

An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features

Knut Baumann[†]

Department of Pharmacy and Food Chemistry, Am Hubland, D 97074 Wuerzburg, Germany

Received July 2, 1999

A molecular descriptor based upon a count statistic of the topological distance matrix is described and evaluated for use in QSAR studies. Encoding a molecule is done by computing many selective count statistics (histograms) reflecting the distribution of different atom types and bond types in the molecule. The descriptor was also extended to incorporate geometric features of molecules by weighting the topological distance counts with the geometric distance. It is invariant to both translation and rotation. As a result, it does not require the alignment of the structures under study. The method was applied to several QSAR data sets and performed equally well or better than CoMFA and the EVA descriptor. Compared to the latter two methods, it is computationally easier.

1. INTRODUCTION

During the past decades, numerous theoretical molecular descriptors such as graph-theoretic, geometric, electronic, and combined molecular descriptors have been developed for the analysis and prediction of physical, chemical, environmental, and biological properties of molecules.^{1–5} Put boldly, descriptors can be subdivided into two main classes. The first class consists of descriptors characterizing a molecule by a single number (e.g. topological indices, log P, Taft's E_s , Hammett's σ , molar refractivity, surface area, etc.). Several of these indices are then combined to give the final molecular descriptor. Although these descriptors are highly successful, a treatment of this class is beyond the scope of this contribution. The second class comprises descriptors that use a single algorithm to compute a multivariate descriptor which characterizes a molecule as a whole. This second class can be further subdivided into two classes. Well-known three-dimensional multivariate structure descriptors in the first subclass are comparative molecular field analysis⁶ (CoMFA) and comparative molecular similarity analysis (CoMSIA).⁷ In CoMFA and CoMSIA the molecules under scrutiny need to be aligned in order to compare them. The alignment determines to what extent the descriptors differ from one molecule to the next. Consequently, it substantially influences the results of the evaluation. Hence, significant and relevant results can only be expected if the alignment was carried out properly. The second subclass comprises holistic structure descriptors such as autocorrelation descriptors,^{8–11} EVA,^{12,13} WHIM,¹⁴ MS-WHIM,¹⁵ the molecular transform,^{16–18} and GRIND¹⁹ which are translationally and rotationally invariant descriptors (TRI-descriptors). TRI-descriptors do not need an alignment of the structures. While the downside of TRI-descriptors is their more difficult and sometimes even impossible interpretability (exception: GRIND), their handling is far easier. This is especially true for heterogeneous data

sets. In this contribution a new TRI-descriptor is described which is versatile and at the same time quick and easy to compute. It is based on a previously published graph-theoretic descriptor termed SE-vector²⁰ (for Start-End-vector). The original algorithm is modified in such a way that only the information contained in the topological distance matrix is needed to encode a molecule. Moreover, a geometric extension of the originally two-dimensional descriptor is proposed. The usefulness of the modified and extended descriptors is evaluated with several QSAR data sets of different diversity.

2. MATERIAL AND METHODS

2.1. Algorithm. The mathematical notation in the remainder of the paper is as follows. Scalars are represented by lowercase italicized letters (*a*). Vectors will be denoted by lowercase bold letters (**a**), matrices by uppercase bold letters (**A**). All vectors are column vectors if not indicated otherwise. Single vector elements are represented by lowercase italicized symbols with one index (*a_i*), single matrix elements by lowercase italicized symbols with two indices (*a_{ij}*). Transposed matrices and vectors are assigned the 'T' superscript (**a**^T).

The graph-theoretic part of the descriptor will be outlined first. In graph-theoretic approaches the constitutional formula of a molecule is perceived as a planar mathematical graph where vertices represent atoms and edges covalent bonds. The descriptor outlined is a topological distance counting descriptor¹ (pp 107–110), where each variable is associated with a topological distance occurring in the molecular graph. In the simplest version of the descriptor the value of each variable p_l is equal to the total number of *shortest* paths of length l ($l = 0, 1, \dots, \text{maximum distance (maxD)}$) in the molecular graph. A path of length one being defined as an alternating sequence of an edge (bond) and a vertex (atom). A path of length zero is defined as a vertex in the graph. Thus, the variable p_0 equals the number of atoms in the molecule. The variable p_1 equals the number of bonds.

[†] Corresponding author phone: ++49 931 8885473; e-mail: knut.baumann@mail.uni-wuerzburg.de.

Scheme 1. Pseudocode for Computing a Shortest Path Count Vector (**p**-Vector)

```

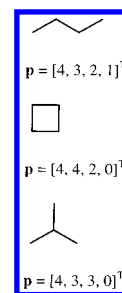
1. Let vector p with elements  $p_0, p_1, \dots, p_{maxD} = 0$ 
2. For  $i = 1$  to  $numatm$  (number of atoms)
3.   Select atom  $i$  as start atom
4.   Let vector d with elements  $d_1, d_2, \dots, d_{numatm} = 0$ 
5.   Perform a breadth first search23 starting at atom  $i$  to determine the shortest
       paths (distances) to all other atoms (end atoms) and store the distance to
       atom  $j$  in vector d at index  $d_j$  for all  $j$ 's ( $j = 1, \dots, numatm$ ).
6.   For  $j = 1$  to  $numatm$ 
7.     If  $d_j \leq maxD$ 
8.        $p_{d_j} = p_{d_j} + 1$  // Update distance counts  $p_0, \dots, p_{maxD}$ 
9.     End If
10.  Next  $j$ 
11. Next  $i$ 
12. Since every path of length  $> 0$  is counted twice (from both ends),  $p_1$  to  $p_{maxD}$  are
    divided by 2 to correct for this.
End.

```

Variables p_2 to p_{maxD} characterize the branching of the molecule. The maximum distance to be considered in the encoding step is a user-defined parameter that depends on the problem. A reasonable range of $maxD$ is 4–7. A slightly different version of this simple distance counting descriptor has first been published by Randic and Wilkins.^{21,22} However, these authors and the authors of the original SE-vector consider all paths of a given length and not only the shortest paths. For acyclic compounds the two differently defined descriptors yield the same code. However, for cyclic compounds the two descriptors differ. Since the shortest paths descriptor is easier to compute and since it is a straightforward count statistic of the topological distance matrix the descriptor outlined here is defined in terms of shortest paths. This modification renders the descriptor to a histogram of the distance matrix. The computational procedure to determine the shortest path counts is shown in Scheme 1 for a simple **p**-vector.

The algorithm in Scheme 1 computes the topological distance matrix (breadth first search^{23,24} in step 5) and counts the numbers 0, 1, ..., $maxD$ (step 8) occurring in the distance matrix. The counts of the distance 1, 2, ..., $maxD$ are divided by two because the algorithm counts these paths twice (from both ends; atom $i \rightarrow$ atom $j =$ atom $i \leftarrow$ atom j). Note that the topological distance matrix **D** is obtained if the i th **d**-vector (step 5) is stored as the i th column of the matrix **D**. The topological distance matrix **D** is a symmetric matrix (i.e. $d_{i,j} = d_{j,i}$) of the size $numatm \times numatm$ with zeros on its diagonal, where $numatm$ is the number of atoms in the molecule. It will be computed only once in the final algorithm to avoid the recalculation of the breadth first search.

Figure 1 shows the **p**-vectors for butane, cyclobutane, and 2-methylpropane (hydrogens are suppressed). The **p**-vector is the basic ingredient of the SE-descriptor, yet a single **p**-vector does not carry enough information for complex QSARs. Since mathematical graphs represent molecular connectivity and lack information about different atom types and bond types the plain graph is “colored” in subsequent encoding steps. Different atoms get different colors (different attributes, ‘O’, ‘N’, ‘Cl’, ‘Br’, etc.). Double and triple bonds are handled in a similar fashion. The atoms taking part in double or triple bonds are assigned the attribute ‘2’ or ‘3’, respectively, and are treated like heteroatoms (they are referred to as pseudoheteroatoms). As a result of the bond

**Figure 1.** Vectors for butane, cyclobutane, and 2-methylpropane (hydrogens are suppressed).

encoding scheme different arrangements of alternating bonds in aromatic systems generate the same descriptor. Double bonds and aromatic bonds may also be differentiated but this option was not explored yet. After all atoms have been assigned their respective attributes, selective distance count statistics for all combinations of different attributes are computed. A selective distance count statistic ‘XY’ (e.g. ‘OCl’) counts the distance between start and end atom only if it starts at an atom exhibiting attribute ‘X’ (e.g. ‘O’) and ends at an atom exhibiting attribute ‘Y’ (e.g. ‘Cl’).

To characterize the entire molecule every atom gets at least one and at most three attributes. The first attribute is ‘T’ for topology. Every atom in the molecule is assigned the T-attribute to thoroughly characterize the topology of the molecule. The second attribute is the atom type. The atom symbol is used here. Note that all atom types may be included in the encoding step which is another difference to the original SE-algorithm of Clerc and Terkovichs. In the original algorithm hydrogens are always suppressed and carbons only get the T-attribute. It was found, however, that incorporation of extra attributes for hydrogens and carbons is sometimes beneficial. By default, hydrogens and carbons are ignored. Additionally, more detailed atom types such as ‘N⁺’ or ‘O⁻’ for charged atoms may be included. The third attribute is assigned to atoms taking part in a double or triple bond. Scheme 2 shows the algorithm for computing all combinations of selective distance count statistics.

The two do-loops generate all combinations of the attributes for the selective count statistics. Before starting the iterations set₁ and set₂ hold all attributes to be processed. During the first iteration of the outer do-loop an attribute is removed from set₁, and this attribute is combined with all attributes of set₂ in the inner do-loop. On completion of the inner do-loop set₂ is empty and is immediately reinitialized with the current set₁ (i.e. set₁ with the first attribute already removed) in step 23. This ensures that every combination of attributes is generated only once. Put differently, if the combination ‘T2’ has been processed in the inner do-loop, then the combination ‘2T’ will never be generated. This is reasonable and lowers the computational burden, since ‘2T’ would yield the same count statistic than ‘T2’ and would count the same paths a second time. The for-loop (steps 9–19) computes the path count statistic. The difference to Scheme 1 being that paths are counted only if they start on an atom with the property determined by the current attribute selected from set₁ (step 10) and end on an atom with the property exhibiting the current attribute selected from set₂ (step 14, 15). Recall that every atom gets the attribute ‘T’. Consequently, for the TT-vector both if-statements in the inner do-loop are always true, and the resulting **p**-vector is

Scheme 2. Pseudocode for Computing a Shortest Path SE-Vector


```

1.  set1 contains all attributes to be considered (e.g. set1 = {T, N, O, S, 2}).
2.  k = 1 // counter for the storage of p-vectors
3.  Let set2 = set1.
4.  Do until set1 is empty
5.      Remove an attribute from set1. Let the removed attribute be A.
6.      Do until set2 is empty
7.          Remove an attribute from set2. Let the removed attribute be B.
8.          Let vector p with elements  $p_0, p_1, \dots, p_{maxD} = 0$ 
9.          For  $i = 1$  to  $numatm$  (number of atoms)
10.             If atom  $i$  shows attribute A then
11.                 Select atom  $i$  as start atom
12.                 Let the vector d be the  $i$ th column of the topological
13.                 distance matrix D.
14.                 For  $j = 1$  to  $numatm$ 
15.                     If atom  $j$  shows attribute B and  $d_j \leq maxD$ 
16.                          $p_{d_j} = p_{d_j} + 1$  // Update distance counts
17.                     End If
18.                 Next  $j$ 
19.                 End if
20.                 If attribute A = B then divide  $p_i$  to  $p_{maxD}$  by 2, since these
21.                 paths have been counted twice (from both ends).
22.                 Store vector pT as the  $k$ th row of matrix P.  $k = k + 1$ .
23.             End Do
24.             Let set2 = set1.
25.         End Do
26.         Flatten the matrix P by chaining all rows of P yielding the final descriptor for the
27.         molecule at hand.
28.     End.

```

the same as that computed by Scheme 1. All other combinations of attributes will generate new **p**-vectors carrying additional information on the molecule. If the attributes selected from set₁ and set₂ are identical, the paths are again counted twice (from both ends). This is accounted for in step 20 where paths of length > 0 are divided by 2. On completion, all selective distance count statistics (**p**-vectors) are chained to give the final descriptor. As mentioned before, the algorithm for computing the SE-vectors can be made more efficient by computing the topological distance matrix **D** in advance. In that case, the breadth first search (step 5 of Scheme 1) to determine the shortest path has to be carried out only once and not over and over again. Then step 12 of the final algorithm simplifies to **d** = j th column of **D** (**d** = **D**(:, j) in MATLAB notation). Figure 2 shows the complete SE-descriptor for acetic acid and isoxazole. Figure 3 shows that the count statistics can also be viewed as radial distribution functions. While computing the descriptor each atom acts as the center point once. The distribution of atoms and bonds around this center point is evaluated depending on the distance from the center up to a maximum distance. The contributions of identical center points (i.e. same atom type, same bond type, or same atom and bond type) are summed over the whole molecule. Note that the topological part of the descriptor is related to the atom pairs descriptor^{24,1} (ref 1: p 428). However, the definition of the atom types and handling of different bond types is different.

2.2. Geometric Extension. The topological (graph-theoretic) part takes different atom types and hybridization into account, but all stereochemical and conformational information is lost. Consequently, cis- and trans-isomers or boat- and chair-conformers yield the same code. Introducing stereochemistry is simple and straightforward. Define the

				
TT ₀	TT ₁	TT ₂	TT ₃	
4	3	3	0	
T2 ₀	T2 ₁	T2 ₂	T2 ₃	
2	4	2	0	
TO ₀	TO ₁	TO ₂	TO ₃	
2	2	4	0	
22 ₀	22 ₁	22 ₂	22 ₃	
2	1	0	0	
2O ₀	2O ₁	2O ₂	2O ₃	
1	2	1	0	
OO ₀	OO ₁	OO ₂	OO ₃	
2	0	1	0	


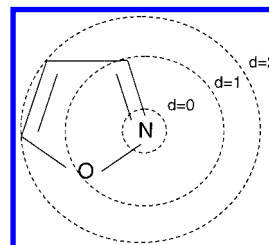
				
TT ₀	TT ₁	TT ₂	TT ₃	
5	5	5	0	
T2 ₀	T2 ₁	T2 ₂	T2 ₃	
4	8	8	0	
TN ₀	TN ₁	TN ₂	TN ₃	
1	2	2	0	
TO ₀	TO ₁	TO ₂	TO ₃	
1	2	2	0	
22 ₀	22 ₁	22 ₂	22 ₃	
4	3	3	0	
2N ₀	2N ₁	2N ₂	2N ₃	
1	1	2	0	
2O ₀	2O ₁	2O ₂	2O ₃	
0	2	2	0	
NN ₀	NN ₁	NN ₂	NN ₃	
1	0	0	0	
NO ₀	NO ₁	NO ₂	NO ₃	
0	1	0	0	
OO ₀	OO ₁	OO ₂	OO ₃	
1	0	0	0	

Figure 2. Complete SE-descriptor considering only shortest paths for acetic acid and isoxazole.**Figure 3.** The (weighted) path counts can be viewed as radial distribution functions of atoms and bonds. Taking nitrogen as the center atom we find for distance $d = 1$ that the (sp^2 -hybridized) nitrogen has two neighbors (irrespective of atom type and hybridization) which yields a TN_1 -count of 2. It is surrounded by one oxygen ($NO_1=1$) and a sp^2 -hybridized carbon ($2N_1=1$). These are the entries nitrogen exclusively determines. Furthermore, nitrogen contributes to TT_1 (+2), $T2_1$ (+2), TO_1 (+1), 22_1 (+1), and $2O_1$ (+1) as either "T" (any atom) or "2" (sp^2 -hybridized atom).

geometric distance matrix as Euclidean distance matrix in three-dimensional space

$$DG_{j,k} = \sqrt{\sum_{i=1}^3 (x_{j,i} - x_{k,i})^2}$$

where $x_{j,i}$ and $x_{k,i}$ are the coordinates of the j th and k th atom in the i th dimension. The only change in the statistic as outlined before is the incremental value (step 15 of Scheme 2). Now, the incremental value is not fixed to 1 for a given topological distance, but becomes the Euclidean distance $DG_{j,k}$ divided by the topological distance $D_{j,k}$ (i.e. the number of separating bonds). That means that the new incremental value is the geometric distance weighted by the topological distance. For instance, let the j th and the k th atom be separated by three bonds ($D_{j,k} = 3$) with a Euclidean distance $DG_{j,k}$ of 4. Hence, the increment for the geometric distance



	3D-coordinates			
	#	x	y	z
	1	0.0000	0.0000	0.0000
	2	1.0000	0.0000	0.0000
	3	1.6124	0.5000	0.6124
	4	1.0000	1.0000	0.0000
	5	0.0000	1.0000	0.0000
	6	-0.6124	0.5000	-0.6124
	1 - 5	identical		
	6	-0.6124	0.5000	0.6124
Geometric extension:	TT ₀	TT ₁	TT ₂	TT ₃
Chair	6.0000	6.0000	4.5915	1.7893
Boat	6.0000	6.0000	4.5915	1.6844

Figure 4. Geometric descriptor for a hypothetical boat and chair conformation. The descriptor differs in TT₃ reflecting the different geometric distances between the third and sixth atom.

count statistic p_3 is 4/3 (as opposed to 1 in the topological descriptor). The weighting scheme does not work for the first vector entry at distance 0 since this would result in a division by 0 ($D_{ii} = 0$, for $i = 1, \dots, \text{numatm}$). Keeping the information about the number of atoms and the number of heteroatoms is essential for many applications, thus the entry at distance 0 is computed like in the topological part. With this weighting scheme, molecules with the same connectivity but smaller geometric distances lead to a molecular descriptor with numerically smaller values. This is shown in Figure 4 for a hypothetical boat and chair conformation. The coordinates have been generated to yield a geometric distance of 1 for all adjacent atoms. The only difference for the two geometrical objects is the geometric distance between atom 3 and atom 6, with the distance for the boat conformation being smaller. This is reflected in the fourth vector entry ('TT₃', topological distance 3) of the descriptor, which is smaller for the boat conformation.

The incremental value could not only be weighted by the geometric distance but could also be computed as a product of atom properties such as van der Waals volume, electronegativity, or partial atomic charge. This leads to the related approach of Moreau and Broto⁸ who termed this descriptor autocorrelation vector. However, geometric information is lost in the autocorrelation vectors. Later, these authors extended the autocorrelation vectors to three-dimensional molecular models by replacing the topological distance by the geometric distance.⁹ A similar approach could be used to incorporate geometric information about the molecule into the SE-vectors as well. The algorithm would be as follows: define a maximum distance (maxD) and a number of bins (numbins). Bin the elements of the lower triangular of the geometric distance matrix into numbins equally spaced containers ranging from 0 to maxD and use the number of elements in each container as the molecular descriptor. That means a histogram of the molecule's interatomic distances is computed where the width of the containers in the histogram is defined by maxD and numbins . Histograms can be computed for all combinations of attributes such as in the topological part of the SE-descriptor. This algorithm was also used and compared to the aforementioned weighting scheme. For a number of applications the results were very similar (results not reported here). However, getting good results needs a comparatively high number of bins in the histogram-like descriptor (2–3 per distance unit). As a result, the number of variables doubles or triples as compared to

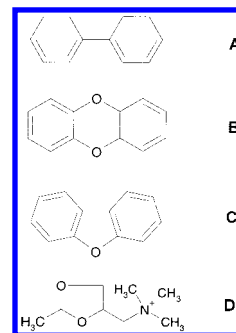


Figure 5. Skeletons of the molecules studied in data sets 1–3 (A–C) and the most potent muscarinic agonist (D).

the weighted version which is the reason the weighting scheme is usually to be preferred. The difference of the 3D-autocorrelation vector⁹ to the histogram-like descriptor just outlined is the computation of histograms for all combinations of attributes. Only one histogram with incremental values weighted by atomic properties is computed in case of the 3D-autocorrelation vector.

2.3. Reduction in Dimensionality. Generating the final molecular code is done by chaining the set of \mathbf{p} -vectors of dimension $\text{maxD} + 1$ in a constant way (Scheme 2, step 25). The resulting vector is of dimension $m = (h(h + 1)/2) (\text{maxD} + 1)$. The dimensionality m of the code depends on the number of attributes h that are included in the encoding step and the maximum distance maxD . The code is highly redundant. For example, the first element of a TX-vector and an XX-vector is always the same (cf. Figure 2, TN₀ and NN₀). Thus, it is advantageous to reduce the high dimensional molecular code in dimensionality by means of a principal component analysis (PCA).²⁵ If PCA is used for compressing the molecular code, the final structure descriptor is simply the a -dimensional score vector. a is the number of principal components to be retained which can be determined by cross-validation.²⁶ The scores can be computed by singular value decomposition (SVD).²⁷ The compressed code can now be used for modeling, classification, or similarity searching. If principal component regression (PCR) or partial least squares (PLS)²⁸ is used for modeling, then the reduction step can be skipped since PCR and PLS will generate a lower rank approximation of the data matrix anyway. In this paper no application of the compressed code is given, but it was used in two previously published papers for similarity searching²⁹ and hierarchical model building.³⁰

2.4. Data Sets. The first four data sets were previously analyzed by CoMFA and the EVA descriptor.^{13,31,32} The fifth data set was taken from refs 12 and 35 where results for the EVA descriptor and for an artificial neural network approach are published. The sixth data set is the well-known steroid benchmark data set.^{6,33,34} Details about the structures, structure descriptors, and the biological activity are given in the cited references. Briefly, the first three data sets deal with the ability of 25 polychlorinated and polybrominated dibenzo-*p*-dioxins, 39 polychlorinated dibenzofurans, and 14 polychlorinated biphenyls to bind to the cytosolic Ah (dioxin) receptor in "in-vitro" rat hepatocyte assays. The skeletons of the structures are shown in Figure 5 (A–C). The biological activity is expressed as pEC_{50} and ranges from 4 to 9.35 (mean = 6.99; $\text{SYY} = 50.08$), 3.00 to 7.82 (mean = 5.90; $\text{SYY} = 67.58$), and 3.85 to 6.89 (mean = 4.95; $\text{SYY} = 9.86$)

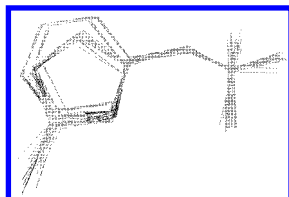


Figure 6. Superimposition of the muscarinic agonists.

for the dioxins, the benzofurans, and the biphenyls, respectively (for the definition of *SY* see below). The fourth data set deals with predicting the binding affinity of a set of 39 noncogeneric muscarinic agonists. The most potent structure is shown in Figure 5D. The compounds were tested on the M3 receptor subtype. The biological activity is expressed as pD_2 determined on isolated rat jejunum and ranges from 4.97 to 7.56 (mean = 5.92; *SY* = 20.92). The fifth data set deals with the prediction of the octanol–water partition coefficient (log *P*) and was employed to demonstrate the usefulness of the EVA descriptor.¹² Several errors were reported by Devillers³⁵ whose corrected data were used here. The data set consists of 185 objects split into 135 training objects (mean = 1.59; *SY* = 215.64) and 50 test objects (*SY*_{Test} = 110.40). The sixth data set studies the binding affinity of 31 steroids to the corticosteroid-binding globulin (CBG). It consists of 21 training objects (mean = 6.15; *SY* = 27.65) and 10 test objects (*SY*_{Test} = 9.27). All data sets are available upon request from the author.

2.5. Structure Descriptors. Two different kinds of SE-vectors were computed. On one hand, the originally published SE-vectors by Clerc and Terkovic are designated as SE-vectors, which count all paths of a given length. On the other, hand SE-vectors were calculated based on the shortest path algorithm as outlined before. These are designated as SESP-vectors (SP for shortest paths). Both versions were calculated to show that there is no loss of information if only shortest paths are considered. If not stated otherwise, all heteroatoms occurring within the data set were considered for generating the descriptor. Carbons and hydrogens were ignored, i.e., they were only assigned the ‘T’ attribute. The maximum distance parameter (*maxD*) was set to 7 after a coarse optimization for the first three data sets and the sixth data set. For the log *P* data (sixth data set) it was set to 5 to avoid an overly complex descriptor. *maxD* has been varied for each of the descriptors for the fourth data set (muscarinic agonists).

The three-dimensional coordinates for the geometric version of the SE-vectors were computed with Alchemy 2000.³⁶ The procedure to compute the coordinates was as follows: first two-dimensional connection tables were converted with Alchemy’s 3-D Builder (Concord) which were then submitted to a molecular mechanics geometry optimization using the Tripos force field and default parameter settings (exceptions: MM3 for the log *P* data, CORINA for the steroids¹⁰). The resulting conformations from geometry optimization were accepted as computed. In case of the muscarinic agonists which show rotatable bonds in their common skeleton, three different sets of conformers were generated. The first set was generated like the descriptors for the other data sets, i.e., the conformations of the 3D-BUILDER were further refined. The second set was generated from the first by a Monte Carlo conformational search (termination criterion: 10 000 valid structures), and the third

set was obtained by superimposing all molecules in the data set. The superimposition was carried out so that the quaternary nitrogen and the maximum common substructure of all molecules matched best under the constraint that the hydrogen-bond donor and acceptor properties are as similar as possible. The superimposition is shown in Figure 6. It was computed with the function “multifit” of the Tripos package Sybyl (version 6.6).³⁶ The template molecule was the most potent one and is shown in Figure 5D.

The proposed structure descriptor was compared to CoMFA and the EVA descriptor. In a CoMFA⁶ analysis molecules are represented by steric and electrostatic interactions between the compound under scrutiny and a probe atom placed at the intersections of a regular three-dimensional lattice. The separation of the lattice points is usually chosen to be 2 Å, and the lattice must be large enough to surround all of the compounds in the training data set. The resulting three-dimensional array of interaction energies is then unfolded in a constant way to yield a vector for each compound. Statistical model building is done by partial least squares (PLS). The results of a CoMFA analysis are not independent of the orientation of the molecules. Consequently, CoMFA needs an alignment of the structures of interest. The EVA descriptor¹³ is derived on the basis of the fundamental vibrations of a molecule and thus looks roughly like an IR-spectral profile. Fundamental vibrations are calculated by means of a normal coordinate analysis from the fully optimized geometry which may be calculated with a molecular mechanics method or with semiempirical/ab initio quantum mechanics methods. The EVA descriptor is translationally and rotationally invariant. Statistical models are also computed with PLS.

2.6. Statistical Modeling and Validation. All QSAR models in this study were calculated by PLS (SimPLS³⁷) if not stated otherwise. SimPLS automatically centers the predictors. By default the descriptors were not scaled. Cross-validation was carried out to select the model complexity (number of latent variables) and to assess the predictive ability of the final model. During resampling which was done randomly without replacement the data were recentered. Cross-validation was computed as a leave-*k*-out procedure ($k \approx n \cdot 1/3$ of the data) for selecting the number of latent variables which was chosen to minimize the estimator of the prediction error *RMSEP* (root-mean-squared error of prediction).

$$RMSEP = \sqrt{\frac{1}{R} \cdot \sum_{r=1}^R \frac{1}{k} \cdot \sum_{i=1}^k (y_{r,i,obs} - y_{r,i,pred})^2}$$

$y_{r,i,obs}$ is the observed value of the *i*th object of the *r*th cross-validation run that was left out. $y_{r,i,pred}$ is the corresponding predicted value of this object. The number of cross-validation runs *R* was set to 200. Note that the *RMSEP* of the leave-*k*-out procedure ($k \gg 1$) may be biased upward due to the large portion of data that is left out.³⁸ But this procedure tends to select fewer latent variables and hence is expected to yield a more stable model with respect to prediction. If *RMSEP*-values decreased only little from one latent variable to the next, the maximum F_{CV} (same as *F* using Q^2 instead of R^2 , see below) was used to decide when to stop adding another latent variable since F_{CV} penalizes

larger number of latent variables. A leave-one-out (LOO) procedure ($k = 1, R = n, r = 1, 2, \dots, n$) was used to calculate the prediction error $RMSEP$ after the model complexity has been chosen (only these $RMSEP$ values are reported here). This was done to achieve comparability to the original studies. The other statistical parameters to assess the quality of the QSARs were calculated as follows:³⁹

$$R^2 = 1 - RSS/SYY$$

$$RSS = \sum_{i=1}^n (y_{i,fit} - y_{i,obs})^2$$

$$SYY = \sum_{i=1}^n (y_{i,obs} - y_{mean})^2$$

$$s = \sqrt{RSS/(n - a - 1)}$$

$$F = \frac{R^2 \cdot (n - a - 1)}{a \cdot (1 - R^2)}$$

$$Q^2 = 1 - \sum_{i=1}^n (y_{i,pred} - y_{i,obs})^2 / SYY$$

where R^2 is the coefficient of determination, s is the standard error of the fit, F is Fisher's F -value, Q^2 is the leave-one-out cross-validated coefficient of determination, and a equals to the number of latent variables. The subscripts are defined as follows: *obs* refers to the experimental (observed) data, *fit* refers to quantities based on the fitted data, and *pred* refers to the cross-validated data. Assessing model quality is based on the predictive ability of the models, i.e., on $RMSEP$ and Q^2 or on $RMSEP_{Test}$ and R^2_{Test} (see below). Figures of merit based on the fitted data (R^2 , s , F) are given to completely characterize the models. $RMSEP$ and Q^2 are measures of predictive ability that are limited to the training data set. They are derived from an internal cross-validation. They do not necessarily give information about the performance of predicting an external test set, particularly so, if the test set contains structural features that are not contained in the training data set. For the fifth and the sixth data set test objects were available. $RMSEP_{Test}$ and R^2_{Test} were calculated analogously to the cross-validated data.

3. RESULTS AND DISCUSSION

Table 1 shows the results for the polychlorinated biphenyls. Models 1, 2, and 3 of Table 1 are as good as the published CoMFA model³¹ and better than the best EVA model¹³ with respect to Q^2 and $RMSEP$. It does not make a significant difference whether the topological (model 1, $Q^2 = 0.57$) or the geometric (model 2, $Q^2 = 0.54$) version of the shortest path algorithm is used. The same is true for the two different topological versions. The shortest path algorithm (model 1) and the all-path algorithm (model 3, $Q^2 = 0.57$) yield the same results. It was expected that the geometric version of the descriptor would perform different (better) than the two topological versions since the geometry of the biphenyls strongly depends on the chlorine substitution pattern (cf. ref 31 for the different 2-1-1'-2' torsion angles) of the biphenyls. The geometric descriptor has the capability

Table 1. QSAR Results of Ah (Dioxin) Receptor Binding Affinity for Biphenyls

model	RMSEP	Q^2	R^2	s	F	m^a	LV^b	n
1 SESP-Top ^c	0.55	0.57	0.76	0.46	18	32	2	14
2 SESP-Geo ^d	0.57	0.54	0.74	0.48	16	40	2	14
3 SE	0.55	0.57	0.76	0.47	17	16	2	14
4 SESP-Top-AS ^e	0.54	0.58	0.78	0.45	19	32	2	14
5 SESP-Geo-AS	0.44	0.72	0.86	0.36	33	40	2	14
6 SE-AS	0.58	0.52	0.74	0.49	15	16	2	14
7 CoMFA ^f	0.57	0.53	0.82	0.40	26	NA	2	14
8 EVA ^g	0.61	0.47	NA	NA	NA	NA	3	14

^a m : number of unique variables with nonconstant variance. ^b LV : latent variables (PLS factors). ^c Top: 2D-topological descriptor. ^d Geo: geometric descriptor. ^e AS: autoscaled. ^f Data taken from ref 31. ^g Data taken from ref 13; Table 5, model: MOPAC PM3.

Table 2. QSAR Results of Ah (Dioxin) Receptor Binding Affinity for Dibenzo-*p*-dioxins

model	RMSEP	Q^2	R^2	s	F	m	LV	n
1 SESP-Top	0.57	0.84	0.93	0.42	53	51	5	25
2 SESP-Geo	0.57	0.84	0.93	0.42	53	73	5	25
3 SE	0.63	0.80	0.90	0.51	43	59	4	25
4 CoMFA ^a	0.76	0.72	0.92	0.45	57	NA	4	25
5 EVA ^b	0.69	0.76	0.91	0.44	116	NA	2	25

^a Data taken from ref 31. ^b Data taken from ref 13; Table 3, model: MOPAC PM3, autoscaled.

Table 3. QSAR Results of Ah (Dioxin) Receptor Binding Affinity for Dibenzofurans

model	RMSEP	Q^2	R^2	s	F	m	LV	n
1 SESP-Top	0.73	0.70	0.83	0.59	40	35	4	39
2 SESP-Geo	0.74	0.68	0.82	0.60	38	48	4	39
3 SE	0.69	0.73	0.82	0.59	53	41	3	39
4 CoMFA ^a	0.67	0.74	0.86	0.54	40	NA	5	39
5 EVA ^b	0.62	0.78	0.97	0.25	274	NA	4	39

^a Data taken from ref 31. ^b Data taken from ref 13; Table 3, Model: MOPAC AM1, autoscaled.

to represent these changes in geometry as opposed to the topological descriptors. Since conformational changes alter the descriptor only slightly (cf. Figure 4), it was autoscaled. Autoscaling does not improve the results of two topological descriptors (models 4 + 6) but does improve the model using the geometric descriptor (model 5). The cross-validated squared multiple correlation coefficient (Q^2) increases from 0.54 (model 2) to 0.72 (model 5). In summary, the conformation of the biphenyls changes depending on the chlorine substitution pattern. The geometric SESP-descriptor contains information about the conformational changes, and this information is related to the receptor binding affinity. As a result, the quality of the QSAR could remarkably be improved.

The SESP-descriptor also performs very well in case of the dibenzo-*p*-dioxins (Table 2). Large differences between the geometric ($Q^2 = 0.84$) and topological ($Q^2 = 0.84$) descriptors are unlikely since the skeleton does not change its conformation depending on the substitution pattern. The SESP-descriptors (topological and geometric) yield slightly better results than the original SE-descriptor ($Q^2 = 0.80$). All three descriptors perform better than the published ones (EVA, $Q^2 = 0.76$).

The opposite is true for the polychlorinated dibenzofurans (Table 3). None of the distance counting descriptors (SESP-

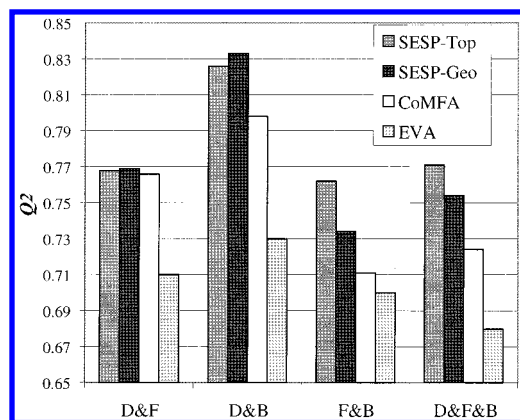


Figure 7. Cross-validated squared multiple correlation coefficient (Q^2) for all combinations of the biphenyl (B), dibenzo-*p*-dioxin (D), and dibenzofuran (F) data sets. The data for CoMFA were taken from ref 31, Table 4, for EVA from ref 13, Table 6, autoscaled. SESP-Top and SESP-Geo refers to the topological and geometric shortest path distance counting descriptor.

Table 4. QSAR Results of Ah (Dioxin) Receptor Binding Affinity^a

model	RMSEP	Q^2	R^2	s	F	m	LV	n
1 T: dioxins and furans	0.70	0.77	0.86	0.59	56	66	6	64
2 T: dioxins and biphenyls	0.66	0.83	0.92	0.48	77	65	5	39
3 T: furans and biphenyls	0.62	0.76	0.85	0.54	37	46	7	53
4 T: all	0.70	0.77	0.85	0.60	57	66	7	78
5 G: dioxins and furans	0.70	0.77	0.85	0.59	55	74	6	64
6 G: dioxins and biphenyls	0.65	0.83	0.92	0.49	75	74	5	39
7 G: furans and biphenyls	0.66	0.73	0.84	0.56	34	50	7	53
8 G: all	0.73	0.75	0.84	0.62	52	74	7	78

^a Combination of the three data sets of Tables 1–3 for the SESP-descriptor (topological (T) and geometric (G) mode).

Geo, $Q^2 = 0.68$) succeeds in modeling the binding affinity better than CoMFA ($Q^2 = 0.74$) or EVA ($Q^2 = 0.78$). Differences between the topological and geometric descriptor were not expected since the skeleton is rigid.

The biphenyl, dibenzo-*p*-dioxin, and dibenzofuran data sets are each of little diversity. The skeleton remains constant, and only the number and position of the halogen substituents are changed. To check whether the SESP-descriptor also works well with heterogeneous data sets, all four combinations of the three data sets were calculated and compared to CoMFA and EVA. The models for the individual data sets were separately optimized, and the performance was seen to depend on scaling options. For the combinations of the data sets default settings were used, i.e., no scaling was carried out. Getting molecular codes of uniform length for all three data sets required some zeropadding since the dioxins are polychlorinated and polybrominated (the variables involving bromines were filled with zeros for the biphenyls and dibenzofurans). Figure 7 shows the predictive ability (Q^2) of the different techniques. Both SESP-descriptors show better predictive ability than CoMFA and EVA, with CoMFA being better than EVA. This indicates that SESP-descriptors can achieve good results with heterogeneous sets of structures. Table 4 gives the complete characterization of the SESP models.

Every 3D-QSAR technique is sensitive to the conformations used for calculating the structure descriptor, otherwise it would not be a 3D-technique. The influence of different conformations is demonstrated with the muscarinic agonists. As already mentioned three different sets of conformers were

Table 5. QSAR Results for the Muscarinic Agonists

model	RMSEP	Q^2	R^2	s	F	m	LV	n
1 SESP-Top	0.39	0.72	0.85	0.29	32	32	7	39
2 SESP-Geo-Df	0.38	0.74	0.88	0.28	34	35	7	39
3 SESP-Geo-Co	0.37	0.75	0.89	0.28	34	35	7	39
4 SESP-Geo-Su	0.37	0.75	0.89	0.28	34	35	7	39
5 SE	0.38	0.73	0.88	0.29	31	35	7	39
6 CoMFA ^a	0.39	0.72	0.90	0.30	22	897	4	39
7 EVA ^b	0.36	0.76	NA	NA	NA	NA	5	39

^a Data taken from ref 32; model 5. ^b Data taken from ref 13; Table 5, model: AMBER.

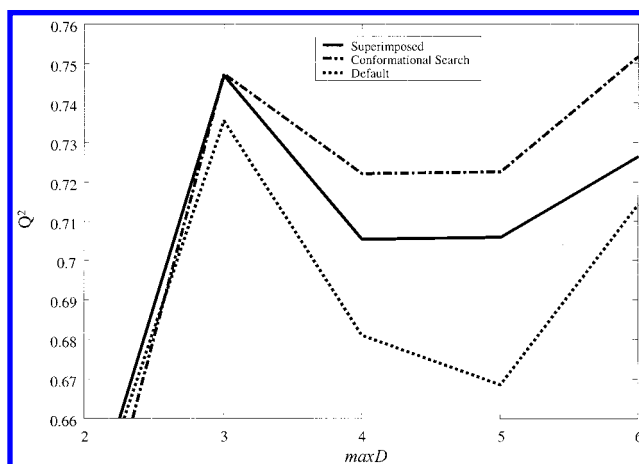


Figure 8. Muscarinic agonists: cross-validated squared multiple correlation coefficient (Q^2) for the three different sets of conformers (SESP-Geo) depending on the maximum distance parameter ($maxD$).

generated: (1) the conformations of the 3D-Builder were refined (Df), (2) the results of a conformational search were used (Co), and (3) the molecules were superimposed (Su). The atom types included in the analysis were T, N, N+, O, and S as well as sp^2 -hybridized-attribute. These amounts ignoring the individual contributions of fluorine, chlorine, bromine, and iodine which were also present as ring substituents. Preliminary calculations showed that the models improved if the halogens were omitted. Hence, only the presence of a substituent and not the nature of the substituent was important to explain biological activity here. The presence of substituents is already encoded by the T-attribute. Moreover, the nature of the substituent is implicitly encoded in the geometric descriptor due to different bond lengths of the carbon–halogen bond. The results for the two topological descriptors and the three geometric descriptors are given in Table 5. The optimal path length ($maxD$) for all five descriptors was $maxD = 3$. The differences between the five sets of descriptors are only slight and range from $Q^2 = 0.75$ for the geometric descriptors to $Q^2 = 0.72$ for the SESP-Top descriptor. The differences are not significant. An optimal path length of three is too short to reveal large differences. Figure 8 shows the results for the three geometric descriptors for $maxD$ ranging from 2 to 6. The number of latent variables was 7 for $maxD = 2, 3, 4$, and 8 for $maxD = 5, 6$. It can be seen that the differences get larger for larger $maxD$. The default approach (converting the structures from 2D to 3D and refining them by molecular mechanics calculations) is inferior to the more sophisticated preprocessing techniques. The largest differences occur at $maxD = 5$ where the Q^2 -values range from 0.67 to 0.72. Note that these

Table 6. Results for Log P Data Set

model	RMSEP _{Test}	R ² _{Test}	RMSEP	Q ²	R ²	s	m	LV	n	n _{Test}
1 SESP-Top	0.50	0.89	0.47	0.86	0.94	0.33	90	12	135	50
2 SESP-Geo	0.46	0.90	0.45	0.87	0.94	0.34	101	11	135	50
3 SE	0.46	0.90	0.51	0.84	0.93	0.36	88	11	135	50
4 EVA ^a	0.88	0.65	0.70	0.68	0.96	0.25	800	6	135	50
5 AC/ANN ^a	0.30					0.26	35		7200	50

^a Data taken from ref 35.**Table 7.** Results for the Steroid Benchmark Data Set

model	RMSEP _{Test} ^a	R ² _{Test} ^a	RMSEP	Q ²	R ²	s	m	LV	n	n _{Test} ^a
1 SESP-Geo	0.45 (>1)	0.81 (<0)	0.41	0.87	0.94	0.32	8	4	21	9 (10)
2 SESP-Geo	0.46 (>1)	0.79 (<0)	0.62	0.70	0.90	0.40	3	3	21	9 (10)
3 EVA ^b	0.51 (0.56)	0.74 (0.69)	0.55	0.80	0.96	0.24	NA	2	21	9 (10)
4 CoMFA ^c	0.40 (0.81)	0.84 (0.35)	0.47	0.85	0.93	0.32	NA	2	21	9 (10)

^a Values in brackets are for all 10 test set compounds, i.e., with the fluorine containing singleton structure. ^b Data taken from ref 12; Table 2, $\sigma = 4 \text{ cm}^{-1}$. ^c Data taken from ref 12; Table 4, Grid interval: 2 Å.

differences are only weakly significant. Since these differences are only present for suboptimal models, there is no reason to worry in this case. For future models incorporating flexible ligands the following strategy for the geometric version is recommended. If information about the bioactive conformation of the molecules under study is available, incorporate it. If there is no information about the bioactive conformation, use a low-energy conformation. Finally it should be noted that the bioactive conformation of ligands and the alignment of a data set are two different subjects. Geometric descriptors that are invariant to translation and rotation can avoid the need for an alignment, but they still need sensible conformations of the molecules under study. Invariance to different conformations can, to the best of our knowledge, only be achieved by 2D-techniques or by 4D-QSAR which samples a large number of conformations.⁴⁰

The fifth data set (log P) was chosen to demonstrate the ability of the SESP-descriptors to handle structurally diverse data sets. The data set contains molecules ranging from methanol to anthracene. To produce comparable results the corrected data by Devillers were used.³⁵ The results are summarized in Table 6. Briefly, the SE-descriptor ($R^2_{\text{Test}} = 0.90$) as well as the SESP-descriptors (SESP-Geo, $R^2_{\text{Test}} = 0.90$) outperform EVA ($R^2_{\text{Test}} = 0.65$) and yield very good models. Test set prediction is almost better by a factor of 2 (SESP-Geo, $RMSEP_{\text{Test}} = 0.46$; EVA, $RMSEP_{\text{Test}} = 0.88$). Neither the SE-descriptor, nor the SESP-descriptors are as good as the artificial neural network using autocorrelation vectors (ANN/AC, $RMSEP_{\text{Test}} = 0.30$) of ref 35. However, the ANN/AC approach uses 7200 molecules for training, whereas the other approaches are trained by only 135 molecules. Given this difference the SESP/PLS approach compares very well with the ANN/AC approach.

As already mentioned in the Introduction, a downside of TRI-descriptors is the often harder interpretability. Some of the TRI-descriptors are impossible to interpret (e.g. molecular transform, autocorrelation vectors). The SESP-descriptor is in principle interpretable, but interpretation is often cumbersome and complicated. If interpretation is desired, a variable selection procedure has to be applied since the entire descriptor carries too much (correlated) information. For the sixth data set (steroid data) a newly developed variable selection routine for PCR and PLS was applied. The

Table 8. Mean-Centered Structure Descriptor and the Respective Regression Coefficients for Three Selected Steroid Molecules

variable	cortisol	cortisone	dehydro- epiandrosterone	coeff $\hat{\beta}$
TT ₅	12.503	12.861	-4.611	0.107
T2 ₀	1.238	3.238	-0.762	0.215
TO ₁	2.915	2.694	-1.156	-0.113
22 ₇	0.059	0.923	0.050	-0.301
2O ₀	0.714	1.714	-0.286	0.201
2O ₂	2.049	2.048	-1.527	0.286
2O ₄	0.348	1.400	0.484	-0.414
2O ₇	0.693	2.403	-0.202	-0.404
\bar{Y}	6.146	6.146	6.146	
\hat{Y}	7.709	7.019	4.991	
\bar{Y}	7.881	6.892	5.000	
$\hat{Y} - \bar{Y}$	-0.172	0.127	-0.009	

technique is a combination of tabu search^{41,42} and leave-multiple-out cross-validation.⁴³ Over a broad range of parameter settings for the cross-validation (leave-7-out to leave-11-out) the selection procedure returns an eight-parameter model with four latent variables. In this case the regression technique was PCR not PLS. Owing to the computationally expensive calculations the results in Table 7 are only given for the geometric SESP descriptor. It can be seen that the final model is very good ($R^2_{\text{Test}} = 0.81$) and can compete with other 3D-QSAR techniques.³³ Prediction of the test set also yields very good results provided the only fluorine containing steroid (molecule 31 in ref 33) is deemed an outlier as was done in most previous studies.³³ Actually, prediction of the fluorine containing steroid cannot work in case of an atom-type based descriptor since no variable in the training data set accounts for the influence of fluorine (there is no fluorine in the training set). The regression coefficients along with the mean-centered structure descriptors of three representative molecules are given in Table 8. The discussion is restricted to three molecules for the sake of brevity. It is not the aim of this study to provide a thorough explanation of the mode of binding of steroids to CBG. The fitted value (\hat{Y}) for the molecules can be obtained by matrix multiplication of the structure descriptor times the regression coefficient and adding the mean binding constant (\bar{Y}).

Four of the variables (TT₅, T2₀, 2O₀, 2O₂) are related to increased binding and the remaining four (TO₁, 22₇, 2O₄, 2O₇)

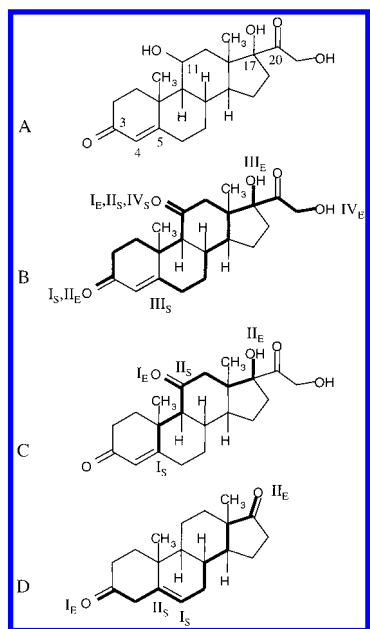


Figure 9. A: cortisol. B: cortisone. The paths for variable $2O_7$ are shown bold. Start (subscript S) and end (subscript E) atoms are indicated by Roman characters. Paths I and III are identical to cortisone, and paths II and IV are due to the ketone in position 11. C: cortisone with paths for variable $2O_4$. Path I can also be found in cortisol; path II is only present in cortisone. D: dihydroepiandrosterone. Path I resembles $2O_4$; path II resembles $2O_7$.

to decreased binding to CBG. TT_5 is mainly a size and shape encoding variable since it does not depend on a particular atom type and encodes the branching of side chains and the presence of substituents on the steroid skeleton. Within this series of steroids TT_5 is especially large if a side chain in position 17 is present (like in cortisol and cortisone) since far more paths of length 5 become possible with this side chain. Within this series of molecules a side chain in position 17 is always identical to the presence of a ketone at position 20 with the ketone being important for high binding constants. Unfortunately, the variable TT_5 is too generic to pinpoint such structural requirements. T_2 encodes the number of sp^2 -hybridized atoms (ketones, double bonds within this series). Taken literally, that means the more ketones and the more double bonds are present the better the binding. This is also expressed in variable $2O_0$ which encodes the number of sp^2 -hybridized oxygen atoms and also has a positive contribution to binding. These unspecific statements cannot be true in general terms but need to be interpreted in a comparative fashion. With the information that T_2 and $2O_0$ are related to increased binding, it is easily revealed that all highly active compounds show ketones in positions 3 and 20 and a double bond in position 4,5. The exact arrangement of ketones and double bonds positively contributing to binding is controlled by the penalizing variables 22_7 , $2O_4$, and $2O_7$, all involving sp^2 -hybridized atoms and in particular sp^2 -hybridized oxygen. If the ketones and double bonds do not obey a certain pattern, the positive contribution to binding of T_2 and $2O_0$ is quickly annihilated by the penalizing variables. This is shown in Figure 9 (A–C) for cortisone and cortisol for the variables $2O_4$ and $2O_7$ since they mainly determine the difference between these two molecules. In cortisone more paths of type $2O_4$ and $2O_7$ than in cortisol are possible owing to the ketone in position 11. The same is true for paths of type 22_7 (not shown). The

effect of the penalizing variables is also shown for dehydroepiandrosterone (D). Variable $2O_2$ has a strong increasing effect on binding to CBG. It encodes the eneone-group ($O=C-C=C$; start and end atom are printed bold) in ring A of the steroids and is an important pharmacophoric element for CBG binding. In summary it can be said that a side chain in position 17 (which also means a ketone in position 20) and an eneone-group in ring A positively influence binding to CBG sp^2 -hybridized oxygens and double bonds placed in the right pattern controlled by the remaining variables explain the rest of the binding constant. However, the way these variables control the binding constants is not immediately clear. Visualization of the paths is necessary to grasp the meaning of the single variables (cf. Figure 9). A visualization module for the descriptor is currently under development. Interpretation is slightly obscured by high correlations between some variables (e.g. T_2 and $2O_0$). If the selection criterion is tightened (leave-13-out cross-validation), then a model with three variables and three latent variables results. The selected variables are as follows: TT_5 , T_2 , and $2O_4$ with regression coefficients $[0.154, 0.176, -1.152]^T$. In this model branching of the molecules and sp^2 -hybridized atoms are positively correlated to CBG binding, and the variable $2O_4$ controls the pattern of oxygens and double bonds by penalizing unfavorable substructures. This is certainly one of the simplest models of this quality ($R^2_{test} = 0.79$) for the steroid data set. However, insight into the mechanism of binding is limited.

4. CONCLUSION

A modification and extension of a previously published molecular descriptor has been described. The modification is to consider only shortest paths in distance counting which makes the descriptor easier to compute. The descriptor modified in this way is a straightforward statistic of the distance matrix and can be viewed as distribution function (histograms). It has also been extended to incorporate information about the molecule's geometry by weighting the increment of the count statistic by the geometric distance. This extension proved useful in cases where the conformation of the molecules strongly depends on their substituents. Only in rather extreme cases such as the biphenyls were the differences between the topologic and the geometric SESP-descriptor significant. So we think of the descriptor of being 2.5D rather than 3D. The descriptor has two tuning parameters: the maximum distance to be considered in the encoding step and the atom and bond types to be included in the descriptor. These parameters are easy to optimize. Although only simple atom and bond types were used in most cases in this study more detailed atom types such as charged species (N^+ , O^-) and aromatic bonds instead of alternating double bonds can be used. On the other hand, general attributes such as 'hydrogen-donor', 'hydrogen-acceptor', 'polar atoms', or 'nonpolar atoms' could also be used.⁴⁰ To this end we do not consider the descriptor as being final. Characterizing molecular features by distribution functions (histograms) can be extended to CoMFA fields, CoMSIA fields, molecular interaction fields, and molecular surface properties. These options will be explored in the future. Although the descriptor is difficult to interpret without some kind of variable selection, it provides an efficient way to predict activities of existing and hypothetical candidate

molecules. The efficiency is mainly due to the fact that no alignment of the molecules during the search for new compounds is necessary.

Summing up, the descriptor is versatile, conceptually easy, interpretable, and easy to compute. It bears a lot of potential for building QSAR models. This has been shown for previously published data sets. The new descriptor performed either equally well or better than the published CoMFA and EVA models.

ACKNOWLEDGMENT

I wish to thank Professor Dr. J. T. Clerc (University of Berne, Switzerland, deceased) for many fruitful discussions about the representation of chemical structures. Moreover, I would like to thank Professor E. Pretsch (ETH Zurich, Switzerland), and Dr. Iain McLay (GlaxoSmithKline, Stevenage, UK) for comments on an earlier draft of the paper.

REFERENCES AND NOTES

- Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
- Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph-theoretic and geometrical descriptors in structure-activity relationships. In *From chemical topology to three-dimensional geometry*; Balaban, A., Ed.; Plenum Press: New York, 1997; Chapter 4, pp 73-116.
- Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure-activity correlations. *Top. Curr. Chem.* **1983**, *114*, 21-55.
- Jurs, P. C.; Dixon, S. L.; Eglolf, L. M. Representations of molecules. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH Verlagsgesellschaft: Weinheim, Germany, 1995; Chapter 2.1, pp 15-38.
- Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835-857.
- Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
- Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *Perspect. Drug Discovery Des.* **1998**, *12-14*, 87-104.
- Moreau, G.; Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 357-358.
- Broto, P.; Moreau, G.; Vandyke, C. Molecular structures: perception, autocorrelation descriptor and SAR studies - Perception of molecules: topological structure and 3-dimensional structure. *Eur. J. Med. Chem. - Chim. Ther.* **1984**, *19*, 61-65.
- Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticoid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.
- Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2674-2677.
- Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Philips, L.; Rogan, J.; Snaith, P. J. EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Design* **1997**, *11*, 143-152.
- Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409-422.
- 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.
- Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR methods: MS-WHIM and autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 293-306.
- Soltzberg, L. J.; Wilkins, C. L. Molecular transforms: a potential tool for structure-activity studies. *J. Am. Chem. Soc.* **1977**, *99*, 439-443.
- Csorvassy, I.; Tözsér, L.; Kárpáti, L.; Náray-Szabó, G. The molecular transform as a similarity measure. *J. Math. Chem.* **1993**, *13*, 343-357.
- Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its applications to structure-spectra correlations and studies of biological activities. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334-344.
- Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRIND-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233-3243.
- Clerc, J. T.; Terkovich, A. L. Versatile topological structure descriptor for quantitative structure-property studies. *Anal. Chim. Acta* **1990**, *235*, 93-102.
- Randic, M.; Wilkins, C. L. On a graph theoretical basis for ordering of structures. *Chem. Phys. Lett.* **1979**, *63*, 332-336.
- Wilkins, C. L.; Randic, M.; Schuster, S. M.; Markin, R. S.; Steiner, S.; Dorgan, L. A graph-theoretic approach to quantitative structure-activity/reactivity studies. *Anal. Chim. Acta* **1981**, *133*, 637-645.
- Gallo, G.; Pallotino, S. Shortest path algorithms. *Ann. Oper. Res.* **1988**, *13*, 3-79.
- Carhart, R. E.; Smith, D. H.; Venkataraghavan, R., Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Sys.* **1987**, *2*, 37-52.
- Krzyszowski, W. J. Cross-validation in principle component analysis. *Biometrics* **1987**, *43*, 575-584.
- Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1993; Chapter 2.6, pp 59-70.
- Geladi, P.; Kowalski, B. R. Partial Least-Squares. *Anal. Chim. Acta* **1985**, *185*, 1-17.
- Baumann, K.; Clerc, J. T. Computer-assisted IR spectra prediction - Linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348*, 327-343.
- Junghans, M.; Pretsch, E. Estimation of partition coefficients of organic compounds: local database modeling with uniform-length structure descriptors. *Fresenius J. Anal. Chem.* **1997**, *359*, 88-92.
- 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.
- Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. Comparative molecular field analysis on a set of muscarinic agonists. *Quant. Struct.-Act. Relat.* **1991**, *10*, 289-299.
- Coats, E. A. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discovery Des.* **1998**, *12-14*, 199-213.
- Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 2. Model validation using a benchmark steroid dataset. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271-296.
- Devillers, J. EVA/PLS versus autocorrelation/neural network estimation of partition coefficients. *Perspect. Drug Discovery Des.* **2000**, *19*, 117-131.
- Tripos Inc. 1699 S. Hanley Rd., St. Louis MO, U.S.A.
- de Jong, S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Sys.* **1993**, *18*, 251-263.
- Davison, A. C.; Hinkley, D. V. *Bootstrap methods and their application*; Cambridge University Press: Cambridge, 1997; p 294.
- Kubinyi, H., Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285-294.
- Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.
- Glover, F. Tabu Search - Part I. *ORSA J. Comput.* **1989**, *1*, 190-206.
- Glover, F. Tabu Search - Part II. *ORSA J. Comput.* **1990**, *2*, 4-32.
- Baumann, K.; Albert, H.; von Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Submitted to *J. Chemom.*

CI990070T