Quantitative Structure−Property Relationships (QSPRs) for the Estimation of Vapor Pressure: A Hierarchical Approach Using Mathematical Structural Descriptors

Subhash C. Basak* and Denise Mills

Natural Resources Research Institute, University of Minnesota−Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received December 18, 2000

A set of 379 molecular descriptors was calculated for use in hierarchical quantitative structure−property relationship (QSPR) modeling of vapor pressure for a structurally diverse database consisting of 469 chemicals. The hierarchical approach utilizes topostructural, topochemical, geometrical, and quantum chemical descriptors in a stepwise fashion to develop QSPR models. In this way, the relative roles of the various levels of descriptors can be examined. The results show that the easily calculated topological descriptors explain the majority of the variance and that the addition of geometrical and quantum chemical descriptors does not result in a significantly improved model.

1. INTRODUCTION

One of the major interests in environmental sciences and ecotoxicology is the assessment of fate, transport, and distribution of chemicals. The distribution of chemicals among water, air, and soil is governed primarily by some key physicochemical properties such as vapor pressure, water solubility, air−water partition coefficient, soil sorption, Henry's law constant, etc.[1−3]

Vapor pressure (VP) plays a critical role in the transport, fate, and distribution of pollutants in the ecosystem and hence determines the acceptability of chemical substances and processes that produce them. For environmental pollutants, a knowledge of their VP gives an idea about their distribution between the atmosphere and the soil. In the case of a toxic chemical spill, its VP can be used to estimate the rate at which the chemical evaporates to the atmosphere. VP data are also used in the estimation of liquid viscosity, enthalpy of vaporization, air−water partition coefficient, and flash points.[4]

The vapor pressure of the ever increasing number of chemicals cannot be determined in the laboratory due to the lack of resources and facilities. Another problem area is the VP of low-volatility chemicals where there are analytical difficulties in their determination. Estimation of vapor pressure can be accomplished utilizing models where one uses other physicochemical properties. For example, Mackay et al.[5] developed a VP model using boiling point (T_b) and melting point (T_m), and Banerjee et al.[6] developed a model using molar volume (V), polarizability/dipolarity (π^*), and melting point (T_m). In such models, physicochemical properties are the independent variables which, in turn, either have to be determined in the laboratory (which is again costly and time-consuming) or derived from various estimation methods. Quantitative structure−property relationship (QSPR) models provide a viable alternative by estimating VP using parameters which can be calculated directly from molecular structure of chemicals without any input of experimental data.[7−11]

Research in many scientific fields including toxicology and drug design as well as regulation of chemicals have become increasingly dependent on predictive quantitative structure−property/activity relationships (QSPRs/QSARs). While many studies have been carried out using property values rather than structural information in order to develop predictive models, the limitations of this approach due to a profound lack of experimental data has become painfully obvious. For example, there is no property data at all available for approximately 50% of the roughly 80 000 chemicals listed currently in the Toxic Substances Control Act (TSCA) Inventory.[12] With the advent of combinatorial chemistry, where it is possible to quickly create libraries containing thousands of real or hypothetical chemicals, the number of chemical structures without accompanying property data is increasing dramatically. The cost involved in the laboratory testing of such large numbers of chemicals is prohibitive. Theoretical molecular descriptors, on the other hand, can be calculated algorithmically for any chemical—real or hypothetical. These descriptors can be calculated for large numbers of chemicals at a minimal cost.

We have recently formulated a hierarchical approach where calculated parameters of increasing complexity, viz., topostructural, topochemical, geometrical, and quantum chemical parameters, have been used in the formulation of QSARs/QSPRs for the prediction of biomedicinal, physicochemical, and toxicological properties.[7,13−20] In an earlier paper, Basak et al.[7] used a set of calculated molecular descriptors to estimate VP of a set of 476 chemicals with good results. In this paper, we have formulated QSPR models on a set of 469 compounds with an expanded set of calculated parameters which led to a significant improvement in the predictive power of the model.

2. MATERIALS AND METHODS

2.1. Database. The set of 469 chemicals used in this study is a subset of the 476 chemicals used in the previous study

* Corresponding author phone: (218)720-4230; fax: (218)720-4328; e-mail: sbasak@nrri.umn.edu.

10.1021/ci000165r CCC: $20.00
Published on Web 04/11/2001

Table 1. Chemical Class Composition of the Vapor Pressure Data Set

compd classification	no. of compds	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.

by Basak et al.[7] which was obtained from the Assessment Tools for the Evaluation of Risk (ASTER) database[21] and represents a subset of the Toxic Substances Control Act (TSCA) Inventory[12] for which vapor pressure (p_{vap}) was measured at 25 °C with a pressure range of approximately 3−10 000 mmHg. Of the 476 chemicals in the original data set, seven were two-atom compounds and thus not included in the present study due to the fact that the triplet indices, which are described in the following section, cannot be calculated for two-atom compounds. The molecular weights of the compounds in this data set range from 40 to 338, and the chemical diversity is described in Table 1.

2.2. Calculation of Molecular Descriptors. The majority of the topological descriptors were calculated using software developed by Basak et al., including POLLY 2.3[22] and H-Bond.[23] The topological descriptors include Wiener number,[24] molecular connectivity indices developed by Randić[25] and Kier and Hall,[26] frequency of path lengths of varying size,[26] information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[27] Roy et al.,[28] Basak et al.[29−32] as well as those of Raychaudhury et al.,[33] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[29−33] a simple hydrogen bonding parameter,[23] and Balaban's J indices[34] as well as the triplet indices.[35,36] 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, and 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, $\sum_{ij}(x_i x_j)^{-1/2}$; and (5) product, $N(\sum_i x_i)^{1/N}$.

Additional topological descriptors, including an extended set of molecular connectivity indices, electrotopological state descriptors, general polarity descriptors, and hydrogen bonding descriptors, were calculated by Molconn-Z 3.50.[37] A total number of 363 topological descriptors was calculated for use in the current study.

Ten geometrical descriptors were used, including six kappa shape indices which were also calculated by Molconn-Z. van der Waals volume, V_W, was calculated using Sybyl 6.2.[38] 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 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 semiempirical Hamiltonian using MOPAC 6.0 in the Sybl interface.[39]

A complete list of the 379 parameters calculated for use in the current study, including brief descriptions, is provided in Table 2.

2.3. Hierarchical QSPR. QSPR modeling was performed in a hierarchical fashion, utilizing descriptor classes of increasing complexity. The set of topological descriptors was partitioned into two distinct subsets: topostructural (TS) and topochemical (TC). Topostructural descriptors encode information about the adjacency and distances of atoms in molecular structures irrespective of the chemical nature of the atoms, while topochemical descriptors encode information regarding the connectivity as well as specific chemical properties of the atoms making up the molecule. The topostructural descriptors are at the lowest level of the hierarchy, followed by the topochemical, the geometrical (3D), and finally the quantum chemical (QC). Initially, a model is developed utilizing only TS descriptors, after which TC descriptors are added to the parameters in the TS model, and the regression analysis is repeated to obtain a TS + TC model. This process is repeated, resulting in a TS + TC + 3D model and finally a TS + TC + 3D + QC model.

2.4. Data Reduction and Statistical Analysis. Prior to analysis, all calculated descriptors with nonnegative values were transformed by ln(descriptor value + 1) due to the fact that their scales differed by several orders of magnitude, and some had a value of 0. A constant was added to the descriptors with negative values, such that their sum was greater than 0, before taking the natural logarithm. Any descriptor with a value of 0 for all compounds was removed from the pool and not used in subsequent analyses. Perfectly correlated descriptors, i.e. those having a correlation coefficient of 1.0, were identified using the CORR procedure[40] of the SAS statistical software package. In each case, only one of the perfectly correlated descriptors was retained and used in subsequent analyses.

The number of descriptors remaining (268) was still too large with respect to the number of observations (469) and had to be decreased in order to reduce the probability of chance correlations. To this end, the VARCLUS[40] procedure of the SAS software package was used. In this procedure, each descriptor is assigned to one, and only one, cluster. From each cluster, we selected the one descriptor most highly correlated and all descriptors poorly correlated ($R^2 < 0.7$) with that cluster. In the event that there were two most highly correlated descriptors for a given cluster, the descriptor which was least correlated with the next closest cluster was selected.

Table 2. Symbols, Definitions, and Classification of Molecular Descriptors

	Topostructural (TS)
I^{W}_{D}	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}^{W}_{D}$	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
$DN^{2}S_{y}$	triplet index from distance matrix, square of graph order, and distance sum; operation $y = 1-5$
$DN^{2}1_{y}$	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 $y = 1-5$
$DN^{2}N_{y}$	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 $y = 1-5$
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 $h = 0-6$
${}^{h}\chi^{v}_{C}$	valence cluster connectivity index of order $h = 3-6$
${}^{h}\chi^{v}_{Ch}$	valence chain connectivity index of order $h = 3-6$
${}^{h}\chi^{v}_{PC}$	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 $y = 1-5$
$DN^{2}Z_{y}$	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
fw	molecular weight
${}^{h}\chi^{v}$	valence path connectivity index of order $h = 7-10$
${}^{h}\chi^{v}_{Ch}$	valence chain connectivity index of order $h = 7-10$
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

Table 2 (Continued)

	Topochemical (TC)
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 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, -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
$NHBint_y$	count of potential internal hydrogen bonders ($y = 2-10$)
$SHBint_y$	E-state descriptors of potential internal hydrogen bond strength ($y = 2-10$)
	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, SssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
	Geometrical (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

Table 3. Descriptors Used in Hierarchical Model Development (Eqs 1−3) and the Type I Nonhierarchical Model (Eq 4) for the Set of 469 Diverse Chemicals[a]

descriptor class	descriptors
TS	$^6\chi\ ^7\chi\ ^8\chi\ ^4\chi_C\ ^6\chi_C\ 4\chi_{PC}\ 6\chi_{PC}\ 3\chi_{Ch}\ 5\chi_{Ch}\ 8\chi_{Ch}\ P_6\ P_7\ P_9\ J\ W\ \overline{IC}\ ASN_2\ DN^2N_5\ DN^2S_4\ ASV_1\ DSV_1\ ANV_2\ ANV_5\ NRINGS$
TC	$^0\chi^v\ ^3\chi^v\ ^4\chi^v\ ^8\chi^v\ ^5\chi^b\ ^3\chi^v_{Ch}\ ^6\chi^v_{Ch}\ ^3\chi^b_C\ ^5\chi^b_C\ ^5\chi^b_{PC}\ J^Y\ O_{orb}\ IC_1\ IC_5\ CIC_1\ CIC_4\ SIC_0\ ASZ_2\ AZV_1\ AZV_4\ AZN_4\ TOTOP\ TETS_2$ IDCBAR SHOTHER NELEM HMIN $SSCH_3$ SSBR SUMI PHIA HMAX SHCHNX SSCL SSSSN $SSNH_2$ SAACH SAAS SSSNH HB_1 STCH STSC SSF SHHBA SHVIN SHTVIN SHSSH SSSS SDDC SDSN SDS $SHBINT_3$ SDO SSSO NUMHBD $NHBINT_2$ STN SHCSATS SSI SAAO SHSOH
3D	$V_W\ ^{3D}W\ ^{3D}W_H\ KP_0-KP_3\ KA_1-KA_3$
QC	$E_{HOMO}\ E_{HOMO-1}\ E_{LUMO}\ E_{LUMO+1}\ \Delta Hf\ \mu$

[a] TS and TC descriptors were clustered independently. All 3D and QC descriptors were used.

The VARCLUS[40] procedure was used in two types of variable selection methodologies:

(1) Variable clustering was performed independently on each of the TS and TC classes. The TS and TC descriptors selected by variable clustering, the 10 3D descriptors, and the six QC descriptors (a total of 101 descriptors) were used in regression analysis. A list of these descriptors can be found in Table 3.

(2) The entire pool of descriptors (268), regardless of hierarchical class, was clustered together, and the selected descriptors (68) were used in regression analysis. A list of these descriptors can be found in Table 4.

For the sake of brevity, the actual clustering results, which indicate cluster membership and correlation coefficients for each descriptor with its own cluster and with the next closest cluster, have not been included in this paper for all three of the clustering procedures which were performed in this study (TS, TC, and all descriptors). However, the clustering results for methodology (2) above, in which all descriptors were clustered together, is available in Table 5.

The descriptors selected via methodology (1) above were used both in the hierarchical model development and in the development of one type of nonhierarchical model (Type I). The descriptors selected via methodology (2) above were

Table 4. Descriptors Used in Type II Nonhierarchical Model Development (Eq 5) for the Set of 469 Diverse Chemicals, Selected by Variable Clustering of the Entire Pool of Descriptors

descriptor class	descriptors
TS	$^{7}\chi$ $^{8}\chi$ $^{3}\chi_{Ch}$ $^{6}\chi_{Ch}$ $^{8}\chi_{Ch}$ W H^{V} $\overline{IC}$ P_5 DN^2S_4 ASV_5 DSV_2 $DS1_3$ ANV_3
TC	$^{3}\chi^{b}_{C}$ $^{5}\chi^{b}_{C}$ $^{5}\chi^{b}_{Ch}$ $^{6}\chi^{v}$ $^{5}\chi^{b}_{PC}$ J^{Y} IC_1 IC_5 SIC_1 CIC_4 O_{orb} AZN_4 AZV_4 ASZ_2 NELEM HMIN $SSCH_3$ QV SSSSN SSBR NUMHBD HB_1 TOTOP SHNHCX SSCL SHSOH HMAX STCH STSC SUMDELI SDO SSSO SSSH SAACH SAAS SDDC SDSN SDS SSSNH SHTVIN SUMI SSSS SHOTHER $SHBINT_3$ SSF STN $NHBINT_2$ SHCSATU SHVIN SSI SAAO
3D	KP_3
QC	E_{HOMO} E_{HOMO-1} E_{LUMO} μ

used in the development of a second type of nonhierarchical model (Type II). In addition, TC-only, 3D-only, and QC-only models were developed for comparison purposes.

Two regression procedures from the SAS[40] statistical software package were used to develop the models. When the number of independent variables was less than approximately 20−25, the all possible subset regression option was feasible. However, this option is very time intensive when dealing with a large number of independent variables; in which case, the stepwise regression option which optimizes the improvement of the explained variance (R^2) was used.

In addition to constructing models based on the entire data set, the 469 chemicals were partitioned into a training set ($N = 351$) and a test set ($N = 118$), with an approximate 25/75 split. Models developed using the training set were then used to predict the vapor pressure of the chemicals in the test set. The partitioning was performed using the RANUNI function (a random number generator) of the SAS statistical package.[40]

3. RESULTS

3.1. Regression Analysis of the Complete Set of 469 Chemicals. The topostructural descriptors obtained through variable clustering were used in regression analysis, yielding the following five-parameter topostructural model. The inclusion of additional descriptors does not result in a significantly improved model. For example, even the best 12-parameter TS model explains only 49.8% of the variance with a standard error of 0.53.

$$\log (p_{vap}) = -0.64\ ^{6}\chi + 3.56\ ^{8}\chi + 0.73\ ^{6}\chi_{c} - 0.47\ W - 0.92\ ANV_5 + 4.03 \quad (1)$$

$$n = 496, \quad R^2 = 48.7\%, \quad s = 0.53, \quad F = 87.9$$

Regression analysis utilizing descriptors from the above topostructural model and those topochemical descriptors obtained through variable clustering results in a 12-parameter TS + TC model, with only one of the topostructural descriptors retained. There is a significant increase in the explained variance as well as a significant decrease in the standard error.

$$\log (p_{vap}) = 0.59\ ^{3}\chi^{b}_{C} + 3.39\ ^{8}\chi - 2.70\ AZV_4 - 0.85\ HMIN - 0.78\ SHSOH + 0.26\ SSCL + 0.56\ SSF + 0.44\ SSSO - 0.28\ STN + 0.73\ SAAO - 0.25\ SHCSATS - 0.28\ SHHBA + 11.0 \quad (2)$$

$$n = 496, \quad R^2 = 90.3\%, \quad s = 0.23, \quad F = 354$$

When the geometrical descriptors are added to those in the TS + TC model and regression repeated, we do not find an improved model. Including as many as 19 descriptors in the model results in an increase of only 0.2% in R^2 and no decrease in the standard error.

Again repeating the regression, now adding the quantum chemical descriptors to those included in eq 2, we do find a model with slightly improved R^2 and F values. Nine of the descriptors from the previous model have been retained, and three have been replaced by quantum chemical descriptors.

$$\log (p_{vap}) = 3.65\ ^{8}\chi + 0.61\ ^{3}\chi^{b}_{C} - 2.79\ AZV_4 - 0.91\ SHSOH + 0.14\ SSCL + 0.46\ SSF + 0.31\ SSSO - 0.32\ STN - 0.18\ SHHBA - 0.14\ E_{HOMO-1} - 0.09\ E_{LUMO} - 0.14\ \mu + 9.51 \quad (3)$$

$$n = 496, \quad R^2 = 90.6\%, \quad s = 0.23, \quad F = 364$$

In addition to performing hierarchical modeling, we developed models using each descriptor class independently in order to examine the performance of each class. The topochemical out-performed all other classes of descriptors, followed by the geometrical, the topostructural, and finally the quantum chemical, which produced a very poor model.

Two nonhierarchical models were developed; the first of which (Type I) was obtained utilizing those TS and TC parameters selected by the independent variable clustering of each respective descriptor class as well as all geometrical and all quantum chemical descriptors (Table 3):

$$\log (p_{vap}) = -2.59\ ^{0}\chi^{v} - 0.80\ ^{6}\chi + 0.86\ P_9 - 3.08\ SIC_0 - 2.11\ ANV_5 - 1.00\ SHSOH - 0.34\ STN + 0.20\ SSSO + 0.21\ SSF - 0.19\ HB_1 - 0.13\ E_{HOMO-1} - 0.17\ \mu + 7.40 \quad (4)$$

$$n = 496, \quad R^2 = 91.3\%, \quad s = 0.22, \quad F = 398$$

The second nonhierarchical model (Type II) was developed utilizing parameters selected by variable clustering of the entire set of 268 descriptors used in this study (Table 4):

$$\log (p_{vap}) = 0.61\ ^{3}\chi^{b}_{c} + 3.87\ ^{8}\chi - 2.81\ AZV_4 - 0.94\ SHSOH + 0.26\ SSSO - 0.38\ STN + 0.32\ SSF - 0.13\ HB_1 - 0.19\ \mu - 0.15\ E_{HOMO-1} - 0.08\ E_{LUMO} + 9.35 \quad (5)$$

$$n = 496, \quad R^2 = 90.3\%, \quad s = 0.23, \quad F = 388$$

It is interesting to note the similarity between the two nonhierarchically developed models (eqs 4 and 5), as well as the similarity between these models and the final hierarchical model given in eq 3. Although each of these models was derived using a different method, they each contain two connectivity indices, the same four E-state

Table 5. Variable Clustering Results Using All 268 Descriptors and the Complete Set of 469 Diverse Compounds[a]

	R^2 with			R^2 with			R^2 with			R^2 with	
variable	own cluster	next closest	variable	own cluster	next closest	variable	own cluster	next closest	variable	own cluster	next closest
						Cluster 1					
I^W_D	0.9954	0.8526	$AS1_1$	0.9265	0.7925	ANS_2	0.9622	0.7241	AZS_1	0.9820	0.7904
$\overline{I}^W_D$	0.9537	0.8964	$AS1_2$	0.8823	0.8571	$AN1_1$	0.8096	0.6803	AZS_2	0.9660	0.7565
W	**0.9985**	**0.8282**	$AS1_4$	0.9854	0.8762	$AN1_3$	0.9902	0.8877	AZN_1	0.9719	0.8518
I^D	0.9665	0.8821	$AS1_5$	0.9227	0.7993	$AN1_5$	0.9395	0.7170	AZN_2	0.9570	0.8343
H^D	0.9515	0.9082	$DS1_1$	0.9355	0.8547	ANN_1	0.9867	0.8923	AZN_3	0.9804	0.8693
$^0\chi$	0.9576	0.8465	$DS1_4$	0.9836	0.8502	ANN_2	0.9625	0.8823	AZN_5	0.9590	0.8364
$^1\chi$	0.9720	0.8886	$DS1_5$	0.9046	0.7970	ANN_3	0.9894	0.8950	IDC	0.9877	0.7906
P_0	0.9874	0.9030	ASN_4	0.9758	0.8588	ANN_5	0.9920	0.8788	V_W	0.8813	0.7039
DN^2S_1	0.9638	0.7365	DSN_4	0.9673	0.8512	ANV_4	0.9712	0.9264	$^{3D}W_H$	0.8048	0.7523
DN^2S_3	0.9968	0.8154	DN^2N_2	0.9643	0.8219	$^0\chi^b$	0.9178	0.8190	^{3D}W	0.9934	0.8210
DN^2S_5	0.9675	0.7323	DN^2N_3	0.9683	0.9367	$^1\chi^b$	0.8944	0.8068	KP_0	0.7703	0.6782
DN^21_1	0.9666	0.8406	DN^2N_4	0.9718	0.9138	AZV_1	0.9263	0.9230	KP_1	0.8260	0.6061
DN^21_4	0.9786	0.9123	ANS_1	0.9911	0.7833	AZV_3	0.9385	0.9136	KA_1	0.7994	0.6081
						Cluster 2					
$^4\chi_{PC}$	0.8064	0.5571	$^4\chi^b_{PC}$	0.8437	0.5277	$^6\chi^b_{PC}$	0.8492	0.4938	$^5\chi^v_{PC}$	0.8848	0.5229
$^5\chi_{PC}$	0.9276	0.6291	$^5\chi^b_{PC}$	**0.9531**	**0.5770**	$^4\chi^v_{PC}$	0.7810	0.4675	$^6\chi^v_{PC}$	0.8301	0.4806
$^6\chi_{PC}$	0.8316	0.5458									
						Cluster 3					
H^V	**0.9261**	**0.7521**	ASV_1	0.8431	0.7371	ANV_2	0.7456	0.4718	KP_2	0.8774	0.7028
$\overline{IC}$	**0.6972**	**0.4680**	DSV_1	0.9188	0.7510	IDCBAR	0.9094	0.7586	KA_2	0.8585	0.7339
DN^2S_2	0.9210	0.7384									
						Cluster 4					
DN^21_3	0.9970	0.4954	DSN_1	0.9698	0.5553	DN^2N_1	0.9975	0.4919	DN^2N_5	0.9968	0.4985
$\mathbf{DS1_3}$	**0.9983**	**0.5080**	DSN_5	0.9817	0.5149						
						Cluster 5					
IC_0	0.8695	0.5622	$\mathbf{SIC_1}$	**0.9427**	**0.4774**	CIC_2	0.7504	0.5960	**SSCH3**	**0.6198**	**0.4079**
$\mathbf{IC_1}$	**0.5056**	**0.4396**	CIC_0	0.8571	0.6424	**NELEM**	**0.6896**	**0.5685**	**QV**	**0.6970**	**0.6569**
SIC_0	0.9094	0.5620	CIC_1	0.9318	0.6235	**HMIN**	**0.2793**	**0.1171**			
						Cluster 6					
$^5\chi$	0.8628	0.5713	$\mathbf{P_5}$	**0.8720**	**0.6079**	$^5\chi^b$	0.8498	0.5711	$^5\chi^v$	0.7990	0.5403
$^6\chi$	0.8326	0.4912	P_6	0.8636	0.4437	$^6\chi^b$	0.8346	0.3437	$^6\chi^v$	0.8207	0.3331
						Cluster 7					
SIC_2	0.8262	0.4502	SIC_4	0.9190	0.5810	CIC_3	0.9198	0.4286	CIC_5	0.9534	0.3260
SIC_3	0.9503	0.4860	SIC_5	0.9142	0.5872	$\mathbf{CIC_4}$	**0.9549**	**0.3344**	**SSSSN**	**0.0250**	**0.0141**
						Cluster 8					
ASZ_1	0.9601	0.6180	ANZ_2	0.9621	0.4275	DSZ_2	0.9341	0.5766	DN^2Z_2	0.9097	0.4748
$\mathbf{ASZ_2}$	**0.9761**	**0.6299**	DSZ_1	0.8855	0.7642	DN^2Z_1	0.9579	0.7026	**SSBR**	**0.3885**	**0.0527**
ANZ_1	0.7821	0.2697									
						Cluster 9					
$\mathbf{^3\chi_{Ch}}$	**0.9014**	**0.1369**	$^4\chi_{Ch}$	0.8785	0.2030	$^3\chi^v_{Ch}$	0.8059	0.0884	$^4\chi^v_{Ch}$	0.8369	0.1538
						Cluster 10					
$SHSNH_2$	0.7182	0.0902	**NUMHBD**	**0.8165**	**0.4173**	SHHBD	0.7277	0.5997	$\mathbf{HB_1}$	**0.6580**	**0.3809**
$SSNH_2$	0.7145	0.0677									
						Cluster 11					
M_1	0.9876	0.8646	$\mathbf{DN^2S_4}$	**0.9920**	**0.8572**	DSN_3	0.9875	0.8606	AZV_2	0.9580	0.8049
M_2	0.9770	0.8394	$AS1_3$	0.9330	0.7042	$AN1_4$	0.9666	0.9324	**TOTOP**	**0.6857**	**0.6143**
$^2\chi$	0.9322	0.7448	ASN_1	0.9378	0.7103	ANN_4	0.9759	0.8281	$TETS_2$	0.7302	0.6607
P_1	0.9394	0.9323	ASN_3	0.9738	0.9237	$^2\chi^b$	0.8155	0.6918	PF	0.9830	0.7942
P_2	0.9876	0.7925	ASN_5	0.9250	0.7051						
						Cluster 12					
$^5\chi_C$	0.9257	0.3132	$\mathbf{^5\chi^b_C}$	**0.9510**	**0.3177**	$^5\chi^v_C$	0.7425	0.2871	$\mathbf{^6\chi^v_C}$	**0.6941**	**0.1294**
$^6\chi_C$	0.8816	0.2849	$^6\chi^b_C$	0.8774	0.2860						
						Cluster 13					
P_8	0.9826	0.3434	P_9	0.7317	0.1831	$\mathbf{^8\chi}$	**0.9910**	**0.3258**	$^8\chi^v$	0.7443	0.3100
						Cluster 14					
SHCHNX	**0.9636**	**0.1916**	**SSCL**	**0.6372**	**0.2053**	NUMWHBD	0.9593	0.1721	SHWHBD	0.9590	0.1950
						Cluster 15					
$^3\chi_C$	0.8900	0.4731	$\mathbf{^3\chi^b_C}$	**0.9193**	**0.4575**	$^3\chi^v_C$	0.8373	0.3612	$^4\chi^v_C$	0.8234	0.1478
$^4\chi_C$	0.9040	0.2337	$^4\chi^b_C$	0.9036	0.2299						
						Cluster 16					
O	0.7798	0.4153	IC_2	0.7441	0.5218	IC_4	0.9767	0.4277	SI	0.9500	0.3881
I_{ORB}	0.9399	0.3578	IC_3	0.9532	0.3922	$\mathbf{IC_5}$	**0.9776**	**0.4325**	NCLASS	0.9099	0.4585
$\mathbf{O_{orb}}$	**0.5996**	0.3202									

Table 5 (Continued)

variable	R^2 with own cluster	R^2 with next closest	variable	R^2 with own cluster	R^2 with next closest	variable	R^2 with own cluster	R^2 with next closest	variable	R^2 with own cluster	R^2 with next closest
					Cluster 17						
SHSOH	**0.9328**	**0.2594**	**HMAX**	**0.5149**	**0.2227**	SSOH	0.9323	0.2313			
					Cluster 18						
SHTCH	0.9404	0.0213	**STCH**	**0.9405**	**0.0214**	**STSC**	**0.6613**	**0.0660**			
					Cluster 19						
SUMDELI	**0.6720**	**0.4506**	**SDO**	**0.5226**	**0.1359**	NUMHBA	0.7570	0.5451	μ	**0.5602**	**0.4499**
GMAX	0.8308	0.4514	**SSSO**	**0.3197**	**0.0594**	SHHBA	0.8705	0.5595			
					Cluster 20						
$^6\chi_{Ch}$	**0.9497**	**0.3529**	NCIRC	0.8488	0.3683	$^6\chi^b_{Ch}$	0.8951	0.3689	$^6\chi^v_C$	0.8763	0.3917
NRINGS	0.8542	0.3672									
					Cluster 21						
SHSSH	1.0000	0.0131	**SSSH**	**1.0000**	**0.0129**						
					Cluster 22						
SAACH	**0.9007**	**0.1528**	**SAAS**	**0.4891**	**0.0907**	SHAROM	0.8770	0.1530	**E_{HOMO-1}**	**0.5498**	**0.1764**
					Cluster 23						
SDDC	**0.6835**	**0.0289**	**SDSN**	**0.6631**	**0.0298**	**SDS**	**0.8060**	**0.0216**			
					Cluster 24						
SHSSNH	0.9998	0.0563	**SSSNH**	**0.9998**	**0.0562**						
					Cluster 25						
$SDCH_2$	0.9983	0.1113	**SHTVIN**	**0.9983**	**0.1096**						
					Cluster 26						
J	0.8335	0.4565	J^X	0.9280	0.4131	**SUMI**	**0.5338**	**0.3821**	**SSSS**	**0.0959**	**0.0272**
J^B	0.9629	0.3456	**J^Y**	**0.9685**	**0.3495**						
					Cluster 27						
$^5\chi_{Ch}$	0.9748	0.4084	**$^5\chi^b_{Ch}$**	**0.9829**	**0.3917**	$^5\chi^v_{Ch}$	0.9763	0.4082			
					Cluster 28						
DN^21_2	0.7763	0.7531	$AN1_2$	0.9131	0.8918	**DSV_2**	**0.9555**	**0.6925**	**SHOTHER**	**0.5780**	**0.5006**
$DS1_2$	0.8466	0.7536	ASV_2	0.9330	0.6714	ANV_5	0.7834	0.7401			
DSN_2	0.8833	0.7849	**ASV_5**	**0.6139**	**0.3636**						
					Cluster 29						
$^3\chi$	0.8863	0.7660	P_4	0.8928	0.6461	$^3\chi^b$	0.8587	0.7353	$^4\chi^v$	0.7941	0.5111
$^4\chi$	0.8774	0.6275	ANV_1	0.9428	0.8426	$^4\chi^b$	0.8609	0.6109	AZV_5	0.8775	0.7835
P_3	0.9302	0.7942	**ANV_3**	**0.9444**	**0.9038**	$^3\chi^v$	0.7634	0.6028	WT	0.9438	0.8706
					Cluster 30						
$NHBINT_3$	0.9968	0.1259	**$SHBINT_3$**	**0.9968**	**0.1088**						
					Cluster 31						
SSF	**0.7701**	**0.1202**	ΔHf	0.7701	0.2161						
					Cluster 32						
P_7	0.9466	0.3481	**$^7\chi$**	**0.9610**	**0.4520**	**$^8\chi_{Ch}$**	**0.1308**	**0.0296**	$^7\chi^v$	0.8400	0.3162
					Cluster 33						
$^0\chi^v$	0.8822	0.7630	$^2\chi^v$	0.8719	0.5619	**AZN_4**	**0.4838**	**0.3515**	**STN**	**0.0614**	**0.0327**
$^1\chi^v$	0.8562	0.6655	**AZV_4**	**0.9029**	**0.7507**	FW	0.8401	0.3607			
					Cluster 34						
ASN_2	0.8685	0.7028	PHIA	0.7772	0.7160	**KP_3**	**0.8758**	**0.4559**	KA_3	0.8684	0.3948
					Cluster 35						
$NHBINT_2$	**0.9999**	**0.2325**	$SHBINT_2$	0.9999	0.2325						
					Cluster 36						
SHCSATS	0.7107	0.3141	**E_{LUMO}**	**0.8669**	**0.3778**	E_{LUMO+1}	0.7256	0.4721			
					Cluster 37						
SHCSATU	**0.6978**	**0.0793**	**SHVIN**	**0.6238**	**0.0999**	**E_{HOMO}**	**0.5703**	**0.2009**			
					Cluster 38						
SSI	**1.0000**	**0.1167**									
					Cluster 39						
SAAO	**1.0000**	**0.0178**									

[a] Variables selected for regression analysis are indicated in bold type.

descriptors, one descriptor which encodes information about hydrogen bonding, the same 2−3 quantum chemical descriptors, and one triplet descriptor derived from the adjacency matrix and V (vertex degree) as the free term column vector. These three models are also very similar with respect to statistical measures.

A summary of all models developed on the full set of 469 compounds is given in Table 6. A scatter plot of the observed

Table 6. Summary of Regression Results for the Complete Data Set Containing 469 Diverse Chemicals

descriptor class	complete data set ($N = 469$) descriptors	F	R^2	s
TS	$^{6}\chi_c$, $^{6}\chi$, $^{8}\chi$, W, ANV_5	87.9	48.7	0.53
TS + TC	$^{3}\chi^{b}_c$, 8χ, AZV_4, HMIN, SHSOH, SSCl, SSF, SSSO, STN, SAAO, SHCSATS, SHHBA,	354	90.3	0.23
TS + TC + QC	$^{3}\chi^{b}_c$, $^{8}\chi$, AZV_4, SHSOH, SSCl, SSF, SSSO, STN, SHHBA, E_{HOMO-1}, E_{LUMO}, μ	364	90.6	0.23
TC	$^{0}\chi^{v}$, $^{1}\chi^{v}$, SSF, SDO, SDSN, HMIN, SHSOH, STN, QV, NUMHBD	342	88.2	0.26
3D	^{3D}W, $^{3D}W_H$, V_W, KP_1, KA_2	120	56.4	0.49
QC	ΔHf, μ, E_{LUMO}, E_{HOMO-1}	46.4	28.6	0.63
all classes (Type I) (nonhierarchical)	$^{0}\chi^{v}$, $^{6}\chi$, P_9, SIC_0, ANV_5, SHSOH, STN, SSSO, SSF, HB_1, E_{HOMO-1}, μ	398	91.3	0.22
all classes (Type II) (nonhierarchical)	$^{3}\chi^{b}_c$, $^{8}\chi$, AZV_4, SHSOH, STN, SSSO, SSF, HB_1, E_{HOMO-1}, μ, E_{LUMO}	388	90.3	0.23

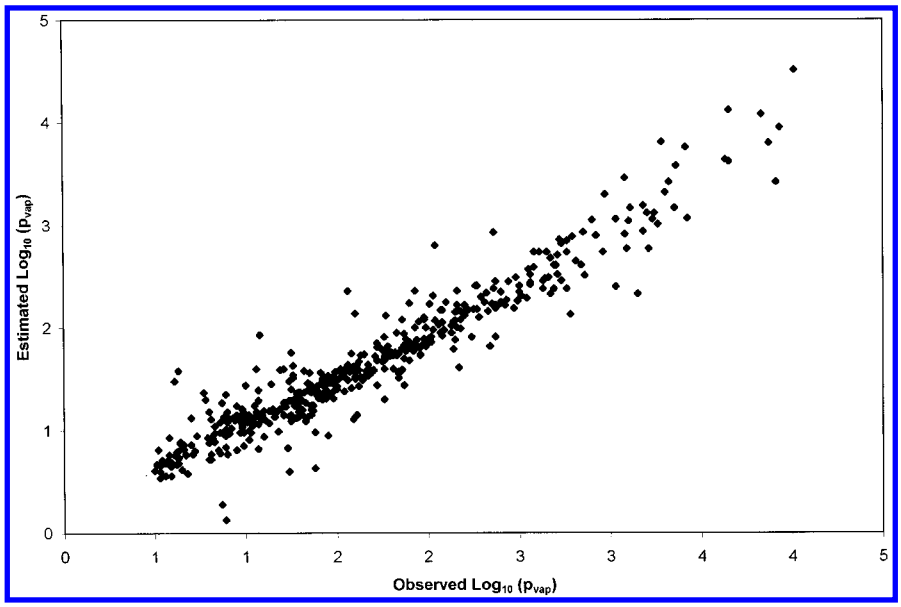

Figure 1. Scatter plot of observed log (p_{vap})vs estimated log (p_{vap}) using eq 4 for the set of 469 diverse compounds.

Table 7. Summary of Regression Results for Training and Test Sets

descriptor class	training set ($N = 351$) descriptors	F	R^2	s	test set ($N = 118$) R^2	s
TS	$^{8}\chi$, J, ANS_1	111	48.9	0.53	39.1	0.58
TS + TC	$^{3}\chi^{b}_c$, $^{4}\chi^{v}$, $^{8}\chi$, AZV_4, $TETS_2$, SSF, SDO, SDSN, STN, SHHBD	228	87.0	0.27	85.9	0.28
TS + TC + QC	$^{8}\chi$, $^{3}\chi^{b}_c$, AZV_4, SSF, SDSN, STN, SHHBD, E_{HOMO-1}, E_{LUMO}, μ	250	88.0	0.26	84.8	0.29
TC	$^{3}\chi^{b}_c$, $^{4}\chi^{v}$, AZV_4, $TeTS_2$, SSF, SDO, SDSN, STN, SHHBD	234	86.1	0.28	85.9	0.28
3D	^{3D}W, $^{3D}W_H$, V_W, KA_1, KA_2	88.3	56.1	0.49	54.0	0.50
QC	E_{HOMO}, E_{HOMO-1}, ΔHf	54.7	32.1	0.61	11.7	0.78
all classes (Type I) (nonhierarchical)	$^{3}\chi^{b}_c$, $^{6}\chi_{PC}$, $^{8}\chi$, $^{8}\chi^{v}$, AZV_4, SSF, STN, $NHBINT_2$, SHHBD, E_{HOMO-1}, E_{LUMO}, μ	250	89.9	0.24	81.0	0.32
all classes (Type II) (nonhierarchical)	$^{3}\chi^{b}_c$, $^{8}\chi$, AZV_4, SSF, STN, SSSSO, SHSOH, $NHBINT_3$ E_{HOMO-1}, E_{LUMO}, μ,	300	90.7	0.23	85.4	0.29

versus the estimated log (p_{vap}) values using eq 4 is shown in Figure 1.

3.2. Regression Analysis of the Training and Test Sets. Summary statistics for all models developed on the training set and prediction results for the test set are provided in Table 7. A topostructural model consisting of three parameters explains 48.9% of the variance with a standard error of 0.53. The addition of topochemical descriptors improved the model significantly, increasing R^2 to 87.0% and decreasing s to 0.27. The improvement due to the addition of the topochemical descriptors is also evident in the test set statistics, with R^2 increasing from 39.1% to 85.9%, and s decreasing from 0.58 to 0.28. When the geometrical descriptors are added, we find again, as was the case in analysis of the complete data set, an improved model is not obtained. The addition of quantum chemical descriptors yields a model showing only slight improvement with a slight decrease in the effectiveness of that model as evidenced by the test set results. The topochemical class independently produced a good model with good predictive ability. Neither the geometrical nor the quantum chemical descriptor classes produced acceptable models, with the quantum chemical model notably poor.

As was done with the full data set, two nonhierarchical models were developed using the training set chemicals. Again, the first of these models (Type I) was derived using TS and TC parameters selected by independent variable clustering of the respective classes, in addition to all geometrical and all quantum chemical descriptors, and the

Table 8. QSPR Vapor Pressure Prediction Models

model type	no. of parameters	N	R^2	s	F	investigators	ref no.
CNN	8	65/420[a]	NR[b]	0.37	NR[b]	McClelland & Jurs (2000)	10
CNN	10	65/420[a]	NR[b]	0.33	NR[b]	McClelland & Jurs (2000)	10
CNN	7	52/352[a]	NR[b]	0.209	NR[b]	Goll & Jurs (1999)	11
MLR	7	479	0.960	0.534	NR[b]	Liang & Gallagher (1998)	8
MLR	5	411	0.949	0.331	1511.4	Katritzky et al. (1998)	9
MLR	10	476	0.843	0.29	249.5	Basak et al. (1997)	7
MLR	12	469	0.913	0.22	398	Basak et al.	current study

[a] Reported error is on external prediction set. Notation indicates: $N_{\text{prediction set}}/N_{\text{total}}$. [b] NR = not reported.

second (Type II) was derived using parameters selected by variable clustering of the entire set of descriptors used in this study. While both methods resulted in acceptable models, the Type II model was the best overall training set model, with good predictive ability as indicated by the test set statistics. The Type I is very similar in terms of statistical measures; however, its predictive ability is slightly diminished as compared to that of the Type II model, the TC model, the TS + TC model, or the TS + TC + QC model.

4. DISCUSSION

The primary objective of the present study was to investigate the usefulness of an expanded set of theoretical molecular descriptors consisting of the major classes of topostructural as well as topochemical parameters in the prediction of vapor pressure. In an earlier study, Basak et al.[7] used a set of 97 indices to predict VP for a set of 476 molecules. In this paper, we have used a group of 268 theoretical parameters in the development of QSPR models for a set of 469 diverse chemicals.

We have developed QSPR models using our hierarchical approach as well as nonhierarchical methods (Tables 6 and 7). It was found that the hierarchical model and those developed from the full set of descriptors using nonhierarchical methods have similar predictive quality. It is interesting to note that topostructural and topochemical parameters explained most of the variance in the VP data, while the addition of geometrical and quantum chemical indices provided marginal improvement in model quality, at best. Our previous hierarchical QSAR/QSPR studies with various types of physicochemical, biomedicinal, and toxicological properties found the same pattern for different and diverse data sets.

A few distinct features are evident when examining the best QSPR models given in Tables 6 and 7: (a) The majority of the models contain path connectivity or count of path length terms which encode information about paths of lengths six or greater. Such terms might represent information regarding the generalized shape and size of molecules which are important for intermolecular van der Waals type interaction. Another connectivity index, ${}^3\chi^b_c$, is present in the majority of the models. This index, a quantifier of molecular branching, also shows the importance of molecular shape in determining vapor pressure. (b) Each of the models has at least one parameter that characterizes hydrogen bonding. The coefficients of all hydrogen bonding terms are negative, indicating that the values of VP are inversely correlated with the strength of hydrogen bonding which determines the tendency of molecules to remain in the liquid state as opposed to escaping to the gaseous phase. (c) Another group of parameters which have an important role in the VP QSPRs is the electropological indices. These indices collectively represent the various types of electronic interactions among molecules.

It is of interest to compare the results of the current study with those of the major published studies in which VP QSPR models were developed utilizing large data sets (Table 8). The enhanced set of theoretical descriptors used in the current study led to improved models as compared to the earlier study by Basak et al.[7] The data sets utilized by McClelland and Jurs,[10] Liang and Gallagher,[8] and Katritzky et al.[9] are all structurally quite diverse. The best QSPR models in the current study are superior to the models reported by these investigators in terms of model error. It should be noted, however, that the results reported by the Jurs group[10,11] are based on external prediction sets. Of the results reported in Table 8, only the model obtained by Goll and Jurs[11] has a lower error than the best model developed in the current study; however, it must be noted that the data set they used was not as diverse, consisting strictly of hydrocarbons and halohydrocarbons, and the total number of compounds in their data set was approximately 25% less than the data set used by our group. The models developed in this study are based on easily calculable molecular descriptors which can be used in the routine estimation of vapor pressure of chemicals.

ACKNOWLEDGMENT

This is contribution number 290 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by Grant F49620-98-1-0015 from the U.S. Air Force. The authors are thankful to Brian D. Gute and Gregory D. Grunwald for technical support.

REFERENCES AND NOTES

(1) (a) Mackay, D.; Shiu, W. Y.; Ma, K. C. In *Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals*; Lewis Publishers: Boca Raton, FL, 1992; Vol. 1−4. (b) Lyman, W. J. In *Environmental Exposure from Chemicals*; Neely, W. B., Blau, G. E., Eds.; Chemical Rubber: Boca Raton, FL, 1985; Vol. I.

(2) (a) Mackay, D. Finding Fugacity Feasible. *Environ. Sci. Technol*. **1979**, *13*, 1218−1223. (b) Mackay, D.; Paterson,S. Calculating Fugacity. *Environ. Sci. Technol*. **1981**, *15*, 1006−1014. (c) Mackay, D.; Paterson, S. Fugacity Revisited. *Environ. Sci. Technol*. **1982**, *16*, 654A−660A.

(3) Mackay, D.; Shiu, W. Y. A Critical Review of Henry's Law Constants for Chemicals of Environmental Interest. *J. Phys. Chem. Ref. Data* **1981**, *10*, 1175−1199.

(4) Reinhard, M.; Drefahl, A. *Handbook for Estimating Physicochemical Properties of Organic Compounds*; Wiley & Sons: New York, 1999.

(5) Mackay, D.; Bobra, A.; Chan, D. W.; Shiu, W. Y. Vapor Pressure Correlation for Low-Volatility Environmental Chemicals. *Environ. Sci. Technol.* **1982**, *16*, 645−649.
(6) Banerjee, S.; Howard, P. H.; Lande, S. S. General Structure-Vapor Pressure Relationships for Organics. *Chemosphere* **1990**, *21*, 1173−1180.
(7) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometrical Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.
(8) Liang, C.; Gallagher, D. A. QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321−324.
(9) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.
(10) McClelland, H. E.; Jurs, P. C. Quantitative Structure−Property Relationships for the Prediction of Vapor Pressures of Organic Compounds from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967−975.
(11) Goll, E. S.; Jurs, P. C. Prediction of Vapor Pressures of Hydrocarbons and Halohydrocarbons from Molecular Structure with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1081−1089.
(12) Auer, C. M.; Nabholz, J. V.; 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.
(13) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals. In *Quantitative Structure−activity Relationships in Environmental Sciences VII.;* Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, FL, 1998; Chapter 17, pp 245−261.
(14) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Hierarchical Approach to the Development of QSAR Models Using Topological, Geometrical and Quantum Chemical Parameters. 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.
(15) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.
(16) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060.
(17) Basak S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2000**, in press.
(18) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.
(19) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.
(20) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of Mutagenicity of Chemicals from Theoretical Structural Parameters: A Hierarchical Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 117−129.
(21) Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An Integration of the AQUIRE Data Base and the QSAR System for Use in Ecological Risk Assessments. *Sci. Total Environ.* **1991**, *109/110*, 667−670.
(22) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY 2.3; Copyright of the University of Minnesota, 1988.
(23) Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.
(24) Wiener, N. *Cybernetics*; Wiley: New York, 1948.
(25) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(26) (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976. (b) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Studies*; Research Studies Press: Letchworth, 1986. (c) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State;* Academic Press: New York, 1999.
(27) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(28) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; 4th International Conference Zurich; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1983; pp 745−750.
(29) (a) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: a QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609. (b) Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. A Quantitative Structure Activity Relationship Study of Tumor Inhibitory Triazenes Using Bonding Information Content and Lipophilicity. *ICRS Med. Sci.* **1982**, *10*, 933−934. (c) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. A Quantitative Structure−Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneimitt.-Forsch. Drug Res.* **1983**, *33*, 501−503.
(30) Basak, S. C.; Magnuson, V. R. Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discr. Appl. Math.* **1988**, *19*, 17−44.
(31) Balasubramanian, K.; Basak, S. C. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367−373.
(32) (a) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogues. *Math. Modelling Sci. Comput.* In press. (b) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modelling Sci. Comput.* In press. (c) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogues of Chemicals. *Math. Modelling Sci. Comput*. In press.
(33) Raychaudhury, C.; Ray, S. K.; Roy, A. B.; Ghosh, J. J.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Indices. *J. Comput. Chem.* **1984**, *5*, 581−588.
(34) (a) Balaban, A. T. Highly Discriminating Distance-Based Topological Index, *Chem. Phys. Lett.* **1982**, *80*, 399−404. (b) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206. (c) Balaban, A. T. Chemical Graphs. 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH),* **1986**, *21*, 115−122. (d) Balaban, A. T.; Filip, P. Computer Program for Topological Index J (Average Distance Sum Connectivity). *Math. Chem. (MATCH)* **1984**, *16*, 163−190.
(35) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61−83.
(36) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.
(37) (a) MOLCONN-Z; Hall Associates Consulting: Quincy, MA. (b) Hall, L. H.; Kier, M. B. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.
(38) Tripos Associates, Inc.: St. Louis, MO, 1994.
(39) Stewart J. J. P. MOPAC Version 6.00; QCPE #455; Frank J. Seiler Research Laboratory: U.S. Air Force Academy, CO, 1990.
(40) SAS/STAT. *User's Guide*, Release 6.03 ed.; SAS Institute: Cary, NC, 1988.

CI000165R