

Flexophore, a New Versatile 3D Pharmacophore Descriptor That Considers Molecular Flexibility

Modest von Korff, Joel Freyss, and Thomas Sander*

Department of Research Informatics, Actelion Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

Received October 2, 2007

A novel pharmacophore descriptor Flexophore is presented, which considers molecular flexibility when comparing descriptor similarities. The descriptor is a complete reduced graph of the underlying molecule. Its nodes are represented by enhanced MM2 atom types, while the edge descriptions encode the molecular flexibility by means of a histogram of node distances in a diverse conformer distribution. For comparing two descriptor nodes, a statistical function derived from the Cambridge Crystallographic Database is implemented. To assess the capability of the descriptor to describe the bioactivity space, 350 test data sets with 1000 molecules each are compiled. The data sets were spiked with molecules active on one of 18 different targets. In 175 of the 350 data sets, all molecules chemically similar to the query molecules were removed. Virtual screening on these data sets showed that the Flexophore descriptor detects active molecules despite chemical dissimilarity, whereas the results for the screening of the complete data sets show enrichments comparable to chemical fingerprint descriptors. The diversity analysis of the enriched compounds demonstrates that the Flexophore descriptor describes the chemical space orthogonal to chemical fingerprint descriptors.

INTRODUCTION

Comparing and selecting molecules from chemical databases is one of the major issues in chemo-informatics. Virtual screening of databases for molecules with potential for activity on a certain target protein is a common task in drug discovery.^{1–4} If the target protein structure is unknown, a geometrical query has to be deduced from known ligands of the protein (ligand-based virtual screening). The principal of virtual screening is simple; it relies on the comparison of database molecules with one or many query structures. The representation of the database and query molecules and the comparison algorithm strongly influence the success of the search. Thus, the search for appropriate molecule representations⁵ and appropriate comparison metrics has involved enormous efforts in chemo-informatics. The definitions for molecular similarity are manifold⁶ and depend on the goal. For ligand-based virtual screening, a query descriptor is derived from one or more active molecules or from the target under consideration. The descriptors similarly derived from the database molecules are compared to the query. The database is then sorted according to the similarity values. Molecules that are similar to the query are more likely to be active than molecules with a lower similarity. The underlying assumption is that similarity in the descriptor space indicates similarity in the “bioactivity space”.^{7,8} Typically, however, one aims to discover substances which are substantially different from known actives to avoid either patented scaffolds or unwanted side effects. The ultimate goal is a new chemical entity active on the target protein. Active means acting as an agonist, an antagonist, or an allosteric modulator of the protein. The term “scaffold hopping” became common for the detection of new active scaffolds

during the past decade.⁹ In ligand-based screening, “scaffold hopping” implies being able to describe a chemical structure from the target protein point of view.¹⁰ Imaginary ligand features called “pharmacophore points” can do this.^{11,12} The pharmacophore concept has become a well-accepted technique to describe ligand features responsible for the target–ligand interactions. These interactions are often classified in basic interaction types as hydrogen-bonding, charge-transfer, electrostatic, aromatic, and hydrophobic interactions. Analogously, pharmacophore types are often classified as hydrogen-bond donors or acceptors, charge centers, hydrophobic centers, and aromatic centers. One of the first approaches in computational chemistry to describe pharmacophore points in ligands was the substructure-based definition with SMARTS¹³ in the ALADDIN¹⁴ program. A combination of substructures with Boolean logic to define pharmacophore points was realized by Greene et al.¹⁵ A new description of ligand pharmacophoric features was introduced with the GRIND descriptor¹⁶ and the VolSurf descriptor.¹⁷ They applied the molecular interaction field (MIF) from Goodford¹⁸ to derive the pharmacophoric features. A further development of a force-field-based pharmacophore description is the 3D descriptors derived from molecular field extrema.¹⁹ The Cambridge Crystallographic Database²⁰ contains information about thousands of ligand–protein interactions. The experimental data are summarized by interaction tables in the Isostar software.²¹ The Superstar²² algorithm, derived from Isostar, was used to define ligand-based pharmacophores for QSAR.²³

After the definition of the ligand interaction points, the arrangement of the pharmacophore points in the ligand will be considered. The description of a molecule as a static arrangement of pharmacophoric features is not sufficient to represent its bioactivity. Since the introduction of the induced fit theory,²⁴ molecular flexibility has been recognized as an

* Corresponding author phone: +41 61 565 65 23, fax: +41 61 565 65 00, e-mail: Thomas.sander@actelion.com.

important issue in drug optimization. The recognition process for ligand–protein interactions is of high complexity and includes conformational changes of the ligand and protein.²⁵ The conclusions of the enhanced induced fit theory imply a new definition of molecular similarity. Similarity of bioactive ligands can no longer be defined as a comparison of pharmacophores arranged as static molecular conformations. The definition of similarity must now include the dynamic conformational changes of molecules during the recognition process. A molecule is similar to a ligand if it is able to mimic its role in the recognition process. The challenge is to develop a molecular descriptor reflecting this dynamic behavior. The necessity of flexible 3D molecular descriptions has been discussed since the early 1990s^{26,27} and resulted in manifold descriptors.¹ Descriptors such as ALADDIN,¹⁴ 3DSEARCH,²⁸ CATALYST,¹⁵ and DISCO²⁹ represent the three-dimensional structure of molecules as pharmacophore points. They consider molecular flexibility by definition of flexible bond lengths between pharmacophore points. The Quasar³⁰ concept considers the induced fit by generating “envelopes” which mimic the binding cavity of the protein.

Another group of 3D pharmacophore descriptors represents the molecules as vectors of a certain length. Examples are the 3D pharmacophore descriptors of Mason et al.,³¹ the MIF descriptors from Cruciani,³² the MIP descriptors from Baumann and Stiefl,³³ the MOLPRINT descriptors,³⁴ the 4D descriptors of Hopfinger et al.,³⁵ and a recent reimplementation of Mason's 3D pharmacophores.³⁶ The advantage of the vector-based descriptors is their high performance for similarity calculations. A disadvantage is the loss of information by reducing the molecular information into a descriptor vector. Three-dimensional graph-based pharmacophore models enable a more realistic molecule description, but the comparison of three-dimensional graphs is much slower than the comparison of vector-based descriptors. While three-dimensional pharmacophore models are still a limited approximation of reality, the most successful virtual screening would presumably result from the simulation of the complete molecular recognition process, from the flexible movement of the molecule through the protein cavity to its final location where it docks. The algorithm has to consider the flexibility of the ligand as well as the flexibility of the protein. Moreover, it must consider the solvent and solvated salt ions. This approach would be so time-consuming that its application for virtual screening on a typical molecule library with several hundred thousand molecules is far beyond the computational resources available for computational chemistry at this time. It also requires reliable three-dimensional target protein structures. For the largest group of interesting drug targets, the G-protein coupled receptors (GPCR), however, almost no crystallographic structure data are available. The discrepancy between this “perfect model” and the model implemented should be a constant consideration. A descriptor is always a tradeoff between the model accuracy and the desired performance. After reviewing the literature covering both disappointments^{37–40} and successes^{41–43} in 3D ligand-based virtual screening, the challenge became to develop a graph-based pharmacophore descriptor (Flexophore). The aim is automatic descriptor generation, without the need of interactively defining molecule geometries or refining parameters. It was considered that a virtual screening tool, intended for routine use, should be automatic and

Table 1. Data Sets Used^a

target protein	# ligands	# test data sets	source
5-HT1A	185	10	GPCR-lib
adenosine 1 (A1)	94	10	GPCR-lib
activated factor X (FXa)	343	10	IDDB
angiotensin converting enzyme (ACE)	155	10	IDDB
angiotensin II (At2)	705	10	IDDB
benzodiazepine (Bzd)	131	10	IDDB
Ca channel (Ca)	630	10	IDDB
cannabinoid CB1	198	10	GPCR-lib
chemokine receptor 5 (CCR5)	144	10	GPCR-lib
cyclin-dependent kinase 2 (CDK2)	148	10	IDDB
cyclooxygenase 2 (Cox2)	128	10	Stahl
dopamine 3 (D3)	132	10	GPCR-lib
endothelin A (ETA)	76	10	GPCR-lib
gonadotropin releasing hormone (GNRH)	93	10	GPCR-lib
melanin concentrating hormone (MCH)	149	10	GPCR-lib
muscarinic acetylcholine receptor 1 (M1)	163	10	Burden
phosphodiesterase type 5	156	10	IDDB
thrombin (Thr)	67	5	Stahl
sum	3697	175	

^a # ligands: number of ligands, in total 3697. # test data sets: number of data sets generated from the ligands, in total 175 each for D_{filtered} and D_{all} .

unbiased and give a reproducible generation of queries. Constructing one query from several active molecules is prone to fail if these molecules bind in different locations of the target pocket. The resulting query would average unrelated information and contain a high level of noise. A recent publication showed that mixing the information of several molecules into one descriptor is of limited use when aiming for good results in virtual screening.⁴⁴

The descriptor described here represents the molecule by a complete graph. This is a simple graph where each pair of vertices is connected by an edge. A vertex is derived from one atom or from a substructure fragment containing several atoms. The vertices are labeled with enhanced MM2 atoms types⁴⁵ of the corresponding atoms. The edges are histograms of the vertex distances resulting from diverse molecule conformations. The similarity function introduces the target point of view into the descriptor comparison. Interaction statistics for the enhanced MM2 atom types were derived from the crystallographic data in the PDB.⁴⁶ A similarity matrix of these atom types, concerning their interaction behavior, was created. A simple reference to the generated table reveals the similarity between any two vertices. Edge similarities are calculated from the fraction of overlap between two distance histograms. To represent the maximum common substructure of two descriptor graphs, an overall similarity score is calculated for the best match of all vertices and edges.

The 3D descriptor was developed to explore the chemical space beyond topological similarity. To assess its performance for each of 18 target proteins, a set of active molecules from different sources were collected (Table 1). Depending on its size, every set was split into five or 10 subsets, each containing between 10 and 50 molecules active on one target. One molecule from every subset was randomly selected to

serve as the query structure. Inactive molecules were added to each subset to yield 1000 molecules each. The enrichment rate for the query molecule served as a benchmarking function. The chemical fingerprint from ChemAxon⁴⁷ (ChemAxonFp) and Actelion's in-house-developed chemical fingerprint descriptor ActelionFp⁴⁸ were used to assess chemical similarity. Actelion's topological pharmacophore histogram descriptor (Act2DHist) was used to compare 2D- and 3D-pharmacophore descriptors.

METHODS

Data. The enrichment rates for 18 different targets analyzed were used to assess the performance of the descriptor. For each target t , a data set A_t ($t = 1-18$) with between 67 and 705 active molecules was used (Table 1). The data were compiled from different sources: GPCR-lib is Actelion's in-house GPCR data set containing around 4500 molecules; IDDB stands for Thomson investigational drugs database;⁴⁹ the data set labeled with Burden was published in Orlek et al.⁵⁰ and data sets labeled with Stahl were published in the Journal of Medicinal Chemistry.⁵¹ The data sets Burden and Stahl were acquired over the cheminformatics.org Web site from Bender.⁵² From any of these molecule sets A_t , five or 10 subsets, depending on the set's size, were built by applying the following procedure either five or 10 times: A query molecule Q was randomly chosen and removed from A_t . Then, all molecules chemically similar to the query were also removed from the set. As chemical similarity criteria, the Tanimoto coefficients of the ActelionFp as well as of the ChemAxonFp were used. Both descriptors are binary chemical fingerprints. The ActelionFp is a fragment dictionary-based molecular descriptor.⁴⁸ The ChemAxonFp is derived from walks on the molecular graph. The results of the walks are hashed into the descriptor vector. A molecule was considered similar to the query if at least one of the Tanimoto coefficients was above 0.4. This rigorous threshold was chosen to ensure that the final set would not contain any molecule chemically similar to the query. If a set still contained more than 50 molecules, then molecules were randomly removed until 50 remained. If less than 10 molecules remained, then this set was eliminated and the procedure repeated with another query molecule. Afterward, all data sets were complemented with bioactive molecules from either a diverse subset of the IDDB database⁴⁹ or with molecules from Actelion's in-house GPCR database to reach 1000 molecules in total. Only supposedly inactive molecules from these databases, that is, molecules not labeled active on the target under consideration, were selected. When this method was used, 175 test sets of 1000 molecules denominated $B_{t,q}$, with t being the target index and q being a query index, were selected. Every set contained between 10 and 50 molecules which are active on a specific target and all chemically dissimilar to an associated ligand Q_q of the same target. Data sets selected by this algorithm are henceforth referred to as "filtered" or D_{filtered} .

Another 175 test sets $C_{t,q}$ were created analogously with the exception of the step removing compounds chemically similar to the query. Every one of the $C_{t,q}$ sets contained 50 molecules active on target t diluted in 950 supposedly inactive molecules. These data sets are, in the following, referred to as "unfiltered" or D_{all} . From the original data set,

all records were removed that contained structures without C atoms; structures containing heavy metal atoms U, Pd, Pb, Pt, Au, Ag, Hg, Fe, and Cu; molecules outside a molecular weight range of 140–850; and records containing the string "radio". Around 45 000 structures remained after this procedure.

DESCRIPTORS

ChemAxonFp. For comparison reasons, a chemical fingerprint descriptor from ChemAxon⁴⁷ was used. This descriptor encodes the topological information between atoms of a molecular graph into a binary vector. The descriptor length, and therefore the resolution, is user-defined. For creating the fingerprint, any possible walk up to a predefined length is performed on the molecular graph. The result of a walk is a collection of types of atoms and bonds visited during the walk. The atom types in the ChemAxonFp descriptor are defined by the atomic numbers, while the bonds are defined through their bond orders. For any path, a hash value is calculated and binned into a binary vector with a logical OR operation. Hence, each possible walk is associated to one bit, which may represent other walks at the same time. The result is a hashed chemical fingerprint. In this exercise, a fingerprint length of 512 bits was used with a maximum walk length of four bonds. In-house experiments confirmed that calculated similarities based on this descriptor typically correlate with the perception of medicinal chemists.

Actelion Topological Pharmacophore Histogram (Act2DHist). The Act2DHist descriptor is an in-house development quite similar to a descriptor developed together with ChemAxon.⁵³ The Act2DHist descriptor is closely related to the atom pair descriptor⁵⁴ and to the binding property pair's descriptor.⁵⁵ For the detection of the pharmacophore points, a combination of the defined substructures and logic expressions was used. The substructures for the pharmacophore point definitions were chosen according to Greene et al.¹⁵ According to preliminary experiments, it was decided to use the following pharmacophore types: hydrogen bond donor (d), hydrogen bond acceptor (a), hydrophobic (h), positive charged (+), negative charged (−), and aromatic (r). The pharmacophore points are put into a relationship by the histograms of the topological distance counts. To generate the histogram, the topological distance between each pair of atoms belonging to a certain pharmacophore type is counted. The maximum topological distance was set to 12 bond lengths. The counts are added to the histogram for the pharmacophore point pair combination. This is performed for each two-point pharmacophore point combination, and all resulting histograms from one molecule are written into a descriptor vector.

Flexophore. This descriptor is an adaptation of the assumption that the recognition process between the ligand and the protein can be explained with the induced fit theory. The descriptor consists of a representation of pharmacophore interaction points (PIPs) and their conformational flexibility. For the creation of the descriptor, each atom in a molecule is labeled with its enhanced MM2 atom type (Table 2), which is an in-house extension of the MM2 atom types. These were developed earlier as part of an in-house force-field-minimization algorithm. A more detailed description of the MM2

Table 2. Enhanced MM2 Atom Types Used for the Interactions (Additional Types Are in Bold)

C alkane	C alkene	C carbonyl	C alkyne
C cyclopropane	C cyclopropene	C guanidine	
N amine	N ammonium	N enamine	N amide
N imine	N nitrile	N guanidine	N sulfonamide
N aromatic (N attached to an aromatic ring)			
O alcohol	O carboxyl	O carbonyl	O ether
O enol	O furan		
S thiophene	S thiol	S thioether	S sulfone
P	Cl	Br	I

atom types used is given in the Pharmacophore Point Interaction Statistics section. For performance reasons, ring systems with up to seven members are grouped and then represented by one PIP only. A scheme for the descriptor generation is given in Figure 1. Therefore, the center of gravity of all group members is calculated and taken as the position of the vertex (Figure 1, II). The vertex is labeled with all atom types existing within the group, which is related to the superatom types introduced by Zheng et al.⁵⁶ Likewise, these frequent fragments are represented by one PIP only: sulfonamide, amide, carboxylic acid and ester, guanidine, and Schiff bases. Aliphatic linker chains are not considered, while aliphatic end chains are also summarized into a PIP. End chains are methyl, ethyl, propyl, and butyl. The resulting n_{vert} vertices are stored as a complete graph, which means that every possible combination of any two vertices is considered an edge (Figure 1, III). Therefore, the number of edges is $n_{\text{ed}} = (n_{\text{vert}}^2 - n_{\text{vert}})/2$. The n_{ed} edges are defined as distance histograms, which have a length of 20 Å with a resolution of 0.5 Å, and thus consist of 40 bins (Figure 1, IV). In order to populate the histograms with frequency values, a diverse and representative set of n_{conf} conformers, using the Actelion Conformation Sampler, was generated as described below. For each conformation, the coordinates of all vertices are determined. For vertices, which represent multiple atoms, the centers of gravity are calculated. For every pair of vertices, the Euclidean distances in all conformers are calculated and the frequency values of the corresponding histogram bins increased accordingly. This results in a total of $n_{\text{ed}} n_{\text{conf}}$ distance values. Since the protonation state of the molecules can be drawn from the MM2 definitions and from the description of the conformation sampler, they are not needed to calculate the Flexophore descriptor. This eases the descriptor generation, as no assumptions have to be made concerning the protonation state of the ligand at the target binding site.

Flexophore Similarity Calculation. Henceforth, a description of the comparison between two Flexophore descriptors D_A and D_B will be used, D_A with n_A pharmacophore points and D_B with n_B pharmacophore points, with $n_A \geq n_B$. The Flexophore descriptor is a complete graph where each pair of pharmacophore points in the descriptor is connected with a distance histogram. An edge in the Flexophore descriptor is defined by its atom types, and an edge between two pharmacophore points is defined by a distance histogram. The comparison of two Flexophore descriptors D_A and D_B is related to the maximum common subgraph-isomorphism problem, which often occurs in cheminformatics.^{57,58} Two complete graphs are defined to

be isomorphic if there is a one-to-one correspondence between their vertices and a one-to-one correspondence between their edges. In the Flexophore descriptor, the vertices are represented by the pharmacophore points and the edges are the distance histograms. The similarity between two pharmacophore points is calculated from interaction statistics derived from entries in the Protein Data Bank (PDB). A detailed description of the similarity calculation for the pharmacophore points follows. The algorithm for locating an isomorphic subgraph with n_{iso} vertices of two descriptors considers two pharmacophore points and matches if their similarity value s_{pp} lies above a predefined threshold t_{pp} . The similarity of the corresponding distance histograms is the squared fraction of overlap s_{hist}^2 between them. Analogously to the vertex similarity, a threshold t_{hist} is defined for the distance histograms. Valid results from the comparison of two Flexophore descriptors are all isomorphic subgraphs whose pharmacophore points and distance histograms are fulfilling the threshold criteria t_{pp} and t_{hist} . From the quantity of isomorphic subgraphs, the “best matching subgraph” has to be found. Score calculations for the isomorphic subgraphs are outlined in the following. A histogram score sc_{hist} for a histogram comparison is the reciprocal of the maximum index; a common bit is set in two corresponding histograms. The overall histogram score $sc_{\text{hist,av}}$ is the sum of all sc_{hist} values divided by the number of histograms. Overall, pharmacophore point similarity $sc_{\text{pp,sum}}$ is the sum of all s_{pp} values in an isomorphic subgraph. From $sc_{\text{hist,av}}$ and $sc_{\text{pp,sum}}$, an isomorphic subgraph score sc_{iso} is calculated: $sc_{\text{iso}} = sc_{\text{pp,sum}} / (sc_{\text{hist,av}}^2 n_{\text{iso}}^2)$. To consider the pharmacophore points not included in the isomorphic graph, a score $sc_{\text{sub,max}}$ is calculated. The overlap score is calculated as $sf_{\text{match,A}} = [(n_A - n_{\text{iso}})/n_A \times 10]^2 + 1$ and $sf_{\text{match,B}} = [(n_B - n_{\text{iso}})/n_B]^2 \times 10]^2 + 1$. The dissimilarity score for an isomorphic subgraph, given by $sc_{\text{iso,A,B}} = sc_{\text{iso}} sf_{\text{match,A}} sf_{\text{match,B}} sc_{\text{iso,A,B}}$, is calculated for all isomorphic subgraphs from two descriptors D_A and D_B . The highest score is taken, scaled between 0 and 1, and converted into a similarity score. For the sake of consistency, a descriptor comparison results in similarity values between 0 and 1 for any of the four descriptor types used in this publication.

Conformation Generator. The 3D-pharmacophore relies on a broad diversity of reasonable conformations. Rule-based algorithms, however, tend to miss areas of the conformation space because they construct conformers in a stepwise fashion while considering allowed bond lengths, bond angles, and torsions and because of the incorporation of ring systems and fragments with predefined conformations. Therefore, the generation of conformers presented here is based on a stochastic self-organization algorithm, originally suggested by Agrafiotis et al.⁵⁹

The algorithm starts by positioning all of the molecule's atoms at random coordinates within a reasonably sized cube. Then, a set of constraints is calculated which serves as guidance for the self-organization process. The majority of these fall into the class of distance constraints that define for any pair of atoms either an exact distance, a set of preferred distances, or an allowed distance range. Exact distance constraints are created for any pair of atoms, of which the shortest connection path consists of one, two, or, in some cases, three bonds. In the case of one bond, the distance is determined by averaging lengths of similar bonds

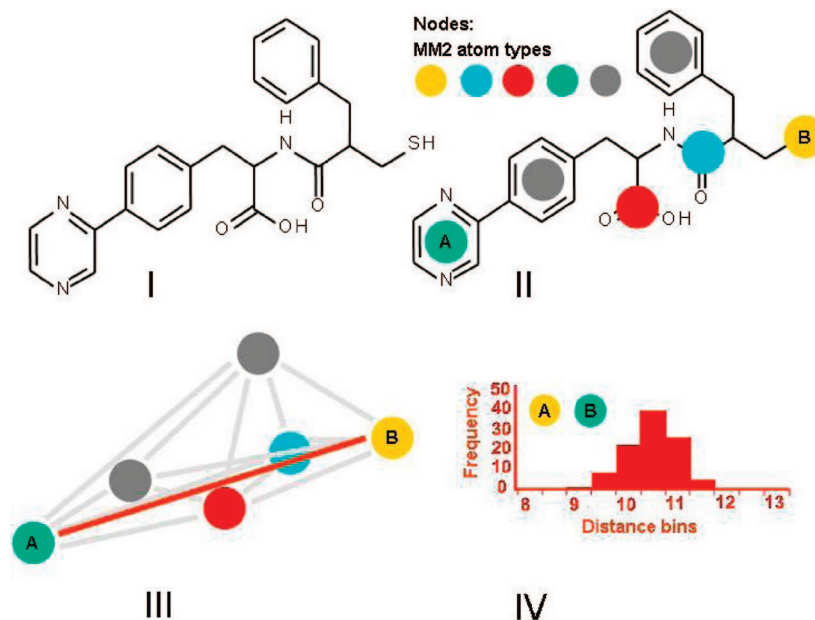


Figure 1. Scheme for descriptor generation. I: Starting molecule. II: Colored circles represent MM2 atom types which encode molecular substructures. III: All edges for the complete graphs added, with one edge highlighted in red for IV. IV: The information connected to the edge is the histogram of all distances between nodes A and B.

available in the Cambridge Structural Database. In the case of two bonds, an assessment of the angle at the bridging atom from its type, hybridization, and, if applicable, strains from small ring membership is performed. If there is a connecting chain of three bonds, then an exact distance can only be calculated when the central bond is not freely rotatable. Otherwise, a calculation of a set of preferred distances from bond lengths, bond angles, and a set of preferred torsions, also extracted from the CSD, is used. In addition to the distance constraints, plane constraints are created by grouping atoms that need to end up in the same plane, for example, members of an aromatic ring and their direct neighbors. Stereo constraints encode the configuration of stereocenters or double bonds. After the definition of all constraints, the self-organization runs by repeatedly picking a constraint randomly. If the constraint is violated, atoms are moved in order to reduce the extent of the violation. In the case of distance constraints, atoms are moved toward or away from each other. Whereas, in the case of plane constraints, atoms are moved toward that plane closest to all atoms being involved. The extent of the movement is guided by an adaptation factor that is reduced during the entire self-organization process. A breakout mechanism avoids trapping atoms in high-strain situations.

Pharmacophore Point Interaction Statistics. To calculate the similarity between two pharmacophore points, their probability of binding to the same functional group at a protein is determined, as with the Isostar²¹ program. A total of 12 000 high-resolution protein–ligand complexes were extracted from the PDB database. Only nonpeptide ligands that counted between 10 and 50 atoms and buried more than 50% of its volume inside the cavity were considered. The interaction statistics were made to create a dissimilarity matrix that shows how different any atom type is from any other in terms of interactions with the protein. Two atoms are considered similar if they bind to the protein in a similar way, that is, if they show similar optimal distance and a similar strength of interaction with the same atoms of the

protein. Around 130 atom types were considered, on the basis of the enhanced MM2 atom type (Table 2), its valence, its aromaticity, and whether it belongs to a ring or not.

Given a protein atom i and a ligand atom j , the number of occurrences $N_{ij}(r)$ were extracted from those complexes. After distance normalization ($g_{ij}(r)$) and after comparing to the reference normalized distribution ($g(r)$), as described by Gohlke et al.⁶⁰

$$g_{ij}(r) = \frac{N_{ij}(r)/4\pi r^2}{\sum_r N_{ij}(r)/4\pi r^2} \quad (1)$$

we obtained a pair potential ΔW_{ij} between each pair of atoms:

$$g(r) = \frac{\sum_{ij} g_{ij}(r)}{i \times j} \quad (2)$$

$$\Delta W_{ij}(r) = -\ln[g_{ij}(r)/g(r)] \quad (3)$$

From the potential function, the optimal distance R_{ij} and the strength S_{ij} of each protein–ligand interaction was established. The dissimilarity d_{ij1j2} between two ligands atoms $j1$ and $j2$ to a protein atom i was defined as the squared Euclidian distance between the two minima of the potential functions (Figure 2).

$$d_{ij1j2} = \left[(R_{ij1} - R_{ij2})^2 + \left(\frac{S_{ij1} - S_{ij2}}{2} \right)^2 \right] \quad (4)$$

To give more weight to the most frequently occurring atoms and less weight to rarer atoms, the overall dissimilarity matrix D_{j1j2} is equal to the weighted average across all protein atoms, where the weights are equal to the number of occurrences in the PDB database.

$$D_{j1j2} = \frac{\sum_i (N_{ij1} + N_{ij2}) \times d_{ij1j2}}{\sum_i (N_{ij1} + N_{ij2})} \quad (5)$$

An example of a dissimilarity matrix is given in Table 3. Summarizing, any ligand atom type with a binding prob-

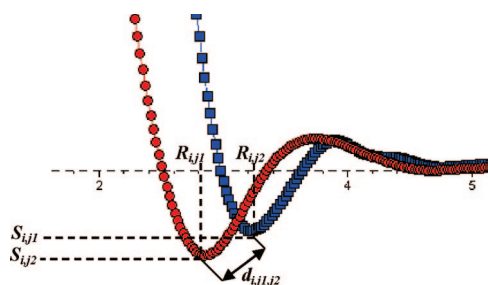


Figure 2. Potential function for protein–ligand interaction. The potential between O-alcohol and C-carbonyl is displayed in red, and the potential between O-alcohol and O-alcohol is displayed in blue.

Table 3. Example Matrix for Pharmacophore Interaction Statistics^a

	1.	2.	3.	4.	5.	6.
1. C Alkane (primary)	1.00	0.72	0.83	0.36	0.37	0.30
2. C Alkane (secondary)	0.72	1.00	0.51	0.38	0.40	0.38
3. C Carbonyl	0.83	0.51	1.00	0.41	0.45	0.45
4. O Carbonyl	0.36	0.38	0.41	1.00	0.98	0.88
5. O Ether	0.37	0.40	0.45	0.98	1.00	0.85
6. O Alcohol	0.30	0.38	0.45	0.88	0.85	1.00

^a The column headers are only indexed since the matrix is symmetric. A value close to one (green) indicates a high probability that two ligand atoms, described by their atom types, can be found in an identical environment at the target protein.

ability profile toward all protein atom types was associated. Two ligand atom types can be compared via their probabilities to bind to the same protein atom type. The result of the atom-type comparison is a bioisoteric score which is used to calculate the Flexophore descriptor similarity. A pharmacophore point in the Flexophore descriptor can consist of one or several atom types. The comparison of two pharmacophore points, each containing one atom type, is simply the reference value from the dissimilarity table. For pharmacophore points containing more than one atom type, the similarity score is taken as the maximum similarity from a total comparison of all atom types. A further limitation in case of present heteroatoms is that the similarity of carbon-based atom types is not considered. Preliminary experiments showed that this method is superior to other approaches.

Enrichment. A relative enrichment factor e_{rel} at 1.0% of the total data set ($\text{frac} = 0.01$) was calculated in order to compare enrichment values of data sets with different numbers of active molecules. Often, higher fractions are applied for calculating enrichment values; however, the authors consider values from 0.001 to 0.01 to reflect the reality of high-throughput screening, where rates of true positives are frequently below 0.001. The relative enrichment e_{rel} is the achieved enrichment e_{abs} normalized by the maximum possible enrichment $e_{\text{abs,max}}$.

$$e_{\text{abs}} = n_{\text{hits, fraction}} / (n_{\text{hits, total}} \times \text{frac}) \quad (6)$$

$$n_{\text{max-poss-hits}} = \text{minimum}(n_{\text{mols-in-frac}}, n_{\text{hits-total}}) \quad (7)$$

$$e_{\text{abs,max}} = n_{\text{max-poss-hits}} / (n_{\text{hits, total}} \times \text{frac}) \quad (8)$$

$$e_{\text{rel}} = 100e_{\text{abs}} / e_{\text{abs,max}} \quad (9)$$

In order to assess the similarity S_D of two descriptors concerning their compound similarity perception, the fol-

lowing procedure was applied. First, a data set $C_{t,g}$ was sorted according to the similarity score to its query molecule calculated by descriptor 1. The first fraction (0.01) of molecules was flagged as set 1. This procedure was repeated for descriptor 2, yielding a selected set 2. The number of molecules found in set 1 which were not a member of set 2 was then divided by the total number of molecules in the fraction.

Selection Overlap for Hits. The calculation of the diversity for the enriched hits was derived from the previous one. Since the number of hits can be between 0 and the number of active molecules in the test data set, an additional rule was added: The diversity is set to zero when both descriptors failed. The number of different molecules is divided by the number of hits for the descriptor with the highest enrichment rate. Again, a fraction of 0.01 was applied.

Computational Details. With the exception of the implementation of the ChemAxonFp descriptor, all descriptors and their similarity metrics were implemented by the authors. For this implementation, Java 1.5 was used. The processing for the 350 test data sets took approximately 14 min on a workstation with an Intel Dual Core 2.13 GHz processor, equivalent to 400 similarity calculations per second. While the comparison of Flexophore descriptors is a relatively swift process, its generation is rather time-consuming due to the necessity of generating a diverse set of conformers. This takes around 10 s per compound. For this reason, the descriptor generation was distributed to multiple computers on Actelion's PC GRID to harness unused processor capacity. The similarity calculation can also be distributed onto this GRID, which is practiced for the purpose of searching commercial libraries of screening compounds.

RESULTS

The results of the virtual screening experiments are summarized in Figure 3. The distribution of relative enrichment rates in percentages are shown as histograms for each descriptor. The x axis represents the bins for the enrichment. A result was only taken into the histogram if the enrichment was better than the value reachable by random screening. Below the x axis, the total number of data sets with an enrichment rate above random enrichment is given. The y axis shows the frequency of hits into a bin. Figure 3A and C show the results for the filtered data sets screened with the Flexophore descriptor and the Act2DHist descriptor, respectively. There was almost no enrichment for the two fingerprint descriptors due to the structure of the data set. The Flexophore descriptor enriched bioactive molecules in 57 out of 175 data sets. The performance of Act2DHist was very similar, with enrichment in 58 data sets. For both descriptors, the majority of enrichment rates are localized in the first quartile of the histograms.

The results for the unfiltered data sets are fairly comparable for all four descriptors (Figure 3B,D,E,F). ActelionFp shows the highest enrichment for the unfiltered data sets (147 out of 175 data sets), closely followed by the Flexophore descriptor (146 data sets) and the ChemAxonFp descriptor. The Act2DHist descriptor is a little less efficient, with enrichment in 135 out of 175 data sets. Patterns in the area distribution show no significant differences between descriptors. For all

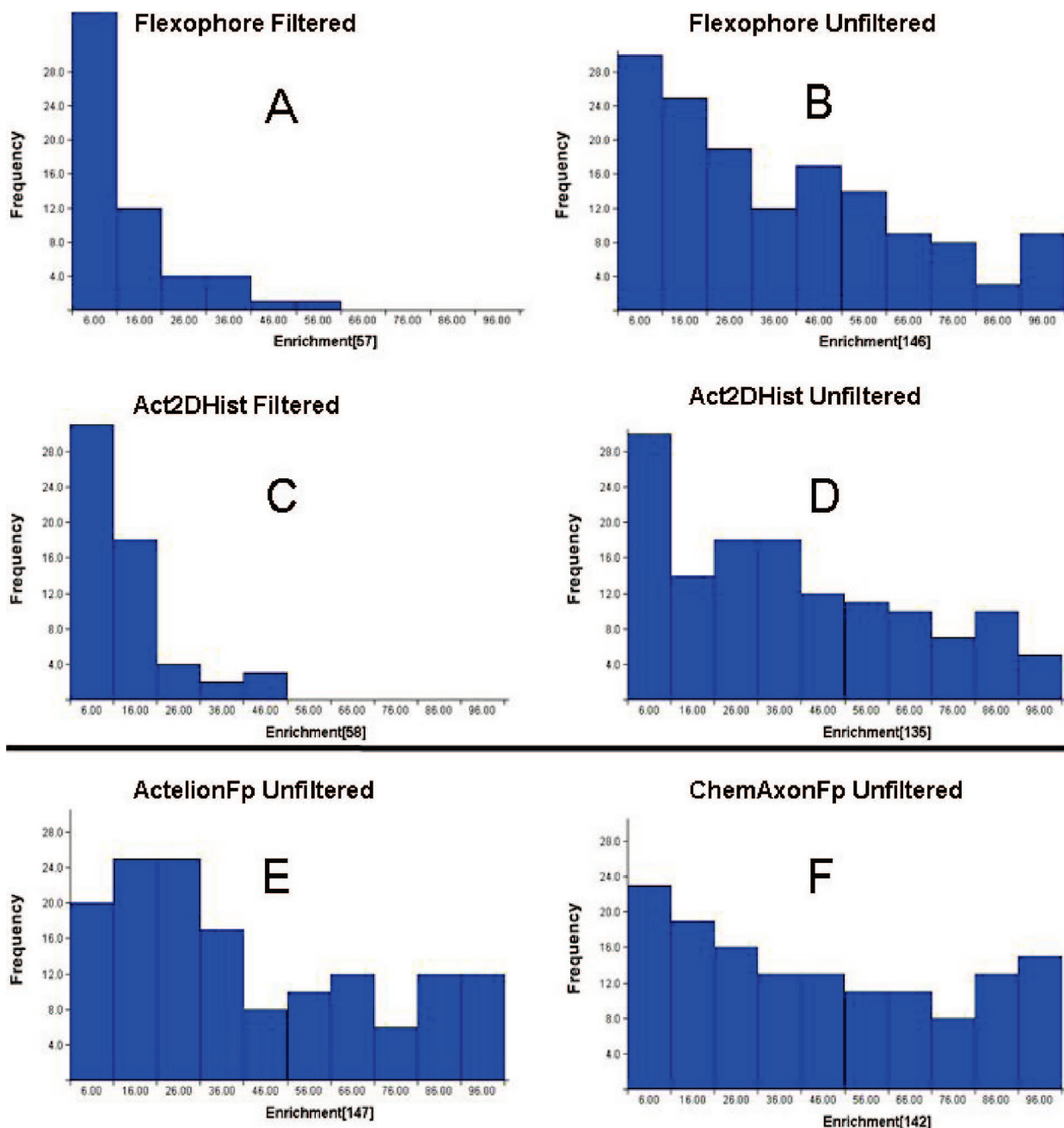


Figure 3. Relative enrichment-factor histograms for each descriptor. The enrichment factor is calculated from the data set fraction at 0.01. The histogram header consists of the descriptor name and the data set. The filtered and the unfiltered data set each contain 175 subsets, consisting of a query molecule with ligands and decoys. The number in brackets at the x-axis caption indicates the number of data sets with an enrichment rate above random.

descriptors, the enrichment rates are better in the first quartile of the histograms, where the lower enrichment rates are located. The enrichment rates in the fourth quartile of the histograms are slightly better for the Flexophore descriptor (Figure 3B), the ActelionFp descriptor (Figure 3E), and the ChemAxonFp descriptor (Figure 3F). An analysis of the cumulative enrichment rates for each target is given in Figure 4. For the unfiltered data set, it was observed that the Flexophore descriptor performed best for six targets: 5-HT1A, CB1, CCR5, D3, GNRh, and Thrombin. The Act2DHist descriptor performed best for CDK2. The ActelionFp descriptor enriched the most data sets for A1, ACE, Angiot2, MCH, and PDE5. The ChemAxonFp descriptor performed best for Bzd, CaBlock, ETA, and Musc. For the filtered data set, the Act2DHist descriptor performed best for nine targets,

and the Flexophore descriptor was better for the remaining nine targets. This balance of performance for the two descriptors on the filtered data sets is remarkable.

The analysis of the enrichment rates showed that all four descriptors perform comparably in terms of active molecule enrichment for the unfiltered data set. However, the remaining question is: do the different descriptors select a different diversity of molecular structures? The result of the selection overlap analysis is summarized in Tables 4–7. Table 4 shows the selection overlap for the D_{all} data set. The lowest overlap (14.1%) can be observed between the descriptors Flexophore and Act2DHist. The highest overlap (33.3%) is given by the descriptor pair ActelionFp and ChemAxonFp. This was expected since both descriptors are chemical fingerprints. The overlap of the enriched hits is again lowest for the descriptor

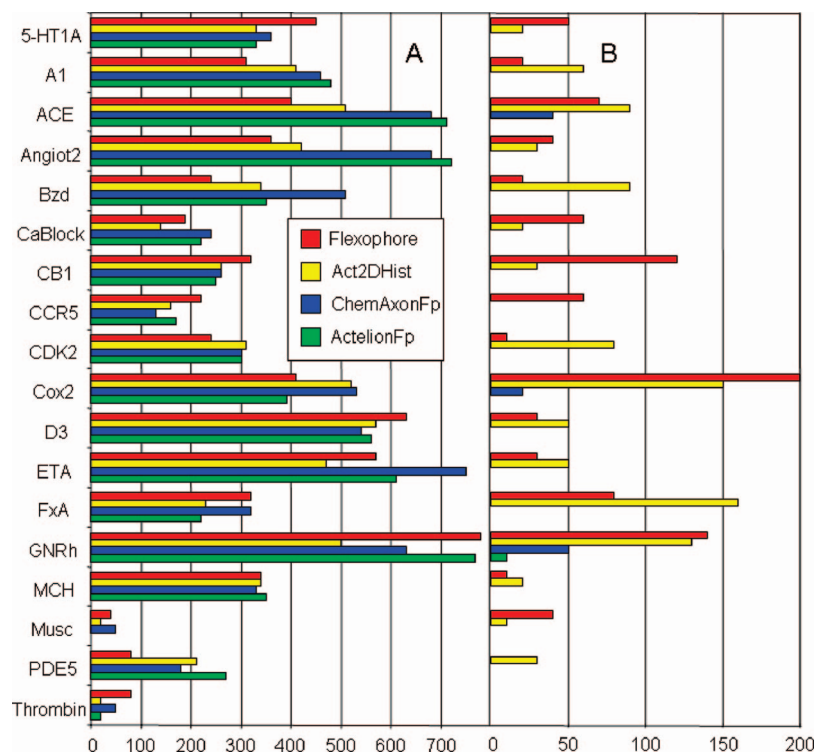


Figure 4. Cumulative relative enrichment factors [%] for the different descriptors. (A) Unfiltered data sets. (B) Filtered data sets.

Table 4. Selection Overlap Given As a Percentage of the Unfiltered Data Sets (D_{all}) in the 0.01 Fraction of Enriched Molecules^a

descriptor	ActelionFp	Act2DHist	Flexophore	ChemAxonFp
ActelionFp	100	17.2	20.6	33.3
Act2DHist	17.2	100	14.1	17.8
Flexophore	20.6	14.1	100	21.8
ChemAxonFp	33.3	17.8	21.8	100

^a All molecules in the chosen fraction were considered.

Table 5. Selection Overlap for the Enriched Hits Given As a Percentage of the Unfiltered Data Sets (D_{all}) in the 0.01 Fraction^a

descriptor	ActelionFp	Act2DHist	Flexophore	ChemAxonFp
ActelionFp	100	21	25.5	39.1
Act2DHist	21	100	17.3	19.8
Flexophore	25.5	17.3	100	24.6
ChemAxonFp	39.1	19.8	24.6	100

^a Only active molecules in the chosen fraction were considered.

Table 6. Selection Overlap Given As a Percentage of the Filtered Data Sets (D_{fil}) in the 0.01 Fraction of Enriched Molecules^a

descriptor	ActelionFp	Act2DHist	Flexophore	ChemAxonFp
ActelionFp	100	6.5	3.2	16.4
Act2DHist	6.5	100	3.3	6.3
Flexophore	3.2	3.3	100	4.6
ChemAxonFp	16.4	6.3	4.6	100

^a All molecules in the chosen fraction were considered.

pair Flexophore and Act2DHist. When taking into consideration the selection overlap in the filtered data sets in Table 6, the situation changes; the overlap decreases for all descriptor pairs. Here, the combination of Flexophore and Act2DHist shows an overlap which is almost equal to the combination Flexophore and ActelionFp. Table 7 shows that

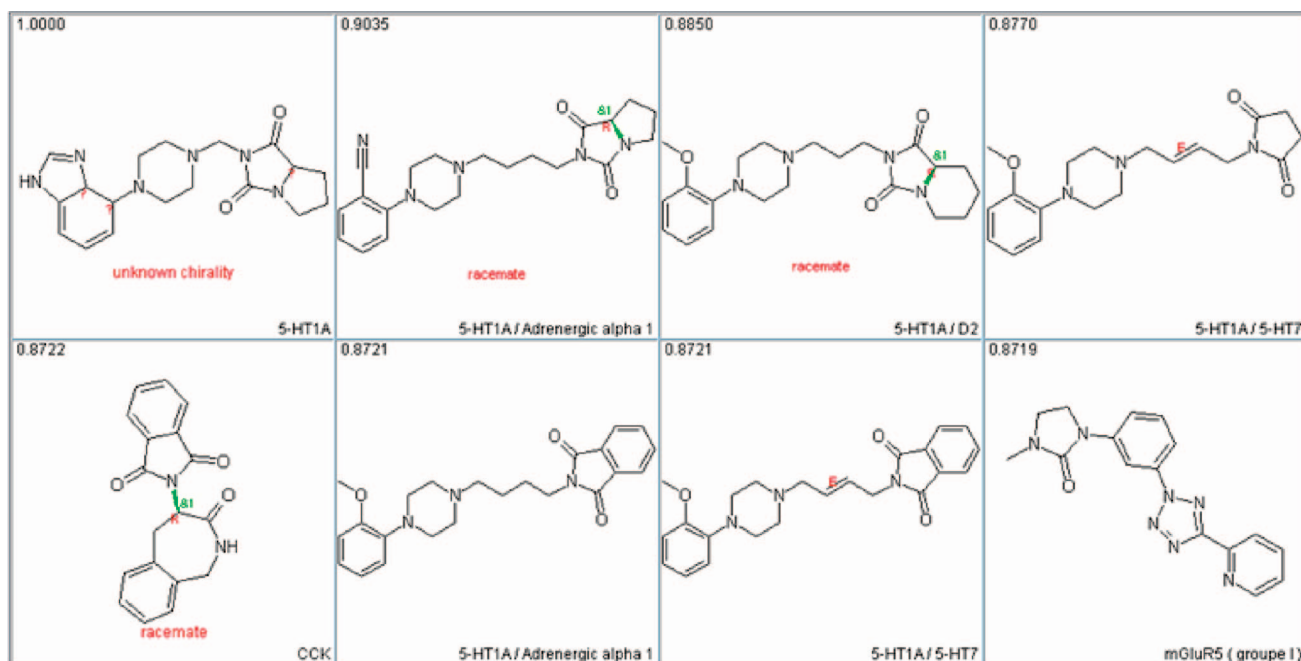
Table 7. Selection Overlap As a Percentage for the Enriched Hits of the Filtered Data Sets (D_{fil}) in the 0.01 Fraction^a

descriptor	Act2DHist	Flexophore
Act2DHist	100	5.9
Flexophore	5.9	100

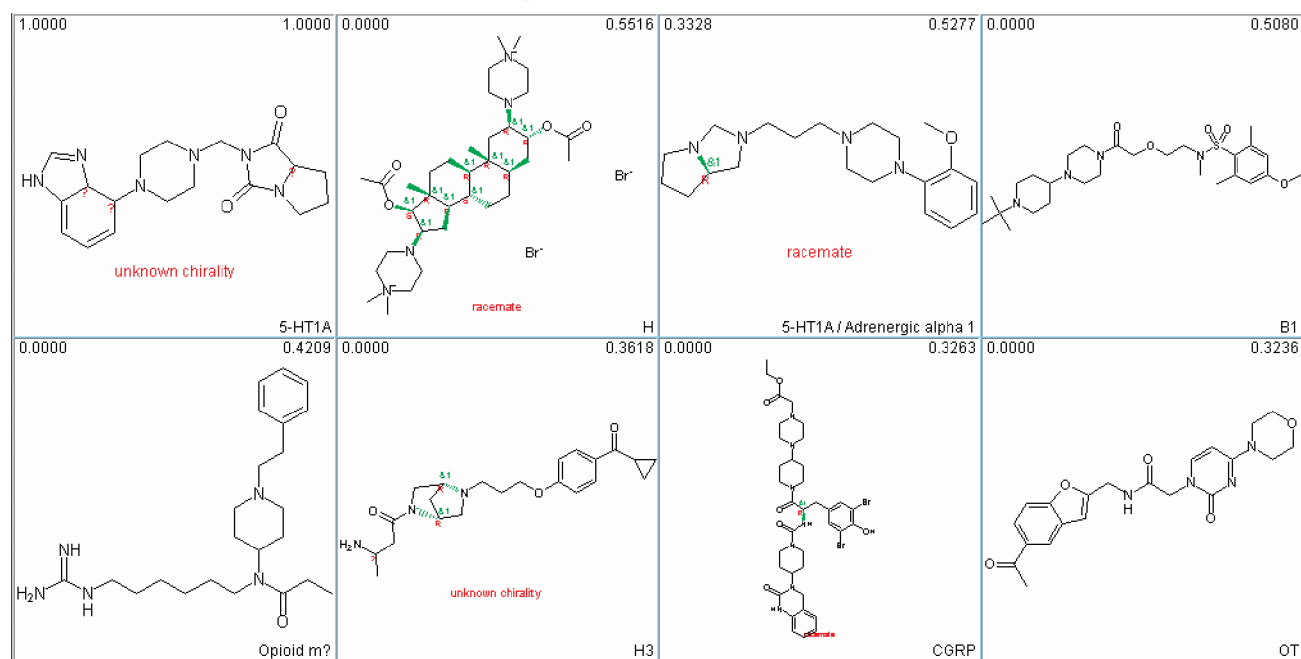
^a Only active molecules in the chosen fraction were considered.

the overlap of the enriched hits in the filtered data sets is quite low at 6%. In this table, only the results for the descriptors Flexophore and Act2DHist are shown, due to the lack of enrichment for the other two descriptors.

To provide a medicinal chemistry “feeling” for the Flexophore descriptor, the structures of several enriched molecules are presented in Charts 1–7. Charts 1–3 show the results for one of the 5-HT1A data sets. The molecules are sorted according to the similarity score of the descriptor comparison between the query molecule and the test-set molecule. The sorting is in descending order from left to right and from the first to the second row. In Chart 1, the molecules were sorted according to the Flexophore descriptor score. In Chart 2, the sorting is according to the Act2DHist descriptor score and, in Chart 3, according to the ActelionFp descriptor score. The query pharmacophore pattern was a linker with five atoms. A heterocyclic unsaturated ring system was attached at one end; an aliphatic/polar substructure was attached at the other end. The linker itself contains a basic amine in the middle. Flexophore and ActelionFp both selected the same first molecule (m_1) from the test data set. One difference to the query was the linker length of eight atoms. The difference in linker length between the query and selected molecule was unimportant for ActelionFp. For the Flexophore descriptor, the flexibilities of the two linkers were similar. The hydroimidazole ring in the query molecule was replaced by a nitrile group in the molecule m_1 . These two different pharmacophore points are similar for the PDB-based metric used to compare the two Flexophore descrip-

Chart 1. Result of Virtual Screening for One 5-HT_{1A}_{all} Data Set, Sorted According to Flexophore Descriptor Score^a

^a Upper-left corner contains the Flexophore descriptor score. Lower-right corner contains receptor information. The first molecule is the query for this test data set.

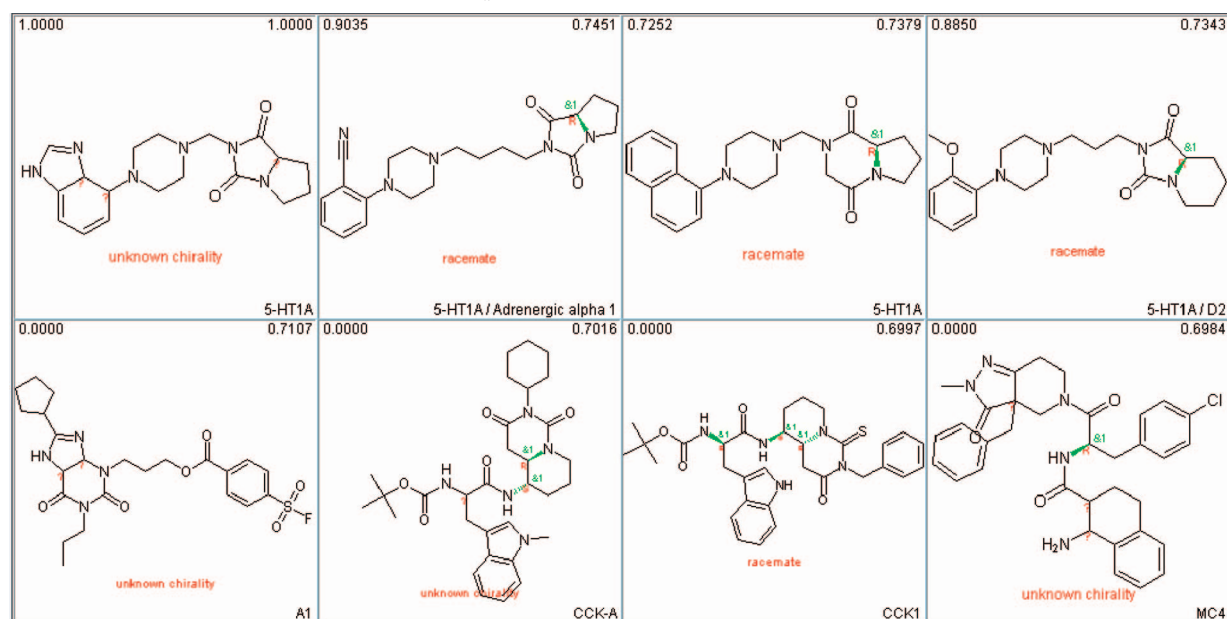
Chart 2. Results of Virtual Screening for One 5-HT_{1A}_{all} Data Set, Sorted According to Act2DHist Descriptor Score^a

^a Description according to Chart 1. The score for the Act2DHist descriptor is given in the upper-right corner.

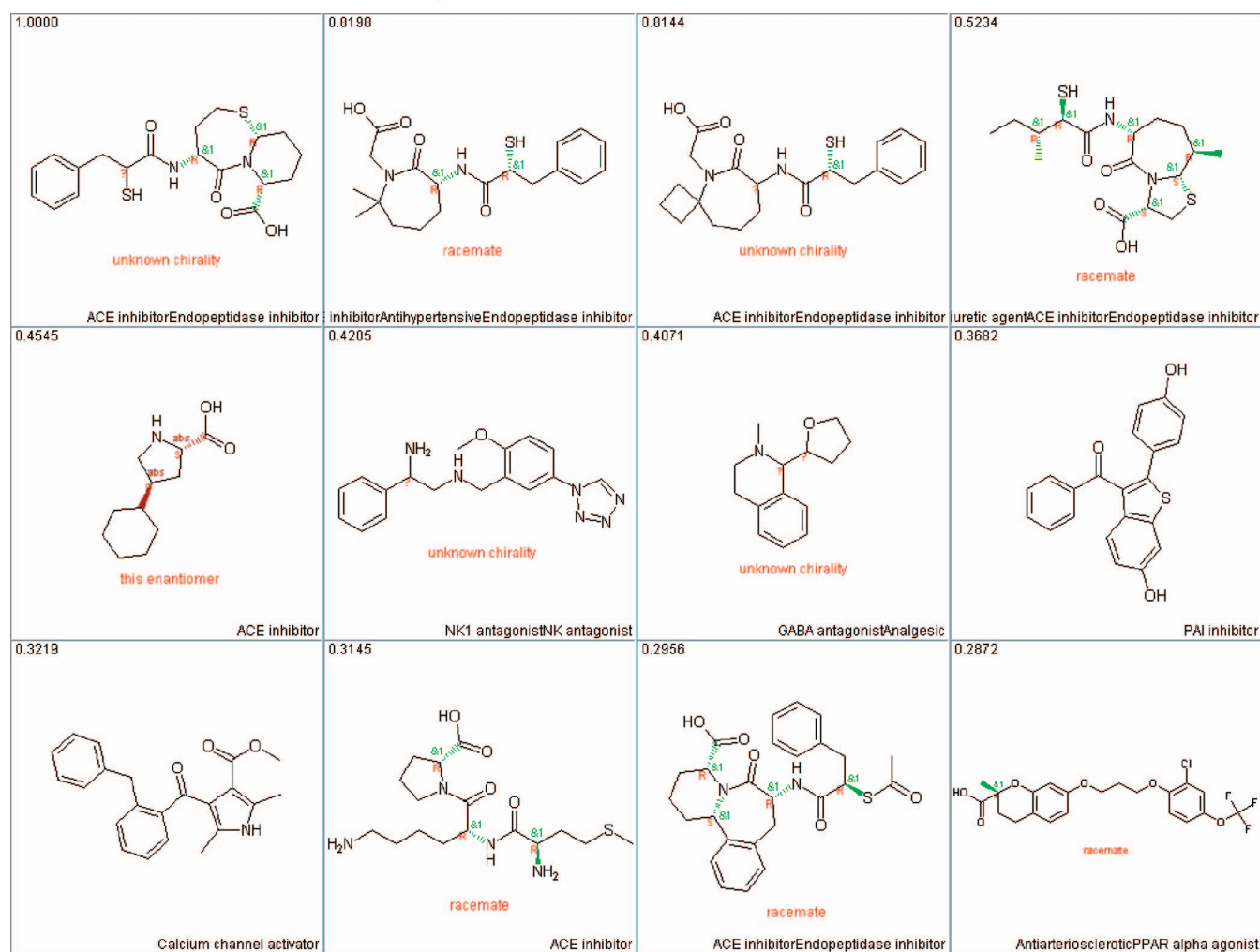
tors. From the seven data set molecules in Chart 2, which were enriched by the Act2Ddescriptor, only one was active on the 5-HT_{1A} receptor. ActelionFp was able to pick out three active molecules (Chart 3). A Flexophore similarity score of zero for some of the inactive test data set molecules in Charts 2 and 3 indicates that the Flexophore similarity metric detected no similarities between query and test data set molecules.

Flexophore performed best for this 5-HT_{1A} data set by enriching four molecules (Chart 1). Chart 4 shows the results of enrichment using the Flexophore descriptor on an ACE test data set, the descriptor enriched for molecules, the last one with a similarity score of 0.3145. The result

for the filtered data set is shown in Chart 5 and demonstrated that the descriptor had a problem in its perception of pharmacophores. Although the third molecule obviously does not fit, it was given a score of 0.63, a score within the range where reasonable molecules were placed in the chart with enriched ACE molecules. The last molecule in this row is a penicillin derivative, which at a glance fits nicely into the query pharmacophore. An assay to confirm some biological activity on the ACE receptor was not available. In Chart 6, the query molecule was taken from the CB1 data set. In the third row, it is shown that even with a similarity of 0.14 the descriptor was able to perform

Chart 3. Results of Virtual Screening for One 5-HT_{1A}_{all} Data Set, Sorted According to ActelionFp Descriptor Score^a

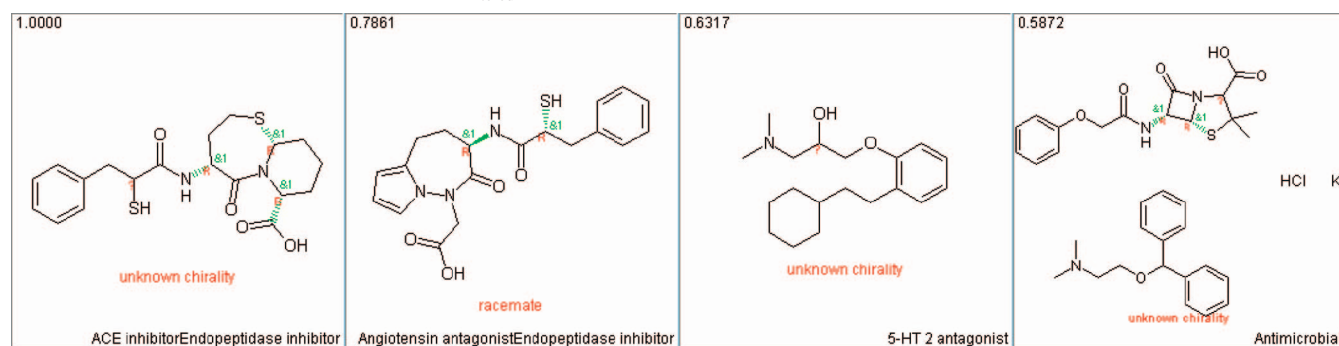
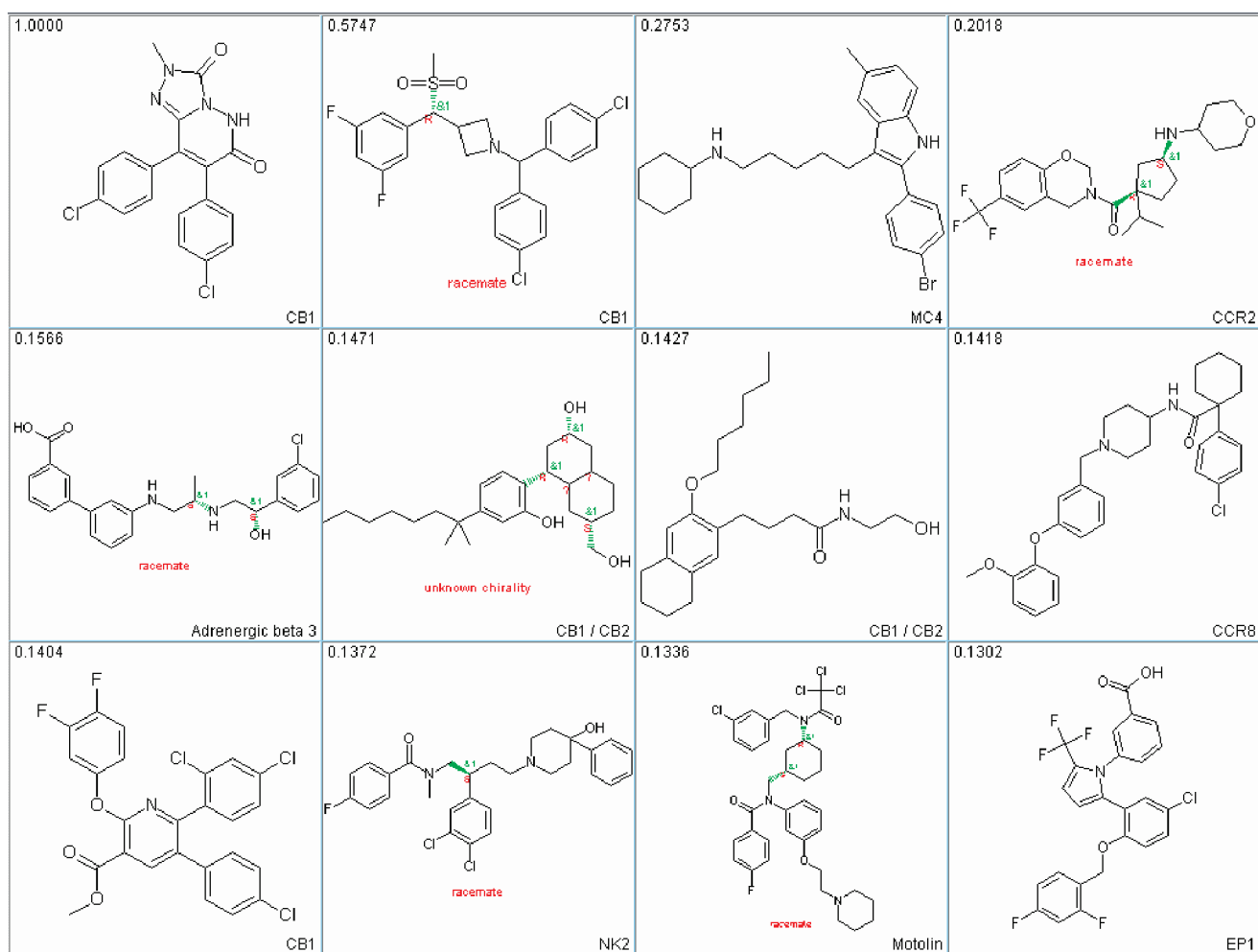
^a As described in Chart 1. The score for the ActelionFp descriptor is given in the upper-right corner.

Chart 4. Results of Virtual Screening for ACE_{all} Data Set, Sorted According to Flexophore Descriptor Score^a

^a Description according to Chart 1.

a reasonable discrimination between molecules. In Chart 7, the results for a filtered Cox2 data set are presented. The Flexophore descriptor found five active molecules

among the 11 data set molecules in the chart. All of these actives have a different core (scaffold), whereas the two substituents remain very similar.

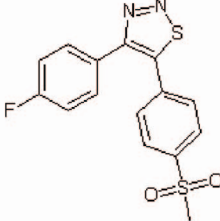
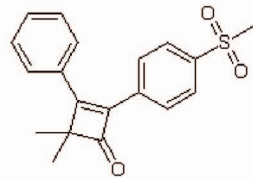
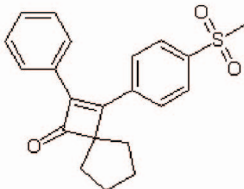
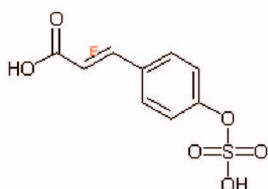
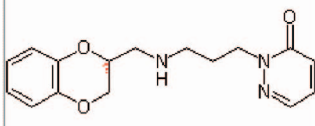
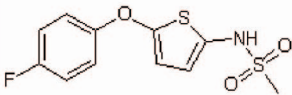
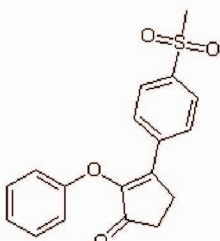
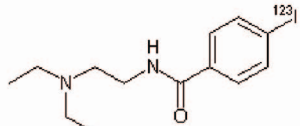
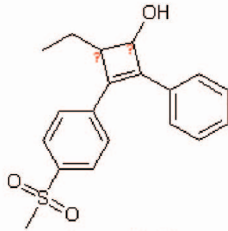
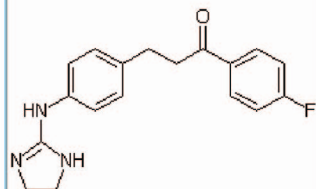
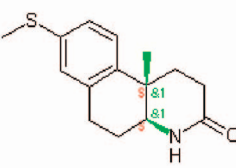
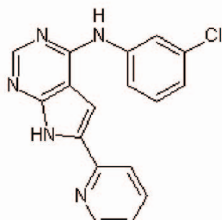
Chart 5. Results of Virtual Screening for ACE_{filtered} Data Set, Sorted According to Flexophore Descriptor Score^a^a Description according to Chart 1.**Chart 6.** Results of Virtual Screening for CB1_{filtered} Data Set, Sorted According to Flexophore Descriptor Score^a^a Description according to Chart 1.

CONCLUSION

In terms of enrichment rates for the unfiltered test data sets, the Flexophore descriptor showed enrichment rates equal to the chemical fingerprint descriptors. After removing from the test sets those molecules that are chemically similar to the query molecules, the Flexophore descriptor was still capable of enriching for active molecules, where chemical similarity-based descriptors completely failed. This proves that the Flexophore descriptor successfully encodes three-dimensional protein-binding behavior rather than chemical similarity. The Flexophore descriptor described the chemical space in a different way compared to the chemical finger-

prints; this feature is underlined by the selection overlap analysis of the enriched molecules. For the filtered data sets in particular, a structural similarity overlap of less than 10% between the chemical fingerprints and the Flexophore descriptor was observed. The selection overlap in the filtered data sets was below 10% for the Flexophore and Act2DHist descriptors. Both descriptors are pharmacophore-orientated; however, the fixed description of the pharmacophore in the Act2DHist descriptor and the interaction statistics-based pharmacophore description of the Flexophore descriptor, as well as the different distance descriptions, lead to a totally different description of the chemical space. Both descriptions

Chart 7. Results of Virtual Screening for Cox2_{filtered} Data Set, Sorted According to Flexophore Descriptor Score^a

1.0000  Cox2_Filt10(46)	0.5405  Cox2_Filt10(46)	0.3987  Cox2_Filt10(46)	0.2585  B_onlyC_DescriptorsNoCyclooxygenase
0.2397  HCl unknown chirality B_onlyC_DescriptorsNoCyclooxygenase	0.1967  Cox2_Filt10(46)	0.1838  Cox2_Filt10(46)	0.1626  B_onlyC_DescriptorsNoCyclooxygenase
0.1301  unknown chirality Cox2_Filt10(46)	0.1044  B_onlyC_DescriptorsNoCyclooxygenase	0.0913  racemate B_onlyC_DescriptorsNoCyclooxygenase	0.0689  B_onlyC_DescriptorsNoCyclooxygenase

^a Description according to Chart 1.

are correct in that they resulted in successful descriptors which performed very well on the virtual screening data sets. These findings contradict the results of Sheridan et al. for the 3D Geometric Atom Pair (3DGAP) descriptor.⁶¹ They found that the ranking of molecules in virtual screening was quite similar for 2D and 3D approaches. The discrepancy can be explained by the different approaches taken to consider geometric information in the descriptors. The 3DGAP descriptor considers only one conformation generated by CONCORD,⁶² while here a representative set of conformers is described. As outlined in Sheridan et al.,⁶¹ geometric distance often correlates with bond distances. In a recent publication, Agrafiotis et al.⁶³ argued that stochastic proximity embedding, which was developed from Xu et al.⁵⁹ and adapted from using this publication, is well-suited for sampling in bioactivity-relevant conformational space. Their findings indicate that one of the reasons for the good performance of the Flexophore descriptor is its appropriate description of molecular flexibility. From a medicinal chemist's perspective, the molecules enriched by the Flexophore descriptor contain pharmacophores similar to the query molecule pharmacophore. This demonstrates that the pharmacophore comparison with its basis on the molecular interaction table represents the medicinal chemists "gut feeling". To summarize, the capacity of the Flexophore descriptor to model biological similarity not only outperforms chemical fingerprints, but Flexophore comparisons can also identify biologically active compounds where topological

pharmacophore comparisons fail. Thus, they are a useful aid in the search for nonpatented islands of active molecules in chemical space directly relevant to a specific biological target. The first application in an in-house medicinal chemistry program resulted in confirmed active molecules, which would neither have been discovered by chemical similarity searches nor by applying the topological pharmacophore descriptor Act2DHist.

ACKNOWLEDGMENT

Christian Rufener implemented Actelion's computing grid. Andrew Jones and Richard Moon reviewed this manuscript.

Supporting Information Available: The complete similarity table for the MM2 atom-type interactions is available. This material is available free of charge via the Internet at <http://pubs.acs.org>. A Java applet to explore the Flexophore descriptor is available at <http://www.cheminformatics.ch/flexophore>.

REFERENCES AND NOTES

- (1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (2) Martin, Y. C. Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J. Comb. Chem.* **2001**, *3*, 231–250.
- (3) Sheridan, R.; Kearsley, S. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.

- (4) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
- (5) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
- (6) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
- (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) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffoldhopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- (10) Strader, C. The view from inside the receptor. *J. Med. Chem.* **1996**, *39*, 1.
- (11) Gund, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117–143.
- (12) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
- (13) Daylight Chemical Information Systems. Smiles ARbitrary Target Specification (SMARTS). <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed Jan 18, 2008).
- (14) van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–251.
- (15) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297–1308.
- (16) 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.
- (17) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach. *J. Mol. Struct.* **2000**, *503*, 17–30.
- (18) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (19) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, *46*, 665–676.
- (20) Lommerse, J. P.; Taylor, R. Characterising non-covalent interactions with the Cambridge Structural Database. *J. Enzyme Inhib.* **1997**, *11*, 223–243.
- (21) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: a library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.
- (22) Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. SuperStar: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.* **2001**, *307*, 841–859.
- (23) Böhm, M.; Klebe, G. Development of New Hydrogen-Bond Descriptors and Their Application to Comparative Molecular Field Analyses. *J. Med. Chem.* **2002**, *45*, 1585–1597.
- (24) Koshland, D. E. Enzyme flexibility and enzyme action. *J. Cell. Comp. Physiol.* **1959**, *54*, 245–258.
- (25) Bosshard, H. Molecular recognition by induced fit: how fit is the concept. *News Physiol. Sci.* **2001**, *16*, 171–3.
- (26) Martin, Y. C.; Bures, M. G.; Willett, P. Searching databases of three-dimensional structures. In *Reviews in Computational Chemistry*; Lipkowitz, K., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 213–256.
- (27) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric pattern matching in files of three-dimensional chemical structures: use of bounded distance matrices for the representation and searching of conformationally flexible molecules. *J. Mol. Graphics* **1992**, *10*, 194–204.
- (28) Sheridan, R. P.; Nilakantan, R.; Rusinko, A.; Bauman, I. N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.
- (29) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83–102.
- (30) Dobler, A. V. a. M. 5D-QSAR: The Key for Simulating Induced Fit. *J. Med. Chem.* **2002**, *45*, 2139–2149.
- (31) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
- (32) Cruciani, G. *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*; Wiley: New York, 2006.
- (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) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569–6583.
- (35) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526–1539.
- (36) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.
- (37) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (38) 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.
- (39) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211–1225.
- (40) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.
- (41) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723–1740.
- (42) Makara, G. M. Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem.* **2001**, *44*, 3563–71.
- (43) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004**, *47*, 144–59.
- (44) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–85.
- (45) Allinger, N. L.; Zhou, X.; Bergsma, J. Molecular Mechanics Parameters. *J. Mol. Struct.* **1994**, *312*, 69–83.
- (46) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (47) ChemAxon. GenerateMD. <http://www.chemaxon.com/jchem/doc/user/fingerprint.html> (accessed Nov 8, 2007).
- (48) von Korff, M.; Sander, T. Toxicity-Indicating Structural Patterns. *J. Chem. Inf. Model.* **2006**, *46*, 536–544.
- (49) Thomson Investigational Drugs Database (IDDB). <http://scientific.thomson.com/products/iddb/> (accessed Sep 2004).
- (50) Orleak, B. S.; Blaney, F. E.; Brown, F.; Clark, M. S. G.; Hadley, M. S.; Hatcher, J.; Riley, G. J.; Rosenberg, H. E.; Wadsworth, H. J.; Wyman, P. Comparison of azabicyclic esters and oxadiazoles as ligands for the muscarinic receptor. *J. Med. Chem.* **1991**, *34*, 2726–2735.
- (51) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
- (52) Bender, A. Cheminformatics. <http://www.cheminformatics.org/menu.shtml> (accessed Nov 8, 2005).
- (53) Korff, M. v.; Steger, M. GPCR-Tailored Pharmacophore Pattern Recognition of Small Molecular Ligands. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1137–1147.
- (54) 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.
- (55) Baumann, K. An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26–35.
- (56) Yuan, S.; Zheng, C.; Zhao, X.; Zeng, F. Identification of maximal common substructures in structure/activity studies. *Anal. Chim. Acta* **1990**, *235*, 239–241.
- (57) McGregor, J. J. Backtrack search algorithms and the maximal common subgraph problem. *Software—Pract. Exper.* **1982**, *12*, 23–34.

- (58) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–33.
- (59) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1186–1191.
- (60) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–56.
- (61) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- (62) Rusinko, A. I.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *CONCORD: A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; The University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988.
- (63) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007**, *47*, 1067–1086.

CI700359J