Development of quantitative structure-activity relationship models for vapor pressure estimation using computed molecular descriptors

Subhash C. Basak* and Denise Mills

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

Dedicated to Dr. Alexandru T. Balaban on the occasion of his 75th birthday

(received 2 Aug 05; accepted 17 Nov 05; published on the web 19 Nov 05)

Abstract

Vapor pressure is an important property which is an indicator of chemical volatility, along with transport, partitioning, fate and distribution of environmental pollutants. Various models have been developed for the prediction of vapor pressure of chemicals using physicochemical and calculated structural properties. We have used different classes of graph theoretic indices, e.g., topostructural indices, topochemical indices, geometrical (3D) indices and, quantum chemical descriptors, for the development of predictive models for vapor based on a structurally diverse set of 469 chemicals. Initially, a set of 379 molecular descriptors was calculated using the software *POLLY*, *Triplet*, *Sybyl*, *MOPAC*, and *Molconn-Z*. Comparatively, three linear regression methodologies were used to develop hierarchical QSAR (HiQSAR) models, namely ridge regression (RR), principal components regression (PCR), and partial least squares (PLS) regression. The results indicate that, in general, RR outperforms PCR and PLS, and that the easily calculated topological descriptors are sufficient for the prediction of vapor pressure based on this large, diverse set of chemicals.

Keywords: Hierarchical QSAR, ridge regression, principal components regression, partial least squares regression, topological indices, vapor pressure

Introduction

The assessment of fate and distribution of environmental pollutants in various phases including air, water, and soil is important for the risk assessment of chemicals.[1] The partitioning of chemicals among different phases is usually assessed using a critical list of physical properties including vapor pressure (VP), aqueous solubility, air: water partition coefficient, and octanol: water partition coefficient.

Pollutants with high vapor pressure tend to concentrate more in the vapor phase as compared to soil or water. Therefore, VP is a key physicochemical property essential to the assessment of chemical distribution in the environment. This property is also used in the design of various chemical engineering processes.[2] Additionally, VP can be used for the estimation of other important physicochemical properties. For example, one can calculate Henry's law constant, soil sorption coefficient, and partition coefficient from VP and aqueous solubility.

Therefore, it is not surprising that various authors have attempted to model this important physicochemical property using quantitative structure-property relationships (QSPRs) based on calculated molecular descriptors. Katritzky et al used descriptors calculated by CODESSA in the formulation of QSPRs for a diverse set of 411 chemicals.[1] Engelhardt et al used topological descriptors and computational neural networks (CNNs) in the formulation of QSPRs for the estimation of VP for a diverse set of 420 organic compounds.[3] Liang and Gallagher,[4] along with Staikova et al,[5] used quantum chemically derived indices, polarizability in particular, in the development of QSPRs for vapor pressure estimation.

Basak et al formulated the hierarchical quantitative structure-activity relationship (HiQSAR) approach for the estimation of properties, biomedicinal activities, and toxicities of chemicals from computed descriptors.[6-18] The objective of this HiQSAR/ HiQSPR research has been two-fold: description and prediction. The HiQSPR formalism uses progressively more complex indices in the development of models. The type of parameters important for the estimation of a property at each level, e.g., topological, geometrical, and quantum chemical, can be determined and used in order to understand the molecular and submolecular basis of the property (description), and good quality models based on algorithmically derived descriptors can be used for the estimation of the property of interest for any chemical, real or hypothetical (prediction).

Basak et al have used the HiQSPR approach previously in the development of VP prediction models.[12,15] However, the current study utilizes an expanded set of descriptors along with three statistical modeling approaches, namely ridge regression (RR), principal components regression (PCR), and partial least squares (PLS) regression, which are appropriate for data sets wherein the number descriptors is large with respect to the number of chemical compounds and when the molecular descriptors are highly intercorrelated.

Methods and Materials

Experimental Data

The set of 469 chemicals used in this study was obtained from the Assessment Tools for the Evaluation of Risk (ASTER) database[19] and represents a subset of the Toxic Substances Control Act (TSCA) Inventory[20] for which vapor pressure (p_{vap}) was measured at 25 °C with a pressure range of approximately 3 –10 000 mm Hg. The molecular weights of the compounds in this data set range from 40 to 338, and the chemical diversity is described in Table 1.

Table 1. Chemical class composition of the vapor pressure data set

Compound classification	No. of compounds	Pure	Substituted
Total Data Set	469		
Hydrocarbons	253		
Non-Hydrocarbons	216		
Nitro compounds	4	3	1
Amines	20	17	3
Nitriles	5	4	1
Ketones	7	7	0
Halogens	97	92	5
Anhydrides	1	1	0
Esters	18	16	2
Carboxylic acids	2	2	0
Alcohols	10	6	4
Sulfides	38	37	1
Thiols	4	4	0
Imines	2	2	0
Epoxides	1	1	0
Aromatic compounds[a]	15	10	4
Fused-ring compounds[b]	1	1	0

[a] The 15 aromatic compounds are a mixture of 11 aromatic hydrocarbons and four aromatic halides.

[b] The only fused-ring compound was a polycyclic aromatic hydrocarbon.

Reproduced with permission from *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 692-701. Copyright 2001 Am. Chem. Soc.

Structural descriptors and hierarchical QSAR

In general, a wide variety of molecular descriptors based on chemical structure have been formulated and used in QSAR and QSPR studies.[21,22]

In the present study, the majority of the topological descriptors were calculated using software including *POLLY v. 2.3*[23] and *Triplet*[24]. The topological descriptors obtained from these programs include Wiener number[25] molecular connectivity indices developed by Randić[26] and Kier and Hall,[27] frequency of path lengths of varying size,[27] information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[28] Roy et al.,[29] Basak et al.,[30,31] as well as those of Raychaudhury *et al*.,[32] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[10,11,12] and Balaban's J indices[33-35] as well as the triplet indices.[24] The triplets result from a matrix, main diagonal column vector, and free term column vector which are converted into a system of linear equations. The notation used to represent the vectors and matrices is as follows:

A = Adjacency matrix
V = Vertex degree
S = Distance sum
N = Total number of vertices in the graph
Z = Atomic number
D = Distance matrix
1 = Unity matrix.

After the system of N linear equations is solved, the local vertex invariants, x_i, are assembled into a triplet descriptor based on one of the following operations:

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

Additional topological descriptors, including an extended set of molecular connectivity indices, electrotopological state descriptors,[36,37] general polarity descriptors, and hydrogen bonding descriptors, were calculated by *Molconn-Z v. 3.50*.[38] An additional hydrogen bonding parameter was obtained from software developed by Basak et al.[39]

In addition, ten geometrical descriptors were calculated, including six kappa shape indices which were also obtained by *Molconn-Z v. 3.50*. Van der Waals volume, V_W, was calculated using *Sybyl v. 6.2*.[40] In addition, two variants of the 3-D Wiener number, ^{3D}W and $^{3D}W_H$, based on the hydrogen-suppressed and hydrogen-filled geometric distance matrices, respectively, were also calculated by *Sybyl v 6.2* using a SPL (Sybyl Programming Language) program developed by our group.

The six quantum chemical descriptors included in the study; namely, E_{HOMO}, E_{HOMO-1}, E_{LUMO}, E_{LUMO+1}, ΔHf, and μ, were calculated for the AM1 semi-empirical Hamiltonian using *MOPAC v. 6.0* [41] in the *Sybl* interface.[40]

A complete list of the 379 parameters calculated for use in the current study, including brief descriptions, is provided in Table 2. Note that the topological descriptors are partitioned into two classes: topostructural, which are based solely on the connectivity of atoms within a molecule, and topochemical, which encode chemical as well as topological information. From the initial set of descriptors listed in Table 2, the following descriptors were removed and not used in the subsequent analyses: 1) Any descriptor with a constant value for all of the 469 chemicals in the data set, 2) one descriptor of each perfectly correlated pair (i.e., $r = 1.0$), as determined by the CORR procedure of the *SAS* statistical package,[42] and any descriptors with undefined values. A total of 268 descriptors were retained and subsequently used in model development.

Table 2. Symbols, definitions and classification of calculated molecular descriptors

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
I_D^W	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
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order h = 0–10
${}^h\chi_C$	Cluster connectivity index of order h = 3–6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4–6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3–10
P_h	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
DN^2S_y	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation y = 1–5
DN^21_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 = 1–5
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
DN^2N_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 = 1–5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1–5

Table 2. Continued

ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1–5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1–5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1–5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1–5
	Topochemical (TC)
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r 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
${}^h\chi^b$	Bond path connectivity index of order h = 0–6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3–6
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 3–6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4–6
${}^h\chi^v$	Valence path connectivity index of order h = 0–10
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3–6
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 3–10
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order h = 4–6
J^B	Balaban’s J index based on bond types
J^X	Balaban’s J index based on relative electronegativities
J^Y	Balaban’s J index based on relative covalent radii
HB_1	Hydrogen bonding parameter
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 = 1–5
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 = 1–5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1–5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1–5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule

Table 2. Continued

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 (kp1* kp2/nvx)
IdCbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp
Pf	Plattf
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
numwHBd	Number of weak hydrogen bond donors
numHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 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, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH–, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH,–NH_2, –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,–NH_2, –NH–, >N–, –O–, –S–, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2–10)
$SHBint_y$	E-State descriptors of potential internal hydrogen bond strength (y =2–10)

Table 2. Continued

	Electrotopological State index values for atoms types: SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss,Bem, SssBH, SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC,StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
	Geometrical / Shape (3D)
kp0	Kappa zero
kp1–kp3	Kappa simple indices
ka1–ka3	Kappa alpha indices
V_W	Van der Waals volume
^{3D}W	3D Wiener number based on the hydrogen-suppressed geometric distance matrix
$^{3D}W_H$	3D Wiener number based on the hydrogen-filled geometric distance matrix
	Quantum Chemical (QC)
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO-1}	Energy of the second highest occupied molecular
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO+1}	Energy of the second lowest unoccupied molecular orbital
ΔHf	Heat of formation
μ	Dipole moment

We have used the hierarchical QSAR (HiQSAR) approach to model development in which increasingly more complex and computer-resource intensive classes of structural descriptors are used in a graduated manner, first utilizing the topostructural (TS) descriptors alone, followed by the addition of the topochemical (TC) descriptors, the 3-dimensional (3D) descriptors, and finally the quantum chemical (QC) descriptors. The predictive ability of the resulting models, based on the cross-validated R^2 values, are compared in order to determine whether or not the more complex descriptors are necessary in order to predict the property or activity of interest, or if the easily calculable descriptors are sufficient. For comparative purposes, predictive models based on each descriptor class, alone, were also developed.

Statistical methodology

Prior to analysis, all calculated descriptors were transformed by: ln (x + c), where x represents the original descriptor value and c is a constant added to avoid possible arithmetic error. In most cases, c = 1, as the original descriptor values are generally greater than -1. A small number of descriptors, however, have minimum values less than or equal to -1, in which case the constant added was the smallest natural number that would provide a positive sum in the equation above.

For comparative purposes, three regression methodologies were used for the development of predictive models, namely ridge regression (RR),[43] principal components regression (PCR),[44] and partial least squares (PLS).[45] Each of these methodologies makes use of all available descriptors, as opposed to subset regression, and is useful when the number of descriptors is large with respect to the number of compounds and when the descriptors are highly intercorrelated. Formal comparisons have consistently shown that using a subset of available descriptors is less effective than using alternative regression methods that retain all available descriptors, such as RR, PCR, and PLS.[45,46] RR is similar to PCR in that the independent variables are transformed to their principal components (PCs). However, while PCR utilizes only a subset of the PCs, RR retains them all but downweights them based on their eigenvalues.[43] With PLS, a subset of the PCs is also used, but the PCs are selected by considering both the independent and dependent variables. For each model developed, the cross-validated R^2 was obtained using the leave-one-out approach and can be calculated as follows (eq. 1):

$$R_{cv}^2 = 1 - \frac{PRESS}{SSTotal} \qquad 1$$

where *PRESS* is the prediction sum of squares and *SSTotal* is the total sum of squares. For the sake of brevity, the highly parameterized models are not included in this paper, rather the cross-validated R^2 and *PRESS* statistic are reported for each model. Another useful statistical metric is the *t* value associated with each model descriptor, defined as the descriptor coefficient divided by its standard error. Descriptors with large $|\, t \,|$ values are important in the predictive model and, as such, can be examined in order to gain some understanding of the nature of the property or activity of interest.

It should be strongly stated that ordinary least squares (OLS) regression is inappropriate when the number of descriptors is large with respect to the number of chemical compounds in the data set, and that the conventional R^2 metric is without value in this situation. Unlike R^2, which tends to *increase* upon the addition of any descriptor, the cross-validated R^2 tends to *decrease* upon the addition of irrelevant descriptors and is a reliable measure of model predictability.[47] Unlike ordinary least squares regression, the number of descriptors is not an issue with the regression methodologies used in the present study. The number of descriptors included in the regression models developed in this study is as follows: TS (99), TS+TC (252), TS+TC+3D (262), TS+TC+3D+QC (268), TC (99), 3D (10), QC (6). RR, PCR, and PLS are appropriate methodologies when the number of descriptors exceeds the number of observations, and they are designed to utilize all available descriptors, as opposed to subset regression, in order to produce an unbiased model whose predictability is accurately reflected by the R^2_{cv}, regardless of the number of independent variables in the model. The distinction between these methods and OLS regression is important and cannot be overemphasized.

Results and Discussion

The major objective of this paper was the estimation of vapor pressure of chemicals using molecular descriptors that can be computed directly from molecular structure without the input of any other experimental data. To this end, we used topostructural, topochemical, geometrical, and quantum chemical descriptors in the formulation of HiQSPR models for $\log_{10}(p_{\text{vap}})$. All models developed in this study are based the complete set of 469 structurally diverse chemicals. Results in Table 3 indicate that the combination of TS and TC descriptors resulted in a highly predictive RR model ($R^2_{c.v}$ = 0.895).; the addition of three dimensional and quantum chemical indices to the set of independent variables did not result in significant improvement in model quality. It may be noted that we have observed such results for various other physicochemical and biological properties including mutagenicity,[10,48] boiling point,[49] blood:air partition coefficient,50 tissue: air partition coefficient,[51] etc.[6,14,16,18] Only in limited cases, e.g., halocarbon toxicity,[9] the addition of quantum chemical indices after TS and TC parameters resulted in significant improvement in QSAR model quality.

It is interesting to note that of the three linear regression methods used, viz. RR, PCS, and PLS, ridge regression outperformed the other two methods significantly. This is in line with our earlier observations with HiQSARs using the three methods.[14,50-52]

Table 3. Ridge regression (RR), principal components regression (PCR), and partial least squares (PCR) regression model metrics

	RR		PCR		PLS	
Model Type	$R^2_{c.v.}$	*PRESS*	$R^2_{c.v.}$	*PRESS*	$R^2_{c.v.}$	*PRESS*
TS	0.444	135	0.451	133	0.445	134
TS+TC	0.895	25.3	0.479	126	0.480	126
TS+TC+3D	0.902	23.7	0.481	125	0.468	129
TS+TC+3D+QC	0.906	22.8	0.488	124	0.465	129
TS	0.444	135	0.451	133	0.445	134
TC	0.851	35.9	0.473	127	0.524	115
3D	0.552	108	0.453	132	0.556	107
QC	0.201	193	0.189	196	0.203	193

It is instructive to look at the top ten molecular descriptors, based on *t* value, in the ridge regression VP model derived from TS + TC indices (Table 4). They can be looked upon as representing the following features: a) size (totop, DN^2Z_1), hydrogen bonding (HB_1, SHHBa), c) polarity (Qv), d) heterogeneity of atom types (IC_0), and e) presence of various types of hetero atoms and functional groups (SssO, SsF, $SsNH_2$, SaaO).

In the LSER approach, a combination of molecular size, hydrogen bonding, and polarity are used to estimate partitioning behavior of chemicals.[53,54] The presence of specific hetero atoms, functional groups and different atom types, as encoded by information theoretic, triplet, and electrotopological indices, will probably be related to dipole-dipole interactions among the molecules and also specific regional interactions such as hydrogen bonding. Such factors have been found to be useful in predicting VP by Liang and Gallagher,[4] Katritzky et al,[1] Engelhardt et al,[3] and Staikova et al.[5]

Table 4. Important topological descriptors for the prediction of vapor pressure, based on *t* value, from the TS+TC ridge regression model

Descriptor label	Description	| *t* |
SssO	Sum of the E-states for –O–	10.07
SsF	Sum of the E-states for –F	8.58
HB_1	General hydrogen bonding descriptor	7.76
$SsNH_2$	Sum of the E-states for $–NH_2$	6.83
IC_0	Mean information content or complexity of a hydrogen-filled graph based on the 0 order neighborhood of vertices	6.57
SaaO	Sum of the E-states for :O:	6.56
SHHBa	Hydrogen bond acceptor index	6.21
DN^2Z_1	Triplet index from distance matrix, square of graph order (number of vertices), and atomic number	6.13
Qv	General polarity descriptor	6.06
totop	Total topological index	5.87

In conclusion, the VP prediction model for a diverse set of organic chemicals derived from easily calculated molecular descriptors gave very good results. Such a model could be useful in the estimation of vapor pressure of chemicals that fall within the structural types considered in this paper.

Acknowledgements

This paper represents contribution number 393 from the Center for Water and the Environment of the Natural Resources Research Institute. Research was supported by Grant F49620-02-1-0138 from the United States Air Force and Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

References

1. Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720.
2. Daubert, T. E.; Jones, D. K.; AIChE Symp. Ser. 86. 1990; pp 29-39.
3. Engelhardt, H.; McClelland, H. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967.
4. Liang, C.; Gallagher, D. A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321.
5. Staikova, M.; Wania, F.; Donaldson, D. J. *Atmospheric Environment* **2004**, *38*, 213.
6. Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Topological Indices and Related Descriptors in QSAR and QSPR*, Devillers, J.; Balaban, A. T. Eds; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 675-696.
7. Basak, S. C.; Gute, B. D.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1999**, *10*, 117.
8. Basak, S. C.; Mills, D.; Gute, B. D.; Hawkins, D. M. In *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, Benigni, R. Ed; CRC Press: Boca Raton, FL, 2003; pp 207-234.
9. Basak, S. C.; Balasubramanian, K.; Gute, B. D.; Mills, D.; Gorczynska, A.; Roszak, S. J. *Chem. Inf. Comput. Sci.* **2003**, *43*, 1103.
10. Basak, S. C.; Mills, D. *SAR QSAR Environ. Res.* **2001**, *12*, 481.
11. Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671.
12. Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 692.
13. Basak, S. C.; Natarajan, R.; Mills, D. *WSEAS Transactions on Information Science and Application*, **2005**, 958.
14. Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. *Indian J. Chem.* **2003**, *42A*, 1385.
15. Basak, S. C.; Gute, B. D.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651.
16. Gute, B. D.; Basak, S. C. *SAR QSAR Environ. Res.* **1997**, *7*, 117.
17. Gute, B. D.; Balasubramanian, K.; Geiss, K.; Basak, S. C. *Environ. Toxicol. Pharmacol.* **2004**, *16*, 121.
18. Gute, B. D.; Grunwald, G. D.; Basak, S. C. *SAR QSAR Environ. Res.* **1999**, *10*, 1.
19. Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. *Sci. Total Environ.* **1991**, *109/110*, 667.
20. United States Environmental Protection Agency, What is the TSCA Chemical Substance Inventory? http://www. epa.gov/opptintr/newchems/invntory.htm, 2004.
21. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Weinheim (GER), 2000.
22. Topological Indices and Related Descriptors in QSAR and QSPR. Devillers, J., Balaban, A. T., Eds; Gordon and Breach Science Publishers: The Netherlands, 1999.
23. POLLY, version 2.3, Copyright of the University of Minnesota, 1988.
24. Filip, P. A.; Balaban, T. S.; Balaban, A. T. *J. Math. Chem.* **1987**, *1*, 61.
25. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
26. Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.

27. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.
28. Bonchev, D.; Trinajstic, N. *J. Chem. Phys.* **1977**, *67*, 4517.
29. Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathl. Modelling Sci. Tech.* Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds; Pergamon Press: 1983; pp 745-750.
30. Basak, S. C. *Med. Sci. Res.* **1987**, *15*, 605.
31. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Appl. Math.* **1988**, *19*, 17.
32. Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581.
33. Balaban, A. T. *Math. Chem. (MATCH)* **1986**, *21*, 115.
34. Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
35. Balaban, A. T. *Pure and Appl. Chem.* **1983**, *55*, 199.
36. Hall, L. H.; Mohney, B.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76.
37. Kier, L. B.; Hall, L. H.; Frazer, J. W. *J. Math. Chem.* **1991**, *7*, 229.
38. Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA, 2000.
39. Basak, S. C. H-Bond, *Copyright of the University of Minnesota,* 1988.
40. SYBYL v. 6.2, Tripos Associates, Inc., St. Louis, MO, 1995.
41. Stewart, J. J. P. MOPAC Version 6.00, QCPE #455, Frank J Seiler Research Laboratory, US Air Force Academy, CO, 1990.
42. SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03, Cary, NC, 1988.
43. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 55.
44. Massy, W. F. J. *Am. Statistical Assoc.* **1965**, *60*, 234.
45. Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109.
46. Rencher, A. C.; Pun, F. C. *Technometrics* **1980**, *22*, 49.
47. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.
48. Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Quantitative Structure-activity Relationships in Environmental Sciences VII,* Chen, F.; Schuurmann, G. Eds; SETAC Press: Pensacola, FL, 1998; pp 245-261.
49. Basak, S. C.; Gute, B. D.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054.
50. Basak, S. C.; Mills, D.; El-Masri, H. A.; Mumtaz, M. M.; Hawkins, D. M. *Environ. Toxicol. Pharmacol.* **2004**, *16*, 45.
51. Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. *SAR QSAR Environ. Res.* **2002**, *13*, 649.
52. Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. *Risk Analysis* **2003**, *23*, 1173.
53. Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877.
54. Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Marcus, Y.; Taft, R. W. *J. Phys. Chem.* **1988**, *92*, 5244.