Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D)

Andreas Bender, Hamse Y. Mussa, Gurprem S. Gill, and Robert C. Glen*

Unilever Centre for Molecular Science Informatics, Chemistry Department, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Received May 27, 2004

A novel method (MOLPRINT 3D) for virtual screening and the elucidation of ligand−receptor binding patterns is introduced that is based on environments of molecular surface points. The descriptor uses points relative to the molecular coordinates, thus it is translationally and rotationally invariant. Due to its local nature, conformational variations cause only minor changes in the descriptor. If surface point environments are combined with the Tanimoto coefficient and applied to virtual screening, they achieve retrieval rates comparable to that of two-dimensional (2D) fingerprints. The identification of active structures with minimal 2D similarity ("scaffold hopping") is facilitated. In combination with information-gain-based feature selection and a naïve Bayesian classifier, information from multiple molecules can be combined and classification performance can be improved. Selected features are consistent with experimentally determined binding patterns. Examples are given for angiotensin-converting enzyme inhibitors, 3-hydroxy-3-methylglutaryl−coenzyme A reductase inhibitors, and thromboxane A2 antagonists.

1. Introduction

Molecular similarity searching[1−4,59] attempts to relate differences between observed properties of a set of molecules to their differences in descriptor space, also known as chemical space. A property in this context may be any physical, chemical, biological, or other property that can be attributed to the underlying chemical structure. In the following, the property we focus on is bioactivity which is mediated by ligand−target interaction.

The "molecular similarity principle" [1,59] states that small changes to molecular structure, and thus changes of the position of a structure in descriptor space, usually have small effects on the property under consideration. This is generally true, although cases are known where minor changes of the structure lead to major changes of the property considered.[5,6] This principle leads to the definition of "neighborhood behavior",[7,8] which evaluates descriptors with respect to their ability to associate active compounds with each other in descriptor space.

The concept of chemical space, which is employed to describe molecules, is an often used but not a thoroughly defined idea. In practical terms, positioning of compounds in chemical space generally means the generation of abstract descriptors for a molecule. This is achieved by means of an algorithm, which in most cases is of an empirical nature, and it is not known from first principles whether the algorithm used for descriptor generation is useful. In addition to good performance and general applicability on several data sets, medicinal chemists additionally expect a descriptor to confer "meaning" by being interpretable (which is often not the case). This is achieved if the descriptor can be projected back on the molecular structure in order to identify favorable and unfavorable regions for binding.

Similarity between two items, in this case the comparison of molecules, generally involves generation of representative features for each item (which have to conserve as much relevant information as possible). Selection of those features deemed to be important may be performed (this step is optional) and finally the actual similarity metric is applied to define the distance of items in descriptor space.

A variety of descriptors for molecular structures exist, which are commonly classified according to the dimensionality of data used to calculate them. One-dimensional descriptors use overall properties such as volume and log P,[9] two-dimensional descriptors may be derived from the connectivity table,[10] and three-dimensional descriptors employ geometrical information from points in 3D space.[11,12]

Three-dimensional descriptors are generally created in a more complex and more computationally demanding process than two-dimensional descriptors. Compared to 2D descriptors, they have to deal with problems of translational and rotational variance as well as coping with the information overload resulting from a possible conformational explosion in 3D space. Still they possess the advantage of being able to identify molecules which exhibit similar properties (e.g., pharmacophores) in three-dimensional space without sharing 2D (connectivity table) similarity.

Descriptors that are invariant to both rotation and translation are known as TRI (translationally and rotationally invariant) descriptors. Translational invariance can be achieved by using a coordinate system relative to the molecule and by centering the molecule with respect to it. Rotational invariance can be achieved by using distances between features instead of measuring coordinates in absolute space. This is the basis of

* Corresponding author: phone +44 (1223) 336 432; fax +44 (1223) 763 076; e-mail rcg28@cam.ac.uk.

10.1021/jm049611i CCC: $27.50 © 2004 American Chemical Society
Published on Web 11/16/2004

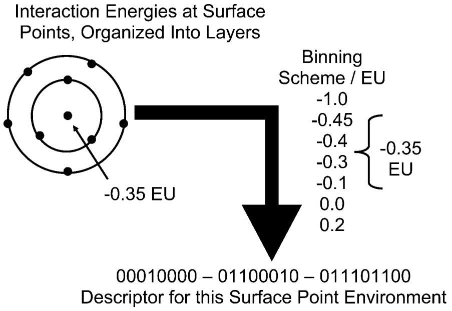

Interaction Energies at Surface Points, Organized Into Layers
-0.35 EU
Binning Scheme / EU
-1.0
-0.45
-0.4
-0.3
-0.1
0.0
0.2
-0.35 EU
00010000 – 01100010 – 011101100
Descriptor for this Surface Point Environment

Figure 1. Illustration of the descriptor generation step. The surface point environment in the upper left is created for every point on the molecular surface. Interaction energies (given in EU, energy units) at every surface point are binned according to a binning scheme, which is calculated to give equally occurring bit frequencies in a random selection of molecules from the MDDR database and which is constant for every probe used. Bits are set in the final descriptor if interaction energies within a bin range are given in the particular layer. Hyphens separate parts of the descriptor created from different layers of the surface point environment.

autocorrelation approaches, which are well-known in both two dimensions[13,14] and three dimensions.[15,16]

“Surface point environments”, the descriptors introduced in this paper, are constructed in a three-step process (see Figure 1). First, points on a molecular surface are computed. Second, interaction energies at surface points are calculated by use of hypothetical probes with varying parameters corresponding to different interaction types. Third, interaction energies are encoded into descriptors, encoding only local information about interaction profiles in binary presence/absence features. In the Experimental Section, we will describe the descriptor in detail and also briefly summarize some of the descriptors that are most similar to surface point environments.

When descriptors are calculated for a molecule, its representation in (this) chemical space is defined. In particular, in the world of 2D fingerprinting, feature vectors (fingerprints) are first calculated, followed by similarity calculation using agreements and discrepancies between all the computed features. Following the idea that most of the features calculated are (for our purposes) noise, a feature selection method is advisable. Here, we employ information-gain based feature selection as introduced by Quinlan.[17] (Details are given in the Experimental Section.)

As mentioned in the previous paragraph, comparison of molecules usually employs similarity or dissimilarity coefficients,[2,18] of which the Jacquard/Tanimoto coefficient[19] and the Manhattan distance[2] are well-known. The Tanimoto coefficient as well as the representation of information in binary bitstrings possess inherent properties.[20−22] One of those properties of the Tanimoto coefficient is its size dependence,[20,21] which was recently addressed by a size-modified Tanimoto coefficient.[23] To allow comparison to other algorithms from previously published results, we also employ the Tanimoto coefficient (which is usually applied to 2D fingerprints).

If information from more than one molecule is given, the problem of merging information can be difficult. Similarity coefficients have a shortcoming in that they, by nature, are only able to deal with single fingerprints. A simple method to combine information from multiple molecules is to define minimum cutoff frequencies for a feature to enter the merged fingerprint.[24] Another recent approach is data fusion,[18,25] which can be based on carrying out a series of single query similarity searches.

Here we follow a different route in calculating similarity. In a fashion similar to binary kernel discrimination,[26,27] a type of data fusion is performed prior to scoring by using the naïve Bayesian classifier. This means that from all the information given in the representation step (and selected in the information-gain based feature selection step), a model is created that incorporates knowledge from all active structures. This contrasts with data fusion, as in data fusion single fingerprints are used for similarity searching and information from multiple searches is only fused after ranking. In contrast, by using the naïve Bayesian classifier, a unified model of active compounds is constructed prior to scoring.

To test MOLPRINT 3D we have utilized a data set containing 957 structures[28] derived from the MDDR[29] that contains active compounds from five different activity classes. The set contains 49 5-hydroxytryptamine (5HT3) receptor antagonists, 40 angiotensin-converting enzyme (ACE) inhibitors, 111 3-hydroxy-3-methylglutaryl−coenzyme A reductase inhibitors (HMG), 134 platelet activating factor antagonists (PAF), and 49 thromboxane A2 antagonists (TXA2). An additional 574 compounds were selected randomly from the MDDR database and did not belong to any of these activity classes. This data set was previously examined[28] by a variety of similarity searching methods and therefore serves as a suitable benchmark.

A number of tasks have been performed on this data set. First, parameters of the algorithm were optimized to achieve good performance in similarity searching. Performance in similarity searching tasks was then compared to other algorithms. As described above, translational and rotational variances as well as conformational tolerance are important points where 3D descriptors are used. The tolerance of this descriptor with respect to conformational variance was investigated by use of a sample from the MDDR data set of all groups of active compounds. Finally, selected features were projected back onto molecular space to investigate agreement with experimentally determined binding patterns. This was performed to check that results from similarity searching were not random patterns of surface points and in addition to investigate the use of this method for elucidating binding patterns in ligand−receptor complexes.

Section 2 presents details of the method and puts it into context with other algorithms. Section 3 gives the results obtained and discusses them fully. This section also gives a comparison of the performance of the algorithm to those of established methods. Conclusions are presented in section 4.

2. Materials and Methods

(a) Descriptor Generation/Molecular Representation. The generation of surface point environments comprises four major steps, which are summarized in Table 1. First, 3D coordinates are calculated from the two-dimensional representation of the structure and saved in hydrogen-depleted SD format. This step is performed by use of Concord 4.0.7[30,31] with standard settings.

Table 1. Main Steps in Descriptor Generation, Listing Programs Currently Used in Each Step and Exemplary Important Parameters[a]

algorithm step	currently used program	selected important parameters
generation of 3D coordinates	Concord	
calculation of surface points	msms	sphere radius, probe size, triangulation density
calculation of interaction energies	GRID	probe (and various others)
transformation of interaction energies into descriptors	Perl script	binning, number of bins, threshold levels

[a] In principle, most parts of the algorithm are replaceable by a wide variety of programs.

The three-dimensional structure of the molecule is used as input for the triangulation of the molecular surface. The program msms[32] was used to calculate the solvent-excluded surface with default radii multiplied by a factor of 2.0. This gives a representative (although not uniform) interaction surface. Exemplary radii used are 3.08 Å for nitrogen, 2.8 Å for oxygen, 3.48 Å for carbon, and 2.4 Å for hydrogen. The probe radius for the triangulation of the surface is set to 1.5 Å, approximately corresponding to the radius of a water molecule. Triangulation densities are set to 0.5/Å^2 and 2.0/Å^2, giving about 400 and 2000 points for an average-sized molecule, respectively.

It is known that the algorithm implemented by msms does not create equidistant points on the molecular surface.[33] To achieve equidistant points on the surface, algorithms such as GEPOL[34] may be employed instead. One should keep in mind, though, that there is no molecular microscopic equivalent of the macroscopic concept of a “surface”, so that parameter choices in this step are by and large arbitrary. In addition, liquid systems are governed by dynamic changes of angles, distances, and charges (by proton transfer), which underlines the fact that a molecular “surface” is only a crude approximation on a microscopic scale.

The SD file of the molecule is converted to hydrogen-added mol2 format by use of OpenBabel 1.100.2.[35] The mol2 file is converted to PDB format containing GRID atom types by the utility gmol2 that accompanies GRID.[36,37] The three-dimensional coordinates calculated in the previous step are fed into the GRID input file grid.in employing the POSI directive in order to calculate interaction energies at the calculated surface points. The maximum energy (EMAX) is set to 5.0 in grid.in. The LEVL −1 directive is used to write GRID output in ASCII format; otherwise standard settings for GRID are used. Currently C3, DRY, N1+, N2, O, and O− probes are used, which we expect to cover a variety of possible interactions between ligand and target.

The energy values calculated at the points on the molecular surface are binned by use of a Perl script. Binning of energy values is illustrated in Figure 1. For each point on the molecular surface, its topologically adjacent neighbors (as given by msms) are calculated and arranged in layers. Points on the surface that are adjacent to the central point (“level 0”), for which the descriptor is generated in this particular step, belong to layer 1. Points that are adjacent to points in layer 1 belong to layer 2, excluding the central point. Points in layer *n* generally are those that are adjacent to points in layer *n* − 1 and that have not been assigned to a layer of lower order.

To create binary presence/absence interaction energy ranges, equifrequent bin ranges have been calculated for the discretization of continuous interaction energies. A random selection of 53 structures from the MDDR database was chosen, and surface points and interaction energies were determined with a triangulation density of 0.5/Å^2 and the C3, DRY, N1+, N2, O, and O− probes. Cumulative frequencies of interaction energies were calculated. The seven bin thresholds were set to give equal populations to all eight bits. The resulting cutoff energies are given in Table 2. All bits corresponding to interaction energies present in a given layer are set in the bitstring.

Overall, for each point on the molecular surface a separate surface point environment vector is calculated. This vector encodes interaction energies at each point of the surface and its neighboring points. Thus, it describes a local surface point environment that potentially facilitates (or reduces) ligand−target binding. No long-distance information is included in our descriptor, which paves the way for a conformationally tolerant description of the molecular surface. On the other hand, it neglects information about overall shape of the molecule. The whole molecule is described by a set of surface point environment vectors.

Table 2. Energy Cutoff Values for Binning of Interaction Energies at Surface Points into Bits[a]

	probe type					
bin cutoff	C3	DRY	N1+	N2	O	O−
cutoff 1	−1.45	−1.12	−4.30	−5.20	−1.90	−2.80
cutoff 2	−1.08	−0.72	−3.20	−3.70	−1.20	−2.10
cutoff 3	−0.85	−0.40	−2.30	−2.45	−0.95	−1.75
cutoff 4	−0.65	−0.08	−1.70	−1.80	−0.80	−1.45
cutoff 5	−0.50	−0.05	−1.30	−1.38	−0.65	−1.22
cutoff 6	−0.35	−0.01	−0.90	−1.00	−0.52	−0.90
cutoff 7	0.72	−0.001	−0.55	−0.75	−0.42	−0.60

[a] Values are calculated to give equifrequent bits in a random selection of molecules from the MDDR database.

The surface point environment descriptor described here is the surface equivalent of the two-dimensional atom environment descriptor, which has been published earlier and which is also known as MOLPRINT 2D.[38,39,44] We will now briefly compare it to approaches that are similar to it.

Some of the best-known TRI descriptors are the GRIND[37] (GRid INdependent Descriptor) and the MaP (Mapping of atomic Properties) descriptor.[33] Three-dimensional autocorrelation[16] also shows resemblance to the method presented here.

The GRIND[37] descriptor is based on interaction energies of the molecule with a probe, which is positioned on a regularly spaced grid. Interaction energies are calculated on a continuous scale with the program GRID.[36] The probes used are typically O and N1 for hydrogen-bonding interactions and the DRY probe for lipophilic regions, but several dozen probes are predefined in GRID and cover a range of possible ligand−target interactions. All interaction energies at grid points are then clustered to simplify the descriptor. Distance ranges (“bins”) are defined and auto- and cross-correlations between interaction energies are calculated. Because only the maximum product of interaction energies enters the descriptor, backprojectability is achieved. In contrast to the GRIND descriptor, our approach explicitly uses points on the molecular surface and bins are replaced by neighbor/nonneighbor relationships between points in space. In the method presented here, encoding is stopped at a fixed number of layers of adjacent surface points, only covering about 3−8 Å in diameter (depending on the surface point density chosen). Interaction energies resulting from different probes are treated independently, in that a separate fingerprint for a given surface point is created for each individual probe used.

MaP[33] also uses points on the surface of the molecule. Employing a modification of the GEPOL algorithm,[34] equally spaced points on the molecular surface are calculated and categorical putative interaction properties of the underlying atom type are assigned. Fuzzy counts are used to increment the bin corresponding to the given triplet of two properties and the distance between them as well as, to a lesser extent, neighboring bins. The calculation of equally spaced surface points provides information about surface interaction properties as well as a size description. In contrast to the MaP descriptor, continuous variables from the GRID force field are employed (which are subsequently binned). Bins are replaced by neighbor/nonneighbor relationships between points in space, and only small parts of the molecule are encoded in each feature in the fingerprint. Fingerprints resulting from different probes are treated independently.

Surface autocorrelation[16] constructs a spatial autocorrelation vector on the molecular surface. The electrostatic potential is calculated and assigned to surface points, which is then encoded in a single surface autocorrelation vector for the whole molecule. In contrast, we construct individual descriptors for each point on the surface by use of different probes, which only cover part of the molecular surface.

(b) Feature Selection. Feature selection is only employed in combination with the naïve Bayesian classifier and multiple query structures. This step is skipped where the Tanimoto coefficient is employed.

The information content of individual surface point environments is calculated from the information gain measure of Quinlan.[17,40] Higher information gain is related to lower information entropy of the subsets defined by presence and absence of a particular feature. Features with higher information gain are expected to allow better classification than features with lower information gain. One of the shortcomings of this method is that no overall optimization of the selection of features is performed; only the next single best feature is chosen in each selection step (which can result in complex solutions).

The information gain, I, is given by

$$I = S - \sum_{v} \frac{|D_v|}{|D|} S_v$$

where

$$S = -\sum_{i} p_i \log_2 p_i$$

S is the information entropy (which is defined analogously to entropy in real mixtures); S_v is the information entropy in data subset v; $|D|$ is the total number of data sets; $|D_v|$ is the number of data sets in subset v; and p is the probability that a randomly selected molecule of the whole data set (or subset in the case of D_v) belongs to each of the defined classes. The probability p can also be seen as the normalized size of each part of the data set. Those parts are denoted by the index i. If a split of the whole data set with respect to presence and absence of a feature is performed, i counts from 1 to 2. Then, p_1 and p_2 represent the size of the data set containing the feature under consideration and the size of the data set not containing the feature under consideration, respectively.

Those features possessing highest information gain I are selected. In order for I to be maximal, and given that the overall entropy of the data set is constant, the information entropy of the subsets has to be minimized. The ideal feature creates pure subsets (of zero entropy) so that all data entries in one subset (e.g., active molecules) possess the particular feature and all data entries in the other subset (e.g., inactive molecules) do not possess it, or vice versa. In practice, features that are as close as possible to this ideal case are selected.

(c) Classification. Two methods were employed for classification: the conventional Tanimoto coefficient and the naïve Bayesian classifier.

The Tanimoto coefficient[2] is a symmetrical similarity coefficient, which takes both similar and dissimilar properties of two items to be compared into account. In the case of binary feature vectors (which are given here; surface point environments are either present or absent in each molecule), the Tanimoto coefficient T_C can be written as

$$T_C = \frac{AND}{OR}$$

where AND is the number of features that are present in both feature vectors to be compared and OR is the number of features that are present in only one of the feature vectors. Features that are present in none of the vectors are neglected by this coefficient.

On the other hand, a naïve Bayesian classifier[41] was employed as a classification tool. Its underlying assumption is the independence of features, although it appears to perform surprisingly effectively where features are not strictly independent.[41,42] Because descriptors calculated from adjacent surface points in this method are often highly correlated, this tolerance toward dependent features is important for our method to work.

The classifier is trained with training data sets that consist of known feature vectors (F) and their associated known classes (CL_v). A Bayesian classifier predicts the class that a new feature vector belongs to as the one with the highest probability of $P(CL_v|F)$, which is given by

$$P(CL_v|F) = \frac{P(CL_v)P(F|CL_v)}{P(F)} \quad (1)$$

where

$P(CL_v)$ is the probability of class v,
$P(F)$ is the feature vector probability, and
$P(F|CL_v)$ is the probability of F given CL_v.

For two data sets, after the assumption of independence of features is applied, the resulting binary naïve Bayesian classifier is given by

$$\frac{P(CL_1|F)}{P(CL_2|F)} = \frac{P(CL_1)}{P(CL_2)} \prod_i \frac{P(f_i|CL_1)}{P(f_i|CL_2)} \quad (2)$$

This equation is used to perform relative scoring; that is, all molecules are represented by their feature vectors F and the resulting ratios $P(CL_1|F)/P(CL_2|F)$ are sorted in decreasing order. Molecules with the highest probability ratios are most likely to belong to class 1 (e.g., the class of active molecules). Molecules with the lowest values are most likely to belong to class 2 (e.g., the class of inactive molecules). The prior, $P(CL_1)$ and $P(CL_2)$ in formula 2, is set to the relative training set sizes.

(d) Data Set Preprocessing. Salts and solvent were removed, if present. Structures were converted to SD format by use of OpenBabel[35] 1.100.2 with the −h option to add all hydrogen atoms. Only the neutral forms of molecules were considered. Surface fingerprints were calculated directly from SD files. The 49 structures of the 5HT3 data set and the 40 structures from the ACE data set were converted correctly. From the PAF data set, two out of the original 134 structures were not converted, leaving 132 structures. One of the 49 structures from the TXA2 data set and 14 out of 574 structures from the "inactive" data set were not converted, leaving 48 and 560 structures, respectively. Overall, descriptors for 937 of 957 structures were calculated. Failure was in all cases due to msms, which produced core dumps. Replacement by a different algorithm might reduce the failure rate.

3. Results and Discussion

To better understand behavior of the surface point generation step, the coordinates of surface points generated by msms[32] were analyzed: For point densities of 0.5/Å^2, 1.0/Å^2, and 2.0/Å^2, distances to all nearest neighbors of each individual surface point were calculated and density functions of nearest-neighbor distances were plotted for corticosterone (Figure 2 and Table 3). The mean distance between points decreases from 1.74 to 1.37 to 1.09 Å if surface point densities are increased from 0.5/Å^2 to 1.0/Å^2 to 2.0/Å^2. Median distances decrease from 1.57 to 1.12 to 0.79 Å in this case. Point densities for other compounds show comparable distributions.

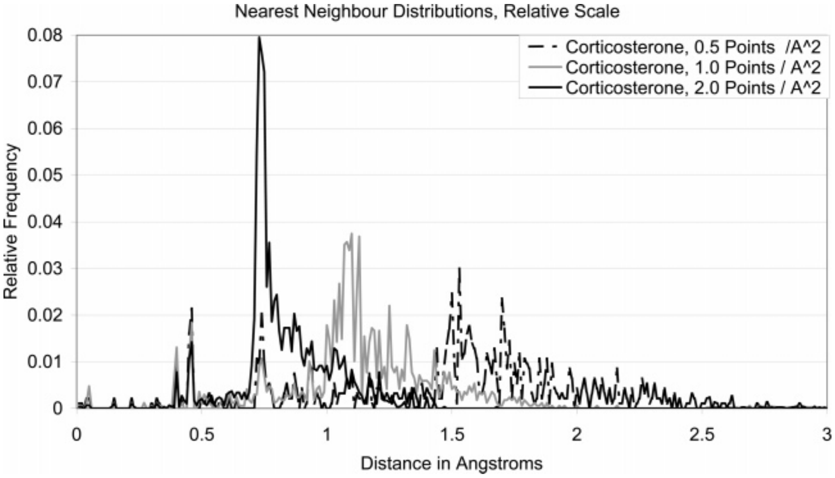


Figure 2. Distribution of surface points generated by msms. Displayed are distributions of nearest-neighbor distances on the surface of corticosterone at triangulation densities of 0.5, 1.0, and 2.0 points/Å^2. The higher the point density chosen, the smaller the nearest-neighbor distances become. In each case, considerable spread of distances is observed. Distributions show comparable means and distributions for other molecules as well.

Table 3. Mean and Median Distances and First and Third Quartiles of Interpoint Distances between Points on the Molecular Surface

point distance (Å)	0.5/Å^2	1.0/Å^2	2.0/Å^2
mean	1.74	1.37	1.09
first quartile	1.17	1.02	0.73
median	1.57	1.12	0.79
third quartile	1.88	1.33	0.95

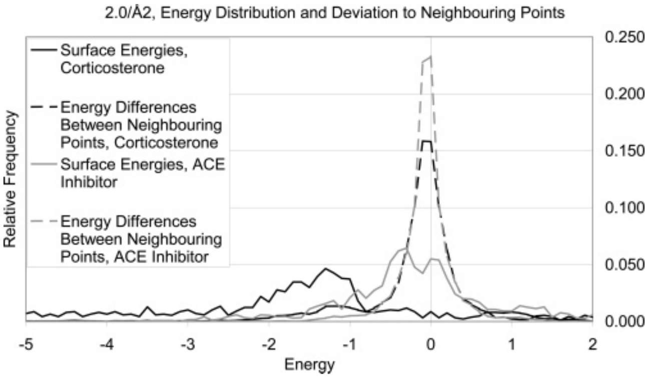


Figure 3. Energy distribution for corticosterone and an ACE inhibitor (both structures are given in Figure 4) at 2.0 points/Å^2 for the O− probe. Distributions of nitrogen probes shift the ACE distribution more to the left (compared to the ACE inhibitor). Dashed lines show energy differences between nearest neighbors and show smooth energy transitions from a point to its nearest neighbor.

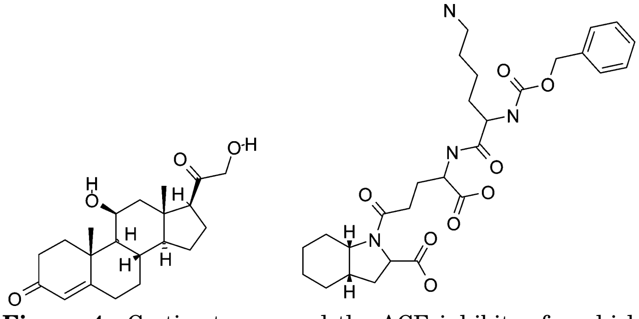


Figure 4. Corticosterone and the ACE inhibitor for which energy distributions with point spacing of 2.0/Å^2 and an O− probe are shown in Figure 3.

The distribution of points on the molecular surface is not equidistant but shows considerable spread, in particular in the case of smaller point densities. The distributions are remarkably similar among different individual compounds (data not shown). The average distance between points of around 1 Å at point densities of 2.0/Å^2 amounts to an area of the molecular surface covered by each individual descriptor (layer 0−4) that spans about 8 Å in diameter.

For all six probes used by GRID, the energy distribution (relative frequencies of surface points within a certain energy range) over the molecular surface was calculated for a rigid molecule, corticosterone, and a more flexible ACE inhibitor at 2.0/Å^2 grid spacing with an O− probe (Figure 3; see Figure 4 for structures). In addition, average energy changes from each point to its neighbors (smoothness of the potential) were calculated.

Absolute energy distributions show a maximum at a higher value for the ACE inhibitor than for corticosterone. Since the O− probe used is negatively charged and the maximum is shifted to higher (more unfavorable) energies for the ACE inhibitor, this is consistent with the larger number of negatively polarized oxygen atoms in this structure, relative to the total molecular surface area. Energy differences between individual points and their next neighbors show a sharp peak around the origin of the coordinate system, indicating that most changes in potential between points occur gradually. This shows that the energy functions are behaving smoothly, which decreases the probability of artifacts in the descriptor generation step. Distributions for nitrogen (N1+ and N2) probes shift energy distributions for both compounds in opposite directions. The C3 probe gives approximately overlapping distributions for both compounds. Different probes show, as expected, different energy distributions.

In the first series of calculations, a 10-fold random selection of single structures from each data set of active compounds was performed. The average hit rate for each class of active molecules (= the number of molecules among the 10 most similar structures belonging to the same activity class as the query structure) was calculated for each compound. Performance, defined as average hit rates, was compared for surface point

Table 4. Comparison of Performance of the Atom Environment Descriptor and the Surface Point Environment Descriptor in Combination with All Probes, Both Used in Combination with the Tanimoto Coefficient[a]

layers used	point density	5HT3	ACE	HMG	PAF	TXA2	average
0	0.5/Å^2	1.80 (1.40)	0.90 (0.88)	3.0 (2.16)	2.10 (1.37)	1.60 (1.46)	1.88 (1.46)
	2.0/Å^2	1.70 (0.67)	1.80 (2.00)	2.80 (1.69)	3.30 (1.25)	1.50 (1.65)	2.22 (1.45)
0−1	0.5/Å^2	4.90 (3.03)	4.80 (2.20)	6.30 (2.36)	7.50 (2.55)	6.90 (1.91)	6.08 (2.41)
	2.0/Å^2	4.30 (2.45)	2.60 (1.96)	6.90 (2.77)	7.10 (3.03)	5.70 (2.26)	5.32 (2.49)
0−2	0.5/Å^2	5.50 (2.37)	5.80 (2.82)	3.60 (2.84)	7.60 (3.24)	7.00 (2.16)	5.90 (2.68)
	2.0/Å^2	4.30 (2.67)	3.90 (2.18)	5.20 (3.08)	7.60 (2.59)	7.50 (2.17)	5.70 (2.54)
0−3	0.5/Å^2	5.60 (2.22)	6.40 (2.95)	4.00 (3.06)	7.60 (1.96)	6.50 (1.96)	6.02 (2.72)
	2.0/Å^2	5.70 (2.16)	5.00 (2.62)	4.70 (2.45)	7.70 (2.79)	7.50 (2.22)	6.12 (2.45)
0−4	0.5/Å^2	5.40 (2.22)	5.40 (2.55)	4.00 (3.02)	7.30 (3.27)	6.30 (2.21)	5.68 (2.65)
	2.0/Å^2	6.10 (1.85)	6.10 (2.42)	4.30 (2.54)	7.60 (3.03)	7.10 (2.47)	6.24 (2.47)
atom environments		7.4 (2.2)	7.8 (2.6)	8.6 (2.1)	7.7 (2.3)	6.6 (2.2)	7.5 (2.3)

[a] Given are mean hit rates among the 10 most similar compounds of a random selection of 10 active compounds of each active data set (standard deviation in parentheses). In the case of surface point environments, the number of layers used for descriptor generation and point densities of 0.5/Å^2 and 2.0/Å^2 is varied.

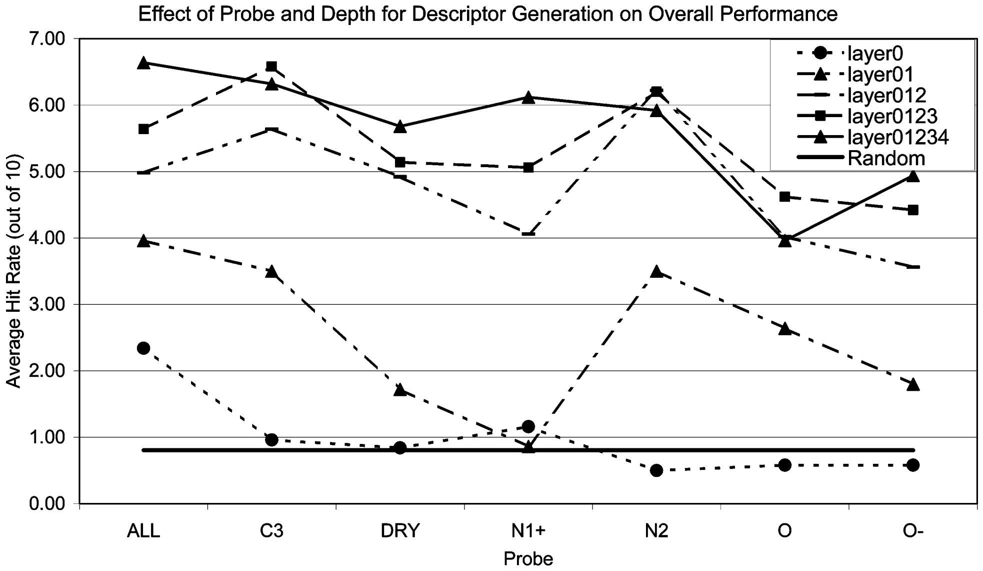


Figure 5. Similarity searching performance upon varying layer depth and choice of probe for descriptor generation. Performance is given for 500 features in each case and employing the Bayesian classifier for classification. Performance depends on the choice of probe mainly in those cases of a small number of layers and levels off between three and four layers.

environments calculated with a triangulation density of 0.5/Å^2 and 2.0/Å^2. In the second calculation, the Tanimoto coefficient was replaced by information-gain-based feature selection and the naïve Bayesian classifier for classification. Ten-fold random sets of five active molecules have been selected and the set of inactive molecules was used in a 50/50 split.[39] Feature selection was set to select 200, 500, or 1000 features for each set of molecules. As in the previous calculation, the hit rate among the 10 highest ranked hits of the sorted library was calculated.

Hit rates for the surface point environment descriptor and atom environments in combination with the Tanimoto coefficient are given in Table 4. Overall hit rates are best for the 2D descriptor (atom environments), which on average retrieves 7.5 active compounds among the 10 structures most similar to the query (bottom of Table 4). Surface point environments created with a point density of 2.0/Å^2 are second in performance with, on average, 6.2 structures retrieved (if layers 0−4 are used for descriptor generation). At a lower point density of 0.5/Å^2, on average 6.1 structures are retrieved (if layers 0−1 are used for descriptor generation). If a point density of 0.5/Å^2 is employed, performance is broadly constant at 5.68−6.08 if the number of layers used to generate the descriptor is varied. This is true with the exception of employing single surface points; in this case, performance drops rapidly to an average hit rate of 1.88. If a point density of 2.0/Å^2 is employed, using layers 0−4 for descriptor generation gives best results, with an average hit rate of 6.2. Performance is again broadly constant at 5.32−6.24 if the number of layers used to generate the descriptor is varied. This is also true with the exception of employing single surface points; in this case, performance drops rapidly to an average hit rate of 2.22.

If single points are used ("level 0"), only a slight increase of performance over random selection can be observed (Table 4 and Figure 5). This means that the number of surface points with a given interaction energy with the probe (a property roughly analogous to measures such as polar surface area, measuring the fraction of the surface with a given property) is not sufficient to achieve classification. Finer point spacing may be better at capturing local properties, although differences are minimal. Overall, significant enrichment is observed for each of the point densities chosen above (except for single points). Performance increases until all points up to layer 4 are incorporated into the descriptor and levels off at that point. This behavior can be explained by the

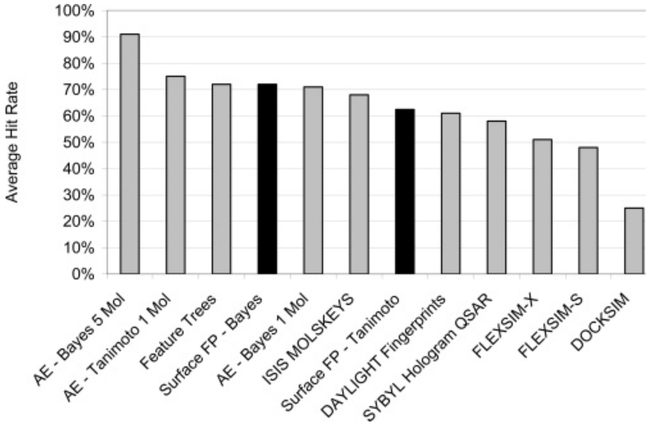


Figure 6. Comparison of similarity searching performance of the surface fingerprint descriptor in combination with other similarity searching methods. Performance of the surface fingerprint descriptor on these dataset is comparable to that of two-dimensional methods. The Bayesian classifier combines information from multiple molecules and is able to increase classification performance.

way surface point environment descriptors are generated. Bits in the feature vector represent presence/absence of interaction energies in a given energy range in a given layer. The further out the layer from the central surface point, the more points are present in that layer. In layer 4, a high number of surface points (typically about 40−50) are present. More often than not, all interaction energies are present in that layer, setting all bits in the vector to 1. Therefore, no new information enters the descriptor in this layer. This is even less likely in layers of higher order (as they contain an even larger number of points).

The superiority of the two-dimensional descriptor compared to its three-dimensional analogue (Table 4) purely with respect to hit rates is in agreement with earlier findings.[43] Still, in relative numbers, surface point environments retrieve only 17% and 24% fewer active compounds than atom environments, which still compares favorably with a number of 2D methods (Figure 6).

Dependence of retrieval performance upon varying parameters of this algorithm is given in Table 5. Results obtained if all field fingerprints are used and a fixed number of 500 features is selected are visualized in Figure 5. Performance is given by use of the Bayesian classifier and sets of five active molecules and a 50/50 split of inactive molecules.

Combining information from all interaction fields used (ALL column in Table 5) improves results over those obtained from only single interaction fields (other columns with probe names) in the case of the best overall retrieval rate, with layers 0−4 and 200 features. Still, performance with only the C3, DRY, N1+, and N2 probes is surprisingly good and, depending on the precise parameters, often comparable to the performance achieved with all probes. A possible explanation is that every probe simply describes the same variance in the data. Although different interaction energies are assigned to the same point in space if different probes are used, the overall variance (which is essential for classification) remains similar. A positive and a negative charge may give the same information, simply with an opposite sign.

Feature selection does not influence results at a small number of layers used for descriptor generation but improves results throughout if more than layer 1 is used for descriptor generation. Feature selection continuously improves classification results if more than layers 0 and 1 are employed for descriptor generation (Table 5), and this is in analogy to atom environments if a large number of feature vectors is employed.[44]

The difference in performance between atom environments and surface point environments depends on the class of active compounds. Performance on the 5HT3, ACE, and PAF data sets is better for the 2D atom environments. Both 5HT3 and ACE data sets give on average 6.1 (2.0/Å^2) and 5.4 (2.0/Å^2) hits for surface point environments, compared to 7.4 hits (5HT3) and 7.8 hits (ACE) for atom environments. On the PAF data set, surface point environments retrieve on average 7.6 (2.0/Å^2) and 7.3 (0.5/Å^2) hits versus 7.7 in case of atom environments. For the TXA2 data set, results are comparable; the hit rates are 7.1 (2.0/Å^2) and 6.9 (0.5/Å^2), where atom environments retrieve on average 6.6 hits. Atom environments retrieve twice as many active compounds from the HMG data set, though; hit rates are 8.6 for atom environments versus 4.3 and 4.0 for surface point environments at high and low point density, respectively. Surprisingly, performance of surface point environments on the HMG data set shows much better performance if only layers 0−1, corresponding to much smaller surface patches, are used for classification. Elimination of 2D similar molecules from the HMG data set did not give the expected result that high connectivity similarity favors the 2D method on this data set in particular. The underlying reason for different performance of surface point environments and atom environments on this data set is as yet unknown.

A comparison of the performance of Tanimoto coefficients and the naïve Bayesian classifier is given in Table 6. The point density is varied between 0.5/Å^2 and 2.0/Å^2, and in the case of the Bayesian classifier, the number of selected features is varied as well. Given the fact that the Tanimoto coefficient uses single queries and no information from inactive structures, it performs surprisingly well. If at least points adjacent to the central surface point are used for descriptor generation, average hit rates of the Tanimoto coefficient are between 5.68 and 6.08, compared to hit rates between 3.96 and 7.16 for the Bayesian classifier with five active structures (all at 0.5/Å^2). At a higher point density of 2.0/Å^2, average hit rates with the Tanimoto coefficient are between 5.32 and 6.24, compared to between 2.64 and 5.04 if the naïve Bayesian classifier is used. The Tanimoto coefficient and the Bayesian classifier thus show opposite tendencies with respect to classification performance if the surface point density is increased (Table 6): Tanimoto performance slightly improves with denser surface points, while performance of the Bayesian classifier decreases. This may be due to assumptions underlying the naïve Bayesian classifier employed here, which is the independence of features. If highly correlated features are present in a given molecule that, for example, classifies a molecule to be active, they are all treated as independent features. Classification of the molecule is thus skewed, because the naïve Bayesian classifier treats them as independently biased toward

Table 5. Similarity Searching Performance upon Varying the Layer Depth Used for Descriptor Generation, Choice of Probe for Generation of Interaction Energies, and Number of Features Selected

		interaction probe						
layers used	no. of features	ALL	C3	DRY	N1+	N2	O	O−
0	all	2.34 (0.38)	0.96 (0.52)	0.84 (0.34)	1.16 (0.72)	0.50 (0.32)	0.58 (0.86)	0.58 (0.74)
0−1	200	4.02 (1.92)	4.06 (1.54)	2.90 (1.78)	2.46 (1.42)	4.32 (1.60)	2.44 (1.22)	2.36 (1.28)
	500	3.96 (1.52)	3.50 (1.84)	1.72 (1.06)	0.86 (0.58)	3.50 (1.84)	2.64 (1.02)	1.80 (0.50)
	1000	4.00 (1.62)	3.18 (1.46)	1.72 (1.20)	0.70 (0.38)	2.18 (1.44)	1.96 (0.94)	1.70 (0.52)
0−2	200	4.96 (2.00)	5.14 (1.36)	4.80 (1.68)	3.60 (1.68)	5.72 (1.64)	5.16 (2.06)	3.40 (1.92)
	500	4.98 (1.50)	5.64 (1.68)	4.92 (1.68)	4.06 (1.16)	6.22 (1.58)	4.02 (1.78)	3.56 (1.44)
	1000	5.38 (1.24)	6.00 (1.48)	4.08 (1.68)	3.16 (1.86)	5.90 (1.90)	3.82 (1.38)	2.74 (1.34)
0−3	200	5.86 (2.14)	6.12 (2.02)	5.38 (1.52)	5.10 (1.88)	5.66 (1.46)	4.80 (1.86)	4.34 (2.40)
	500	5.64 (2.44)	6.58 (1.86)	5.14 (1.68)	5.06 (1.76)	6.20 (1.40)	4.62 (2.02)	4.42 (1.60)
	1000	4.78 (2.06)	6.66 (1.34)	5.02 (1.82)	4.62 (1.52)	6.48 (1.22)	4.26 (1.56)	3.76 (1.84)
0−4	200	7.16 (1.64)	5.64 (2.22)	4.92 (1.88)	5.38 (1.98)	4.68 (1.62)	4.10 (1.46)	4.24 (1.76)
	500	6.64 (2.00)	6.32 (1.86)	5.68 (1.86)	6.12 (1.68)	5.92 (1.76)	3.96 (1.88)	4.94 (2.16)
	1000	6.50 (2.28)	6.76 (1.58)	4.82 (1.74)	5.16 (1.84)	6.24 (1.90)	4.58 (1.66)	4.66 (1.98)

[a] The Bayesian classifier is used for classification. Performance increases continuously from use of only layer 0 and levels off if central points up to points 4 layers apart are used for descriptor generation. Performance is given for a point density of 0.5/Å^2. Values in parentheses are standard deviations from the mean values.

Table 6. Performance of the Similarity Searching Algorithm Using Surface Fingerprints in Combination with the Tanimoto Coefficient upon Varying the Number of Layers Used for Descriptor Generation, Number of Features Selected, and Surface Point Density[a]

Tanimoto Method

point density	performance for the following no. of layers used				
	0	0−1	0−2	0−3	0−4
0.5/Å^2	1.88 (1.46)	6.08 (2.41)	5.90 (2.68)	6.02 (2.71)	5.68 (2.65)
2.0/Å^2	2.22 (1.45)	5.32 (2.49)	5.70 (2.54)	6.12 (2.45)	6.24 (2.47)

Bayes Method, Five Actives

		performance for the following no. of layers used												
point density	no. of features:	0	0−1			0−2			0−3			0−4		
		all	200	500	1000	200	500	1000	200	500	1000	200	500	1000
0.5/Å^2		2.34 (0.38)	4.02 (1.92)	3.96 (1.52)	4.00 (1.62)	4.96 (2.00)	4.98 (1.50)	5.38 (1.24)	5.86 (2.14)	5.64 (2.44)	4.78 (2.06)	7.16 (1.64)	6.64 (2.00)	6.50 (2.28)
2.0/Å^2		1.36 (0.80)	4.08 (1.74)	3.44 (1.44)	2.64 (1.66)	4.84 (1.72)	4.00 (1.56)	4.08 (1.92)	6.08 (1.32)	5.68 (1.44)	4.68 (1.08)	4.60 (1.98)	5.04 (2.72)	4.36 (1.86)

[a] To estimate performance gain if multiple molecules and the naïve Bayesian classifier are used, corresponding values are given for comparison. Values are averaged over all data sets, and numbers in parentheses are standard deviations from the mean values.

activity, although they (in an extreme case) all confer the same information. This result that the naïve Bayesian classifier does not perform particularly well in case of partially correlated features (as it is the case here) was also found earlier.[42] It is still surprising that, overall, performance of single structures with the Tanimoto coefficient is of comparable performance to the Bayesian classifier, although the latter methods has knowledge about multiple active structures as well as about inactive structures.

Overall performance is compared to other methods in Figure 6. Compared are atom environments[39] with the Tanimoto coefficient,[44] feature trees,[45] surface fingerprints with the Bayesian classifier (as described in this work), atom environments with the naïve Bayesian classifier,[39] ISIS MOLSKEYS,[46] surface point environments with the Tanimoto coefficient (as described here), Daylight fingerprints,[47] SYBYL Hologram QSAR,[48] and three virtual affinity fingerprint methods: Flexsim-X,[49] Flexsim-S,[50] and DOCKSIM.[51] Performances of methods other than atom environments and surface point environments are taken from Briem and Lessel.[28] The method presented here outperforms the (3D) virtual affinity fingerprint methods as well as the (2D) Daylight and SYBYL Hologram QSAR fingerprints. One of the reasons for that may be conformational tolerance of this descriptor, as discussed in detail below. Other 3D descriptors, which employ overall distance information between pharmacophores (be it surface points or atom-centered pharmacophores), change considerably if the descriptor is calculated for multiple conformations, while this descriptor is reasonably tolerant to conformational changes.

Three-dimensional descriptors always depend (to a varying degree) on the particular conformation of the molecule to be described; hence, tolerance of the descriptor presented here with respect to conformational changes was examined. Ten molecules from each of the five sets of active compounds were chosen randomly. By use of the genetic algorithm conformational search in Sybyl,[48] a set of 10 random conformations of each molecule was created. Genetic search was favored over the random search option because random searches do not cover conformational space sufficiently well if only a small number of conformations is created. The window size for the genetic search was set to 10° in the case of rigid 5HT3 ligands and 100° in the case of all other data sets (ACE, HMG, PAF, and TXA2), giving highly diverse conformations. Structures were optimized with the Tripos force field for 100 iterations to remove steric strain. All 10 conformations were put into the database containing "inactive" structures as well as all active structures from the five active data sets, excluding the query structure. The query was generated by Concord

Table 7. Percentage of Conformations of the Same Structure Found at the Top of the Sorted Database[a]

top *n* positions	% of conformations found					
	5HT3	ACE	HMG	PAF	TXA2	average
10	70	69	75	56	50	64
20	85	87	91	81	70	82.8
30	89	94	94	90	78	89
40	90	96	96	93	88	92.6
50	90	97	96	95	92	94

[a] In the top 50 positions (corresponding to about 5% of the database), 94% of the conformations are found.

and optimized by the Tripos force field for 100 iterations to remove steric strain. All structures of the database were ranked according to Tanimoto similarity to the query structure. For a truly conformationally invariant descriptor, all 10 conformations should occur at the top of the sorted list because all descriptors were calculated for different conformations of the same structure. For a very sensitive descriptor, considerable spread throughout the database is expected. The number of different conformations of the query structure among the top 10, 20, 30, 40, and 50 positions of the sorted library was calculated to gauge conformational tolerance of the descriptor.

The influence of conformational variance on descriptor generation is given in Table 7. Nearly two-thirds (64%) of all conformations of the same molecule are identified as most similar by the Tanimoto coefficient (placed at the top 10 positions of the sorted list), and 94% of all conformations are found in the top 50 positions (roughly 5%) of the sorted library. Thus, if a molecule that is similar to the query molecule is present in the database, it is likely to be ranked at the top of the sorted database. This leads to the tentative conclusion (based on the five different data sets and diverse sets of conformations employed here) that the descriptor is unlikely to miss an active molecule when it is just not present in the "correct" conformation (e.g., the binding conformation or any other pharmacophoric conformation) in the database.

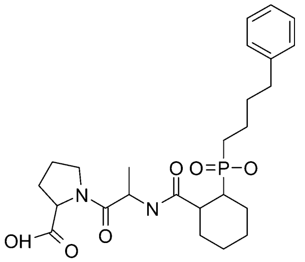
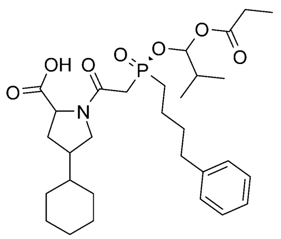
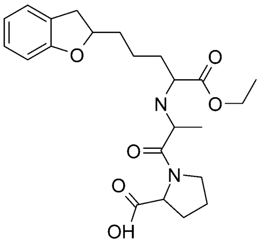
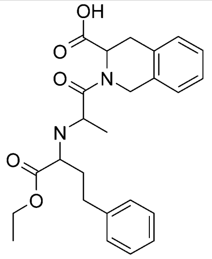
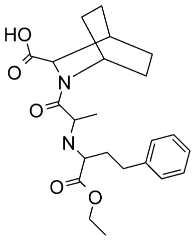
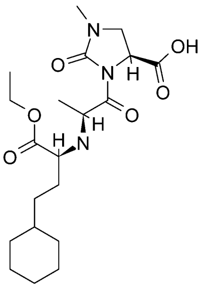
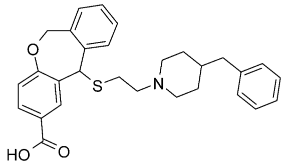
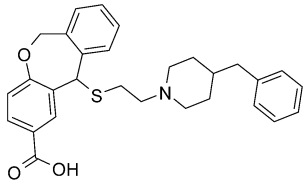
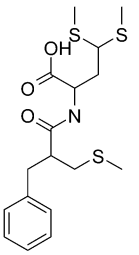
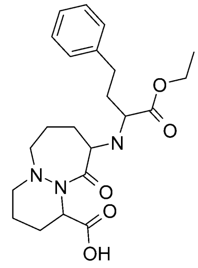
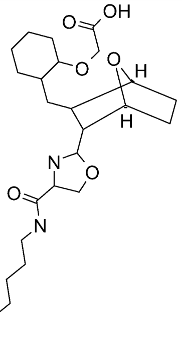
In addition to finding active structures (examined in the preceding calculations), it is one of the superior properties of 3D descriptors over 2D descriptors that they potentially facilitate "scaffold hopping": the finding of structures that possess shape and pharmacophore similarity without being similar with respect to their connectivity tables. This is illustrated for one query from the data set of ACE inhibitors (Table 8) and the data set of thromboxane A2 antagonists (Table 9).

Table 8 shows the query (ACE inhibitor) used to screen the database and the highest ranked structures found. Of the 10 most similar compounds retrieved, all except structures 6, 7, and 10 are classified as being ACE inhibitors in the MDDR database. (One might think that one only needs to identify amide bonds or carboxylic acids to identify ACE inhibitors, but this is not sufficient since there were more than 100 structures in the database containing either of those features without being classified as ACE inhibitors.) To gauge complementarity of our method to established methods, the same query was used to screen the database by seven other methods implemented in MOE:[52] MACCS Keys, 2D-graph based 3-point pharmacophores (GpiDAPH3), typed atom distances (TAD), typed atom triangles (TAT), typed graph triangles (TGT), 3D 3-point pharmacophores (piDAPH3), and 3D 4-point pharmacophores (piDAPH3). The total number of retrieved structures varied between 2 (typed atom distances) and 8 (MACCS keys). Only surface point environments, graph-based 3-point pharmacophores, and MACCS keys retrieved structures that were not retrieved by any other method—five, one, and four compounds, respectively. Five of the active structures found by our method (structures 3−5, 8, and 9 in Table 8) were not found by any of the other seven methods employed.

Table 9 compares the active structures found for a thromboxane A2 antagonist query and different similarity searching methods. Compared are surface point environments, spatial three-point pharmacophores (TAT), and graph-based 3-point pharmacophores (GpiDAPH3, as implemented in MOE). The total number of active structures retrieved is seven for surface point environments, seven for spatial 3-point pharmacophores, and 10 for graph-based 3-point pharmacophores. While graph-based 3-point pharmacophores retrieve the highest number of active structures, all except one of the structures contain the bicyclic ring system also present in the query compound. In contrast, surface point environments retrieve only seven active compounds among the 10 most similar structures, but on the other hand four out of the seven active compounds retrieved do not retain the bicyclic ring system of the query compound. In addition, three different scaffolds are present among this subset of retrieved active compounds without the ring system. Illustrated by two examples with an ACE inhibitor and a TXA2 antagonist as query compounds, the method presented here seems to complement established 2D and 3D methods used currently for similarity searching.

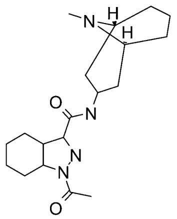
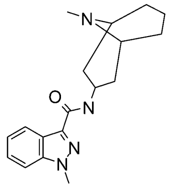
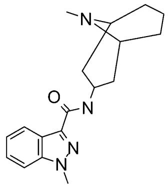
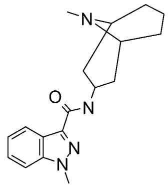
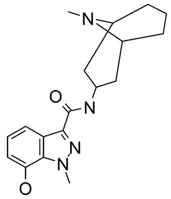
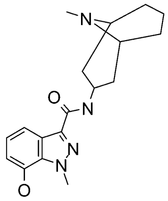
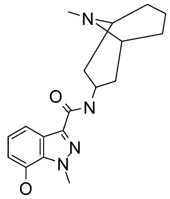
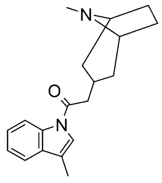
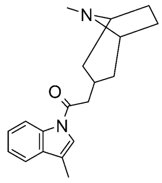
Finally, it is likely that a method captures sensible features for classification (as opposed to randomly finding active compounds) if it performs well on several different data sets. By selecting features that are identified as being important for activity by the algorithm and projecting them back on the molecular surface, it can be verified that they do not constitute incomprehensible sets of features that are only accidentally correlated with activity. (In most data sets, variables such as this exist, which enable classification but are still meaningless.) In addition, the projection of features on the molecular surface may provide insight into ligand features responsible for binding. This is illustrated by projecting features of inhibitors of 3-hydroxy-3-methylglutaryl−coenzyme A reductase, angiotensin-converting enzyme, and features of antagonists of thromboxane A2 back onto the molecular surface. Surface fingerprint descriptors were calculated at point densities of 2.0/Å^2 for all six interaction fields and layers 0−4 for descriptor generation. Information gain feature selection was performed to select those features possessing highest information gain, which were more frequent in the set of active molecules. Those features are shown in Figures 7−9. Figure 7 shows features selected to be characteristic for a 3-hydroxyl-3-methylglutaryl−coenzyme A reductase inhibitor, Figure 8 shows features from an angiotensin-converting enzyme inhibitor, and Figure 9 illustrates the selected features for a thromboxane A2 antagonist.

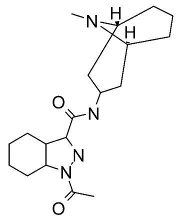
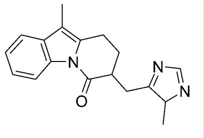
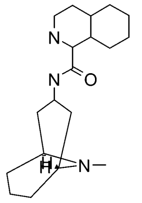
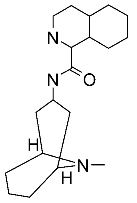
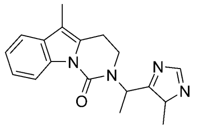
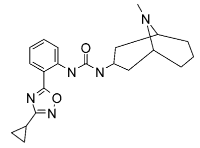
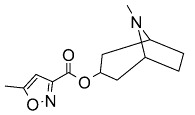
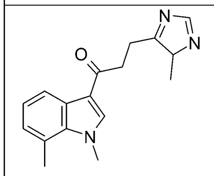
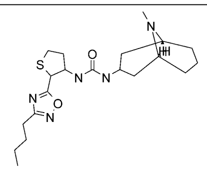
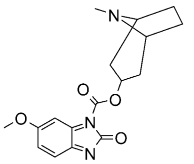
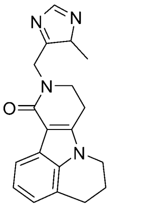
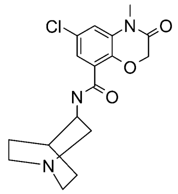
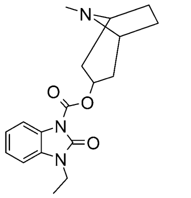
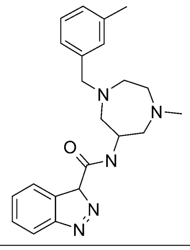
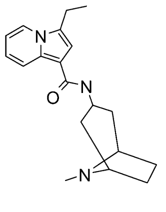
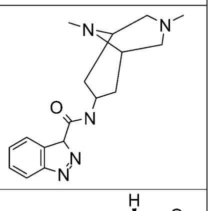
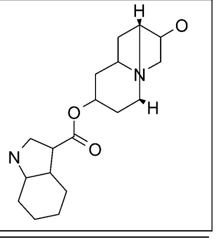
Table 8. Query (ACE Inhibitor) Used to Screen the Database and the Highest Ranked Structures Found[a]

Query			
Ranking position	Structure	Ranking position	Structure
1 (active)		2 (active)	
3 (active)		4 (active)	
5 (active)		6 (inactive)	
7 (inactive)		8 (active)	
9 (active)		10 (inactive)	

[a] Out of these structures, all except 6, 7, and 10 are classified as being ACE inhibitors in the MDDR database. Five of the active structures found (3–5, 8, and 9) were not found by any of the other seven methods employed.

Table 9. Query (TXA2 Antagonist) Used to Screen the Database and Active Structures Found among the Ten Most Similar Compounds[a]

		
Surface Point Environments	Typed Atom Triangles (TAT)	Graph-Based Three-point-pharmacophores (GpiDAPH3)
Identical active structures found (all with bicyclic system / 2D-similar to query)		
		
		
		

		
Surface Point Environments	Typed Atom Triangles (TAT)	Graph-Based Three-point-pharmacophores (GpiDAPH3)
Active structures identified by only one or two of the similarity methods		
		
		
		
		
		
		
		

[a] Compared are surface point environments, graph-based three-point pharmacophores, and spatial three-point pharmacophores. While other methods retrieve in total more active structures, the method presented here retrieves more structures dissimilar to the query in this example.

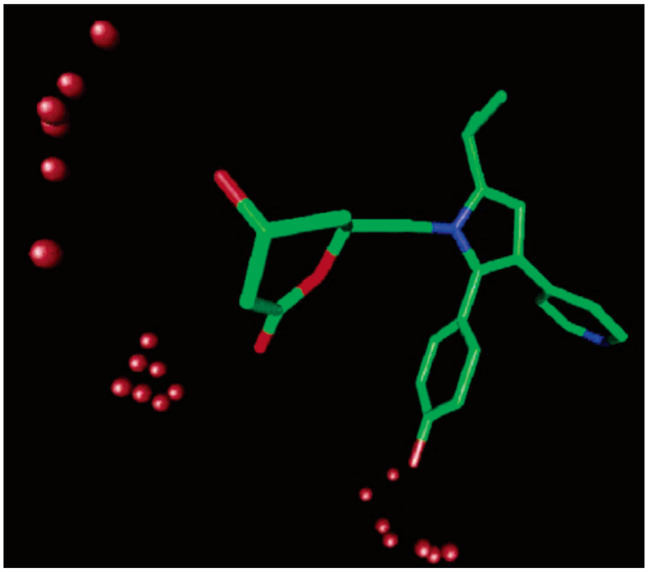


Figure 7. Features identifying the putative pharmacophore of a 3-hydroxyl-3-methylglutaryl−coenzyme A reductase inhibitor. The polar interactions in the upper left corner and the lipophilic interaction of the fluorobenzyl moiety match binding patterns observed in crystal structures of other HMG−CoA inhibitors.

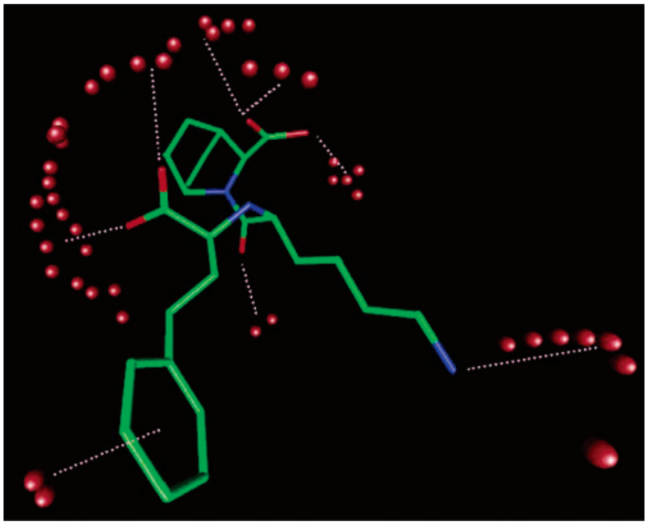


Figure 8. Features identifying the putative pharmacophore of an angiotensin-converting enzyme inhibitor. Both lipophilic interactions (the aromatic ring in the lower left corner) and hydrogen-bonding and charge interactions (in the upper left-hand corner and in the top middle of the figure) are identified by the algorithm, based solely on ligand information.

Selected features of the HMG−CoA inhibitor in Figure 7 are adjacent to oxygen substituents on the left-hand side and the lipophilic ring at the bottom of the figure. Crystal structures of HMG−CoA reductase complexed with statins[53] show a common binding pattern between the carboxylic acid and hydroxyl groups of the HMG moiety and polar side chains of the protein. In addition, a lipophilic cleft perpendicular to the axis of polar interactions is present, which is surrounded by a flexible α-helix that is able to accommodate lipophilic groups of different shapes and sizes. Both features, the oxygen atoms corresponding to the polar interactions of the HMG moiety and the lipophilic fluorobenzene, are correctly identified by the algorithm.

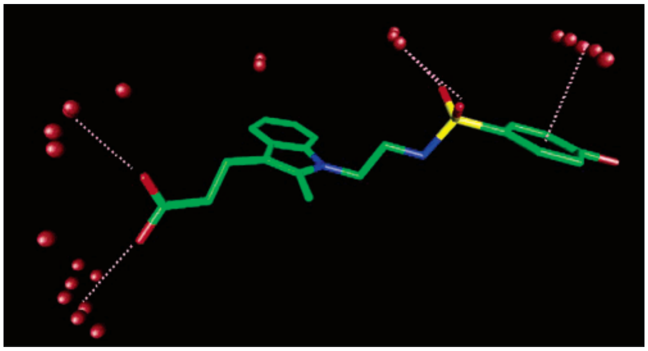


Figure 9. Features identifying the putative pharmacophore of a thromboxane A2 antagonist. Polar interactions of the carboxylic acid group on the left-hand side, hydrogen-bond acceptor potential of the sulfonamide moiety, and lipophilic interaction of the fluorobenzyl ring match binding patterns derived from homology models of the binding site. The bound conformation of the ligand is likely to be bent[56] at an angle of about 90° so that the lipophilic ring points downward.

Selected features of the ACE inhibitor in Figure 8 are assigned to various carbonyl oxygens and lipophilic moieties. The experimentally determined binding site of ACE[54,55] exhibits pairs of hydrogen-bond donors and acceptors as well as lipophilic pockets. The algorithm identifies lipophilic rings and hydrogen-bond-accepting carbonyl groups as well as the carboxylic acid, which was though to interact with a bound Zn^{2+} in the enzyme. Although deemed to be important and successfully used for the design of ACE inhibitors,[55] the recently resolved crystal structure of an angiotensin-converting enzyme/lisinopril complex did not show a zinc binding site[54].

Binding of the TXA2 antagonist in Figure 9 is suggested to be enhanced by interactions of the carboxylic acid on the left-hand side, aromatic interactions and hydrogen-bond acceptor properties in the center of the figure, and a fluoro-substituted benzene ring, shown on the right-hand side of the figure. This binding pattern can be compared to a ligand−target complex derived by homology modeling.[56] An arginine residue of thromboxane A2 is thought to form a charge interaction with a carboxylic acid group of the ligand. A serine residue from the target in this model forms a hydrogen-bond interaction, where a hydroxyl group of the ligand acts as an acceptor. In addition, a large lipophilic pocket is present perpendicular to the arginine-serine axis. All three features, carboxylic acid, hydrogen-bond acceptor (in this case a sulfonamide group), and the fluorobenzene that points in the lipophilic pocket, are identified correctly by the algorithm presented here. This is achieved without having the binding conformation of the ligand available, which is more likely to be bent (the C−C bond between the sulfonamide and indole moiety can be rotated by 180° at both carbon atoms to achieve a bent conformation).

Back-projection of the features deemed to be responsible for binding on different molecular scaffolds should also be able to identify fragments that are similar with respect to their binding properties, despite showing different atom types and/or connectivity (bioisosteres). An example of that capability is shown in Figure 10 for two compounds having antagonistic properties on TXA2. Lipophilic interactions are formed via an aliphatic chain in the case of the first compound and via a halide-

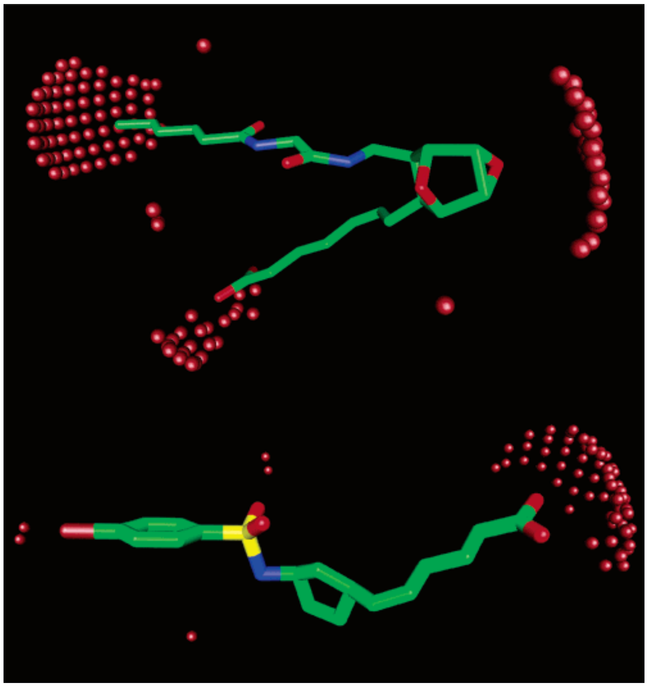


Figure 10. Features identifying the putative pharmacophore of a thromboxane A2 antagonist, back-projected onto two active compounds with different scaffolds. Lipophilic interactions are mediated by an aliphatic chain in the first case and by a halide-substituted benzyl ring in the second case. Hydrogen bonds are formed through a cyclic ether in the first case and a sulfonamide in the second case. Although the structures shown are neither aligned nor binding conformations, similar features are identified in both cases.

substituted benzyl ring in the case of the second compound. Hydrogen bonds are formed via a cyclic ether in the first case and via a sulfonamide in the second case. Charge interactions require a carboxylic acid function in both cases.

Overall, the features selected from the information-gain-based feature selection exhibit similar binding patterns to those observed experimentally or in modeling studies. This is achieved without knowledge of the target nor about the conformation of the ligand in the bound state.

Finally, we would like to comment on the use of local information for descriptor generation. Every descriptor derived from 3D coordinates depends on the particular conformation of the molecule described. There exists a considerable tradeoff: the more focused local descriptors do not include any distance information but are on the other hand invariant to conformational changes. Descriptors that include interdescriptor distance information potentially cover all possible pharmacophore point combinations but are on the other hand very dependent on the particular conformation chosen. It is likely that an optimum descriptor for a particular task lies between those extremes.

To illustrate the dependence of intramolecular distances on conformation, the ACE inhibitor from Figure 4 was subjected to a 10 ps molecular dynamics simulation in vacuo with Sybyl. Standard settings were used in combination with a NTV ensemble at 310K, the Tripos force field, Gasteiger−Huckel charges, and a distance-dependent dielectric function. The distances between the outer carbon of the aromatic ring and the amide oxygen, the nitrogen in the alkyl chain, and the oxygen atoms of the closer and terminal carboxylic acid groups were recorded for the run time of the simulation (illustrated in Figure 11). The time-dependent distribution function of intramolecular separations is given in Figure 12. While the intrafeature distance between close features such as the aromatic ring and the amide oxygen shows a sharp peak, the intrafeature separation between all others shows considerable variation. The distance between the aromatic ring and the distal carboxylic acid moiety varies between 7 and 18 Å with a "forbidden zone" between 12 and 16 Å. This illustrates conformational problems when distances between features are taken into account.

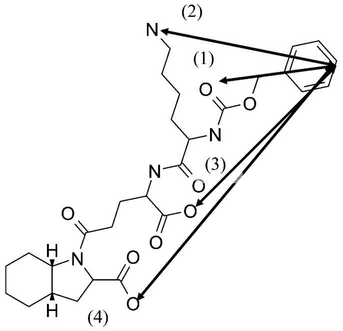


Figure 11. Illustration of the distances measured during the MD simulation of an ACE inhibitor. All distances recorded are taken from the outer atom of the aromatic ring to heteroatoms throughout the rest of the structure.

Conformational analysis of this type will help identify lower energy conformations, particularly where the molecules are predominantly rigid, therefore minimizing conformational flexibility. However in most cases, the receptor-bound conformation is not the in vacuo (or solvated) lowest energy conformation. In a recent study,[57] it was found that 60% of the ligands studied do not bind in a local minimum conformation. Strain energies of at least 9 kcal/mol were found in more than 10% of the bound ligand conformations. Including conformational constraints is often difficult in the absence of experimental or other evidence (e.g., use of the active analogue approach[58]) of the receptor-bound conformation.

The surface point environment descriptor introduced here (coupled with feature selection and a naïve Bayesian classifier) represents part of the molecular surface, spanning about 8 Å in diameter, and seems reasonably tolerant to conformational changes (in small druglike molecules) when used for database searching. Also, it is important to note that the classification of a molecule generally depends on more than one feature (in practice, usually several hundreds). Some features may overlap and therefore they will represent continuous regions considerably greater than 8 Å in diameter. Probability of class membership will thus depend on a number of different features, in effect representing an implicit *AND* of those features. Chosen features may represent continuous or discontinuous regions, allowing flexible representation of important field properties.

4. Conclusions

We present a novel similarity searching algorithm based on surface point environment descriptors in combination with the Tanimoto coefficient and the naïve

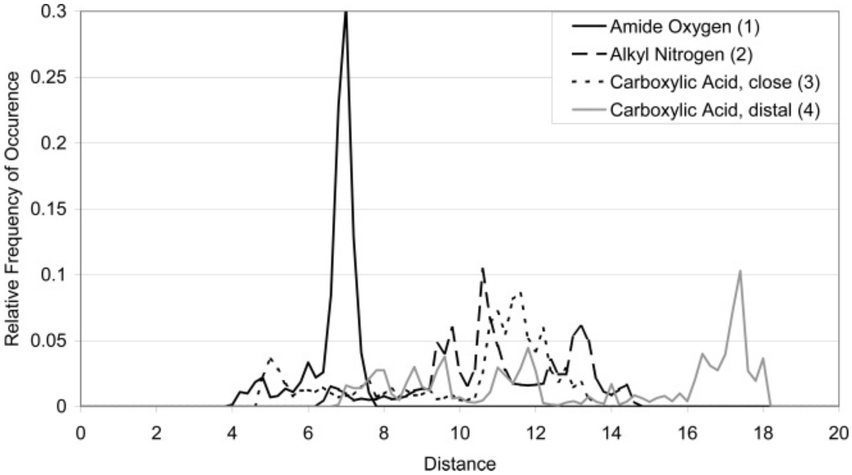


Figure 12. Relative frequencies of distances between the outer carbon of the aromatic ring and the amide oxygen, alkyl nitrogen, and two carboxylic acid groups of the ACE inhibitor shown in Figure 11. Distances are derived from a 10 ps molecular dynamics simulation in Sybyl and illustrate that distances between features depend to a great extent on the particular conformation chosen.

Bayesian classifier. It shows high retrieval rates, the identification of active structures with different scaffolds, and back-projectability of features that can be correlated with experimentally determined binding patterns. Used in combination with Tanimoto coefficients, its performance is comparable to that of commonly used 2D fingerprints. If the Tanimoto coefficient is replaced by a Bayesian classifier, information from multiple structures can be combined.

The descriptor is shown to be tolerant to conformational variations of the ligand structure. On average, two-thirds of randomly generated conformations of sets of 10 structures each from five activity classes are classified as being most similar to one conformation used for querying the whole database. The database in this case contains more than 900 structures in total and 39−131 structures of the same activity class (depending on the particular class chosen). This implies that, in most cases, active structures were not missed where conformations of similar molecules present in the database were not the ones that would give the best (conformational) match to the query.

Active structures retrieved by this approach possess a variety of scaffolds, as illustrated by a database search for an ACE inhibitor and a thromboxane A2 antagonist. Active structures with no apparent 2D similarity to the query and among themselves are identified, showing that the method is capable of “scaffold hopping”. The active compounds retrieved as illustrated by those two cases were also not found by seven other 2D and 3D similarity searching methods. This indicates complementarity of the algorithm presented here to established similarity searching algorithms.

Feature selection is shown to identify important features that, if projected back onto the molecular surface, can be associated with experimentally observed binding patterns. This is achieved without alignment of structures or information about the target structure. Illustrations of feature selection are given for inhibitors of 3-hydroxy-3-methylglutaryl−coenzyme A reductase and angiotensin-converting enzyme as well as antagonists of thromboxane A2. Features contributing to the binding of a thromboxane A2 antagonist are correctly identified, even where the structure is not given in the observed binding conformation. Features responsible for binding are also able to identify bioisosteric fragments of the molecule.

Acknowledgment. We thank Unilever, The Gates Cambridge Trust, and Tripos Inc. for support. Uta Lessel is thanked for providing us with the MDDR data set. A number of participants at the Joint Sheffield Conference on Chemoinformatics 2004 are thanked for input on the method, in particular Stephan Reiling (Aventis). David Patterson at Tripos, Inc., is thanked for numerous discussions and comments. John B. O. Mitchell is thanked for proofreading, and the anonymous reviewers are thanked for valuable input on the manuscript.

References

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(3) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening − an overview. *Drug Discov. Today* **1998**, *3*, 160−178.
(4) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity − A review. *QSAR Comb. Sci.* **2004**, *22*, 1006−1026.
(5) Kubinyi, H. Similarity and dissimilarity: A medicinal chemist’s view. *Perspect. Drug Discov. Des.* **1998**, *9−11*, 225−252.
(6) Kubinyi, H. Chemical similarity and biological activities. *J. Braz. Chem. Soc.* **2002**, *13*, 717−726.
(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(8) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(9) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.
(10) Estrada, E.; Uriarte, E. Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.
(11) 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.
(12) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.
(13) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.
(14) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold-hopping” by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.

(15) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures − perception, auto-correlation descriptor and SAR studies − auto-correlation descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.

(16) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(17) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81−106.

(18) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen* **2002**, *5*, 155−166.

(19) Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547−579.

(20) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.

(21) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.

(22) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(23) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44*, 110−119.

(24) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(25) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Des.* **2000**, *20*, 1−16.

(26) Harper, G. Ph.D. Thesis, Oxford University. The Selection of Compounds for Screening in Pharmaceutical Research, 1999.

(27) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.

(28) Briem, H.; Lessel, U. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discov. Des.* **2000**, *20*, 231−244.

(29) MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.

(30) Concord, Version 4.0.7, Tripos Inc., St. Louis, MO.

(31) Pearlman, R. S. CONCORD: Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Autom. News* **1987**, *2*, 5−7.

(32) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.

(33) Stiefl, N.; Baumann, K. Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure−activity relationship technique. *J. Med. Chem.* **2003**, *46*, 1390−1407.

(34) Pascual-Ahuir, J. L.; Silla, E. GEPOL: An Improved Description of Molecular Surfaces. I. Building the Spherical Surface Set. *J. Comput. Chem.* **1990**, *11*, 1047−1060.

(35) OpenBabel, http://openbabel.sourceforge.net/.

(36) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(37) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(38) Xing, L.; Glen, R. C.; Clark, R. D. Predicting p*K*(a) by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870−879.

(39) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(40) Glen, R. C.; A-Razzak, M. Applications of Rule-Induction in the Derivation of quantitative structure−activity relationships. *J. Comput. Aided Mol. Des.* **1992**, *6*, 349−383.

(41) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.

(42) Domingos, P.; Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* **1997**, *29*, 103−130.

(43) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(44) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(45) Rarey, M.; Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* **1998**, *12*, 471−490.

(46) ISIS, Version 2.1.4, Molecular Design Ltd., San Leandro, USA.

(47) DAYLIGHT, Version 4.62, DAYLIGHT Inc., Mission Viejo, CA.

(48) SYBYL, Version 6.5.3, HQSAR Module, Tripos Inc., St. Louis, MO.

(49) Lessel, U. F.; Briem, H. Flexsim-X: a method for the detection of molecules with similar biological activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246−253.

(50) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(51) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.

(52) MOE (Molecular Operating Environment), Chemical Computing Group Inc., Montreal, Quebec, Canada.

(53) Istvan, E. S. Structural mechanism for statin inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Am. Heart J.* **2002**, *144*, S27−32.

(54) Natesh, R.; Schwager, S. L.; Sturrock, E. D.; Acharya, K. R. Crystal structure of the human angiotensin-converting enzyme−lisinopril complex. *Nature* **2003**, *421*, 551−554.

(55) Cushman, D. W.; Ondetti, M. A. Design of angiotensin converting enzyme inhibitors. *Nat. Med.* **1999**, *5*, 1110−1113.

(56) Yamamoto, Y.; Kamiya, K.; Terao, S. Modeling of human thromboxane A2 receptor and analysis of the receptor−ligand interaction. *J. Med. Chem.* **1993**, *36*, 820−825.

(57) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(58) Marshall, C. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. *The Conformational Parameter in Drug Design: The Active Analogue Approach. Computer Assisted Drug Design*; American Chemical Society: Washington, DC, 1979; pp 205−226.

(59) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

JM049611I