netherlands **1999**, 563–593.
- [8] D. Janežič, A. Miličević, S. Nikolić, N. Trinajstić, *Graph-Theoretical Matrices in Chemistry*, University of Kragujevac and Faculty of Science Kragujevac, Kragujevac, Serbia, **2015**.
<https://doi.org/10.1201/b18389>
- [9] H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
<https://doi.org/10.1021/ja01193a005>
- [10] H. Hosoya, *Bull. Chem. Soc. Jpn.* 1971, *44*, 2332–2339. <https://doi.org/10.1246/bcsj.44.2332>
- [11] L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, **1976**.
- [12] M. Randić, *J. Am. Chem. Soc.* **1975**, *7*, 6609–6615.
<https://doi.org/10.1021/ja00856a001>
- [13] L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotological State*, Academic Press, San Diego, CA, **1999**.
- [14] S. C. Basak, *HYLE – International Journal for Philosophy of Chemistry* **2013**, *19*, 3–17.
- [15] S. C. Basak, A. B. Roy, J. J. Ghosh, *Study of the structure-function relationship of pharmacological and toxicological agents using information theory*, in: (Eds.: X. J. R. Avula, R. Bellman, Y.L. Luke, A. K. Rigler) University of Missouri-Rolla, Rolla, Missouri, USA, **1979**, p. 851.
- [16] A. B. Roy, S. C. Basak, D. K. Harriss, V. R. Magnuson, Neighborhood complexities and symmetry of chemical graphs and their biological applications, In: *Mathematical Modelling in Science and Technology*, (Eds.: X. J. R. Avula, R. E. Kaman, A. I. Lapis, E. Y. Rodin, Pergamon Press, **1984**, pp. 745–750.
<https://doi.org/10.1016/B978-0-08-030156-3.50138-7>
- [17] S. C. Basak, V. R. Magnuson, *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501–503.
- [18] C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
<https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [19] D. Bonchev, *Information theoretic indices for characterization of chemical structures*, Research Studies Press, Letchworth, Hertfordshire, U.K., **1983**.
- [20] D. Bonchev, N. Trinajstić, *J. Chem. Phys.* **1977**, *67*, 4517. <https://doi.org/10.1063/1.434593>
- [21] A. T. Balaban, T. S. Balaban, *J. Math. Chem.* 1991, *8*, 383–397. <https://doi.org/10.1007/BF01166951>
- [22] S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY v. 2.3: **1988**; Copyright of the University of Minnesota, USA.
- [23] MolconnZ, Version 4.05, **2003**; Hall Ass. Consult.; Quincy, MA, USA.
- [24] S. C. Basak, G. D. Grunwald, A. T. Balaban, TRIPLET, Copyright of the University of Minnesota, **1993**.
- [25] M. V. Diudea, I. Gutman, L. Jäntschi, *Molecular topology*, Nova, New York, **2002**.
- [26] M. V. Diudea, M. S. Florescu, P. V. Khadikar, *Molecular Topology and Its Applications*, Eficon, Bucharest, **2006**.
- [27] D. Janežič, A. Miličević, S. Nikolić, N. Trinajstić, *Graph Theoretical Matrices in Chemistry*, *Math. Chem. Monographs*, University of Kragujevac, **2007**.
- [28] M. V. Diudea, M. Topan, A. Graovac, Layer matrices of walk degrees. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1071–1078. <https://doi.org/10.1021/ci00021a006>
- [29] M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 292–299. <https://doi.org/10.1021/ci960037w>
- [30] M. V. Diudea, *J. Chem. Inf. Comput. Sci.* 1994, *34*, 1064–1071. <https://doi.org/10.1021/ci00021a005>
- [31] M. V. Diudea, *MATCH Commun Math. Comput. Chem.* **1997**, *35*, 169–183.
- [32] C. N. Lungu, C-C Chemokine receptor type 3 inhibitors: bioactivity prediction using local vertex invariants based on thermal conductivity layer matrix, *STUDIA UBB CHEMIA*, LXIII, *1*, **2018**, p. 177–188. <https://doi.org/10.24193/subchem.2018.1.13>
- [33] O. Ursu, M. V. Diudea, TOPOCLUJ software program, Cluj, Romania: Babes-Bolyai University, **2005**.
- [34] M. Johnson, S. C. Basak, G. A. Maggiora, *Math Comp. Model.* **1988**, *11*, 630–634.
[https://doi.org/10.1016/0895-7177\(88\)90569-9](https://doi.org/10.1016/0895-7177(88)90569-9)
- [35] S. C. Basak, G. J. Niemi, G. D. Veith, *J. Math. Chem.* **1991**, *7*, 243–272.
<https://doi.org/10.1007/BF01200826>
- [36] *Models and Tools Developed by EPA to Assess Hazard under TSCA*, <https://www.epa.gov/tsca-screening-tools/models-and-tools-developed-epa-assess-hazard-under-tsca> (accessed 30 March 2021).
- [37] S. Majumdar, S. C. Basak, Claudiu N. Lungu, Mircea V. Diudea, G. D. Grunwald, *Mol. Inf.* 2019, *38*, 1800164.
<https://doi.org/10.1002/minf.201800164>
- [38] D. M. Hawkins, S. C. Basak, X. Shi, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663–670.
<https://doi.org/10.1021/ci0001177>
- [39] D. M. Hawkins, S. C. Basak, D. Mills. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.
<https://doi.org/10.1021/ci025626i>
- [40] S. C. Basak, S. Majumdar, *Curr Comput Aided Drug Des* **2015**, *11*, 2–4.
<https://doi.org/10.2174/157340991101150722142144>
- [41] J. Fan, R. Li., *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
<https://doi.org/10.1198/016214501753382273>

- [42] D. Hawkins, S. Basak, D. Mills, *Environ. Toxicol. Pharmacol.* **2004**, *16*, 37–44.
<https://doi.org/10.1016/j.etap.2003.09.001>
- [43] A. K. Debnath, G. Debnath, A. J. Shusterman, C. Hansch, *Environ. Mol. Mutagen.* **1992**, *19*, 37–52.
<https://doi.org/10.1002/em.2850190107>
- [44] S. C. Basak, S. Bertelsen, G. D. Grunwald, *Toxicol. Lett.* **1995**, *79*, 239–250.
[https://doi.org/10.1016/0378-4274\(95\)03375-U](https://doi.org/10.1016/0378-4274(95)03375-U)
- [45] J. V. Soderman, *CRC Handbook of Identified Carcinogens and noncarcinogens: Carcinogenicity-Mutagenicity Database*, Vol. I, CRC Press, Boca Raton, Florida, **1982**.
- [46] S. Majumdar, S. C. Basak, C. N. Lungu, M. V. Diudea, G. D. Grunwald, *SAR QSAR Environ. Res.* **2018**, *29*, 579–590.
<https://doi.org/10.1080/1062936X.2018.1496475>
- [47] H. Li, C. W. Yap, C. Y. Ung, *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384. <https://doi.org/10.1021/ci050135u>
- [48] M. Zeeshan, M. Mukhtar, Q. Ul Ain et al. In: *Pharmaceutical Formulation Design - Recent Practices*, **2019**. <https://doi.org/10.5772/intechopen.83040>
- [49] H. Pajouhesh, G. R. Lenz, *NeuroRx.* **2005**, *2*, 541–553.
<https://doi.org/10.1602/neurorx.2.4.541>
- [50] B. Hemmateenejad, R. Miri, M. A. Safarpour, A. R. Mehdipour, *J. Comput. Chem.* **2006**, *27*, 1125–1135.
<https://doi.org/10.1002/jcc.20437>
- [51] World Health Organization Report - Harmonization Project DRAFT Document for Public and Peer Review, 14 MUTAGENICITY TESTING FOR CHEMICAL RISK 15 ASSESSMENT (Accessed 8 April 2021).
https://www.who.int/ipcs/methods/harmonization/areas/mutagenicity_testing_draft.pdf
- [52] Food and Drug Administration, GUIDANCE DOCUMENT, S2(R1) *Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use* (Accessed 8 April 2021).
<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/s2r1-genotoxicity-testing-and-data-interpretation-pharmaceuticals-intended-human-use>
- [53] G. R. Verheyen, K. Van Deun, S. Van Miert, J. Genet. *Genome Res.* **2017**, *4*, 029.
<https://doi.org/10.23937/2378-3648/1410029>
- [54] H. Pajouhesh, G. R. Lenz, *NeuroRx.* **2005**, *2*, 541–553.
<https://doi.org/10.1602/neurorx.2.4.541>
- [55] W. M. Pardridge, *NeuroRx.* **2005**, *2*, 3–14.
<https://doi.org/10.1602/neurorx.2.1.3>
- [56] Guidelines for Neurotoxicity Risk Assessment - US Environmental Protection Agency (USEPA) (Accessed 8 April 2021) https://www.epa.gov/sites/production/files/2014-11/documents/neuro_tox.pdf
- [57] S. C. Basak, S. Majumdar, *Curr. Comput.-Aided Drug Des.* **2016**, *12*, 84–86.
<https://doi.org/10.2174/157340991202160713190446>