**Current Computer Aided Drug Design**

**Title of the Paper )**

**Estimation of mutagenicity of chemicals from their calculated molecular descriptors: A case study with structurally homogeneous versus diverse data sets**

**Subhash C. Basak 1,\*, Subhabrata Majumdar 2**

1 *Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*; E-Mail: sbasak@nrri.umn.edu

2 *School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA*; E-Mail: majum010@umn.edu

**\*** Author to whom correspondence should be addressed; E-Mail: sbasak@nrri.umn.edu;   
Tel.: +1 (218)720-4230; Fax: +1 (218)720-4328.

**Abstract:** This is the abstract section. One paragraph only (Maximum 200 words).

**Keywords:** Mutagenicity; interrelated two-way clustering; topological indices; ridge regression; hierarchical QSAR

**1. Introduction**

Risk assessment of chemicals is often carried out using very little of no experimental data. In 1984, the National Research Council evaluated the availability of toxicity data on industrial chemicals and concluded that many of these chemicals have very few or no test data [l]. Such a situation is prevalent also with substances identified in industrial effluent, hazardous waste sites and environmental monitoring surveys [2]. In USA, the Toxic Substances Control Act (TSCA) inventory currently has about 86, 000 entries and the list is growing by nearly 3,000 per year [3, 4]. Of these 3,000 chemicals submitted yearly to the United States Environmental Protection Agency (USEPA) for the pre-manufacture notification (PMN) review process, more than 50% have no experimental data at all, less than 15% have empirical mutagenicity data, and about 6% have experimental eco-toxicological and environmental fate data [2].

Mutagenicity is one important toxicity data essential both environmental protection as well as new drug discovery. Identification of potential mutagenicity of environmental pollutants is important for the protection of human and ecological health. In the realm of drug discovery, early detection of mutagenicity of drug candidates can help in the effective allocation of resources for the expensive drug discovery protocol which costs on the average US $ 1.8 billion and takes 10-15 years to discover a new life saving drug Experimental determination of mutagenicity of all possible candidate chemicals, both for environmental protection and drug design, can be very expensive. Therefore, potential mutagenicity of chemicals from quantitative structure-activity relationship (QSAR) models has been accepted for the evaluation of chemicals in lieu of experimental mutagenicity data [7].

QSARs for any biological endpoint of a set of chemicals can be formulated either from other experimental properties or molecular descriptors calculated from molecular structure without the input of any other experimental data. The frequently used descriptors for QSAR are topological indices, substructures, and quantum chemical descriptors [7, 8]. For large sets of structures, high level quantum chemical descriptors can be very resource intensive. Alternately, descriptors derived from topological aspects of chemical structures, e.g. topological indices and different types of substructures, have found successful applications in numerous good quality QSAR studies. For a recent summary of the topic, please see the review by Basak [9, 10]. Ref # 10:. Mathematical descriptors in the prediction of property, bioactivity, and toxicity of chemicals from their structure: A chemical-cum-biochemical Approach, Basak, S. C., Current Computer Aided Drug Design, 2013, 9, 449-462.

There are a few critical needs for the development of QSARs: a) Good quality and sufficiently large property/activity/ toxicity database, b) Computed molecular descriptors which quantify aspects of molecular structure associated with the toxicity or biological activity of interest, and c) Proper methods for model building [10]. In previous studies, Basak *et al* [8] reviewed the results of various QSAR studies where topostructural (TS), topochemical (TC), geometrical (3-D), and quantum chemical (QC) indices were used in QSAR building in a graduated manner. Results of such hierarchical QSARs (HiQSARs) showed that in most cases a combination of TS and TC indices gave the best models. The addition of 3-D or QC descriptors to the set of independent variables did not make much improvement in model quality [10]. The currently available software like PaDEL [10], Dragon [11], MolconnZ [12], POLLY [13], APProbe [14] are capable of calculating a large number of topological and substructural chemodescriptors. But, in many cases the situation becomes “rank deficient” because there are many more available descriptors as compared to the number of data points to be modeled. So, proper statistical methods of variable selection, QSAR formulation, and model validation have to be followed for best results.

This paper has a two-fold objective: 1) Use a relatively new method, called Interactive Two-way Clustering (ITC) in developing QSAR for the prediction of mutagenicity of two data sets, viz., a homogeneous set of 95 aromatic and heteroaromatic amines and a structurally diverse set of 508 chemicals, and 2) Use the HiQSAR approach formulated by Basak *et al* [8] in evaluating the relative effectiveness of TS, TC, 3-D, and QC descriptors in the development of high quality QSARs for these two data sets.

The experimental section below should come here.

**2. Results and Discussion**

The work of Majumdar *et al* focused on the use of ITC as a variable selection method in high-dimensional Chemometrics setup. Here we apply that methodology in a congeneric dataset instead to check for a possible increase of prediction accuracy. Since there are 4 types of variables, we apply a hierarchical approach (See similar studies by Basak *et al* [?,?,?]). Ridge regression is applied for classification purpose on the full and ITC-selected sets of variables to check the effect of variable selection, if any, in this scenario.

*2.1. Analysis*

The analyses involving the ITC methodology was done in MATLAB, version R2010a [matlab], while normal ridge regressions were done in the free statistical softward R [rversion]. In the putputs, there is a clear jump of predictive ability from the TS only model to the TS+TC model. However, adding 3D and QC descriptors do not improve the predictions. This agrees with the findings of Basak *et al* in the several Hierarchical QSAR studies mentioned above [?,?,?] that when TS and TC descriptors are already being used, 3D and QC descriptors seem to add little to the performance of a predictive model to explain activity of chemical compounds (Table 1, top half).

However, the performance of the same models change when we do the analysis on a more homogenous dataset. In general the percentage of correct prediction improves, and these percentages are very close to each other for hierarchical descriptor sets. Moreover, variable selection improves the prediction performance for all the models, but in this case also there is no visible effect of the hierarchy of variables (Table 1, bottom half).

**Table 1.** Comparison of model performances for diverse and congeneric datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset used** | **Predictive model** | **Type of predictor used** | **No. of predictors** | **Correct classification %** | **Sensitivity** | **Specificity** |
| **508 compound diverse dataset** | Ridge regression  [hawk] | TS | 103 | 53.14 | 52.34 | 53.97 |
| TS+TC | 298 | 76.97 | 83.98 | 69.84 |
| TS+TC+3D+QC | 307 | 77.17 | 84.38 | 69.84 |
| ITC + Ridge regrerssion [itc508] | TS+TC+AP | 203 | 78.35 | 84.38 | 72.22 |
| **95 amines congeneric dataset** | Ridge regression | TS | 108 | 83.16 | 75.47 | 88.42 |
| TS+TC | 266 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D | 269 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D+QC | 275 | 84.21 | 77.36 | 92.86 |
| ITC + Ridge regrerssion | TS | 108 | 88.42 | 90.57 | 85.71 |
| TS+TC | 266 | 86.32 | 88.68 | 83.33 |
| TS+TC+3D | 269 | 88.42 | 92.45 | 83.33 |
| TS+TC+3D+QC | 275 | 85.26 | 88.68 | 80.95 |

*Subha, in Table 1 above I see some inconsistency: For 508 set, we used TS+TC+AP. Why not TS+TC+3D+QC + AP? It looks like AP suddenly appears from nowhere.*

*Also, why did we not do TS+TC+3D+QC+ AP for the 95 amine set?*

*2.2. Discussion*

Results of regression models on the two mutagen data sets, viz., 95 aromatic and heteroaromatic amines and 508 diverse chemicals are presented in Table 1. In both the cases hierarchical QSAR analysis was carried out using TS, TC, 3-D, and quantum chemical descriptors in a graduated manner.

For the structurally homogeneous set of amines, ridge regression using the TS descriptors gave 83.16% correct classification with 75.47 % and 88.42 % sensitivity and specificity, respectively. For the same set of descriptors, ITC based variable selection followed by ridge regression yielded 88.42 % correct classification with 90.57 % sensitivity and 85.71 % specificity. In both cases there were minimal improvement in model quality by the addition of TC, 3-D, and quantum chemical descriptors to the set of independent variables.For the 508 set of mutagens, the ridge regression model developed from the TS set of indices gave 53.14 % correct classification with 52.34 % sensitivirty and 53.97 % specificity. Contrary to the case of 95 aromatic amines, there was a singnfficant improvent in model quality by the addition of TC descriptors to the set of independent variables.

ADD NEW DISCUSSION BASED ON THE SUGGESTED DATA ANALYSIS ABOVE.

A lot of QSAR is based on the congenericity principle, which states that molecular structures which are similar or do not differ substantially from one another have similar properties. But in this age of combinatorial chemistry and high throughput screening (HTS) data sets are being generated which are not congeneric structures. The TSCA Inventory also is a collection of chemicals which do not belong to a single or a few chemical or biochemical classes. One needs diverse collection of molecular descriptors to predict the common biochemical/ toxicological action of such diverse sets. This has been called “Diversity begets diversity” principle by Basak [ref]

It is tempting to speculate that the findings in this study are in line with the ‘Diversity Begets Diversity Principle.’ According to this principle, the most efficient (in terms of computational load and number of variables considered) predictive model for a certain set of chemical molecules depends on the composition of the data set and the structural variability within the set sample compounds. While a limited number and type of descriptors are enough to obtain quality QSAR models for homogenous compounds from a particular family of chemical compounds, one needs collection of heterogeneous types of molecular descriptors to obtain a comparable model for a more diverse set of sample compounds. Here we were able to obtain good prediction performance using only one type of descriptor (TS) for the 95-compound dataset consisting of amines, and progressively adding more types of compounds didn’t improve the prediction performance.. But when we consider the larger, more diverse set of compounds, there was a large improvement when TC descriptors are added,. Futher research in this line of reasoning is necessary to validate the utility of these two principles in the fourmulation of QSAR models.

**3. Experimental Section**

*3.1. Data*

The dataset used in our analysis is due to Debnath *et al* [debnath] and concerns mutagenic activities of 95 aromatic and heteroaromatic amines in *S. typhimurium* TA98+S9 microsomal preparations. There are 275 descriptors for each compound: among them 97 topostructural (TS), 162 topochemical (TC), 10 3-dimensional (3D) and 6 quantum-chemical (QC). Number of revertants per nmol of test culture in log scale (log *R*) is the original response variable. For binary classification we take the 0/1 indicator log *R* being > 0 or < 0 as the response variable in our analysis. At the original scale, this amounts to no. of revertants/nmol being greater/less than 1, respectively.

The results obtained are compared with previous studies on a second, and more diverse, dataset records Ames mutagenicity of 508 chemical compounds (256 mutagens and 252 non-mutagens), and is taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [crc]. The set of descriptors for these compounds includes the above type of descriptors, as well as a large number of Atom-pair (AP) descriptors.

Both datasets contain the same 3D and QC variables, while the TS and TC variables are slightly different. The numbers of each type of variables in them are summarized in Table 2. All variable names for both datasets are given in the supplementary material.

**Table 2.** Number of different types of variables in two datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Number of variables for each type** | | | |
| **Topostructural** | **Topochemical** | **3-dimensional** | **Quantum-Chemical** |
| **508 compound diverse dataset** | 103 | 195 | 3 | 6 |
| **95 amines congeneric dataset** | 108 | 158 | 3 | 6 |

*3.2. Descriptors*

<Descriptor calculation details>

*3.2. Methods*

3.2.1. Variable selection

Because the number of samples (*n* = 95) is smaller than the number of variables (*p* = 275), regression-based methods of variable selection like stepwise forward and backward selection are not applicable in our scenario. Here we used the Interrelated Two-way Clustering (ITC) algorithm to do the variable selection. This algorithm takes in a number of predictor groups, which are pre-determined or obtained by some known unsupervised clustering method (like K-means). It then hypothesizes that if we include only the important predictors and classify samples independently for each predictor groups, the classification should be identical. Keeping this in mind, an iterative procedure is used to eliminate predictors until the classifications based on different predictor groups achieve a certain level of similarity.

A detailed description of the algorithm can be found in the original paper [itc], and its implementation in QSAR was done by Majumdar *et al* [itc508].

3.2.2. Hierarchical QSAR and predictive models

We take a hierarchical approach to build the predictive model. Starting from the set of TS variables, we keep on including TC, 3-dimensional and quantum-chemical variables and check predictive performance of all the models. To tackle high collinearity among different predictors, we use ridge regression to build our predictive models. Given *n* samples and *p* variables, the *n* × *p* data matrix of predictors ***X*** and *n* × *1* vector of 0/1 responses ***Y***, the vector of coefficients obtained by ridge regression is defined as:

|  |  |
| --- | --- |
| ***b*** *=* (***X****’****X*** *+ k****I***)-1***Y*** | (1) |

Where *k* > 0 is the ridge constant, chosen by cross-validation [ridge-book].

While assessing the predictive performance of a model through cross validation, it is essential to not do the variable selection beforehand and the build the model, because that uses information from the test set of compounds in the variable selection step. This results in a synthetic increase of the predictive performance of the model. For this reason, while we did use leave-one-out cross-validation to obtain prediction accuracy, we did both the selection of variables and building ridge regression models for every iteration of the cross-validation procedure.

**4. Conclusions**

Main text paragraph.

Main text paragraph.

**Acknowledgments**

Main text paragraph.

**Author Contributions**

Main text paragraph.

**Conflicts of Interest**

The authors confirm that this article has no conflict of interest.

**References and Notes**

1. National Research Council. *Toxicity Testing Strategies to Determine Needs and Priorities*, National Academy Press: Washington, DC, **1984**.

2. Auer, C. M.; Nabholz, J. V. and Baetcke, K.P. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87*, 183-197.

3.

debnath. Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37-52.

crc. Soderman, J.V. *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*, CRC Press: Boca Raton, FL, **1982**.

itc. Tang, C.; Zhang, L.; Zhang, A.; Ramanathan, M. In: *Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis*, Proceedings of BIBE 2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, November 4-5, 2001; Bilof, R.; Palagi, L., Eds.; IEEE Computer Society: Los Alamitos, CA, **2001**; pp. 41-48.

itc508. Majumdar, S.; Basak, S. C.; Grunwald, G. D. Adapting Interrelated Two-Way Clustering Method for Quantitative Structure-Activity Relationship (QSAR) Modeling of Mutagenicity/Non-Mutagenicity of a Diverse Set of Chemicals. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 463-471.

ridge-book. Yang, X.; Su, X.G. *Linear Regression Analysis- Theory and Computing*, World Scientific: Singapore, **2009**.

hawk. Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 37-44.

2. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, Editor 2, Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.

3. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008;   
pp. 154–196.

4. Author 1, A.B.; Author 2, C. Title of Unpublished Work. Journal Abbreviation, phrase indicating stage of publication.

5. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In *Title of the Collected Work* (if available), Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference; Editor 1, Editor 2, Eds. (if available); Publisher: City, Country, Year (if available); Abstract Number (optional), Pagination (optional).

6. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.

7. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name*, Year, Volume, (Page range), doi or other identification number. Available online: http://URL (accessed on Day Month Year).

8. Title of Site. Available online: http://URL (accessed on Day Month Year).

Reference list: We recommend the use of reference management software to prepare the references list (e.g., Endnote, http://www.mdpi.com/files/word-templates/MDPI.ens).

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).

[I ] National Research Council (1984) Toxicity Testing Strategies to Determine Needs and Priorities, National Academy Press, Washington, DC.

[2] Auer, CM., Nabholz, J.V. and Baetcke, K.P. (1990) Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environ. Health Perspect. 87, 183-197.

[3] Toxic Substances Control Act (TSCA) (1976) Public Law 94-469, 90 Stat. 2003, 11 October 1976

[4] Cash, G., United States Environmental Protection Agency (USEPA, Personal communication, 2014.

[5] Adams, C.; Brantner, V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? Health Aff. (Millwood), 2006, 25, 420-428.

[6] DiMasi, J.; Hansen, R.; Grabowski, H. The price of innovation: new estimates of drug development costs. J. Health Eco., 2003, 22, 151-185.

[7] M. Vracko, Mathematical (Structural) Descriptors in QSAR: Applications in Drug Design and Environmental Toxicology, In: Advances in Mathematical Chemistry and Applications, Subhash C. Basak, Guillermo Restrepo and J. L. Villaveces, Editors, Vol. 1,pp. 222-250, Bentham Science Publishers, in press, 2014.

[8] Basak, S. C., Mathematical Descriptors for the Prediction of Property, Bioactivity, and Toxicity of Chemicals from their Structure: A Chemical-Cum-Biochemical Approach, Current Computer-Aided Drug Design, 2013, 9, 449-462.

[9] Basak, S.C. Role of mathematical chemodescriptors and proteomics-based biodescriptors in drug discovery. Drug Dev. Res., 2010, 72, 1-9.

[10] PaDEL-Descriptor: <http://padel.nus.edu.sg/software/padeldescriptor/>

[11] DRAGON - Software for the Calculation of Molecular Descriptors, Version 5.4, 2006; Todeschini, R.; Consonni, V.; Mauri, A. et al.; Talete srl.; Milan, Italy

[12] MolconnZ, Version 4.05, 2003; Hall Ass. Consult.; Quincy, MA.

[13] Basak, S.C.; Harriss, D.K.; Magnuson, V.R. POLLY v. 2.3: 1988; Copyright of the University of Minnesota, USA.

[14] Basak, S.C.; Grunwald, G.D. APProbe. 1993; Copyright of the University of Minnesota, USA.

**Supplementary materials**

Table S1. Symbols, definitions and classification of topological indices

|  |  |  |  |
| --- | --- | --- | --- |
|  | Topostructural (TS) | In 508 compound data? | In 95  Amine data? |
| *IWD* | Information index for the magnitudes of distances between all possible pairs of vertices of a graph | **✓** | **✓** |
| *IWD* | Mean information index for the magnitude of distance | **✓** | **✓** |
| *W* | Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph | **✓** | **✓** |
| *ID* | Degree complexity | **✓** | **✓** |
| *HV* | Graph vertex complexity | **✓** | **✓** |
| *HD* | Graph distance complexity | **✓** | **✓** |
| *IC* | Information content of the distance matrix partitioned by frequency of occurrences of distance *h* | **✓** | **✓** |
| *M1* | A Zagreb group parameter = sum of square of degree over all vertices | **✓** | **✓** |
| *M2* | A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices | **✓** | **✓** |
| *hχ* | Path connectivity index of order *h* = 0-10 | **✓** | **✓** |
| *hχC* | Cluster connectivity index of order *h* = 3-5 | **✓** | **✓** |
| Cluster connectivity index of order *h* = *6* | **✓** | **✗** |
| *hχPC* | Path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *hχCh* | Chain connectivity index of order *h* = 3, 4, 7, 8 | **✓** | **✗** |
| Chain connectivity index of order *h* = 5, 6, 9, 10 | **✓** | **✓** |
| *Ph* | Number of paths of length *h* = 0-10 | **✓** | **✓** |
| *J* | Balaban’s *J* index based on topological distance | **✓** | **✓** |
| *nrings* | Number of rings in a graph | **✓** | **✓** |
| *ncirc* | Number of circuits in a graph | **✓** | **✓** |
| DN2S*y* | Triplet index from distance matrix, square of graph order, and distance sum; operation *y* = 1-4 | **✓** | **✓** |
| DN21*y* | Triplet index from distance matrix, square of graph order, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| AS1*y* | Triplet index from adjacency matrix, distance sum, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| DS1*y* | Triplet index from distance matrix, distance sum, and number 1; operation *y* = 1-2 | **✓** | **✓** |
| Triplet index from distance matrix, distance sum, and number 1; operation *y* = 3-5 | **✗** | **✓** |
| ASN*y* | Triplet index from adjacency matrix, distance sum, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| DSN*y* | Triplet index from distance matrix, distance sum, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| DN2N*y* | Triplet index from distance matrix, square of graph order, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| ANS*y* | Triplet index from adjacency matrix, graph order, and distance sum; operation *y* = 1-2 | **✓** | **✓** |
| Triplet index from adjacency matrix, graph order, and distance sum; operation *y* = 3-5 | **✗** | **✓** |
| AN1*y* | Triplet index from adjacency matrix, graph order, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| ANN*y* | Triplet index from adjacency matrix, graph order, and graph order again; operation *y* = 1-4 | **✓** | **✓** |
| Triplet index from adjacency matrix, graph order, and graph order again; operation *y* = 5 | **✓** | **✗** |
| ASV*y* | Triplet index from adjacency matrix, distance sum, and vertex degree; operation *y* = 1-2 | **✓** | **✓** |
| Triplet index from adjacency matrix, distance sum, and vertex degree; operation *y* = 3-5 | **✗** | **✓** |
| DSV*y* | Triplet index from distance matrix, distance sum, and vertex degree; operation *y* = 1-2 | **✓** | **✓** |
| ANV*y* | Triplet index from adjacency matrix, graph order, and vertex degree; operation *y* = 1-5 | **✓** | **✓** |
| *kp0* | Kappa zero | **✓** | **✓** |
| *kp1-kp3* | Kappa simple indices | **✓** | **✓** |
|  | **Topochemical (TC)** | **In 95 amine data?** | **In 508 compound data?** |
| O | Order of neighborhood when *ICr* reaches its maximum value for the hydrogen-filled graph | **✓** | **✓** |
| O*orb* | Order of neighborhood when *ICr* reaches its maximum value for the hydrogen-suppressed graph | **✓** | **✓** |
| I*ORB* | Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices | **✓** | **✓** |
| IC*r* | Mean information content or complexity of a graph based on the *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| SIC*r* | Structural information content for *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| CIC*r* | Complementary information content for *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| *hχb* | Bond path connectivity index of order *h* = 0-6 | **✓** | **✓** |
| *hχbC* | Bond cluster connectivity index of order *h* = 3, 5 | **✓** | **✓** |
| Bond cluster connectivity index of order *h* = 4, 6 | **✓** | **✗** |
| *hχbCh* | Bond chain connectivity index of order *h* = 4 | **✓** | **✗** |
| Bond chain connectivity index of order *h* = 5, 6 | **✓** | **✓** |
| *hχbPC* | Bond path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *hχv* | Valence path connectivity index of order *h* = 0-6 | **✓** | **✓** |
| *hχvC* | Valence cluster connectivity index of order *h* = 3, 5 | **✓** | **✓** |
| Valence cluster connectivity index of order *h* = 4, 6 | **✓** | **✗** |
| *hχvCh* | Valence chain connectivity index of order *h* = 3, 4, 7, 8 | **✓** | **✗** |
| Valence chain connectivity index of order *h* = 5, 6, 9, 10 | **✓** | **✓** |
| *hχvPC* | Valence path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *JB* | Balaban’s*J* index based on bond types | **✓** | **✓** |
| *JX* | Balaban’s*J* index based on relative electronegativities | **✓** | **✓** |
| *JY* | Balaban’s*J* index based on relative covalent radii | **✓** | **✓** |
| AZV*y* | Triplet index from adjacency matrix, atomic number, and vertex degree; operation *y* = 1-5 | **✓** | **✓** |
| AZS*y* | Triplet index from adjacency matrix, atomic number, and distance sum; operation *y* = 1-5 | **✓** | **✓** |
| ASZ*y* | Triplet index from adjacency matrix, distance sum, and atomic number; operation *y* = 1-2 | **✓** | **✓** |
| Triplet index from adjacency matrix, distance sum, and atomic number; operation *y* = 3-5 | **✗** | **✓** |
| AZN*y* | Triplet index from adjacency matrix, atomic number, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| ANZ*y* | Triplet index from adjacency matrix, graph order, and atomic number; operation *y* = 1-2 | **✓** | **✓** |
| Triplet index from adjacency matrix, graph order, and atomic number; operation *y* = 3-5 | **✗** | **✓** |
| DSZ*y* | Triplet index from distance matrix, distance sum, and atomic number; operation *y* = 1,2 | **✓** | **✓** |
| DN2Z*y* | Triplet index from distance matrix, square of graph order, and atomic number; operation *y* = 1-2 | **✓** | **✓** |
| DN2Z*y* | Triplet index from distance matrix, square of graph order, and atomic number; operation 3-5 | **✗** | **✓** |
| *nvx* | Number of non-hydrogen atoms in a molecule | **✓** | **✓** |
| *nelem* | Number of elements in a molecule | **✓** | **✓** |
| *fw* | Molecular weight | **✓** | **✓** |
| *si* | Shannon information index | **✓** | **✓** |
| *totop* | Total Topological Index *t* | **✓** | **✓** |
| *sumI* | Sum of the intrinsic state values *I* | **✓** | **✓** |
| *sumdelI* | Sum of delta-*I* values | **✓** | **✓** |
| *tets2* | Total topological state index based on electrotopological state indices | **✓** | **✓** |
| *phia* | Flexibility index (*kp*1\* *kp*2/*nvx*) | **✓** | **✓** |
| *Idcbar* | Bonchev-Trinajstić information index | **✓** | **✓** |
| *IdC* | Bonchev-Trinajstić information index | **✓** | **✓** |
| *Wp* | Wiener *p* | **✓** | ✓ |
| *Pf* | Platt *f* | **✓** | ✓ |
| *Wt* | Total Wiener number | **✓** | ✓ |
| *knotp* | Difference of chi-cluster-3 and path/cluster-4 | **✓** | ✓ |
| *knotpv* | Valence difference of chi-cluster-3 and path/cluster-4 | **✓** | ✓ |
| *nclass* | Number of classes of topologically (symmetry) equivalent graph vertices | **✓** | ✓ |
| *NumHBd* | Number of hydrogen bond donors | **✓** | ✓ |
| *NumHBa* | Number of hydrogen bond acceptors | **✓** | ✓ |
| *SHCsats* | E-State of C *sp3* bonded to other saturated C atoms | **✓** | ✓ |
| *SHCsatu* | E-State of C *sp3* bonded to unsaturated C atoms | **✓** | ✓ |
| *SHvin* | E-State of C atoms in the vinyl group, *=CH-* | **✓** | ✗ |
| *SHtvin* | E-State of C atoms in the terminal vinyl group, *=CH2* | **✓** | ✗ |
| *SHarom* | E-State of C *sp2* which are part of an aromatic system | **✓** | ✓ |
| *SHHBd* | Hydrogen bond donor index, sum of Hydrogen E-State values for *–OH*, *=NH*, -*NH2*, *-NH-,-SH*, and *#CH* | **✓** | ✓ |
| *SHwHBd* | Weak hydrogen bond donor index, sum of *C-H* Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded | **✓** | ✗ |
| *SHHBa* | Hydrogen bond acceptor index, sum of the *E*-State values for *–OH*, *=NH*, *-NH2*, -*NH-*, *>N*, *-O-*, *-S-*, along with –F and –Cl | **✓** | ✓ |
| *Qv* | General Polarity descriptor | **✓** | ✓ |
| *NHBinty* | Count of potential internal hydrogen bonders (*y* = 2) | **✓** | ✗ |
| Count of potential internal hydrogen bonders (*y* = 3-10) | **✓** | ✓ |
| *SHBinty* | E-State descriptors of potential internal hydrogen bond strength (*y =*2, 5-9) | **✓** | ✗ |
| E-State descriptors of potential internal hydrogen bond strength (*y =*3, 4) | **✓** | ✓ |
| *ka1-ka3* | Kappa alpha indices | **✓** | ✓ |
|  | Electrotopological State index values for atom types:  *SHdNH, SHsSH, HssNH, SHtCH, SHCHnX, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SdCH2, StCH, SddC, StsC, SdssC, SaasC, SsNH3p, SssNH2p, SdNH, StN, SsssNHp, SdsN, SsssN, SaasN, SssssNp, SaaO, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SaaS, SdssS, SddssS, SssssssS, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb* | **✓** | **✗** |
|  | Electrotopological State index values for atom types:  *SHsOH, SHsNH2, SHssNH, SHother, Hmax, Gmax, Hmin, Gmin, SsCH3, SssCH2, SdsCH, SaaCH, SsssCH, SaaaC, SssssC, SsNH2, SssNH, SaaNH, SaaN, SddsN, SsOH, SdO, SssO, SsF, SssS, SsCl, SsBr,* | **✓** | **✓** |
|  | **Geometrical (3-D)** | **In 95 amine data?** | **In 508 compound data?** |
| *3DW* | 3D Wiener number based on the hydrogen-suppressed geometric distance matrix | **✓** | **✓** |
| *3DW H* | 3D Wiener number based on the hydrogen-filled geometric distance matrix | **✓** | **✓** |
| *VW* | Van der Waal’s volume | **✓** | **✓** |
|  | Quantum Chemical (QC) | **In 95 amine data?** | **In 508 compound data?** |
| *EHOMO* | Energy of the highest occupied molecular orbital | **✓** | **✓** |
| *EHOMO-1* | Energy of the second highest occupied molecular | **✓** | **✓** |
| *ELUMO* | Energy of the lowest unoccupied molecular orbital | **✓** | **✓** |
| *ELUMO+1* | Energy of the second lowest unoccupied molecular orbital | **✓** | **✓** |
| *ΔHf* | Heat of formation | **✓** | **✓** |
| *μ* | Dipole moment | **✓** | **✓** |