Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies

Viviana Consonni,[†] Roberto Todeschini,*[,†] Manuela Pavan,[†] and Paola Gramatica[‡]

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, Milano-Bicocca University, P.za della Scienza 1, 20126 Milano, Italy, and QSAR Research Unit, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy

Received December 21, 2001

In a previous paper the theory of the new molecular descriptors called *GETAWAY* (GEometry, Topology, and Atom-Weights AssemblY) was explained. These descriptors have been proposed with the aim of matching 3D-molecular geometry, atom relatedness, and chemical information. In this paper prediction ability in structure−property correlations of GETAWAY descriptors has been tested extensively by analyzing the regressions of these descriptors for selected properties of some reference compound classes. Moreover, the general performance of the new descriptors in QSAR/QSPR has been evaluated with respect to other well-known sets of molecular descriptors.

INTRODUCTION

The descriptors[1] GETAWAY are recently proposed molecular descriptors derived from a new representation of molecular structure, the *Molecular Influence Matrix* (MIM), denoted by **H** and defined as the following

$$\mathbf{H} = \mathbf{M}\cdot(\mathbf{M}^{\mathbf{T}}\cdot\mathbf{M})^{-1}\cdot\mathbf{M}^{\mathbf{T}}$$

where **M** is the molecular matrix constituted by the centered Cartesian coordinates x, y, z of the molecule atoms (hydrogens included) in a chosen conformation, and the superscript T refers to the transposed matrix.

The diagonal elements h_{ii} of the molecular influence matrix, called *leverages*, encode atomic information and represent the "influence" of each molecule atom in determining the whole shape of the molecule; in fact mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule itself. Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interactions with the *i*th atom or, in other words, the attitude of the two considered atoms to interact themselves. A negative sign for the off-diagonal elements means that the two atoms occupy opposite molecular regions with respect to the center, hence the degree of their mutual accessibility should be low.

Two sets of theoretically closely related molecular descriptors have been devised: H-GETAWAY descriptors have been calculated from the *molecular influence matrix* **H**, while R-GETAWAY descriptors are from the *influence/distance matrix* **R** where the elements of the molecular influence matrix are combined with those of the geometry matrix. With the aim of catching relevant chemical information, these new descriptors have been defined by applying some traditional matrix operators, concepts of the information theory and spatial autocorrelation formulas, weighting the molecule atoms in such a way as to account for atomic mass, polarizability, van der Waals volume, and electronegativity.

This paper reports investigations into the general usefulness of GETAWAY descriptors in QSAR/QSPR problems. The prediction ability of these descriptors has been tested on a wide range of data sets from various literature sources. For each data set there is a comparison of the GETAWAY descriptor based-models and the models obtained from molecular descriptors of other approaches[2] (constitutional, topological,[3,4] WHIM,[5,6] BCUT,[7,8] Moreau-Broto autocorrelation[9−11] descriptors). All of the models were calculated by Multiple Linear Regression (MLR), and Genetic Algorithms[12,13] were used to search for the best predictive subset of variables within each set of descriptors. Finally, a statistical analysis of all the calculated models was performed; this was to evaluate the overall performance of GETAWAY descriptors for selected data sets, allowing comparison with other considered molecular descriptors.

MATERIALS AND METHODS

Data Sets. The present study was performed using seven different data sets taken from the literature and chosen mainly for their environmental importance or relevance to human health. The main features of these data sets, together with the bibliographic references,[14−18] are summarized in Table 1. The classes of compounds the different data sets refer to are six: polycyclic aromatic hydrocarbons (PAHs), *N*,*N*-dimethyl-2-halo-phenethylamines, nitrobenzenes, polychlorinated biphenyls (PCBs), polychlorinated and polybrominated dibenzo-*p*-dioxins (PDDs), and polychlorinated dibenzofurans (PCDFs). The last data set is given by the union of PCBs, PDDs, and PCDFs. For some classes of compounds more than one property was studied, resulting in a total of

* Corresponding author e-mail: roberto.todeschini@unimib.it.
[†] Milano-Bicocca University.
[‡] University of Insubria.

10.1021/ci0155053 CCC: $22.00
Published on Web 04/20/2002

Table 1. Description of the Data Sets Used for the Comparative Study[a]

data set	obj.	var.	responses (obj.)	ref
PAH	82	484	bp (53), mp (80), logK_{OW} (37)	14
phenetylamines	22	479	log(1/ED_{50})	15
nitrobenzenes	47	494	log(1/IC_{50})	16
PCB	209	457	mp (81), logK_{OW} (139), H (20), logY_W (88), pRB (14)	17, 18
PDD	25	484	pRB	18
PCDF	34	484	pRB	18
PCB+PDD+PCDF	73	484	pRB	18

[a] For each data set the total number of compounds (*obj.*) and the total number of calculated molecular descriptors (*var.*) are reported, together with the studied responses and the corresponding numbers of available data.

Table 2. Dimensionality of the Sets of Molecular Descriptors Used for the Comparative Study[a]

descriptor set	var.
constitutional descriptors	56
topological descriptors	69
Moreau-Broto autocorrelations	32
BCUT descriptors	64
WHIM descriptors	99
GETAWAY descriptors	197
GETAWAY + WHIM descriptors	296
all the descriptors together	517

[a] The number of variables for each set is the maximum possible number; the variables that are constant for a class of compounds have been excluded from the data set.

13 studied properties: boiling point (bp), melting point (mp), and octanol−water partition coefficient (logK_{OW}) for PAHs; adrenergic blocking activity (log 1/ED_{50}) for phenetylamines; acute toxicities toward *Tetrahymena pyriformis* (log 1/IC_{50}) for nitrobenzenes; melting point (mp), octanol−water partition coefficient (logK_{OW}), Henry's law constant (H), aqueous water coefficient (log Y_W) for PCBs, and *Ah* receptor binding affinity (pRB) for PCBs, PDDs, and PCDFs.

Molecular Descriptors. The molecular descriptors used to search for the best regressions of the physicochemical and biological properties of the selected classes of compounds were calculated by the *Dragon* program[19] on the basis of the minimum energy molecular geometries optimized by *HyperChem* package[20] (PM3 semiempirical method). *Dragon* is a new, freely available software (by Milano Chemometrics and QSAR Research Group) for the calculation of more than 800 molecular descriptors. In this study only the following sets of molecular descriptors were calculated: constitutional descriptors, topological descriptors,[3,4] Moreau-Broto 2D-autocorrelations,[9−11] BCUT descriptors,[7,8] WHIM descriptors,[5,6] and GETAWAY descriptors. The prediction ability in QSAR/QSPR was evaluated for each of these descriptor sets as well as for the set provided by the GETAWAY plus WHIM descriptors and for the whole set of these molecular descriptors. Table 2 shows the dimensionality of all these sets. The set GETAWAY + WHIM was considered in the comparative study as GETAWAY descriptors mainly encode local information related to molecular fragments and substituent groups, thus it has been observed that their modeling power increases when they are used with descriptors of the whole molecular structure, such as WHIM descriptors.

Computational Method. The correlations of all of the considered properties are estimated by Multiple Linear Regression (MLR) based on the most predictive molecular descriptors. However as an exhaustive search for the best regressions within a wide set of descriptors requires extensive computational resources and is time-consuming, given the extremely high number of possible descriptor combinations, we used the Genetic Algorithm (GA-VSS) approach[12,13] as the variable selection method. Starting from a population of 100 random models with a number of variables equal to or less than a user-defined maximum value, the algorithm explores new combinations of variables, selecting them by a mechanism of reproduction/mutation similar to that of biological population evolution. The models based on the selected subsets of variables are tested and evaluated by the cross-validated explained variance (Q^2); only the models of the best quality are retained in the population undergoing the evolution procedure. After a few iterations, the evolving population is usually composed of different combinations of variables that correlate well with the response.

All of the calculations were performed by our new in-house software *MobyDigs*/*Evolution* for variable selection for WINDOWS/PC.[21] The best correlations were chosen by using the *leave-one-out* procedure of cross-validation. In all the cases, the most predictive model with a specified number of variables was chosen within the model population selected by the Genetic Algorithm. The X-block correlation, checked using the *K* multivariate correlation index,[22] was always compared with the correlation in the X+Y block in order to avoid chance correlation. Only models with a *K* multivariate correlation calculated on the X+Y block of the 5% greater than the *K* correlation of the X-block were considered statistically significant (QUIK rule[23]). All the variables for the obtained models are highly significant, within a 95% confidence level.

STRUCTURE−PROPERTY CORRELATIONS

For each of the 13 considered properties several regression models were calculated by the different sets of molecular descriptors; the best ones are collected in Tables 4−16 where the descriptors of the models are reported along with statistical information, i.e., the *leave-one-out* cross-validated explained variance (Q^2_{LOO}), the determination coefficient (R^2), and the multivariate correlation index for the descriptor block (K_X). All these coefficients are expressed in percentages. To facilitate the comparison of models based on GETAWAY descriptors and other models, the "GETAWAY models" have been highlighted in boldface. Table 3 gives the definitions of the constitutional and topological descriptors selected in the models. For the definitions of WHIM, BCUT, and Moreau-Broto descriptors we address the reader to the bibliographic references[5−11] and *Dragon* help.[19] In each case, for maximum uniformity in comparability, only models with one to four descriptors were calculated, regardless of the number of training compounds and the amount of response variance explained by the best model. Therefore, the best selected models with 1, 2, 3, and 4 descriptors are reported for each response and for each set of molecular descriptors. The models are always arranged in decreasing order of the Q^2 index. It will be seen that some models are missing. These missing models, based mainly on constitutional descriptors, were not significant according to the QUIK rule. Moreover, when two or more models obtained by

Table 3. Constitutional and Topological Descriptors of the QSAR Regressions Reported in This Study

symbol	definition	ref	symbol	definition	ref
AMW	average molecular weight		$^{0}\chi^{v}$	valence connectivity index of 0-order	28
MW	molecular weight		$^{2}\chi^{v}$	valence connectivity index of 2-order	28
Sv	sum of the atomic van der Waals volumes (scaled on carbon atom)		$^{0}\bar{\chi}^{v}$	mean valence connectivity index of 0-order	28
			$^{2}\bar{\chi}^{v}$	mean valence connectivity index of 2-order	28
Se	sum of the atomic Sanderson electronegativities (scaled on carbon atom)		J	Balaban distance connectivity index	29
			$\bar{\sigma}$	average vertex distance degree	2
Sp	sum of the atomic polarizabilities (scaled on carbon atom)		$\bar{W}$	mean Wiener index	30
Ss	sum of the Kier-Hall electrotopological states		B	Balaban centric index	31
Mv	atomic average van der Waals volume (scaled on carbon atom)		$^{V}\bar{I}_{C,R}$	radial centric information index	32
			$^{V}\bar{I}_{D}^{E}$	mean information content on the distance equality	3
Mp	atomic average polarizability (scaled on carbon atom)		$^{V}\bar{I}_{D}^{M}$	mean information content on the distance magnitude	3
Ms	average electrotopological state		$^{V}\bar{I}_{D,\mathrm{deg}}^{E}$	mean information content on the distance degree equality	3
nAT	number of atoms		$^{V}\bar{I}_{D,\mathrm{deg}}^{M}$	mean information content on the distance degree magnitude	3
nSK	number of non-H atoms		$^{V}I_{D}^{M}$	total information content on the distance magnitude	3
nBT	number of bonds		$^{V}\bar{I}_{adj,\mathrm{deg}}^{E}$	mean information content on the vertex degree equality	3
nBO	number of non-H bonds		TIC	neighborhood total information content 1-order	33
nBM	number of multiple bonds		SIC	structural information content 1-order	33
nDB	number of double bonds		CIC	complementary information content 1-order	33
nH	number of hydrogen atoms		BIC	bonding information content 1-order	33
nC	number of carbon atoms		$^{1}\kappa$	1-path Kier alpha-modified shape index	34
nN	number of nitrogen atoms		$^{2}\kappa$	2-path Kier alpha-modified shape index	34
nO	number of oxygen atoms		$^{3}\kappa$	3-path Kier alpha-modified shape index	34
nCl	number of chlorine atoms		$(P/W)^{2}$	path/walk shape index of 2-order	35
nBr	number of bromine atoms		$(P/W)^{4}$	path/walk shape index of 4-order	35
nNH_2	number of NH2 groups		$(P/W)^{5}$	path/walk shape index of 5-order	35
nCO	number of C=0 groups		PCR	ratio of multiple path counts to path counts	2
nNO_2	number of NO2 groups		PCD	difference between multiple path counts and path counts	2
nCIR	number of circuits		Φ	Kier molecular flexibility index	36
nR05	number of five-membered rings		B_L	benzene-likeliness index	37
nR06	number of six-membered rings		AECC	average eccentricity	38
nR09	number of nine-membered rings		DECC	eccentricity deviation	38
nR10	number of ten-membered rings		UNIP	unipolarity	38
I_{CPX}	Bertz molecular complexity index	24	CENT	centralization	38
$\bar{I}_{AC}$	mean information index on atomic composition	25	VAR	variation	39
M_1	first Zagreb index	26	MSD	mean square distance index	40
$^{V}M_1$	first Zagreb index based on valence vertex degrees	26	λ_1^{LP}	Lovasz-Pelikan index	41
$^{V}M_2$	second Zagreb index based on valence vertex degrees	26	^{V}MTI	Schultz MTI by valence vertex degrees	42
$^{2}\chi$	connectivity index of 2-order	27	$^{V}S_G$	Gutman molecular topological index by valence vertexs degree	43
$^{0}\bar{\chi}$	mean connectivity index of 0-order	27	S_G	Gutman molecular topological index	43
$^{1}\bar{\chi}$	mean connectivity index of 1-order	27	LPRS	log of the product of the distance matrix row sums (PRS)	44
$^{2}\bar{\chi}$	mean connectivity index of 2-order	27			

different sets of descriptors were the same, i.e., with the same variables, only the one derived from the largest set of descriptors is reported in the table. For example, for the octanol−water partition coefficient (log K_{OW}) of the PAHs, the best two-dimensional model, selected by using both the set GETAWAY descriptors and the set GETAWAY + WHIM descriptors, is found to be based on the GETAWAY descriptors *RCON* and *RARS*; therefore, only the model derived by the set GETAWAY + WHIM descriptors is reported in the summary.

The discussion of the results is organized into the five following sections, each related to a class of compounds.

Physicochemical Properties of Polycyclic Aromatic Hydrocarbons (PAHs). The scientific community has largely studied polycyclic aromatic hydrocarbons due to their impact on the environment and the public health. In particular, an extensive study has been performed to measure and model some important physicochemical properties that control their fate in the environment. In the present study three different properties have been modeled: boiling point (bp), melting point (mp), and octanol−water partition coefficient (log K_{OW}). Their values for 82 PAHs are reported in ref 14, collected there from different literature sources.

The best regressions with 1, 2, 3, and 4 molecular descriptors, ordered with respect to the decreasing value of the predictive ability (Q^2), are collected in the Tables 4−6, each referring to one of the studied properties. All the regressions of the boiling point are very good starting from those with one descriptor, explaining in validation up to about 98% of the response variance. Moreover, this property seems to be well modeled by simple molecular descriptors such as constitutional and topological descriptors. The regressions of the melting point show greater variability in the Q^2 values, that range from about 62% to 90%. The best models are from topological, WHIM, and GETAWAY descriptors. It is interesting to note that the best three-dimensional model includes the *standardized information content on the leverage magnitude* I_{SH} which is a GETAWAY descriptor encoding, to some extent, information on molecule entropy and symmetry and hence already supposed to be able to model physicochemical properties related to these structural features.[1] The octanol−water partition coefficient is well modeled by all the molecular descriptor sets; however, the best regressions are derived from GETAWAY and topological descriptors.

Table 4. Data Set PAH: Molecular Descriptors and Statistical Information for the Best Regressions of the **Boiling Point (bp)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	nSK G2u G2s Vm	98.9	99.1	65.1
constitutional	4	nBO nC nR05 nR06	98.6	98.8	65.0
all	3	Ss VAR $\boldsymbol{R^+_3}$**(p)**	98.3	98.7	65.0
topological	4	$^0\bar{\chi}$ $^V\bar{I}_{C,R}$ VAR $^VM_1$	98.2	98.5	59.9
topological	3	$^0\bar{\chi}$ CENT $^VM_1$	98.1	98.3	73.8
all	2	$^VM_1$ BELm3	98.1	98.3	84.6
constitutional	3	Mv nBT nR09	98.0	98.3	41.4
topological	2	$^0\bar{\chi}$ $^VM_1$	97.9	98.1	68.1
constitutional	2	MW Mv	97.8	98.1	73.1
Getaway + Whim	4	E3p $\boldsymbol{H_3}$**(v)** $\boldsymbol{R_5}$**(m)** $\boldsymbol{R_5}$**(v)**	97.8	98.3	65.3
all (topological)	1	$^VM_1$	97.8	98.0	0
BCUT	4	BEHv2 BEHe1 BELe1 BELe3	97.7	98.2	71.6
GETAWAY	4	$\boldsymbol{H_{GM}}$ $\boldsymbol{H_0}$**(m)** $\boldsymbol{H_1}$**(v)** $\boldsymbol{H_3}$**(v)**	97.5	98.2	63.9
Getaway + Whim	3	L3e $\boldsymbol{H_3}$**(v)** $\boldsymbol{HATS_5}$**(e)**	97.5	98.0	53.0
constitutional	1	nBO	97.5	97.7	0
BCUT	3	BEHe1 BELe2 BELe3	97.0	97.4	72.0
Getaway + Whim	2	E3m $\boldsymbol{H_3}$**(v)**	96.9	97.4	50.3
GETAWAY	3	$\boldsymbol{H_{GM}}$ $\boldsymbol{H_3}$**(m)** $\boldsymbol{HATS_2}$**(m)**	96.5	97.8	45.8
WHIM	4	L2v L1s Tu Te	96.2	97.0	65.6
Moreau-Broto	4	ATS7m ATS3v ATS4e ATS4p	95.8	96.9	72.8
Moreau-Broto	3	ATS4m ATS7m ATS3v	95.8	96.7	66.5
GETAWAY	2	$\boldsymbol{H_1}$**(m)** $\boldsymbol{H_3}$**(m)**	95.6	96.4	79.2
BCUT	2	BEHm2 BEHm3	95.0	95.6	74.3
Moreau-Broto	2	ATS3m ATS7m	95.0	95.9	52.7
Getaway + Whim	1	$\boldsymbol{H_3}$**(m)**	94.8	95.2	0
WHIM	3	E2e Dp Vs	92.6	93.8	32.6
WHIM	2	E3v Au	91.6	92.6	38.0
WHIM	1	As	86.8	87.7	0
Moreau-Broto	1	ATS5v	86.1	87.2	0
BCUT	1	BELm8	84.5	85.6	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 5. Data Set PAH: Molecular Descriptors and Statistical Information for the Best Regressions of the **Melting Point (mp)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	nSK VAR Gs Ku	89.2	90.5	43.5
Getaway + Whim	4	L2u Gm Kv $\boldsymbol{R^+_2}$**(v)**	88.5	89.8	30.5
WHIM	4	L2u L2e Gm Ku	87.6	89.1	45.4
all (Getaway + Whim)	3	Ku $\boldsymbol{I_{SH}}$ $\boldsymbol{H_3}$**(m)**	85.6	87.0	19.5
WHIM	3	L2u Gm Km	85.5	86.9	26.5
GETAWAY	4	$\boldsymbol{HATS_4}$**(m)** $\boldsymbol{R_3}$**(m)** $\boldsymbol{R^+_1}$**(m)** $\boldsymbol{R^+_6}$**(p)**	84.4	86.0	29.4
topological	4	BAC $^3\kappa$ J $^V\bar{I}^E_{D,\mathrm{deg}}$	84.0	85.9	36.5
topological	3	$^V\bar{I}^E_{D,\mathrm{deg}}$ $(P/W)^2$ J	82.9	84.7	37.3
GETAWAY	3	$\boldsymbol{HATS}$**(m)** $\boldsymbol{R_2}$**(m)** $\boldsymbol{R^+_1}$**(m)**	82.0	83.6	43.4
all	2	$^V\bar{I}^M_{D,\mathrm{deg}}$ $\boldsymbol{R^+_6}$**(p)**	81.0	82.3	12.9
Getaway + Whim	2	$\boldsymbol{H_3}$**(v)** $\boldsymbol{R^+_6}$**(p)**	80.1	81.3	9.9
topological	2	$^V\bar{I}^E_{D,\mathrm{deg}}$ $^V\bar{I}^M_{D,\mathrm{deg}}$	79.8	81.1	31.8
WHIM	2	L2u Ku	76.8	78.8	37.2
constitutional	4	Mp nBT nBM nR09	75.0	77.2	58.0
Moreau-Broto	3	ATS7m ATS8e ATS3p	74.9	77.2	55.2
Moreau-Broto	4	ATS7m ATS2e ATS3e ATS8e	74.4	77.4	54.0
constitutional	3	Mp nH nR10	74.1	76.2	43.6
Moreau-Broto	2	ATS7m ATS3p	73.3	75.4	63.7
constitutional	2	Mp nAT	72.9	74.8	65.8
all (topological)	1	$^2\chi$	72.7	74.4	0
Getaway + Whim	1	$\boldsymbol{H_3}$**(m)**	72.4	73.5	0
constitutional	1	Ss	71.7	73.4	0
BCUT	2	BELm1 BELe2	68.8	71.9	48.4
BCUT	1	BEHm2	65.5	67.6	0
Moreau-Broto	1	ATS6m	64.6	66.7	0
WHIM	1	Tu	61.7	64.5	0

[a] GETAWAY descriptors are highlighted in boldface.

Biological Activities of Phenethylamines. *N*,*N*-dimethyl-2-halo-phenethylamines constitute a set of compounds widely used in QSAR studies. The response is the antagonism of these compounds to epinephrine in the rat (log $1/ED_{50}$). Table 7 shows the statistical information for the best regressions with 1, 2, 3, and 4 molecular descriptors, ordered with respect to decreasing values of predictive ability (Q^2). The WHIM descriptors give the best performance in modeling this

Table 6. Data Set PAH: Molecular Descriptors and Statistical Information for the Best Regressions of the **log K_{OW}** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	TIC CIC Φ ***H*$_6$(m)**	97.6	98.1	49.5
Getaway + Whim	4	***I*$_{TH}$ *H*$_6$(m) *HT*(e) *RARS***	97.2	98.0	56.7
all	3	${}^{V}\bar{I}_D^M$ ***I*$_{TH}$ *R*$_7$(e)**	96.1	97.0	55.1
topological	3	${}^{0}\bar{\chi}$ ${}^{V}\bar{I}_D^M$ TIC	95.7	96.5	63.9
Getaway +Whim	3	***I*$_{SH}$ *RCON RARS***	95.7	96.6	30.3
topological	4	${}^{0}\bar{\chi}^{v}$ ${}^{V}\bar{I}_D^M$ TIC λ_1^{LP}	95.7	96.6	61.3
all (topological)	2	${}^{0}\bar{\chi}^{v}$ ${}^{V}\bar{I}_{D,deg}^M$	95.0	95.9	60.9
Getaway + Whim	2	***RCON RARS***	94.6	95.6	6.6
constitutional	4	AMW nH nR05 nR09	94.4	95.6	33.2
BCUT	3	BELm6 BEHe1 BELp8	94.3	95.6	81.4
constitutional	3	AMW Mv nH	94.2	95.6	49.9
BCUT	4	BELm1 BELm6 BELv8 BELe2	94.1	95.6	72.3
constitutional	2	AMW nH	94.0	95.1	0.2
BCUT	2	BELe5 BELp8	93.9	94.8	85.4
all (topological)	1	${}^{0}\chi^{v}$	92.9	94.1	0
constitutional	1	nBT	92.9	94.1	0
Moreau-Broto	4	ATS4m ATS8m ATS6v ATS2e	92.3	94.6	63.0
WHIM	4	E2p Au Ds Vm	91.8	94.8	64.3
WHIM	3	L2u P2p Ds	90.4	93.2	46.2
Moreau-Broto	3	ATS4m ATS8v ATS2e	90.2	92.4	59.4
BCUT	1	BELm8	89.5	90.8	0
Getaway + Whim	1	***HT*(p)**	88.9	90.7	0
WHIM	2	G3u Au	88.5	90.6	0.5
WHIM	1	Au	87.6	89.8	0
Moreau-Broto	2	ATS3v ATS7v	87.2	89.3	67.5
Moreau-Broto	1	ATS5m	78.3	80.2	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 7. Data Set Phenetylamines: Molecular Descriptors and Statistical Information for the Best Regressions of the **Adrenergic Blocking Activity log(1/ED$_{50}$)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	MSD As ***H*$_2$(v) *R*$_4$(u)**	97.7	98.5	51.0
Getaway + Whim	4	E2v P2s Tv ***HATS*$_1$(m)**	97.4	98.4	32.1
WHIM	4	E1v G2p P1s Tv	97.4	98.2	31.7
all	3	$(P/W)^4$ Tv ***R*$_4$(e)**	96.4	97.6	23.7
Getaway + Whim	3	P1s Tv ***R*$_4$(e)**	95.9	97.2	18.6
WHIM	3	E1v P1s Tv	95.8	97.2	32.8
topological	4	I_{CPX} $\bar{I}_{AC}$ ${}^{3}\kappa$ AECC	94.1	96.4	47.8
BCUT	4	BELm2 BELm5 BEHv5 BEHv6	93.8	96.1	49.1
GETAWAY	4	***HATS*$_3$(u) *HATS*(u) *H*$_4$(m) *H*$_1$(v)**	93.1	95.8	58.7
topological	3	$\bar{I}_{AC}$ $\bar{I}_D^{VE}$ ${}^{3}\kappa$	92.5	94.4	32.3
all	2	Ms MSD	92.2	94.3	24.9
GETAWAY	3	***HATS*$_6$(v) *H*$_4$(e) *HATS*$_7$(p)**	91.2	94.1	11.2
BCUT	3	BEHm5 BEHv6 BEHp5	91.2	93.8	33.8
Getaway + Whim	2	E3u L1v	90.8	93.9	25.4
BCUT	2	BEHv6 BEHp5	89.2	91.5	45.5
Moreau-Broto	3	ATS2e ATS7e ATS6p	86.8	92.5	42.3
GETAWAY	2	***H*$_4$(v) *H*$_3$(p)**	81.0	85.6	24.4
topological	2	MSD ${}^{V}M_1$	80.1	85.1	41.2
all (WHIM)	1	Tv	79.4	83.2	0
Moreau-Broto	4	ATS6v ATS8v ATS5e ATS8e	76.3	91.1	55.4
constitutional	2	Sv Mv	66.7	75.0	8.0
BCUT	1	BEHv6	66.3	70.1	0
constitutional	1	Sp	64.9	71.5	0
constitutional	3	Sv Se nCl	64.5	75.3	38.2
Moreau-Broto	2	ATS3v ATS5e	61.2	70.4	29.9
GETAWAY	1	***H*$_4$(m)**	57.0	65.2	0
topological	1	${}^{V}\bar{I}_D^E$	48.2	56.6	0
Moreau-Broto	1	ATS8v	31.3	44.0	0

[a] GETAWAY descriptors are highlighted in boldface.

biological response. In fact, the best mixed model, given by GETAWAY, topological, and WHIM descriptors, is only slightly better than the best four-dimensional WHIM model. Moreover, most of the calculated models show a better performance than the previous Hansch[45,46] and CoMFA[46] models: the former with three substituent constants (electrophilic constant σ^+, hydrophobic constant π, and van der Waals radius from the *para* position of the substituent r_p) has $Q^2 = 88.4\%$ and the latter with two latent variables $Q^2 = 80.5\%$.

Toxicity of Nitrobenzenes. The congeneric data set constituted by 47 nitrobenzenes was taken from the litera-

Table 8. Data Set Nitrobenzenes: Molecular Descriptors and Statistical Information for the Best Regressions of **Acute Toxicity (log $1/IC_{50}$)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	nDB $^{V}M_2$ $\boldsymbol{R_6(e)}$ logD_{OW}	89.5	91.0	38.6
Getaway + Whim	4	G1m E2s $\boldsymbol{R^+_6(e)}$ logD_{OW}	88.1	90.1	12.4
BCUT	4	BEHm8 BELm4 BELv4 logD_{OW}	87.1	89.5	53.3
constitutional	4	Ss Mv nCO logD_{OW}	86.0	88.5	34.3
WHIM	4	E2s Tu Te logD_{OW}	85.8	88.5	41.0
all	3	LPRS BEHe6 logD_{OW}	85.7	87.7	46.7
constitutional	3	nN nNH_2 logD_{OW}	85.1	87.7	29.8
Getaway + Whim	3	E2s $\boldsymbol{R_7(m)}$ logD_{OW}	84.6	87.3	13.9
GETAWAY	4	$\boldsymbol{RT(m)}$ $\boldsymbol{R^+_6(e)}$ $\boldsymbol{R^+_3(p)}$ logD_{OW}	84.5	87.4	34.4
BCUT	3	BELm4 BELv4 logD_{OW}	84.1	86.8	43.8
WHIM	3	P2s E2s logD_{OW}	84.0	86.3	39.6
Moreau-Broto	4	ATS6m ATS7v ATS5e logD_{OW}	83.6	86.7	31.0
topological	4	$^{V}\bar{I}_{C,R}$ TIC $\bar{W}$ logD_{OW}	83.6	86.7	49.5
topological	3	$^{V}M_1$ $^{V}M_2$ logD_{OW}	82.8	85.3	51.9
all (constitutional)	2	nNO_2 logD_{OW}	82.7	85.1	1.2
Moreau-Broto	3	ATS7v ATS5e logD_{OW}	82.3	85.0	40.7
GETAWAY	3	$\boldsymbol{R_1(m)}$ $\boldsymbol{R^+_6(m)}$ logD_{OW}	81.1	83.1	19.6
Getaway + Whim	2	G1e logD_{OW}	81.0	82.5	2.5
topological	2	CIC logD_{OW}	80.7	82.9	28.7
Moreau-Broto	2	ATS7m logD_{OW}	78.9	81.4	10.3
GETAWAY	2	$\boldsymbol{R_7(m)}$ logD_{OW}	78.6	80.5	0.9
BCUT	2	BEHm8 logD_{OW}	78.3	80.0	18.2
all	1	logD_{OW}	77.0	78.5	0
constitutional	1	nCO	27.2	34.6	0
topological	1	J	24.9	31.3	0
Getaway + Whim	1	$\boldsymbol{HT(p)}$	24.2	28.8	0
BCUT	1	BELe6	21.6	25.6	0
WHIM	1	Dp	21.5	27.4	0
Moreau-Broto	1	ATS5p	17.3	23.0	0

[a] GETAWAY descriptors are highlighted in boldface.

ture,[16,47,48] being a reference QSAR data set already investigated for a comparative study of different molecular descriptors. The data set consists of nitrobenzene, 12 ortho-, 14 meta-, and 20-para-substituted compounds. The response is the negative logarithm of 50% growth inhibition concentration, in millimoles per liter, of *Tetrahymena pyriformis* in a static assay with 48-h exposure (log $1/IC_{50}$). Along with the sets of molecular descriptors calculated by *Dragon*, also the octanol−water partition coefficient corrected for ionization at pH 7.35 (log D_{OW}) was used, providing an additional descriptor accounting for nitrobenzene hydrophobicity.

The statistical information for the best regressions of acute toxicity with 1, 2, 3, and 4 molecular descriptors, ordered with respect to the decreasing value of the predictive ability (Q^2), is shown in Table 8. To be noted is the crucial presence of the hydrophobicity descriptor (log D_{OW}) in all of these models, in agreement with the results of previous QSAR studies. This molecular descriptor alone gives a model with a prediction ability (Q^2_{LOO}) of 77%. Moreover, the GETAWAY descriptors alone underperform, while constitutional, BCUT and WHIM descriptors give fairly good results. The best model, obtained with nDB (number of double bonds), $^{V}M_2$ (second Zagreb index based on valence vertex degrees), R_6(e) (R-GETAWAY index weighted by atomic electronegativities), and log D_{OW}, has a Q^2_{LOO} index value (89.5%) better than the previous models, i.e., the model derived from the Wiener indices[16] (Q^2_{LOO} = 84.3%), the CoMFA model[48] (Q^2_{LOO} = 76.1%), and the model with quantum indices[47] (Q^2_{LOO} = 82.6%).

Physicochemical Properties of Polychlorinated Biphenyls (PCBs). It is well-known that polychlorinated biphenyls are widespread and persistent organic contaminants. Moreover, it has been demonstrated that they are toxic and lipophilic and tend to be bioaccumulated. Four of the physicochemical properties of environmental relevance for PCB congeners have been chosen: melting point (mp), octanol−water partition coefficient (log K_{OW}), Henry's law constant (H), and aqueous water coefficient, expressed as the negative logarithm (log Y_W). The experimental data have been taken from ref 17, where they had been collected from other sources.

The statistical information for the best regressions with 1, 2, 3, and 4 molecular descriptors, ordered with respect to decreasing values of the predictive ability (Q^2), has been collected in the Tables 9−12, each referring to one of the studied properties. In general, the GETAWAY descriptors perform well with all the considered properties, giving the models with the highest prediction ability. It is also interesting to note that the indices I_{TH} and I_{SH} (i.e. *total* and *mean information content on the leverage magnitude*, respectively), encoding information on molecule entropy, are among the molecular descriptors selected by GA as the most correlated with the melting point of PCBs as well as with the melting point of PAHs.

***Ah* Receptor Binding of PCBs, PDDs, and PCDFs.** Polychlorinated biphenyls (PCBs), dibenzo-*p*-dioxins (PDDs), and dibenzofurans (PCDFs) have been identified in almost every environmental compartment. Furthermore, due to their

Table 9. Data Set PCB: Molecular Descriptors and Statistical Information for the Best Regressions of the **Melting Point (mp)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	TIC ATS7e G1s Tu	82.0	84.6	50.8
getaway + whim	4	Gm $\boldsymbol{H_3}$**(u)** $\boldsymbol{R_1}$**(e)** $\boldsymbol{R^+_2}$**(e)**	81.3	83.9	26.8
GETAWAY	4	$\boldsymbol{I_{TH}}$ $\boldsymbol{H_3}$**(u)** ***HATS*****$_1$(u)** $\boldsymbol{R^+_2}$**(e)**	80.8	83.7	37.7
all	3	$^{v}\bar{I}^{E}_{D,deg}$ Tu $\boldsymbol{R^+_2}$**(m)**	80.8	83.2	32.6
WHIM	4	G1m L2v Tu Am	80.4	83.7	49.7
Getaway + Whim	3	Tu Gm $\boldsymbol{R^+_2}$**(v)**	80.0	82.4	29.6
GETAWAY	3	$\boldsymbol{I_{SH}}$ ***HT*****(e)** $\boldsymbol{R_1}$**(p)**	79.0	81.6	51.2
WHIM	3	Tu Am Gm	78.9	81.7	49.4
topological	4	$^{v}\bar{I}^{E}_{D,deg}$ CIC $(P/W)^2$ $(P/W)^5$	77.9	81.2	32.2
all	2	$^{v}\bar{I}^{E}_{D,deg}$ Tu	77.5	79.9	23.0
Getaway + Whim	2	Tu Gs	77.4	79.8	18.1
topological	3	$^{2}\chi$ $^{v}\bar{I}^{E}_{D,deg}$ $\bar{\sigma}$	76.3	79.3	54.2
GETAWAY	2	$\boldsymbol{I_{SH}}$ $\boldsymbol{R_3}$**(e)**	74.4	77.3	10.6
topological	2	TIC UNIP	73.0	75.7	39.1
BCUT	4	BELm5 BELv4 BELp3 BELp8	69.3	73.7	43.1
all (WHIM)	1	Tu	69.0	71.2	0
BCUT	3	BELm5 BELv4 BELp3	68.5	71.9	58.9
BCUT	2	BELm5 BELp3	67.0	69.8	52.7
Moreau-Broto	2	ATS5m ATS7v	65.8	69.1	64.5
Moreau-Broto	1	ATS7e	65.7	68.1	0
topological	1	UNIP	65.0	67.4	0
BCUT	1	BELv2	64.0	66.5	0
GETAWAY	1	$\boldsymbol{R_7}$**(v)**	63.5	66.1	0
constitutional	1	Mv	58.7	61.5	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 10. Data Set PCB: Molecular Descriptors and Statistical Information for the Best Regressions of the **log** $\boldsymbol{K_{OW}}$ with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	λ_1^{LP} $(P/W)^4$ L1m Ts	96.2	96.4	44.5
Getaway + Whim	4	Ts ***HATS*****$_6$(m)** $\boldsymbol{R_5}$**(u)** $\boldsymbol{R_4}$**(m)**	96.0	96.2	66.5
all	3	ATS4m L1m Ts	95.9	96.1	51.6
GETAWAY	4	$\boldsymbol{H_5}$**(m)** $\boldsymbol{H_2}$**(e)** $\boldsymbol{R_6}$**(e)** $\boldsymbol{R^+_4}$**(p)**	95.9	96.2	44.9
Getaway + Whim	3	Ts As $\boldsymbol{R_4}$**(m)**	95.8	96.0	70.5
topological	4	$^{2}\chi^{v}$ BIC λ_1^{LP} PCR	95.7	96.0	36.2
GETAWAY	3	$\boldsymbol{H_2}$**(p)** $\boldsymbol{R^+_4}$**(m)** $\boldsymbol{R_6}$**(e)**	95.7	95.9	64.8
topological	3	$^{2}\chi^{v}$ SIC PCR	95.6	95.9	46.3
BCUT	4	BELm8 BEHp1 BELp2 BELp8	95.6	95.9	63.6
WHIM	4	E1u L2m Ts Av	95.4	95.7	57.2
WHIM	3	E1m Ts Au	95.4	95.7	49.4
all (topological)	2	$^{2}\chi^{v}$ PCR	95.4	95.6	49.2
BCUT	3	BEHp1 BELp2 BELp8	95.2	95.5	69.4
Getaway + Whim	2	L1u $\boldsymbol{H_2}$**(p)**	95.0	95.2	71.7
Moreau-Broto	4	ATS6m ATS6v ATS8v ATS8e	95.0	95.4	71.7
GETAWAY	2	***HATS*****(u)** $\boldsymbol{H_2}$**(e)**	95.0	95.2	23.4
WHIM	2	L1u As	95.0	95.2	67.3
Moreau-Broto	3	ATS6m ATS6v ATS8e	94.7	95.0	73.9
Moreau-Broto	2	ATS4m ATS7e	94.2	94.5	79.7
all (WHIM)	1	Tu	93.9	94.1	0
BCUT	2	BELe2 BELe4	93.7	94.0	71.2
Moreau-Broto	1	ATS7e	93.6	93.8	0
BCUT	1	BELv2	93.1	93.3	0
topological	1	DECC	92.6	92.8	0
GETAWAY	1	$\boldsymbol{R_2}$**(v)**	92.5	92.6	0
constitutional	1	AMW	84.5	84.8	0

[a] GETAWAY descriptors are highlighted in boldface.

strong lipophilic nature, these compounds tend to be bioaccumulated and have already been detected in human body fluids and tissues. Their biological effects are diverse, among these hepatoxicity, porphyria, chloracne, teratogenicity, and carcinogenicity. Thus, because of the evident risk to the environment and human health, they have been the focus of many studies that have indicated that most of the toxic effects of these compounds are mediated by a common (*Ah* or dioxin) receptor mechanism of action.[18,49−54] The response modeled in the present study is the *Ah* receptor binding

Table 11. Data Set PCB: Molecular Descriptors and Statistical Information for the Best Regressions of **Henry's Law Constant (H)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all (Getaway + Whim)	4	Du ***HATS*$_4$(m) *R*$^+_7$(v) *R*$_7$(p)**	95.1	97.0	53.1
GETAWAY	4	***H*$_4$(p) *R*$_4$(m) *R*$_7$(v) *R*$^+_7$(e)**	94.8	97.2	47.9
all (Getaway)	3	***HATS*$_7$(v) *R*$_4$(m) *R*$^+_7$(e)**	93.4	95.5	42.1
topological	4	S_G (P/W)^4 (P/W)^5 $\bar{W}$	88.0	93.1	48.2
all (Getaway)	2	***HATS*$_4$(m) *R*$_3$(e)**	86.4	91.7	42.8
WHIM	4	E1u L2v E1e P1s	85.9	92.2	48.0
topological	3	UNIP (P/W)^4 $\bar{W}$	85.3	91.6	49.3
Moreau-Broto	3	ATS4m ATS6m ATS8m	81.1	89.9	41.1
Moreau-Broto	4	ATS4m ATS3m ATS8m ATS8e	80.1	91.0	57.8
WHIM	3	P1m P2e E1e	78.3	87.7	42.6
topological	2	$^2\bar{\chi}$ (P/W)^4	76.4	84.4	36.9
BCUT	4	BELv6 BEHe4 BEHp7 BEHp8	71.7	85.3	42.1
all (Getaway)	1	***R*$_8$(u)**	66.3	74.3	0
BCUT	3	BEHe4 BEHp7 BELp6	63.8	78.4	38.1
Moreau-Broto	2	ATS4m ATS8m	63.2	76.6	22.0
topological	1	(P/W)^4	61.4	70.4	0
WHIM	2	E1e E3e	58.7	71.2	16.5
WHIM	1	E1e	52.9	64.1	0
BCUT	2	BELm6 BEHp7	24.3	50.2	27.4
Moreau-Broto	1	ATS4m	19.8	36.6	0
BCUT	1	BEHp7	5.4	23.1	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 12. Data Set PCB: Molecular Descriptors and Statistical Information for the Best Regressions of the **Aqueous Activity Coefficient (log Y_W)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all (Getaway + Whim)	4	Tv ***I*$_{TH}$ *R*$_4$(v) *R* $^+_1$(p)**	87.6	89.0	42.6
GETAWAY	4	***I*$_{SH}$ *HATS*$_4$(u) *H*$_2$(e) *R* $^+_1$(p)**	87.3	88.6	43.0
all	3	$^2\chi$ ***I*$_{SH}$ *RT*$^+$(e)**	86.6	87.8	22.4
Getaway + Whim	3	Tp ***HATS*$_5$(e) *RT*$^+$(e)**	86.3	87.7	40.9
GETAWAY	3	***HATS*$_5$(u) *H*$_2$(m) *RT*$^+$(e)**	85.3	86.8	42.3
WHIM	4	L1u P1p Gm Kv	84.8	86.5	44.6
all	2	$^2\chi$ ***R*$^+_5$(e)**	84.7	85.8	37.5
Getaway + Whim	2	***I*$_{SH}$ *R*$_2$(m)**	84.1	85.1	5.0
topological	4	$^2\chi$ $^V\bar{I}^E_{D,deg}$ λ_1^{LP} PCD	84.0	85.8	50.5
WHIM	3	L1u Gu Ke	83.9	85.4	18.6
topological	3	$^1\bar{\chi}$ CIC λ_1^{LP}	83.8	85.3	53.8
topological	2	$^1\bar{\chi}$ λ_1^{LP}	83.5	84.5	37.8
BCUT	4	BEHm7 BEHv3 BEHe4 BEHp5	82.1	84.3	59.0
all (topological)	1	$^2\chi$	81.8	82.6	0
WHIM	2	E2p Tu	81.5	82.6	43.4
BCUT	3	BEHm4 BELm2 BEHp5	81.4	82.8	69.2
BCUT	2	BELv2 BEHp5	81.2	82.4	68.8
Moreau-Broto	2	ATS4v ATS7e	81.0	82.5	77.7
Moreau-Broto	3	ATS8m ATS7v ATS5e	80.7	82.7	63.6
Moreau-Broto	1	ATS1e	80.2	81.0	0
Moreau-Broto	4	ATS4m ATS5v ATS8v ATS8e	80.1	82.6	64.2
constitutional	1	Sv	80.1	80.9	0
Getaway + Whim	1	***R*$_2$(m)**	80.0	80.9	0
BCUT	1	BELe2	79.9	80.8	0
WHIM	1	Ae	77.4	78.3	0

[a] GETAWAY descriptors are highlighted in boldface.

affinity; its experimental values have been taken from the literature.[18,51]

The three classes of compounds were modeled separately, but since many of the members of the three classes have been shown to produce a qualitatively similar toxicity we also performed an additional global model of the three compound classes together, as is usually done. The summaries of the calculated model statistics are given in the Tables 13−16.

From an analysis of the results, it is apparent that the GETAWAY descriptors outperform the other selected sets of molecular descriptors giving, in all cases, the highest predictive regressions of the *Ah* receptor binding affinity. Moreover, the obtained models are fairly good in comparison with previously published models. In fact, the best four-dimensional model for the 14 biphenyls with the Q^2_{LOO} value of 96.7% is much better than the CoMFA model calculated by Waller and McKinney[51] with Q^2_{LOO} of 53.4% and two latent variables as well as the EEVA/PLS model[18] with Q^2_{LOO} of 54.9% and three latent variables. Furthermore, our four-dimensional model for the dioxins has a Q^2_{LOO} of 94.2%, being better than the CoMFA model (Q^2_{LOO} = 71.5%, four latent variables) and the EEVA/PLS model (Q^2_{LOO} = 86.2%, four latent variables). The same considerations follow for

Table 13. Data Set PCB: Molecular Descriptors and Statistical Information for the Best Regressions of the ***Ah* Receptor Binding Affinity (pRB)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	${}^{V}\bar{I}^{E}_{adj,deg}$ ***R*₃(e) *R*⁺₄(p) *R*⁺₇(p)**	96.7	98.9	46.0
Getaway + Whim	4	G2e G2s ***I*_TH *R*⁺₅(e)**	93.1	97.3	45.1
all	3	$\bar{I}_{AC}$ ***R*₇(m) *R*⁺₄(p)**	91.1	96.0	5.9
GETAWAY	4	***HATS*₄(v) *R*⁺₁(m) *R*⁺₇(v) *R*⁺₇(e)**	88.5	94.9	44.3
WHIM	4	P2u G2p G2s De	82.0	91.3	38.0
Getaway + Whim	3	G2e ***R*₈(v) *R*⁺₅(p)**	80.7	88.6	26.5
GETAWAY	3	***HATS*₄(e) *R*⁺₇(e) *R*⁺₇(p)**	74.9	87.8	53.2
BCUT	4	BELm4 BELm6 BELe5 BEHp3	73.7	88.7	35.1
topological	4	${}^{V}I^{M}_{D}$ SIC BIC $(P/W)^4$	71.1	87.7	50.1
all (Getaway)	2	***R*⁺₇(m) *R*₄(v)**	70.6	81.3	31.3
WHIM	3	L3u G2m G2p	66.4	87.7	53.8
WHIM	2	L1v G2v	64.5	79.3	1.8
topological	3	${}^{V}\bar{I}^{M}_{D}$ $(P/W)^4$ B_L	59.9	78.9	56.4
all (WHIM)	1	E1s	59.0	72.0	0
Moreau-Broto	4	ATS5m ATS6m ATS8v ATS8e	58.6	77.1	54.1
topological	2	$(P/W)^4$ $(P/W)^5$	58.5	70.8	36.3
BCUT	3	BELm4 BEHv3 BELp6	58.4	75.2	41.2
Moreau-Broto	3	ATS5m ATS6m ATS8m	57.3	76.2	40.2
Moreau-Broto	2	ATS4m ATS8m	56.3	71.9	11.8
topological	1	$(P/W)^4$	50.5	65.0	0
BCUT	2	BEHm4 BELe3	48.3	64.3	33.2
GETAWAY	1	***HATS*₈(u)**	36.9	55.4	0
Moreau-Broto	1	ATS8m	21.0	37.2	0
BCUT	1	BEHm4	2.5	24.0	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 14. Data Set PDD: Molecular Descriptors and Statistical Information for the Best Regressions of the ***Ah* Receptor Binding Affinity (pRB)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all (Getaway)	4	***H*₆(u) *HATS*₆(p) *R*⁺₄(u) *R*⁺₁(e)**	94.2	95.5	55.4
BCUT	4	BELv1 BEHe3 BEHe7 BEHp3	92.9	95.2	55.5
all	3	^{V}MTI BEHe7 Tm	91.9	93.9	51.2
Getaway + Whim	3	***H*₆(u) *HATS*₆(v) *R*⁺₁(e)**	91.9	94.2	45.5
WHIM	4	E1m E1e Te Vm	91.6	93.8	41.9
BCUT	3	BELv1 BEHe7 BEHp3	90.4	92.6	45.8
WHIM	3	E1u E1e Tm	90.3	93.1	47.7
topological	4	M_1 ${}^2\kappa$ $(P/W)^4$ B_L	90.1	93.0	46.9
all (Getaway + Whim)	2	Tm ***H*₆(e)**	89.2	91.4	31.5
WHIM	2	L2e Tm	87.4	90.4	1.8
topological	3	${}^{2}\bar{\chi}^{v}$ AECC ^{V}MTI	87.0	90.6	47.8
BCUT	2	BEHv3 BELp4	86.7	89.5	37.7
GETAWAY	2	***HATS*₁(v) *H*₅(e)**	86.4	89.9	22.6
all (WHIM)	1	L1m	83.8	86.0	0
topological	2	${}^{2}\bar{\chi}^{v}$ $(P/W)^4$	83.5	87.6	19.5
Moreau-Broto	3	ATS8m ATS8v ATS6e	77.5	84.8	49.4
GETAWAY	1	***R*⁺₃(e)**	68.8	72.6	0
BCUT	1	BEHp3	67.6	72.1	0
Moreau-Broto	2	ATS6e ATS8p	57.0	68.5	35.2
topological	1	$(P/W)^4$	48.4	56.7	0
constitutional	1	nBr	32.6	42.6	0
Moreau-Broto	1	ATS8p	27.6	38.4	0
constitutional	2	Mp nCl	21.9	43.4	8.5

[a] GETAWAY descriptors are highlighted in boldface.

the other two cases: for the furans our four-dimensional model has $Q^2_{LOO} = 86.8\%$, the CoMFA model $Q^2_{LOO} = 74.2\%$ (five latent variables), and the EEVA/PLS model $Q^2_{LOO} = 79.5\%$ (11 latent variables); for all the compounds together (PCBs, PDDs, PCDFs) our four-dimensional model has $Q^2_{LOO} = 85.6\%$, the CoMFA model $Q^2_{LOO} = 72.4\%$ (six latent variables), and the EEVA/PLS model $Q^2_{LOO} = 81.8\%$ (eight latent variables).

COMPARATIVE STUDY

To realize a general comparison of all the QSAR models calculated for the chosen properties using different sets of molecular descriptors, a statistical analysis of all the models was performed using the Principal Component Analysis (PCA) technique.[55,56] This multivariate technique allows the projection and, therefore, an analysis of the studied multivariate objects in a subspace defined by only a few dimensions. Moreover, it enables an interpretation to be made of the relationships observed among the objects, on the basis of the original variables.

In this case, the studied objects are calculated QSAR models characterized by the kind of molecular descriptors and the dimensionality, i.e., from 1 up to 4 variables. Each object is then described by the predictive ability (Q^2) for

Table 15. Data Set PCDF: Molecular Descriptors and Statistical Information for the Best Regressions of the ***Ah* Receptor Binding Affinity (pRB)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	BEHm8 E2u E1p ***HT*(u)**	86.8	89.2	52.0
Getaway + Whim	4	Ap ***HATS*$_5$(u) *R*$_5$(e) R$^+_1$(p)**	84.8	87.5	37.7
GETAWAY	4	***HATS*$_5$(u) *H*$_7$(e) *H*$_5$(p) *HATS*$_6$(p)**	84.6	87.8	47.1
all	3	BEHm5 E2p ***HATS*$_5$(u)**	84.5	87.1	55.1
Getaway + Whim	3	G3v E1s ***R*$^+_7$(u)**	83.7	86.8	18.4
topological	4	$^2\bar{\chi}$ MSD (P/W)^4 ^VS_G	82.7	88.0	49.3
GETAWAY	3	***HATS*$_5$(u) *R*$_6$(v) *R*$_5$(e)**	82.2	85.7	47.0
WHIM	4	G3m G2v G3p E1s	81.7	85.6	40.1
all (Getaway + Whim)	2	E1s ***R*$^+_7$(u)**	81.3	84.1	14.1
topological	3	$^V\bar{I}^M_D$ DECC (P/W)^5	81.0	85.2	57.9
WHIM	3	G2e G3p E1s	81.0	84.0	21.8
WHIM	2	G3p E1s	80.0	83.1	31.8
GETAWAY	2	***HATS*$_5$(u) *R*$^+_5$(p)**	77.7	80.7	27.8
topological	2	$^V\bar{I}^E_D$ (P/W)^5	76.8	80.3	38.7
all (WHIM)	1	L1s	75.6	77.8	0
BCUT	2	BELm8 BEHv3	72.9	76.9	71.9
BCUT	1	BEHp3	71.8	74.2	0
topological	1	$^V\bar{I}^E_D$	70.2	73.1	0
GETAWAY	1	***R*$_8$(e)**	67.8	71.0	0
Moreau-Broto	1	ATS8e	67.8	72.9	62.7
Moreau-Broto	2	ATS6e ATS8e	67.7	73.0	76.9
Constitutional	1	Ms	53.6	58.1	0

[a] GETAWAY descriptors are highlighted in boldface.

Table 16. Data Set PCB, PDD, PCDF: Molecular Descriptors and Statistical Information for the Best Regressions of the ***Ah* Receptor Binding Affinity (pRB)** with 1, 2, 3, and 4 Variables[a]

approach	size	descriptors	Q^2_{LOO}	R^2	K_X
all	4	BELm4 E1v ***HATS*$_4$(u) *HATS*$_7$(u)**	85.6	87.4	31.9
Getaway + Whim	4	G3u E1v G3p ***H*$_{GM}$**	85.0	87.0	41.4
all (Getaway + Whim)	3	E1p ***HATS*$_4$(u) *HATS*$_7$(u)**	84.2	85.8	38.8
WHIM	4	G3u L3e G3p E1p	84.0	86.2	40.5
all (Getaway + Whim)	2	E1m ***HATS*$_3$(e)**	80.9	82.5	8.0
GETAWAY	4	***HATS*$_1$(p) *R*$^+_3$(u) *R*$^+_6$(e) *R*$_5$(p)**	79.5	81.7	55.8
WHIM	3	L3e E1p Du	79.2	81.3	21.9
BCUT	4	BELe3 BELe4 BEH1p BEH3p	79.0	81.6	34.7
topological	4	$^2\bar{\chi}^v$ 1κ ^VMTI J	77.5	80.2	47.4
BCUT	3	BELm1 BEHv3 BEHe7	76.9	79.3	43.0
WHIM	2	P2u E1p	76.3	78.2	39.8
GETAWAY	3	***HATS*$_1$(v) *H*$_5$(e) *R*$_6$(p)**	76.1	78.4	40.1
topological	3	$^1\bar{\chi}$ $^0\bar{\chi}^v$ (P/W)^4	74.6	77.0	43.9
BCUT	2	BEHe7 BEHp3	70.9	73.2	14.1
GETAWAY	2	***HATS*$_1$(p) *R*$_5$(p)**	69.2	71.7	35.1
all (WHIM)	1	E1p	66.2	67.8	0.0
topological	2	$\bar{I}_{AC}$ $^2\chi^v$	60.9	63.6	37.2
BCUT	1	BEHv7	56.4	58.5	0.0
GETAWAY	1	***HATS*$_1$(v)**	56.4	58.4	0.0
constitutional	4	Sv Ms nBr nCIR	53.4	59.8	39.2
topological	1	$\bar{I}_{AC}$	49.5	51.9	0.0
constitutional	2	AMW nO	45.2	50.1	27.8
constitutional	3	AMW nCIR nR10	44.5	50.6	27.0
Moreau-Broto	2	ATS8m ATS2e	39.8	45.8	41.0
Moreau-Broto	1	ATS1m	36.0	39.6	0.0
constitutional	1	AMW	35.2	38.8	0.0

[a] GETAWAY descriptors are highlighted in boldface.

the 13 selected properties. In other words, Principal Component Analysis, performed by the software SIMCA-S,[57] was applied to a data matrix where the rows represent the best models, with 1, 2, 3, 4 variables for each considered set of molecular descriptors, and the columns their prediction power for the selected properties. The symbols used to identify the different sets of molecular descriptors are as follows: C for constitutional descriptors, T for topological descriptors, ATS for Moreau-Broto autocorrelations, B for BCUT descriptors, W for WHIM descriptors, G for GETAWAY descriptors, GW for GETAWAY + WHIM descriptors, and ALL for all these descriptors. Preceding the symbol that represents the molecular descriptor set there is a number, from 1 to 4, that is used to indicate the number of variables in the model. The models with 2, 3, and 4 constitutional descriptors were not included in the analysis because of the lack of statistical significance for many properties.

The Principal Component Analysis gave two significant components with a cumulative explained variance of 83.7%. The first component alone explains 72.3% of the total information contained in the original data matrix. Figure 1 shows the projection of the studied QSAR models in the

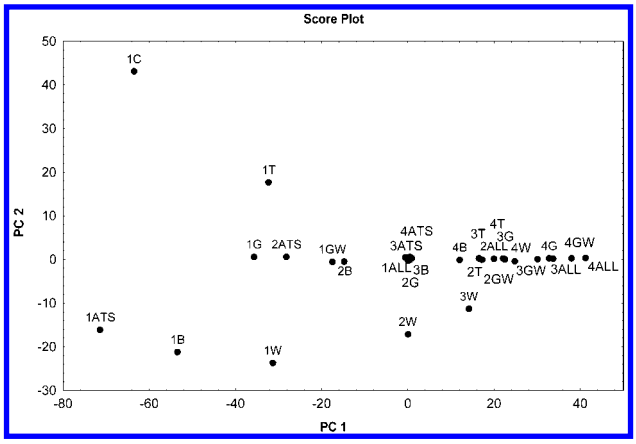

Figure 1. Score plot relative to the first and second principal components.

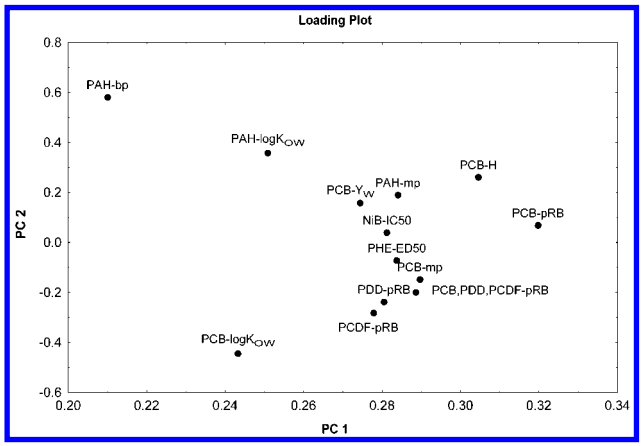

Figure 2. Loading plot relative to the first and second principal components.

space defined by the two principal components (score plot), while Figure 2 depicts the projection of the original variables in the same space (loading plot). The first principal component (PC1) can be interpreted as a quantitative variable, that provides the overall predictive ability of the different sets of molecular descriptors for all the selected properties, the loadings of all the original variables (molecular properties) being positive in this component. The properties to the left in the loading plot (Figure 2), such as the boiling point of PAHs (PAH-bp) and the octanol−water partition coefficient of PAHs and PCBs (PAH-log K_{OW} and PCB-log K_{OW}), are less important than the others in discriminating among the different sets of molecular descriptors, because the variability of the prediction power of the calculated QSAR models for these properties is very small.

In other words, the first principal component can be used to distinguish between sets of molecular descriptors that, in general, perform well with all the considered properties (to the right in the score plot of Figure 1) and those with low overall prediction power (to the left in the score plot). Therefore, it can be concluded that the best models are mixed models, based on descriptors from different approaches, and also models based on GETAWAY descriptors alone or GETAWAY plus WHIM descriptors that perform similarly.

As expected, the models with 1 and 2 variables are, in general, less predictive than those with more variables. The variability of the prediction power of these low-dimensionality models is explained by the second principal component

Table 17. Frequencies of the Different Kinds of Descriptors in the Mixed Models for the 13 Considered Properties

descriptor set	1-ALL	2-ALL	3-ALL	4-ALL	sum
constitutional	0	2	1	3	6
topological	4	9	8	10	31
Moreau-Broto	0	0	1	1	2
BCUT	0	1	3	2	6
WHIM	7	4	8	15	34
GETAWAY	1	9	17	20	47

(11.4% of explained variance). Thus, the one-dimensional model based on constitutional descriptors (1C) is generally not very predictive apart from for the boiling point and the octanol−water partition coefficient of the PAHs (PAH-bp and PAH-log K_{OW}). In the same way, the one-dimensional topological model (1T) performs well with the boiling point and the log K_{OW} of the PAHs as well as with the water solubility coefficient of PCBs (PCB−Y_W). The one-dimensional models based on the Moreau-Broto autocorrelations (1ATS) and the BCUT descriptors (1B) have satisfactory prediction ability for the log K_{OW} of the PCB congeners. Finally, the models with 1, 2, and 3 WHIM descriptors (1W, 2W, and 3W) show a fairly good overall performance but underperform with regard to boiling point and the log K_{OW} of the PAHs and, partially, with Henry's law constant of PCBs.

Further comparison of the predictive ability of the considered sets of molecular descriptors was performed only on the basis of the mixed models, i.e., the models obtained starting from all the descriptors of the different kinds and defined in this work with the label ALL. For each set of molecular descriptors, i.e., constitutional, topological, Moreau-Broto, BCUT, WHIM, and GETAWAY descriptors, the frequencies in these mixed models for the 13 considered properties have been computed. Table 17 shows the calculated frequencies, each entry is the number of descriptors of a given set encountered in the mixed models with 1, 2, 3, and 4 variables. For example, topological descriptors were encountered four times in the one-dimensional mixed models (1-ALL), and, more specifically, they were found for the PAH log K_{OW}, boiling and melting point, and the PCB aqueous activity coefficient. Moreover, these descriptors have a frequency equal to 9 in the two-dimensional mixed models (2-ALL), as there is one topological descriptor in the models for five responses (PAH boiling and melting point, phenethylamines log(1/ED_{50}), PCB melting point and aqueous activity coefficient), and there are two topological descriptors in the models for two responses (log K_{OW} of PAHs and PCBs). The last column of the table summarizes the overall frequencies.

It can be easily observed in Table 17 that GETAWAY, WHIM, and topological descriptors are those most frequently selected by the GA variable selection method, demonstrating that they have the best overall performance in modeling the considered properties. Moreover, the topological and WHIM descriptors are those most selected in the one-dimensional models, while the GETAWAY descriptors have shown preference in the high dimensional models. This probably means that topological and WHIM descriptors give holistic information on the molecular structure, while most of the GETAWAY describe only portions of the molecular structure. In fact, a GETAWAY descriptor, R_8(u), was selected

as the best only in modeling the PCB Henry's law constant, it being well-known that PCBs constitute a very high-congener class of compounds.

CONCLUDING REMARKS

The aim of this work was to realize a significant comparative study of different sets of molecular descriptors on the basis of their predictive ability for some selected properties of relevant classes of chemicals. Particular attention was paid to newly proposed GETAWAY descriptors. From the analysis, we can conclude that the GETAWAY descriptors have an overall good modeling capability, proving their usefulness in QSAR/QSPR studies. These descriptors contain local or distributed information on molecular structure, so in most cases more than one GETAWAY descriptor is needed to reach an acceptable modeling power. Moreover, the joint use of GETAWAY and WHIM descriptors, the latter containing information on the whole molecular structure, seems to provide more predictive models when the property to be modeled depends strictly on the 3D features of the molecule, as in the case of biological activities.

ACKNOWLEDGMENT

The work was sponsored by the Commission of the European Union (R&D project "Beam", EVK1-CT1999-00012).

REFERENCES AND NOTES

(1) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. Part 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.

(2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(3) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983.

(4) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Amsterdam, The Netherlands, 2000.

(5) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D- and 3D-Structures. Theory. *J. Chemom.* **1994**, *8*, 263−273.

(6) Todeschini, R.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−119.

(7) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, The Netherlands, 1998; Vol. 2, pp 339−353.

(8) Pearlman, R. S. Novel Software Tools for Addressing Chemical Diversity. *Internet Communication* 1999; http://www.netsci.org/Science/Combichem/feature08.html.

(9) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.

(10) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures, Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.

(11) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.

(12) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Massachusetts, MA, 1989.

(13) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.

(14) Todeschini, R.; Gramatica, P.; Marengo, E.; Provenzani, R. Weighted Holistic Invariant Molecular Descriptors. Part 2. Theory Development and Applications on Modeling Physico-Chemical Properties of PolyAromatic Hydrocarbons (PAH). *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221−229.

(15) Cammarata, A. Interrelationship of the Regression Models Used for Structure-Activity Analyses. *J. Med. Chem.* **1972**, *15*, 573−577.

(16) Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412−1422.

(17) Gramatica, P.; Navas, N.; Todeschini, R. 3D-Modelling and Prediction by WHIM Descriptors. Part 9. Chromatographic Relative Retention Time and Physico-Chemical Properties of Polychlorinated Biphenyls (PCBs). *Chemom. Intell. Lab. Syst.* **1998**, *40*, 53−63.

(18) Tuppurainen, K.; Ruuskanen, J. Electronic Eigenvalue (EEVA): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. A QSAR Approach to the *Ah* Receptor Binding Affinity of Polychlorinated biphenyls (PCBs), dibenzo-*p*-dioxins (PCDDs) and dibenzofurans (PCDFs). *Chemosphere* **2000**, *41*, 843−4848.

(19) Todeschini, R.; Consonni, V. *Dragon, rel. 1.12 for Windows*; Milano, Italy, 2001. Program for the calculation of molecular descriptors from HyperChem, Sybyl, and SD file formats. Free download at http://www.disat.unimib.it/chm/.

(20) *HyperChem, rel. 4 for Windows*; Autodesk Inc.: Sausalito, CA, 1995.

(21) Todeschini, R. *Moby Digs/Evolution, rel. 1.0 for Windows*; Talete srl: Milano, Italy, 2001. Software for multilinear regression analysis and variable subset selection by Genetic Algorithm.

(22) Todeschini, R. Data Correlation, Number of Significant Principal Components and Shape of Molecules. The K Correlation Index. *Anal. Chim. Acta* **1997**, *348*, 419−430.

(23) Todeschini, R.; Consonni, V.; Maiocchi, A. The K Correlation Index: Theory Development and its Applications in Chemometrics. *Chemom. Intell. Lab. Syst.* **1998**, *46*, 13−29.

(24) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(25) Dancoff, S. M.; Quastler, H. *Essays on the Use of Information Theory in Biology*; University of Illinois: Urbana, IL, 1953.

(26) Gutman, I.; Ruscic, B.; Trinajstic, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.

(27) Kier, L. B.; Hall, L. H. The Nature of Structure-Acitivity Relationships and their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307−312.

(28) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.

(29) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(30) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(31) Balaban, A. T. Chemical Graphs. XXXIV. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.

(32) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Center Concept and Derived Topological Centric Indexes. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 106−113.

(33) Magnuson, V. R.; Harriss, D. K.; Basak, S. C. Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications. In *Studies in Physical and Theoretical Chemistry*; King, R. B., Ed.; Elsevier: Amsterdam, The Netherlands, 1983; pp 178−191.

(34) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(35) Randic, M. Novel Shape Descriptors for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607−613.

(36) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221−224.

(37) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press−Wiley: Chichester, U.K., 1986.

(38) Skorobogatov, V. A.; Dobrynin, A. A. Metric Analysis of Graphs. *MATCH (Comm. Math. Comput. Chem.)* **1988**, *23*, 105−151.

(39) Entiger, R. C.; Jackson, D. E.; Snyder, D. A. Distance in Graphs. *Czech. Math. J.* **1976**, *26*, 283−296.

(40) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.

(41) Lovasz, L.; Pelikan, J. On the Eigenvalue of Trees. *Period. Math. Hung.* **1973**, *3*, 175−182.

(42) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227−228.

(43) Gutman, I. Selected Properties of the Schultz Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1087−1089.

(44) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 4. Graph Theory, Matrix Permanents, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 69−72.

(45) Unger, S. H.; Hansch, C. J. On Model Building in Structure-Activity Relationship: A Reexamination of Adrenergic Blocking Activity of *β*-halo-*β*-arylalkylamines. *J. Med. Chem.* **1973**, *16*, 745−749.

(46) Todeschini, R.; Gramatica, P. New 3D Molecular Descriptors: The WHIM Theory and QSAR Applications. In *3D QSAR in Drug Design*;

Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, The Netherlands, 1998; Vol. 2, pp 355−380.
(47) Dearden, J. C.; Cronin, M. T. D.; Schultz, T. W.; Lin, D. T. QSAR Study of the Toxicity of Nitrobenzenes to *Tetrahymena pyriformis*. *Quant. Struct.-Act. Relat.* **1995**, *14*, 427−432.
(48) Schüürmann, G.; Flemmig, B.; Dearden, J. C. CoMFA Study of Acute Toxicity of Nitrobenzenes to *Tetrahymena pyriformis*. In *Quantitative Structure−Activity Relationships in Environmental Sciences − VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 315−327.
(49) McKinney, J. D.; Darden, T.; Lyerly, M. A.; Pederson, L. G. PCB and Related Compound Binding to the *Ah* Receptor(s). Theoretical Model Based on Molecular Parameters and Molecular Mechanics. *Quant. Struct.-Act. Relat.* **1985**, *4*, 166−172.
(50) Safe, S. H. Polychlorinated Biphenyls (PCBs), Dibenzo-p-Dioxins (PCDDs), Dibenzofurans (PCDFs), and Related Compounds: Environmental and Mechanistic Considerations Which Support the Development of Toxic Equivalency Factors (TEFs). *Crit. Rev. Toxicol.* **1990**, *21*, 51−88.
(51) Waller, C. L.; McKinney, J. D. Comparative Molecular Field Analysis of Polyhalogenated Dibenzo-p-Dioxins, Dibenzofurans, and Biphenyls. *J. Med. Chem.* **1992**, *35*, 3660−3666.
(52) Poso, A.; Tuppurainen, K.; Ruuskanen, J.; Gynther, J. Binding of Some Dioxins and Dibenzofurans to the *Ah* Receptor. A QSAR Model Based on Comparative Molecular Field Analysis (CoMFA). *J. Mol. Struct. (THEOCHEM)* **1993**, *101*, 259−264.
(53) Mekenyan, O.; Veith, G. D.; Call, D. J.; Ankley, G. T. A QSAR Evaluation of *Ah* Receptor Binding of Halogenated Aromatic Xenobiotics. *Environ. Health Persp.* **1996**, *104*, 1302−1310.
(54) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T.W. Evaluation of a Novel Infrared Range Vibration-Based Descriptor (EVA) for QSAR Studies. 1. General Application. *J. Comput. Aid. Mol. Des.* **1997**, *11*, 409−422.
(55) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(56) Basilevsky, A. *Statistical Factor Analysis and Related Methods*; Wiley: New York, 1994.
(57) *SIMCA-S, rel. 6.01 for Windows*; Umetrics AB: Umeå, Sweden, 1995.

CI0155053