A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes

Jeremy L. Jenkins,* Meir Glick, and John W. Davies

Lead Discovery Center, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, Massachusetts 02139

Received May 7, 2004

A primary goal of 3D similarity searching is to find compounds with similar bioactivity to a reference ligand but with different chemotypes, i.e., "scaffold hopping". However, an adequate description of chemical structures in 3D conformational space is difficult due to the high-dimensionality of the problem. We present an automated method that simplifies flexible 3D chemical descriptions in which clustering techniques traditionally used in data mining are exploited to create "fuzzy" molecular representations called FEPOPS (*fe*ature *po*int *p*harmacophore*s*). The representations can be used for flexible 3D similarity searching given one or more active compounds without a priori knowledge of bioactive conformations or pharmacophores. We demonstrate that similarity searching with FEPOPS significantly enriches for actives taken from in-house high-throughput screening datasets and from MDDR activity classes COX-2, 5-HT3A, and HIV-RT, while also scaffold or ring-system hopping to new chemical frameworks. Further, inhibitors of target proteins (dopamine 2 and retinoic acid receptor) are recalled by FEPOPS by scaffold hopping from their associated endogenous ligands (dopamine and retinoic acid). Importantly, the method excels in comparison to commonly used 2D similarity methods (DAYLIGHT, MACCS, Pipeline Pilot fingerprints) and a commercial 3D method (Pharmacophore Distance Triplets) at finding novel scaffold classes given a single query molecule.

Introduction

The search for compounds with similar bioactivity to a reference ligand but with different molecular frameworks has been variously termed "scaffold hopping",[1] "leapfrogging",[2] and "lead-hopping";[3] in silico approaches that seek to systematize this practice have been introduced recently.[1,3−13] The ability to move to new scaffolds can be of interest in situations where the natural ligands or substrates of protein targets are known but synthetic inhibitors are not and structural information about the target protein is not available. An ideal similarity search method could use endogenous ligand structures to discover drug-like mimetics in large databases in an automated manner. Alternatively, a scaffold-hopping method could be used to break out of protected "patent space" around drugs or when lead compounds have intractable chemistry, "flat" structure−activity relationships, or poor pharmacological properties (e.g., molecular weight, solubility, toxicity, membrane permeability). The stratagem can thus be an important tool to identify a structurally diverse set of biologically related hits. The ability to establish diverse hits early in the drug discovery process will help maintain a range of compounds for lead optimization as structural classes are eliminated during later stage development.

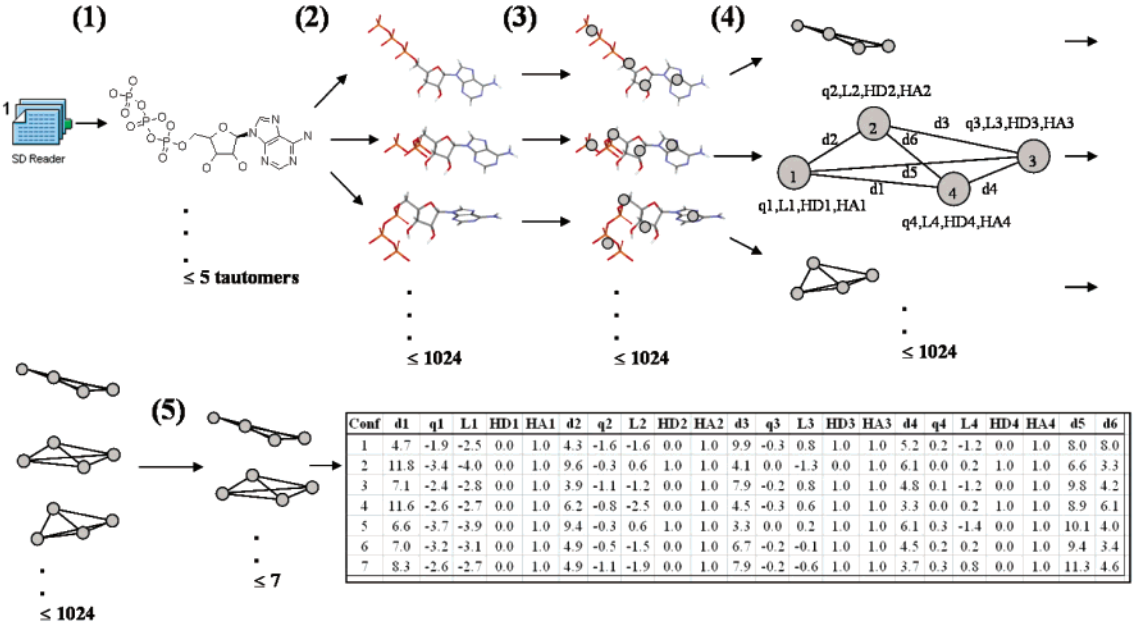
A well-defined criterion for scaffolds is essential for evaluating the diversity of a compound set. Several classifications of chemical structures have been reported.[14,15] A common perception rooted in graph theory is that compound structures may be reduced in a hierarchical manner as a connection of ring systems and linkers that form frameworks or scaffolds, which may contain side chains of functional groups.[16] A topological scaffold may thus be extracted by simple pruning of side chains,[17] and optionally discarding information concerning heteroatoms and bond orders to uncover the graph framework. Ring systems are a further deconstruction with utility in database searching[18,19] and for estimating occupied chemical and drug space.[20−23] While many such objective classifications for scaffolds now exist, the scaffold-hopping ability of 3D-similarity-based or pharmacophore-based search methods is commonly left up to the subjective chemical intuition of the method authors.

All similarity methods are based on the assumption that structurally similar molecules may have similar activity,[24] although the degree of similarity required for similar activity is a matter of dispute.[25] In theory, the increase in structural information from 2D to 3D should provide a more accurate basis for finding new compounds with similar bioactivity. However, 3D-similarity approaches are faced with a number of challenges not faced by topological (2D) methods, such as the generation of flexible molecular conformations and alignment, along with relatively longer computing times. Numerous 3D-similarity methods have been reported,[26] some asserting primacy over 2D methods;[3,27] nevertheless, direct comparisons made in the influential study by Brown and Martin[11] and in more recent reports[28,29] maintain that current 3D methods offer no significant advantages over topological searches in recalling actives or in sampling structural diversity.

Published methods for ligand-based 3D searching commonly take the form of queries based on pharmacophores—pairs, triplets, or quartets of features (hydro-

* Corresponding author. Phone: 617-871-7155. Fax: 617-871-4088. E-mail: jeremy.jenkins@pharma.novartis.com.

10.1021/jm049654z CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/30/2004



Conf	d1	q1	L1	HD1	HA1	d2	q2	L2	HD2	HA2	d3	q3	L3	HD3	HA3	d4	q4	L4	HD4	HA4	d5	d6
1	4.7	-1.9	-2.5	0.0	1.0	4.3	-1.6	-1.6	0.0	1.0	9.9	-0.3	0.8	1.0	1.0	5.2	0.2	-1.2	0.0	1.0	8.0	8.0
2	11.8	-3.4	-4.0	0.0	1.0	9.6	-0.3	0.6	1.0	1.0	4.1	0.0	-1.3	0.0	1.0	6.1	0.0	0.2	1.0	1.0	6.6	3.3
3	7.1	-2.4	-2.8	0.0	1.0	3.9	-1.1	-1.2	0.0	1.0	7.9	-0.2	0.8	1.0	1.0	4.8	0.1	-1.2	0.0	1.0	9.8	4.2
4	11.6	-2.6	-2.7	0.0	1.0	6.2	-0.8	-2.5	0.0	1.0	4.5	-0.3	0.6	1.0	1.0	3.3	0.0	0.2	1.0	1.0	8.9	6.1
5	6.6	-3.7	-3.9	0.0	1.0	9.4	-0.3	0.6	1.0	1.0	3.3	0.0	0.2	1.0	1.0	6.1	0.3	-1.4	0.0	1.0	10.1	4.0
6	7.0	-3.2	-3.1	0.0	1.0	4.9	-0.5	-1.5	0.0	1.0	6.7	-0.2	-0.1	1.0	1.0	4.5	0.2	0.2	0.0	1.0	9.4	3.4
7	8.3	-2.6	-2.7	0.0	1.0	4.9	-1.1	-1.9	0.0	1.0	7.9	-0.2	-0.6	1.0	1.0	3.7	0.3	0.8	0.0	1.0	11.3	4.6

Figure 1. Schematic showing the creation of FEPOPS representations. The steps correspond to those outlined in the first section of the Methods. (1) compound preprocessing, (2) conformer generation, (3) *k*-means clustering of atom coordinates, (4) assignment of features to feature points and sorting by charges, (5) *k*-medoids clustering of FEPOPS conformers, and storage of representative FEPOPS conformers in a lookup table.

phobic group, hydrogen-bond donor or acceptor, etc.) separated from other features by binned distance ranges. For example, a large number of "potential pharmacophores" can be generated automatically for multiple ligand conformers and stored as fingerprints or keys; the overlap between pharmacophore sets for molecules being compared is then calculated.[10−12] Alternatively, pharmacophore alignments can be carried out at run time.[30] Other work has incorporated whole-molecule information into pharmacophores on the basis of geometric distances between defined features and all ligand atoms.[27] Similarity has been measured between 3D topomeric fragments[3,5] and between maximal common substructures.[3,31,32] In contrast to the use of phamacophore points, similarity based on physicochemical property descriptors[33] and molecular fields has been explored.[34−36] Finally, pharmacophoric features can be used in the context of 2D similarity searching[37] or for classification[38,39] and diversity analysis.[40] The references cited above are by no means exhaustive and represent only a sampling of approaches for tackling 3D similarity—a task that clearly has multiple problems with multiple solutions.

The intention of the present work is to create a rapid, automated 3D method that incorporates ligand flexibility. When given a single known drug or natural ligand the method should retrieve from large databases bioactive compounds that are more diverse than those recovered by commonly used 2D methods. In addition to its use as an in silico screening tool, the method should provide pharmacophore-type information about the highly ranked molecules to facilitate the transition from hit discovery to lead optimization. Further, the method may be used as an orthogonal approach in cheminformatics analyses of high-throughput screening (HTS) data to rescue false negatives missed due to low 2D fingerprint similarity—a concept promoted in data fusion strategies.[41] FEPOPS (*fe*ature *po*int *p*harmacophores) evaluates the regional correlation of additive physicochemical and pharmacophoric properties, benefiting from the advantages gained from both field-based similarity and pharmacophore-based queries. Conceptually, the approach bears some resemblance to the 3D-QSAR technique of CoMMA (*co*mparative *m*olecular *m*oment *a*nalysis), where electrostatic or hydrophobic property fields are used for comparing molecules without molecular superposition.[42,43] The defining aspect of FEPOPS is the incorporation of clustering techniques from the field of data mining, which produce a scaled-down representation by *k*-means clustering of atomic coordinates into feature points, followed by the selection of representative conformers by *k*-medoids clustering. The method is fuzzy in three important ways: (i) the decomposition of atoms into feature point representations; (ii) the use of physicochemical descriptors, which are less specific than topological descriptors;[33] and (iii) the reduction of conformational space covered by a molecule to a small number of representative data points. The fuzzy FEPOPS representations retrieve strikingly diverse compounds with similar biological activity to reference queries taken from the MDDR (MDL Drug Database Report) and from real-life HTS datasets (vide infra). Further, we establish a specific, objective criterion for measuring scaffold hopping as well as "ring-system hopping" using molecular equivalence indices.[44,45]

Theory and Methods

Overview of FEPOPS Calculation and Search. The overall strategy for generating FEPOPS representations and similarity searching is shown in Figure 1 and summarized below (followed by a more detailed description). Steps 1−5 for generating FEPOPS apply to both the query molecule and the target database to be searched.

(1) Compounds are preprocessed to generate 3D structures, assign protonation states, enumerate tautomers, and calculate partial charges and atomic log *P* values.

(2) Multiple conformers are generated by systematic rotation of flexible bonds.

(3) Ligand atoms are partitioned into a predetermined number of *k*-mean clusters (typically 4) based on their spatial coordinates.

(4) Atom-type pharmacophoric features—partial charge, log *P*, hydrogen-bond donors and acceptors—of the atoms belonging to each cluster from step 3 are summed and encoded into the centroids to create the "feature points". Distances between feature points are recorded after sorting on the basis of quadrupole directionality.

(5) FEPOPS of the compound conformers are clustered by *k*-medoids to find a small number of representative conformers for each molecule. These are stored in a lookup table.

(6) The similarity of the query molecule FEPOPS to the FEPOPS of each database compound is calculated using Pearson correlation.[46] The rank of the highest scoring conformer of each compound is saved.

Compound Preprocessing. The core FEPOPS programs are implemented in a series of custom scripts and freely available applications. The scripts are launched from an automated data-pipelining protocol in Pipeline Pilot (SciTegic, San Diego, CA) by using batched SOAP (simple object access protocol) technology. The entire protocol processes FEPOPS representations from input 2D structures at ~1.0 compound/s, i.e., ~600K compounds per week. FEPOPS are calculated once and stored in a lookup table.

Initially, a compound library file of type sdf, mol2, or SMILES strings is read into the protocol. Custom filters are applied to remove duplicate compounds; compounds with less than four atoms, more than nine rings, or more than forty rotatable bonds; and salts or counterions associated with compound structures. Three-dimensional coordinates are generated in Pipeline Pilot, followed by addition of hydrogens and a brief minimization using the Clean force field.[47] The protonation states of ionizable groups are set at pH 7.4 on the basis of either lookup tables of pK_a values or partial least squares models. For each compound, all tautomers are enumerated. Finally, Gasteiger–Marsili partial charges[48] are computed.

Atomic log *P* Calculation. log *P* was selected as a feature because it is correlated with hydrophobic binding of receptors and can be calculated atom-wise for a molecule. XlogP, an atom-additive program that predicts octanol/water partition coefficients,[49] is used to calculate atomic log *P* values for each compound.

Conformer Generation. The core computation for conformational searching and atom clustering in FEPOPS is implemented in a C program (for background, see refs 50 and 51). Flexibility is simulated by systematic rotation of bonds at fixed angle increments, followed by eviction of conformations with van der Waals clashes. For compounds with a "drug-like" number of rotatable bonds (less than six), torsional increments between 10° and 120° cover FEPOPS space to a similar degree.[51] In other words, using smaller angle intervals, which increases calculation time linearly, does not lead to a significant increase in FEPOPS conformational information. For compounds with greater than five rotatable bonds, increments of 90° give the optimal trade between speed and accuracy (data not shown); thus 90° intervals are used for the conformational search in the present study. Similar angle intervals have been used by other 3D methods.[52,53] Indeed, the fuzzy representation of FEPOPS is particularly suited to cover conformational space by sampling at larger intervals—since atoms are ultimately partitioned into 3D space on the basis of their coordinates, the atoms of one conformer must be reasonably distant from the atoms of another conformer to yield a unique *k*-means clustering result. The conformers for compounds with five rotatable bonds or less are generated in an exhaustive manner in 90° intervals (maximal number of $4^5 = 1024$ conformers). For more than five rotatable bonds, 1024 conformers are sampled at random. It is worth noting that the objective of conformational sampling in this case is not to find low-energy conformations but to provide a reasonable coverage of the conformational space. Furthermore, low-energy solutions are not necessarily representative of protein-bound ligand geometries.[38] Our similarity method determines biologically relevant conformations by identifying the conformer(s) with the highest correlation to the probe molecule in feature point space.

***k*-Means Clustering of Atoms.** The *k*-means algorithm is an iterative descent clustering method.[54] In FEPOPS, the atomic 3D coordinates of each ligand are partitioned into a designated number of clusters and the geometric centers of the clusters (centroids) are retained to represent the compound (Figure 1). This approach has previously been used for identifying ligand binding sites on proteins[55] and for ligand docking.[51] In the present study, four clusters were used because this allows a reasonable description of molecules of drug-like size while retaining information about chirality, which is lost in triplet-type representations. Additionally, four clusters resulted in FEPOPS representations that performed superior to two or three clusters (not shown). Previous authors have reported approaches to representing molecules in a reduced number of "nodes".[56,57] The algorithm initially guesses the centroid positions. Then, for each atom, the closest centroid is identified, followed by replacement of the centroids with the coordinatewise average of all atoms closest to it. The algorithm minimizes the sum of squared Euclidean distances from centroids to atom cluster members until convergence is achieved. In FEPOPS, each of the four centroids is assigned five pharmacophoric features: (i) the distance to a neighboring centroid (vide infra), (ii) the sum of the partial charges of its cluster-member atoms, (iii) the sum of atomic log *P* values of its cluster-member atoms, and binary flags that indicate occurrence of (iv) hydrogen-bond donors and (v) hydrogen-bond acceptors. Additionally, the distances between juxtaposed feature points 1 and 3 and between feature points 2 and 4 are recorded to determine the chirality of the quartet (Figure 1).

The distribution of electrons in molecules is one of the principle factors in determining their biological, chemical, and physical properties.[58] Prior to recording interpoint distances, the feature points are sorted on the basis of partial charges to enforce a four-point, or quadrupole, directionality. The most negatively charged centroid becomes feature point 1, whereas the most positively charged centroid is assigned to feature point 4. Thus, the descriptor "Distance 1" encoded in feature point 1 will contain the distance between itself, the most negative centroid, and feature point 4, the most positive centroid. "Distance 2" encoded in feature point 2 will contain the distance between feature points 1 and 2, and so on. All FEPOPS configurations are thus "prealigned" by charge distribution, rather than using alignment of shape or geometry during the similarity calculations. Thus, the probe molecule descriptors can easily be compared at search time to the descriptors of target compounds (e.g., d1 of the probe is compared to d1 of the target compound, not d2; also see the discussion on alignment in the "FEPOPS Weaknesses" section). Other "field-based" conformational alignment methods have been reported previously.[35,59,60]

Selection of Representative FEPOPS Conformers. One common strategy for scaling the vast array of potential molecular configurations down to a manageable size is to compute an average or collective fingerprint. An alternative approach is to cluster the conformations and subsequently store a smaller number of explicit conformers.[61] In particular, the creation of fuzzy FEPOPS representations results in a number of conformations that are alike or similar enough to be redundant in terms of describing a given molecule.[51] *k*-medoids is a nonhierarchical crisp clustering method similar to *k*-means. In contrast to *k*-means, which minimizes the sum of squared Euclidean distances from objects to cluster centers, *k*-medoids minimizes the sum of unsquared dissimilarities of objects to their closest representative object (the medoid).[62] The medoid is an actual object or data point that is representative of the structure of the dataset. Selection of medoids is thus less influenced by outliers that might skew the selection of centrotypes. In the present study, medoids are conformers (not conformational averages) that are representative of the total set of a compound's conformers. Importantly, the actual molecular coordinates are not clustered, but rather, the matrix of FEPOPS representations of all calculated conformers for a

molecule are clustered (we will refer to these as "FEPOPS conformers"). The *k*-medoids clustering of FEPOPS conformers is carried out in the R language and environment[63] using the programs PAM (*p*artitioning *a*round *m*edoids) and CLARA (*c*lustering of *lar*ge *a*pplications) (described in detail in ref 62). The data format is an $n \times p$ matrix, consisting of n FEPOPS conformers and p features (or descriptors) for each compound. The features are first scaled by subtracting the mean and dividing by standard deviation, and k representative FEPOPS conformers are selected using the function PAM for compounds with <50 conformers or CLARA for compounds with ≥ 50 conformers. In CLARA, the partitioning is performed on a subset of the data in multiple iterations to speed up calculation; we sample either 100 FEPOPS conformers or 50% of the total number of FEPOPS conformers, whichever is smaller. The number of k clusters designated was seven, based on a thorough analysis of speed and coverage of conformational space. In preliminary studies, we found that, for compounds taken from X-ray structures of protein−ligand complexes, a k value of 7 typically yields a set of medoids containing at least one conformer that is approximately the same as the bioactive conformation (by visual inspection of FEPOPS descriptor space in a modeling package). Additionally, higher k values resulted in redundant conformers for molecules of drug-like sizes (data not shown). If a molecule has less than seven possible conformations, clustering is bypassed and all nonduplicate FEPOPS conformers are saved. We chose not to use a statistical metric[61] to evaluate the goodness of cluster separations during FEPOPS conversions, since calculations at multiple k values would be highly impractical for converting large databases.

FEPOPS Similarity Calculation. The similarity between FEPOPS representations was determined by Pearson correlation. Pearson was suitable for this purpose due to the nature of the continuous, nonbinary FEPOPS descriptors stored in a matrix format. The feature descriptors are first scaled by mean centering (offsetting the values so that their sum is zero) and then divided by a factor so that the variance of the scaled data is equal to one. For all FEPOPS conformers of compounds in the test sets, the Pearson correlations to the FEPOPS conformers of the probe were calculated. The compound conformer with the highest correlation to any probe conformer was then retained for the similarity score (i.e., the "single nearest neighbor" method). In other words, for multiple conformations (i) of the probe and multiple conformations of a database molecule (j), the similarity score is the maximum similarity over all i and j. We used maximum similarity of i and j rather than mean similarity on the basis of the rationale that there is a preferred conformational alignment between any two pairs of molecules that best represents the pharmacophore. The target database is thus reranked in its entirety by correlation to the probe. In searches where more than one probe molecule was used, the maximum similarity among multiple conformations of all probe molecules was kept for each database molecule.

No attempt is made to computationally optimize the search procedure due to the short calculation times. The time required for a FEPOPS single-ligand similarity search and subsequent ranking is on the order of ~8 min for the entire MDDR database or 2 min for a 29 197 compound subset (see Chemical Datasets). Note that Pearson is not appropriate for absolute similarity comparisons between two compounds, but is useful rather for quantifying relative distances between large sets of compounds to a probe. The Pearson coefficient (ranging from −1 to 1) reflects the degree of linear relationship or strength of association between variables; in the present case, the strength of association between test compound features and the reference molecule features with respect to the entire dataset is measured in order to produce a ranked list.

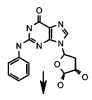
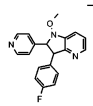
2D Similarity Searches. To assess the scaffold-hopping ability of FEPOPS, its performance was compared with three 2D-descriptor methods used routinely for similarity searches: the 166 publicly available MDL Keys, also known as MACCS (MDL Information Systems, Inc., San Leandro, CA); DAYLIGHT fingerprints (Daylight Chemical Information Systems, Inc., Mission Viejo, CA); and Pipeline Pilot Functional Class Fingerprints with a neighborhood size of four (FCFP_4) (see www.scitegic.com). The Pipeline Pilot fingerprints are based on an extension of the Morgan algorithm.[64] The Tanimoto coefficient was used to calculate similarity for all 2D methods. The searches as implemented required approximately 2 min calculation time for a 29 197 compound database (see Chemical Datasets). For cases where multiple probes were used, the highest Tanimoto similarity of a database compound to any of the probes was used to rank the compound.

Pharmacophore Distance Triplets. In addition to the 2D search methods, we also tested the commercially available 3D method, Pharmacophore Distance Triplets (PDT)[10] from Sybyl 6.9 (Tripos, Inc., St. Louis, MO). A PDT contains three pharmacophoric macro atoms and the three distances separating them, where the distances are binned. Each possible triplet is represented by a single bit in a binary fingerprint; bits are set to "1" in a PDT fingerprint when a triplet is found in a compound during conformational searching. The automated approach used by PDT allows for a fair comparison to be made with FEPOPS, since neither method uses a priori pharmacophore models. "Hand-built" pharmacophore queries could bias the search results if not suitably constructed. We incorporated three pharmacophoric features in the *BinBounds.def* configuration file: hydrogen-bond acceptors, hydrogen-bond donors, and hydrophobes. Distances between features were binned from 3 to 15 Å at 1.5 Å intervals (nine bins). The default 100 conformations was used for both the probe and database molecules. The above parameters result in 3754 bit strings for each compound, as the fingerprint is a composite of all generated conformations. A PDT fingerprint was created for each test case probe in advance of searching. The target database was imported into a UNITY database, and PDT fingerprints for each database compound were generated at search time. Tanimoto similarity was calculated using the UNIX dbmktriplets utility (*evaltype* = similar; *cutoff* = 0) and the hit lists were sorted by similarity scores in Pipeline Pilot. The searches required approximately 8 h of single-processor calculation time for the 29 197 compound database.

Chemical Datasets. Test sets derived from the MDDR (MDL Drug Database Report) were selected for the case studies due to the MDDR classification of database records according to biological activities. FEPOPS typically converts >95% of compounds successfully. (The majority of failures occur during the 3D conversion step; for the purpose of the present work, no further attempt was made to include these compounds.) Of ~128 000 MDDR records, 121 948 structures were successfully converted to FEPOPS representations and stored in 1.32 million explicit FEPOPS conformations with 1.75 tautomers/compound on average. From the entire MDDR set, a diverse subset of 30 000 compounds (hereafter called MDDRds) was selected using the Pipeline Pilot Functional Class Fingerprints in the "Diverse Molecules" filter component. Of the 30 000 compounds, 29 178 successfully converted to FEPOPS. The MDDRds was considered representative of the whole database and therefore used as the "background", because similarity searches using either MDDR or MDDRds yield similar active recall percentages for each search method.

MDDR activity classes were selected to cover the four major ligand−target classifications proposed by Schuffenhauer et al.:[65] enzymes, G-protein-coupled receptors (GPCR), nuclear receptors (NR), and ligand-gated ion channels (LGIC). The ligand targets in the dataset consisted of (i) COX-2 (enzyme), MDDR Activity Index = 78331, 642 records converted; (ii) 5-HT3A (LGIC), MDDR Activity Index = 06233, 788 records converted; (iii) HIV-RT (enzyme), MDDR Activity Index = 71522, 597 records converted; (iv) D2 dopamine receptor (GPCR), 151 agonists (MDDR Activity Index = 11125) and 431 antagonists (MDDR Activity Index = 07701) converted; (v) retinoic acid receptor (NR), MDDR Activity Index = 59505, 331 records converted. Numerous compounds associated with the DA receptor superfamily are present in the MDDR. To avoid false positives due to the various D1, D3, and D4 agonists

Table 1. Reduction of Compound Topology: "Molecule Equivalence Indices" as Criteria for Scaffold Hopping

			Number of Compounds at Topology Level				
Topology	Example 1	Example 2	COX-2	5-HT3	HIV-RT	D2	RAR
Full[a]			**642**	**788**	**597**	**548**	**308**
Atomic Scaffold			298	408	259	337	128
Carbon Scaffold			170	275	177	275	75
Graph Scaffold			138	232	152	245	52
Reduced Scaffold			**58**	**115**	**75**	**106**	**27**
Graph Ring System			45	129	82	108	29
Reduced Ring System			**18**	**44**	**36**	**48**	**15**

[a] The number of actives used for similarity searching (see Theory and Methods).

and antagonists that could resemble dopamine, MDDR records containing dopamine in the biological activity field or as a substructure were eliminated from the MDDRds dataset, with the exception of the D2 actives. In each test case, the actives were "spiked" into the background MDDRds dataset to assess the ability of the similarity methods to enrich for actives above random. As noted previously by others, the assumption that compounds are inactive for a particular biological activity based on activity indices listed in the MDDR is not necessarily valid, since all compounds have not been tested for all activities;[66] however, the assumption is reasonable enough for comparison of similarity methods.

We have further selected four datasets from in-house HTS campaigns. The HTS targets were growth hormone secretagogue receptor (GHS-R; GPCR), γ-secretase (enzyme), matrix metalloprotease-13 (MMP-13; enzyme), and a $T_{315}I$ mutant form of ABL kinase (mABL; enzyme).[67] The database consisted of 60 000 randomly selected inactive compounds from our HTS plated compound library, of which 57 017 structures were converted successfully to FEPOPS. We opted to use multiple probes (5, 10, or 20 compounds) for three of these test cases rather than a single ligand to reflect situations where more than one antagonist is known a priori. The probes were selected from the total set of validated actives using the Diverse Molecules filter component in Pipeline Pilot. We computed the Pearson correlation between each FEPOPS conformer of each database molecule to each FEPOPS conformer of each probe molecule. The highest correlation measured for any database molecule conformer to any conformer from the multiple probes is used as the Pearson score to rank the compound; all other database conformers are discarded. In the mABL study, the Novartis anti-cancer drug STI-571, or Gleevec (imatinib), was used by itself as the reference compound.

Scaffold Hopping Criteria. In principle, 2D-similarity searches are less capable than 3D methods of finding compounds with similar activity to a probe that are yet chemically dissimilar—indeed, the very goal of traditional similarity searching is to find hits that are structurally similar. (We note that some 2D methods not included herein are tailored more for lead-hopping than, say, MACCS or Daylight; however, such methods are proprietary or were not available to the authors at the time of this study.) The primary challenge for 3D methods is to leap from the chemical space describing the probe to distant and diverse chemical spaces. Importantly, unbiased criteria for measuring dissimilarity between molecules should be established that are independent of the fingerprints or descriptors used for similarity searching. One useful strategy is to first define the chemical scaffolds of the actives in the test set and, subsequently, to assess the enrichment of unique frameworks among the top hits selected by the similarity method. In the present study, we classified scaffold classes using molecular equivalence indices (Meqi) developed by Xu and Johnson, which are based on their extension of the Morgan naming algorithm to labeled psuedographs. A molecular equivalence class is defined as an exhaustive subset of molecules that each contain a "recognizable structural feature".[44,45] Table 1 provides an example of how these structural features can be extracted stepwise by reduction of compound topology to create the classes. Molecular equivalences are particularly useful, because the classes are easy to visualize and their memberships do not shift if more compounds are added to the dataset, as in other clustering methods. The total number of compounds used from 5 MDDR activity classes are shown in Table 1 as well as the number of structural classes that occur among the datasets at the reduced topology levels. For example, among the 642 COX-2 inhibitors, 58 "reduced scaffolds" and 18 "reduced ring systems" are present (shown in bold). A reduced scaffold is essentially an ordered ring system containing information about ring connectivity and ring edges, but not ring size. In contrast, a reduced ring system (RRS) is an unordered ring system containing information about the number of rings and internal edges, but not the order of connectivity. RRS is the mostly broadly defined class and produces the smallest number of structural clusters. Reduced scaffolds and RRS representations were chosen in the present study to evaluate scaffold hopping because ring systems are key in determining shape and orienting the functional groups that guide receptor binding and are important in the pharmacological profiles of drugs.[23] Importantly, the Meqi scaffolds and ring systems serve as an impartial and stringent basis on which to evaluate chemotypes, unrelated to the fingerprinting methods herein.

Table 2. Recall of MDDR Actives, Active Scaffolds, and Active Ring Systems by Single-Ligand Similarity Searching

			% recalled in top N%[a]								
			actives[b]			active scaffolds			active ring systems		
target	ligand probe	method	$N = 1$	$N = 0.5$	$N = 0.1$	$N = 1$	$N = 0.5$	$N = 0.1$	$N = 1$	$N = 0.5$	$N = 0.1$
COX-2	SC-558 bioactive	FEPOPS	4.4	2.4	0.3	21	12	3	44	33	11
COX-2	SC-558 flexible	FEPOPS	5.8	3.1	1.6	22	16	7	39	33	22
COX-2	SC-558	FCFP_4	7.2	6.1	1.9	7	7	3	17	17	11
COX-2	SC-558	MACCS	1.6	1.4	0.3	7	7	2	22	22	6
COX-2	SC-558	DAYLIGHT	6.1	5.4	1.9	9	5	2	11	11	6
COX-2	SC-558	PDT	1.6	1.0	0.0	9	7	0	22	22	0
5-HT3	extreg 194584	FEPOPS	5.3	3.6	0.8	17	13	4	27	18	9
5-HT3	extreg 194584	FCFP_4	8.2	7.1	2.9	14	10	3	20	16	5
5-HT3	extreg 194584	MACCS	6.1	3.0	1.3	18	12	7	23	14	11
5-HT3	extreg 194584	DAYLIGHT	6.0	4.4	2.5	13	10	7	16	11	9
5-HT3	extreg 194584	PDT	7.0	4.0	1.3	18	10	3	27	16	7
HIV-RT	extreg 236942	FEPOPS	3.5	2.5	1.2	9	7	4	17	14	6
HIV-RT	extreg 236942	FCFP_4	1.3	1.2	1.0	7	3	1	14	6	3
HIV-RT	extreg 236942	MACCS	1.5	1.2	1.0	3	3	1	6	6	3
HIV-RT	extreg 236942	DAYLIGHT	1.2	1.0	1.0	3	1	1	6	3	3
HIV-RT	extreg 236942	PDT	1.4	0.0	0.0	4	0	0	6	0	0
D2	Dopamine	FEPOPS	1.8	1.3	0.5	8	6	3	10	8	4
D2	Dopamine	FCFP_4	1.0	0.0	0.0	6	0	0	4	0	0
D2	Dopamine	MACCS	0.1	0.1	0.0	3	1	0	2	2	0
D2	Dopamine	DAYLIGHT	0.7	0.7	0.0	3	3	0	2	2	0
D2	Dopamine	PDT	3.9	2.2	1.1	7	6	4	15	13	8
RAR	retinoic acid	FEPOPS	24.0	15.0	3.8	37	33	15	60	60	33
RAR	retinoic acid	FCFP_4	11.0	10.0	4.9	33	30	26	40	40	20
RAR	retinoic acid	MACCS	13.3	9.7	4.5	30	22	15	47	33	27
RAR	retinoic acid	DAYLIGHT	10.4	7.8	3.6	26	22	15	13	13	13
RAR	retinoic acid	PDT	8.6	6.3	2.9	29	25	7	47	40	13

[a] See Table 1 for total number of actives, reduced scaffolds, and reduced ring systems for each target. [b] The % actives possible in the given percentile cutoffs is <100% in all cases.

Results

Active Enrichment. FEPOPS was compared to three topological search methods, FCFP_4, MACCS, and DAYLIGHT, as well as a 3D method, PDT. Traditionally, the success of similarity methods is judged by the "quantity", or the recall, of known actives from a database. To an experimental screener, the important question concerning active recall is how many actives can be retrieved above a designated cutoff containing a practical number of compounds for testing? In the current study, however, we also stress the medicinal chemist's perspective by assessing the "diversity" or "quality" of each hit list based on the recall of active scaffold classes (i.e., scaffold hops). The critical question in terms of scaffold hopping is how well do the methods sample among all possible active chemotypes above a designated cutoff? In other words, does the method find actives dissimilar enough from the probe to be considered novel scaffolds in the eyes of a medicinal chemist? While we acknowledge that many 2D similarity methods were not designed with the intent of scaffold hopping (including the methods herein), we examined how this difference in descriptor design translates in practice. Table 2 provides a comprehensive summary of the percentages of actives, active scaffolds ("Reduced Scaffold" topology from Table 1), and active ring systems ("Reduced Ring Systems" topology from Table 1) retrieved by all three methods using single ligand probes for the five MDDR targets. In practice, the number of compounds that can be tested on the basis of a similarity search is limited. On this basis, we selected a stringent and practicable cutoff of the top 1.0% most similar compounds in order to assess enrichments, corresponding to approximately 296 compounds that would need to be tested (for the MDDRds test set). It is clear that a majority of the actives do not need to be recalled in order to return a sizable percentage of the active scaffolds and ring systems. The following results demonstrate that FEPOPS is ideal for capturing the largest number of scaffolds within a reasonable number of "cherry picks".

COX-2 Inhibitors. COX-2 is an oxidoreductase (EC 1.14.99.1) targeted by numerous therapeutics for pain and inflammation. The COX-2 inhibitor SC-558 was selected as the probe, since an X-ray cocrystal complex structure was available.[68] The 3D coordinates (PDB code 1CX2) were used to obtain a single FEPOP for the bioactive form of the compound by disabling flexibility during the calculation. The results were compared with the results for FEPOPS calculations allowing incremental flexibility in the four rotatable bonds (Table 2). In the top 1% of compounds most similar to SC-558, 4.4% of the 642 COX-2 inhibitors were retrieved using the bioactive form versus 5.8% when flexibility is incorporated (see Discussion). FCFP_4 retrieved the highest percentage of actives (7.2%) given the SC-558 probe. In contrast, FEPOPS identified the largest number of active scaffolds and reduced ring systems (RRS). In the top 1%, 44% of all active ring systems were sampled by FEPOPS (rigid) versus 17% for FCFP_4, 22% for MACCS, 11% for DAYLIGHT, and 22% for PDT, indicating that the most diverse hits are retrieved by FEPOPS. The results are similar for scaffold retrieval. Figure 2 shows cumulative recall curves for the recovery of COX-2 inhibitor scaffolds (top panels) and ring systems (bottom panels). Cumulative recall curves are useful for visual comparison of similarity methods because they are based on ranks rather than scores.[29]

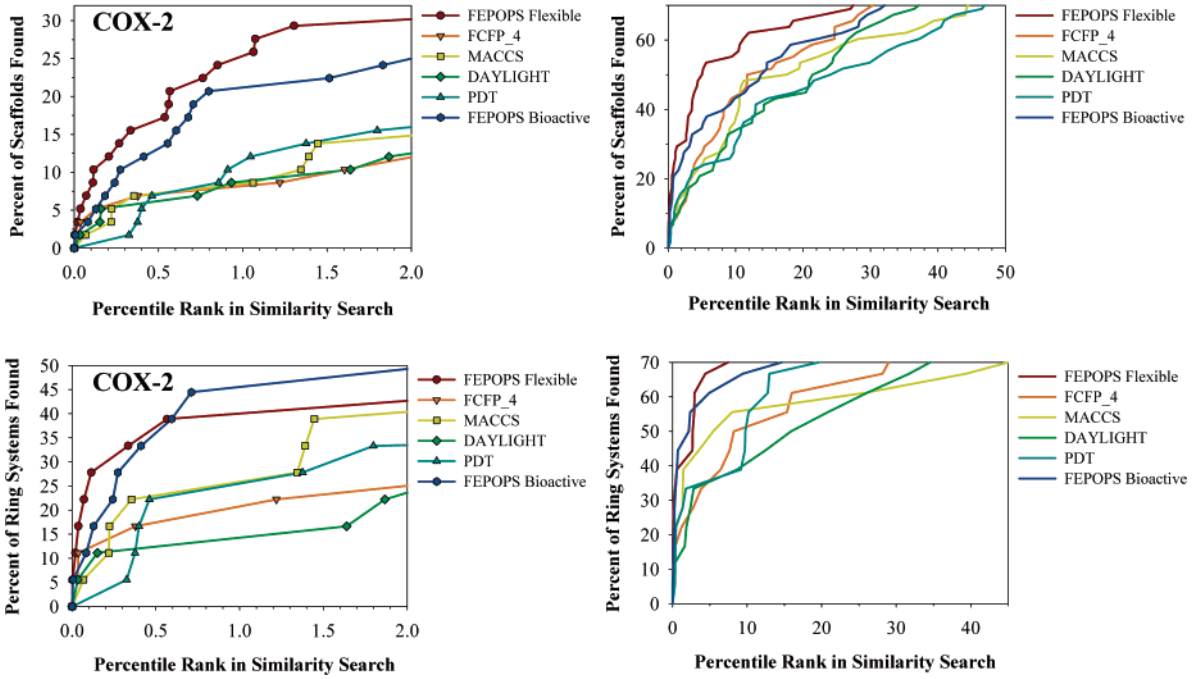


Figure 2. Cumulative recall of the 58 active scaffold classes (top) and 18 reduced ring systems (bottom) from the COX-2 inhibitor set found in the highest ranked 2% of compounds from each method. The points denote the rank of the highest scoring member of a given scaffold class. At right, the percent of the reranked database that would need to be tested to find 70% of the active scaffolds/ring systems.

The recovery of scaffolds and ring systems in the top 2% of the database is displayed (Figure 2, at left). We also assessed what percent of the database would need to be screened in order to find 70% of the active scaffolds and ring systems (Figure 2, at right). For example, FEPOPS (flexible search) found 70% of the active ring systems in the first 7% of the database, versus 20% of the database for PDT.

For each ring system recovered in the top 1%, the highest ranking member along with its actual rank is shown in Figure 3. Several of the active ring systems retrieved by FEPOPS are strikingly divergent from the probe. In certain cases, substituent differences preserve the overall log *P* and charge character of the probe but possess an entirely different "bulkiness", such as the compound from RRS 2 in Figure 3. Notably, ring systems 2, 4, 6, and 8 found by FEPOPS were not found by the 2D-based searches. The compounds shown from these ring systems (Figure 3, bioactive probe conformation) have Tanimoto similarities to the probe of 0.13, 0.06, 0.17, and 0.11, respectively using FCFP_4 descriptors (i.e., they are dissimilar in practical terms). The scaffold hops made by FEPOPS could not have been made by following the chemical approach of isosteric replacement. In contrast, the 2D methods typically find compounds that retain the tricyclic core of SC-558 despite their membership in different RRS classes (Figure 3).

5-HT3A. The 5-hydroxytryptamine receptor (3A) is a ligand-gated ion channel involved in neurotransmitter reuptake and a significant pharmacological target of antidepressant medications. The probe for similarity searching was selected at random by Pipeline Pilot from the 788 MDDR actives (see structure in Figure 5). Enrichment for the total number of actives recovered was again highest for FCFP_4 (Table 2). Although FEPOPS showed the lowest active recovery, its identification of active ring systems was the best at the 0.5% and 1.0% cutoffs (Figure 4, left). Overall, 27% of the active ring systems are recovered in the top 1% by FEPOPS and PDT; however, PDT, MACCS, and FEPOPS all perform competitively at recalling 70% of the active ring systems (Figure 4, right). For each RRS recovered in the top 0.5%, the highest ranking member along with its actual rank is shown (Figure 5). (Note that RRS numbers do not correspond to one another between the different target test cases.) The representative structures shown for the 2D methods generally contain substructures that resemble the core bicyclic ring system from the probe. Half of the ring systems retrieved by the 3D methods were not found by the 2D methods (RRS 8, 10, 12, and 15 for FEPOPS and RRS 1, 15, 24, and 44 for PDT). FCFP_4 and MACCS each found one unique ring system (RRS 35 and RRS 40). This suggests that overall the 3D methods sample the diversity of 5-HT3A antagonist chemotypes more effectively in the top percentiles.

HIV-RT. HIV reverse transcriptase is a nucleotidyl transferase (EC 2.7.7.49) and a significant target of antiretroviral compounds. The probe for similarity searching was selected at random by Pipeline Pilot from the 597 MDDR actives. In this test case, FEPOPS retrieved the highest percentage of actives, scaffolds, and ring systems at any threshold above the top 1% most similar compounds (Table 2 and Figure 6): 21 actives from seven scaffolds and six ring systems were identified in the top 1% when given the single probe. In contrast to its performance in the early portion of the RRS recall curve, DAYLIGHT recalled 70% of the active ring systems in the smallest percentile of the MDDRds (Figure 6). In the top 100 hits from each method (top

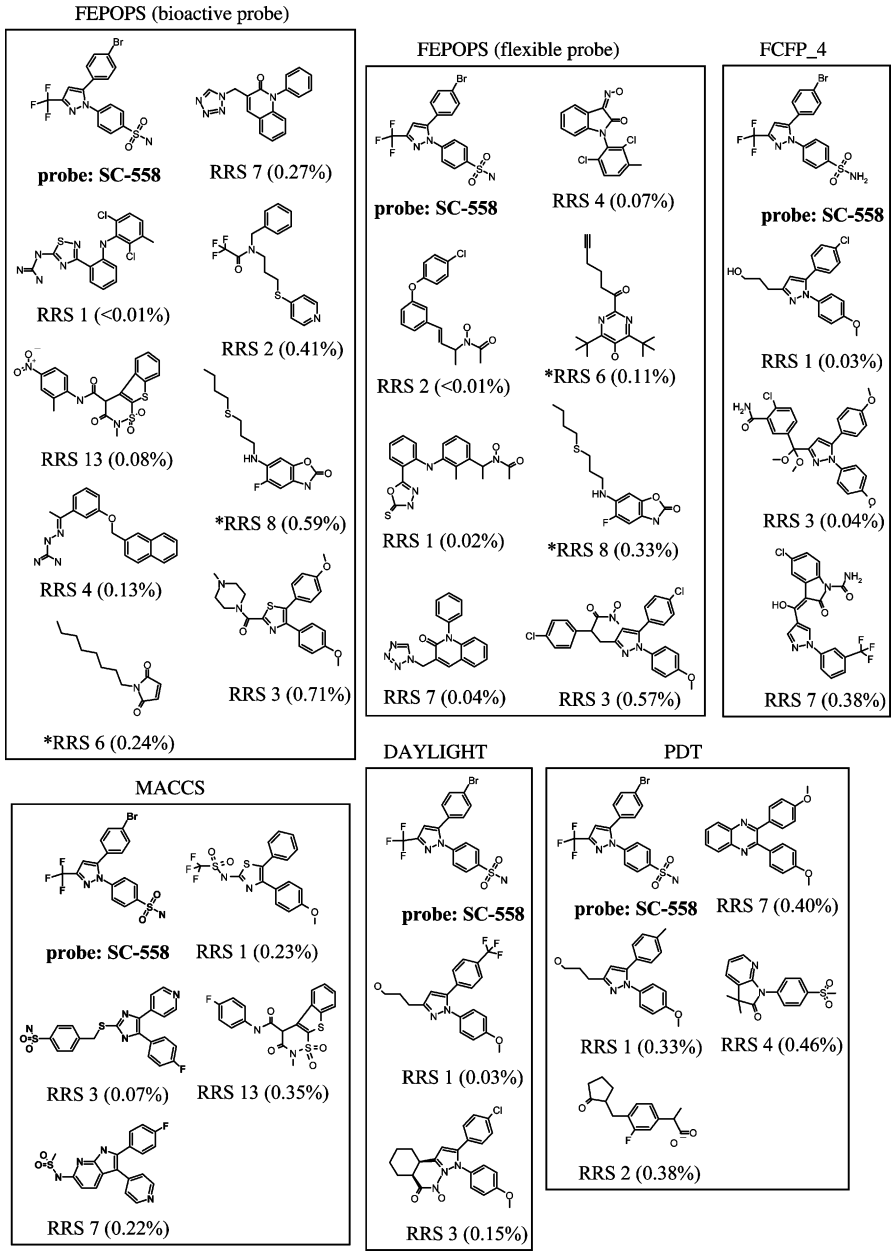


Figure 3. Representatives from COX-2 RRS classes found by the similarity methods in the top 1% using the probe SC-558. Only the highest ranked compound from each RRS is shown. The RRS designation is provided along with the percentile rank in parentheses. RRS classes found uniquely by a method in the top 1% have an asterisk.

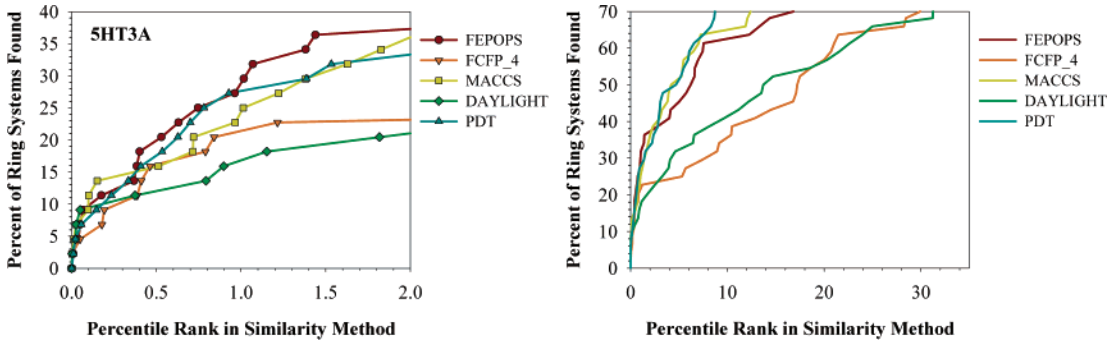


Figure 4. Cumulative recall of the 44 RRS classes in the top 2% for the 5-HT3A dataset (left). The percent of the reranked database that would need to be tested to find 70% of the active scaffolds/ring systems (right).

0.34%), only FEPOPS hops to a ring system that differs from the probe (Figure 7). DAYLIGHT and MACCS each have one RRS unique to their hits in the top 1%, while FEPOPS and FCFP_4 have two.

D2 Agonists and Antagonists. Often endogenous ligands such as peptides, hormones, or cofactors associated with proteins in vivo are known at the beginning of the lead identification process. Ideally, an effective

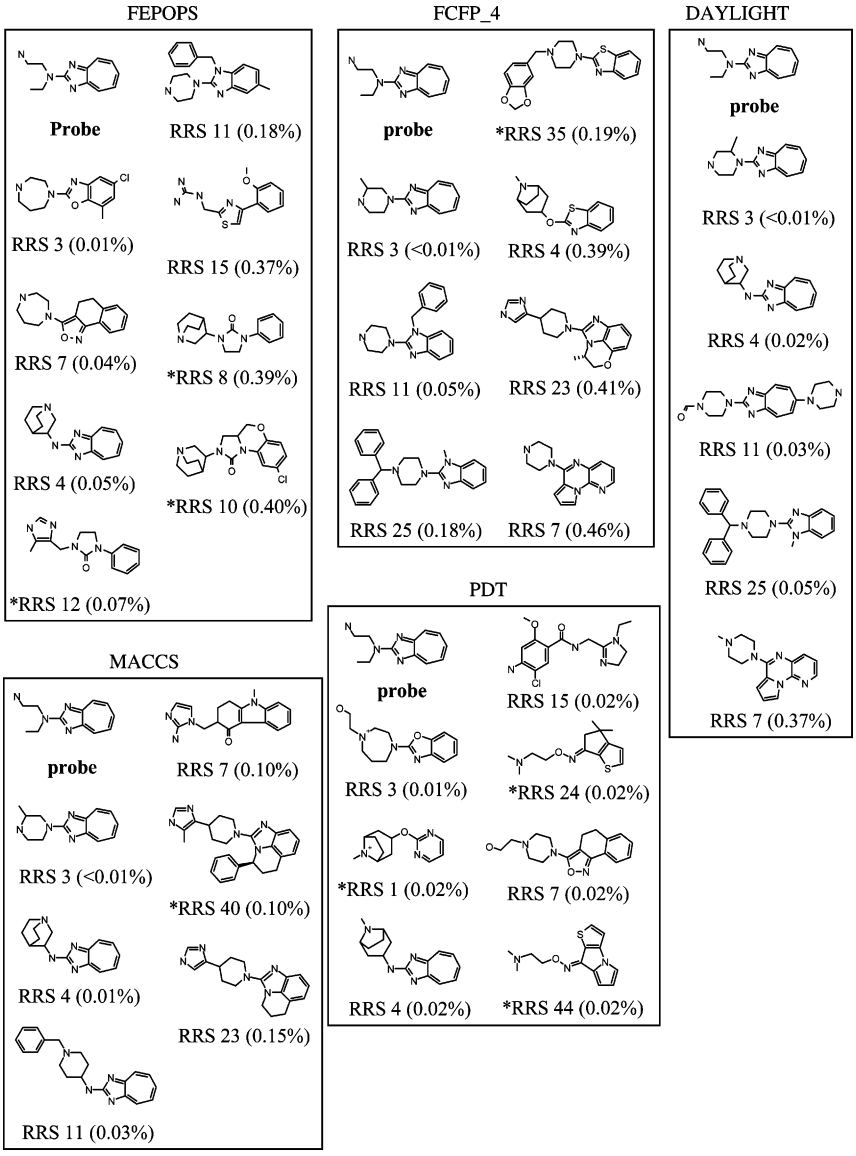


Figure 5. Representatives from 5-HT3A RRS classes found by the similarity methods in the top 0.5% using the probe shown (MDDR extreg 194584). The RRS designation and percentile rank are provided. RRS classes found uniquely by a method have an asterisk.

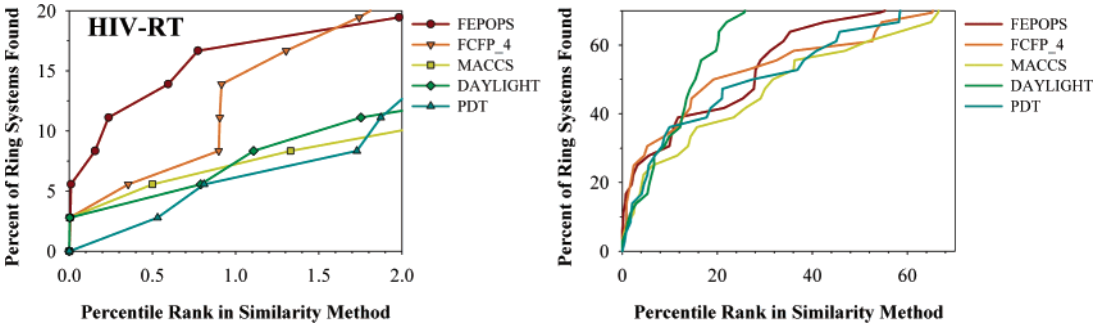


Figure 6. Cumulative recall of the 36 RRS classes in the top 2% for the HIV-RT dataset (left). The percent of the reranked database that would need to be tested to find 70% of the active scaffolds/ring systems (right).

similarity search method could scaffold hop directly from a given endogenous ligand to candidate lead molecules. Consequently, we posed an especially challenging problem to the present similarity methods by supplying dopamine as a probe to search for 548 MDDR D2 receptor agonists and antagonists. D2 receptors belong to the superfamily of seven transmembrane GPCR dopamine (DA) receptors, which have been implicated in neuropsychiatry, cardiovascular, and renal diseases.[69] For example, D2 agonist activity is involved in the action of antiparkinsonian drugs, whereas D2 antagonist activity is associated with the antipsychotic medications used to treat schizophrenia. It is important to stress that a number of compounds in the MDDR

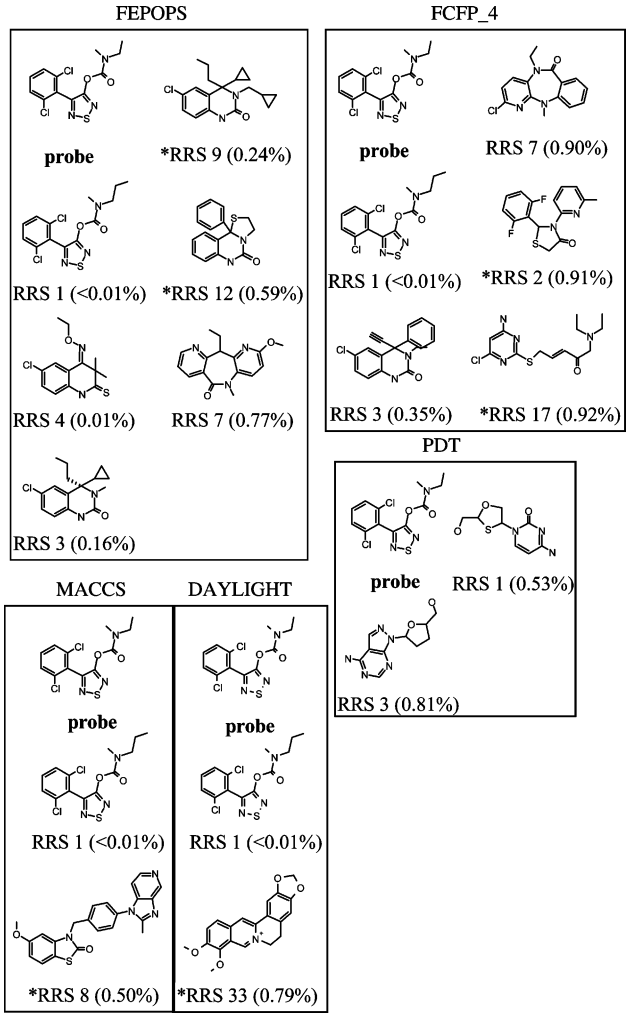


Figure 7. Representatives from HIV-RT RRS classes found by the similarity methods in the top 1% using the probe shown (MDDR extreg 236942). The RRS designation and percentile rank are provided. RRS classes found uniquely by a method have an asterisk.

dataset antagonize binding of dopamine-like compounds, such as serotonin and histamine, adding a level of difficulty to this particular test case.

A small number of the total D2 actives are found in the top 1% by the similarity search methods. Only the 3D methods, PDT and FEPOPS, fare consistently better than random at recovering D2 actives given the structure of DA (Table 2). In terms of hit diversity, FEPOPS finds five different ring systems (10%) in the top 1%, whereas PDT finds seven ring systems. Further, FEPOPS recalls 70% of the RRS classes in the top 20% of the database (Figure 8). Despite the low return on total actives, the enrichment for novel RRS classes by the 3D methods (Figure 9) may afford a suitable starting point for lead optimization in similar real-life cases.

Retinoids. Retinoic acid receptors (RAR) are nuclear receptors and transcription factors critical in the differentiation of various cell types. Both RAR agonists and antagonists have been found to have antitumor activities in several cancers.[70] Members of the RAR family are activated by a number of naturally occurring retinoids, one of which is *all-trans*-retinoic acid (at-RA), a carboxylated form of vitamin A. As a launch point to test for scaffold hopping, we selected at-RA as a probe to assess for recovery of 308 retinoid compounds from the MDDR. FEPOPS showed 24-fold enrichment of actives in the top 1% and 60-fold enrichment for active ring systems (Table 2). Of the 15 active ring systems, FEPOPS found nine in the top 0.5%, versus six found by FCFP_4, five found by MACCS, and two found by DAYLIGHT (Figures 10 and 11). FEPOPS exclusively identified RRS classes 9 and 15 as well as all seven others recalled by FCFP_4, MACCS, DAYLIGHT, and PDT collectively.

HTS Datasets. The MDDR test cases indicated that the Pipeline Pilot functional class fingerprints and MACCS Keys were generally more useful than DAYLIGHT as benchmarks in comparing 2D similarity to the FEPOPS approach. For the HTS dataset test cases, we have simply compared FEPOPS to FCFP_4 by assessing the recall of actives, scaffolds, and reduced ring systems in the top 1% of their ranked lists. The results are summarized in Figure 12. In the GHS-R, γ-secretase, and MMP-13 test cases, FEPOPS recalls more actives than FCFP_4 when given five reference actives, as well as more scaffolds and novel ring systems. However, FCFP_4 recovers more inhibitors of GHS-R and γ-secretase than FEPOPS when given 20 reference actives. This may reflect the fundamental nature of the descriptors used by the different methods. For a 3D method such as FEPOPS, an increase in the number of probe structures may not greatly increase the amount of new information, since the molecular representation is largely based on pharmacophores. In contrast, an increase in new probe structures inherently presents more topological information to guide searches with 2D fingerprints.

The results for mABL are significant from a biological viewpoint, because they involve searching for compounds with similarity to Gleevec that are in fact

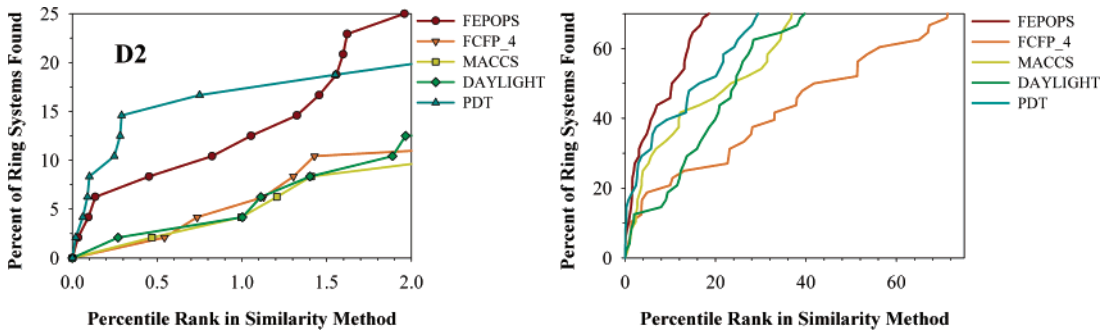


Figure 8. Cumulative recall of the 48 RRS classes in the top 2% for the D2 dataset (left). The percent of the reranked database that would need to be tested to find 70% of the active scaffolds/ring systems (right).

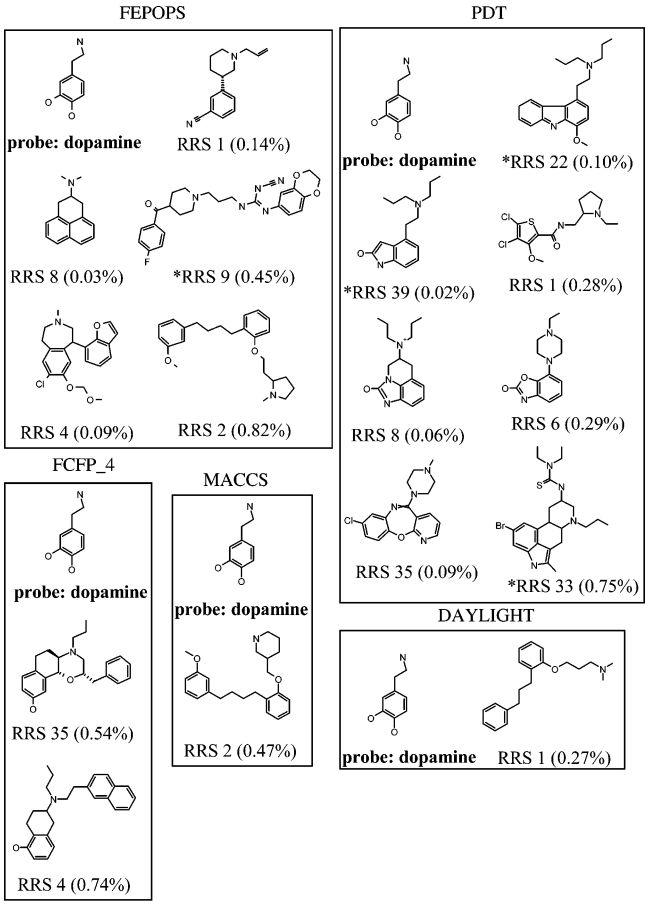


Figure 9. Representatives from D2 RRS classes found by the similarity methods in the top 1% using dopamine as a probe. The RRS designation and percentile rank are provided. RRS classes found uniquely by a method have an asterisk.

inhibitors of a mutant variant of ABL kinase found in vivo in some chronic myelogenous leukemia patients with resistance to Gleevec; thus, the probe molecule itself has dramatically reduced affinity for the target. The challenge is to leap from the ineffective inhibitor to active chemotypes without utilizing structural knowledge about the variant protein target. We found that FEPOPS hopped from Gleevec to 24% of the mABL inhibitors in the top 1% most similar compounds, while sampling from 37% of the active scaffolds and 29% of the active ring systems (Figure 12). This suggests that drug-induced mutations in target proteins could be countered from a pharmacological standpoint by scaffold hopping from impotent drugs to new potent leads.

In addition to our current scaffold-hopping criteria, an alternative way of assessing hit list diversity is to compute the average similarity between the probe molecule and the recalled hits and to compare this value with the average similarity between the probe molecule and the entire activity class. If the similarity of the probe to all possible actives is comparable to its similarity to the recalled actives, then the hit list is an excellent (diverse) sampling of the activity class. We further assessed the diversity of the top ring-system hops made by FEPOPS and FCFP_4 (in the top 1%) by computing their Tanimoto similarity to Gleevec using MACCS as an independent descriptor. The similarity of the FEPOPS ring-system hops to Gleevec was 0.53 on average (0.65, 0.56, 0.52, 0.45, 0.57, 0.53, individually), whereas the Tanimoto similarity of Gleevec to all mABL inhibitors in the dataset was 0.51 on average. These results illustrate that the FEPOPS ring-system hops represented well the existing structural diversity in the dataset. By comparison, the Tanimoto similarity of Gleevec to the FCFP_4 ring-system hops was somewhat higher at 0.58 on average. Table 3 explores this alternative means of assessing hit diversity for the six test cases where a single probe molecule was used. The average similarity is shown for each probe to all actives recalled in the top 1% (not just the scaffold hops). In each case, the average similarity of the probe to the recalled actives was lower for FEPOPS than for FCFP_4.

Discussion

We have introduced a novel means to conduct flexible 3D similarity searches using atom cluster centroids, or feature points, which contain physicochemical (partial charges and atomic log *P*) and pharmacophore (hydrogen-bond donors and acceptors) properties. The feature point pharmacophores are created by electrostatic-based sorting of the four feature points prior to assigning their interpoint distances. We have demonstrated that the FEPOPS representation is applicable for finding leads with novel scaffolds ranked in the top 0.1−1% of the database in cases where one or only a few "binders" are known. The method is extremely robust in terms of scaffold hopping, and can identify novel scaffolds that cannot be identified by isosteric replacements. The fuzzy representation of FEPOPS highly ranks actives from a variety of scaffold classes with minimal reduction in total active recall. Importantly, there is not one best algorithm for all test cases; however, our results suggest that FEPOPS is consistently a strong performer. For example, FEPOPS demonstrated the best ring-system hopping above the top 1% threshold in four of the five MDDR test cases, while recalling 70% of the active ring systems first in three of the five test cases. FEPOPS

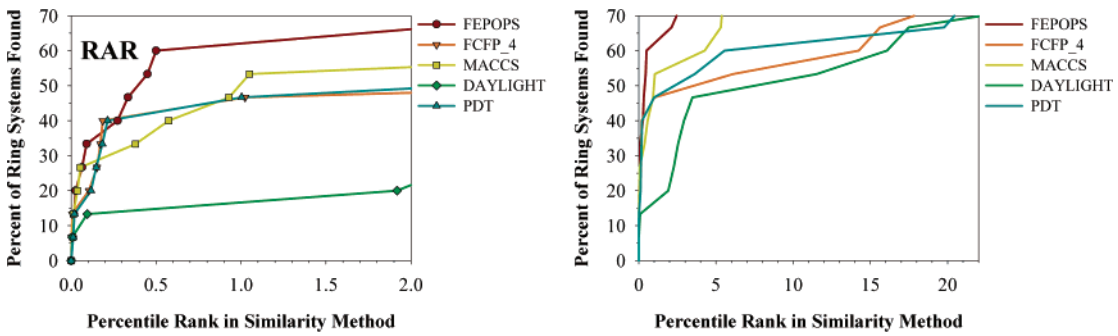


Figure 10. Cumulative recall of the 15 RRS classes in the top 2% for the retinoids dataset. The percent of the reranked database that would need to be tested to find 70% of the active scaffolds/ring systems (right).

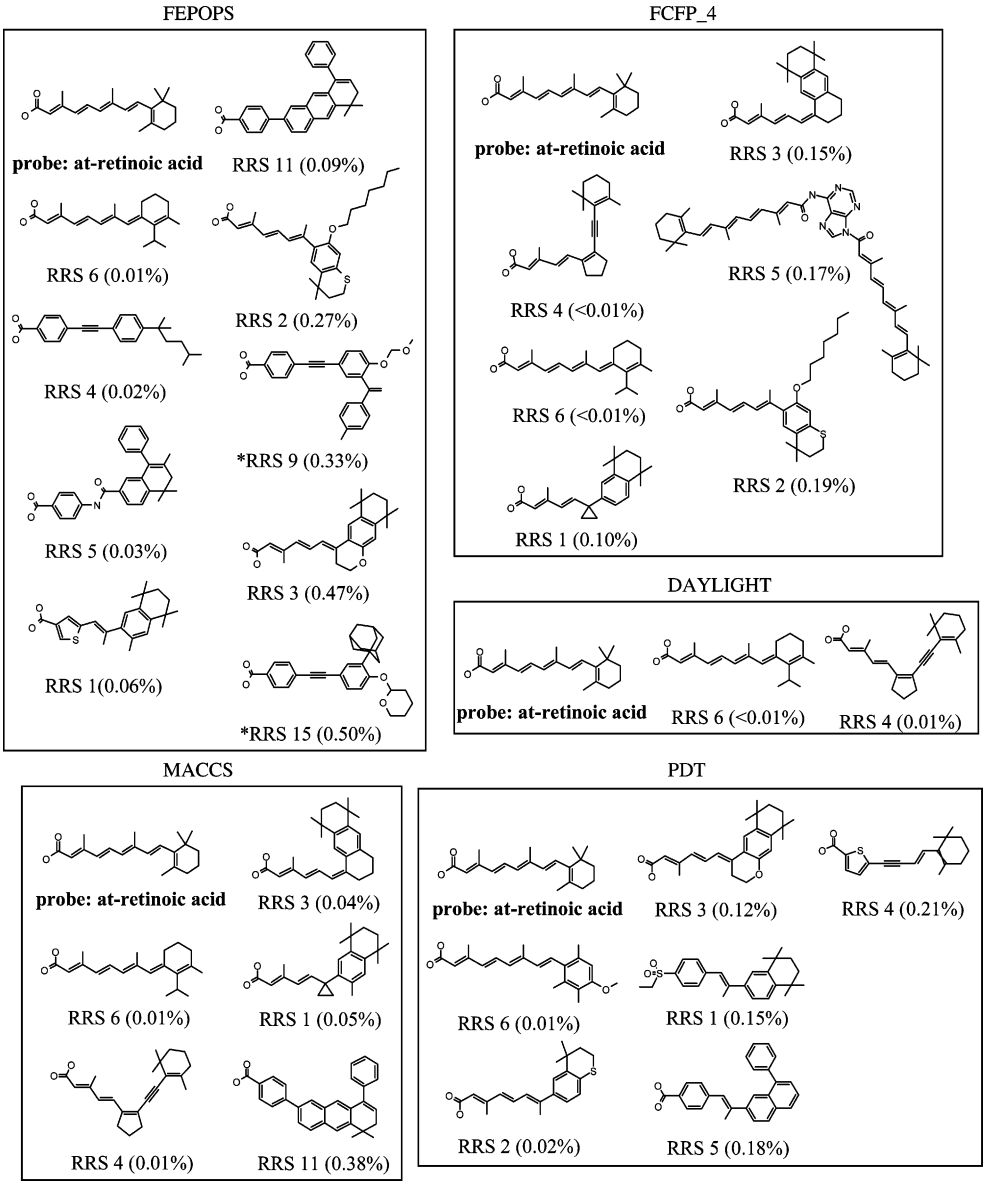


Figure 11. Representatives from retinoid RRS classes found by the similarity methods in the top 0.5% using all-trans retinoic acid as a probe. The RRS designation and percentile rank are provided. RRS classes found uniquely by a method have an asterisk.

performs particularly well compared to topological methods when an endogenous ligand is used as the similarity query. Efforts to leap from natural ligands to lead compounds have been made previously using peptides.[8,66] Natural ligands or substrates may also be used to simply prioritize compounds in a library for testing against a target. Alternatively, FEPOPS may serve as an orthogonal approach to complement 2D methods prior to screening or to recover false negatives missed by HTS, postscreening.

It is clear that 2D methods are capable of scaffold hopping to an extent and/or recalling a set of bioactive compounds with only moderate average similarity to the probe molecule.[29] Further, different 2D similarity methods return different rankings for actives. This same observation made by others has led to the paradigm that using multiple similarity methods benefits the discovery process.[2,41] Nonetheless, in several of the present cases the ring systems recovered by the 2D methods as a collective are not so divergent—if scaffold recall is more desirable than active recall, a better strategy may be to use a reliable 3D method such as FEPOPS in addition to 2D methods,[28] rather than employing consensus 2D methods alone. Additionally, the number of ligands known prior to screening should play a role in the decision of whether to use 2D- or 3D-similarity search algorithms, as a larger number of reference compounds (more than five) more readily benefits 2D methods in the current study (Figure 12).

The scaffold hopping of PDT is competitive with FEPOPS and the topological methods, especially in the D2 test case among the top 1% of ranked compounds (Figure 10). Many of the D2 binders retrieved in the top 1% share a common pharmacophore: a donor—acceptor—hydrophobe triplet, where the acceptor to hydrophobe distance is ~3 Å, as in dopamine itself. These results exemplify how useful explicit pharmacophore similarity methods can be if critical pharmacophores are known in advance of searching. Other 3D methods may perform better still than PDT and further benchmarking against FEPOPS would be useful.

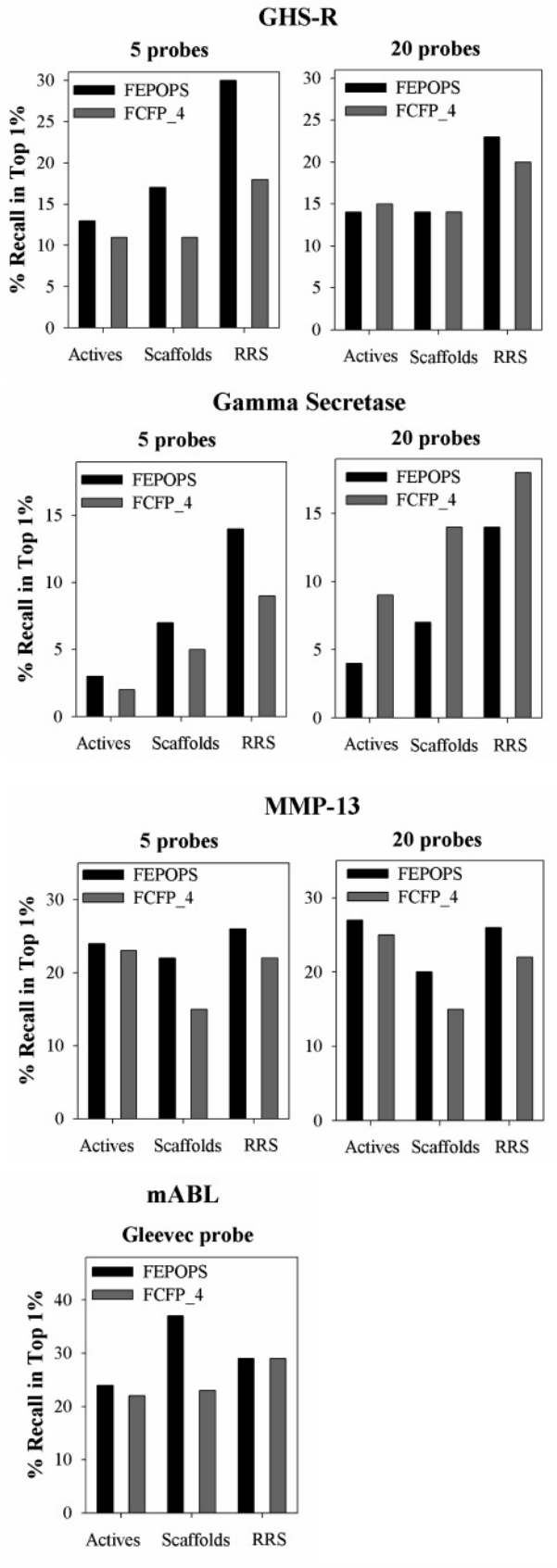


Figure 12. Recall of actives, active reduced scaffolds, and active reduced ring systems (RRS) from HTS hit lists by FEPOPS and by Pipeline Pilot functional class fingerprints. The recall is shown only for the highest ranking 1% of compounds. For cases where multiple probes are used, the similarity of a compound is measured to each of the probe molecules and the highest value is used for ranking.

One of the clear advantages of FEPOPS is the fully automated, unbiased generation of pharmacophore-type information. With published methods, the construction of a pharmacophore model often requires flexible alignment of known active compounds and hand-picking of important features. Typically, pharmacophores are composed of a triplet or quartet of single atoms possessing some critical property (e.g., hydrophobic, hydrogen-bond donor or acceptor). In contrast, each feature point in FEPOPS encodes multiple properties, since each atom with membership in a k-means cluster contributes some information. Thus, the indigenous atomic environment of the cluster members is encoded into the centroids. When performing a pharmacophore-based search, consideration of the atom neighbors of features may prevent the selection of compounds in which the features reside in the context of undesirable neighbors. The tradeoff is that the "whole-molecule pharmacophore" is not explicit and therefore encodes somewhat less specific information than traditional pharmacophores.

Although FEPOPS models contain only a small number of descriptors per compound, the information encodes a layer of complexity that may not be readily apparent. For example, a hypothetical value of >1 for the feature "L1" not only indicates the presence of a hydrophobic region in the molecule, but also reveals that the most negative portion of the molecule (corresponding to feature point 1) is hydrophobic. Conversely, a negative value for "L4" indicates that the most positively charged portion of the molecule is composed of mostly hydrophilic atoms.

A striking result from the COX-2 test case is that use of the bioactive conformation of the probe molecule rather than a representative set of flexible conformers does not improve the results and in fact does worse than the flexible approach at scaffold recovery (Figure 2). It is possible that because SC-558 is not highly rotatable, flexible similarity searching may not deteriorate the pharmacophoric information as much as a ligand with numerous torsions would. Alternatively, a rigid pharmacophore model may simply be too restrictive; for example, it may not take into account flexibility in the active site that enables induced fits of the numerous COX-2 antagonists. We have similarly observed that the bioactive confirmation of lisinopril (PDB code 1O86)[71] does not retrieve ACE inhibitors as well as a flexible lisinopril similarity search (data not shown). There is further unexpected evidence that pharmacophore models based on a single bioactive conformation are limiting: we might anticipate one or two SC-558 conformers from the flexible similarity search are predominantly most similar to all other COX-2 inhibitors (i.e., the probe conformations that best represent a common pharmacophore), but this is not the case. Each of the seven probe conformations have subsets of COX-2 inhibitors to which they are the best correlated. In other words, each explicit probe conformation contributed to the overall enrichment of actives. Intriguingly, this finding holds true for all of the test cases in this study. These results suggest that allowing probe flexibility may encourage scaffold hopping more than incorporating a priori knowledge of bioactive conformations.

FEPOPS Weaknesses. The mechanism of alignment or matching used by 3D methods is both the cornerstone

Table 3. Average Similarity[a] of the Single Probes to Their Entire Activity Class and to Recalled Actives

		average similarity of probe to actives recalled (in top 1%)		
target	ligand probe	average similarity of probe to all target actives[a]	by FEPOPS	by FCFP_4
COX-2	SC-558	0.31	0.33[b], 0.33[c]	0.41
5-HT3A	extreg 194584	0.45	0.48	0.52
HIV-RT	extreg 236942	0.39	0.45	0.66
D2	dopamine	0.26	0.24	0.32
RAR	retinoic acid	0.49	0.58	0.65
mABL	Gleevec	0.51	0.54	0.56
		$\mu = 0.40$	$\mu = 0.44$	$\mu = 0.52$

[a] Tanimoto similarities were calculated using MACCS descriptors. [b] Bioactive SC-558. [c] Flexible SC-558.

and the "Achilles' heel" of active recall. We have incorporated a relatively simple means of sorting feature points to simulate field-based alignment by sequentially assigning feature points from 1 to 4 on the basis of the sum of atomic partial charges in the *k*-means cluster. Precedent for this sorting stratagem can be found in the creation of conformation-dependent chirality descriptors.[15] Additionally, this stratagem was used successfully in parallel studies to dock ligands[51] and to identify binding sites on protein surfaces.[55] Nonetheless, we acknowledge that similarity may be unduly impacted for a congeneric series where various R-group substitutions strongly modify the quadrupole directionality. For example, in the COX-2 test case, substitution of the SC-558 trifluoro group with propionic acid causes feature point 1 to be assigned to the atomic cluster containing the acid; alignment of this analogue with SC-558 is problematic for FEPOPS. Shape-based alignment may offer an alternative way of matching feature points. However, in many cases, shape-based alignment may not reasonably align the critical pharmacophore features. Further, false negatives may also arise in shape-based alignments in a congeneric series when bulky R groups are substituted. In any case, it is important to note that actives from a congeneric series can easily be recovered by 2D methods if they are used to complement 3D approaches. Finally, it may be advantageous to forego presorting of feature points by charge and to align molecules by matching each of their four feature points independently. This strategy would result in better alignments, but a concurrent reduction would occur in the "layer of complexity" of the descriptors as described above [i.e., L1 would not *automatically* describe the hydrophobicity of the most negative portion of the molecule (feature point 1)].

The focus of this work is on the novelty of molecular representation. We have not investigated or attempted to improve on the variety of existing similarity metrics[26] that operate on binary representations of descriptors. Pearson's correlation serves as a useful benchmark in the present study for testing FEPOPS, which, incidentally, preserves the descriptors in a useful form for visual analysis (Figure 1).

Conclusions

There are several noteworthy aspects of our application. FEPOPS is unique in its approach to creating a pharmacophore from all atoms in the ligand, rather than from explicit feature pairings.[72,73] The clustered-atom representation is fuzzy (via *k*-means clustering), which promotes selection of chemotypes with regional similarities but unique frameworks. Compound preprocessing includes vital pK_a and tautomeric information often missing from other similarity methods. FEPOPS finds a small number of diverse, representative conformers that cover FEPOPS space efficiently by selecting conformations with "minimal dissimilarity" to neighboring conformations (*k*-medoids clustering). Training sets of known actives are not required for conducting similarity searches, nor is knowledge about bioactive conformations. The method precludes time-intensive molecular alignments or pharmacophore matching by a simple, yet effective electrostatic sorting of feature points that allows for direct statistical correlations of descriptors. Finally, the calculation time for similarity searches is nominal (~8 min for 100K compound database) and comparable to 2D methods once FEPOPS are precomputed (~600K compounds/week). The representations can optionally be mapped back to parent coordinates; thus, for any test compound, the conformation with the highest correlation to a known active ligand is potentially a bioactive conformation.

Acknowledgment. We would like to thank Dr. Jeff Janes and Dr. Yingyao Zhou at the Genomics Institute of the Novartis Foundation, for assistance with managing descriptors, and Dr. Jon Swanson at Tripos and Drs. Christian Parker, Greg Paris, Peter Gedeck, Mathis Thoma, Anthony Klon, and Bernhard Rohde at NIBRI, for helpful discussions.

References

(1) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem.* **1999**, *38*, 2894−2896.
(2) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.
(3) 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.
(4) Abolmaali, S. F. B.; Ostermann, C.; Zell, A. The Compressed Feature Matrix—A novel descriptor for adaptive similarity search. *J. Mol. Model* **2003**, *9*, 66−75.
(5) Cramer, R. D.; jilek, R. J.; Andrews, K. M. dbtop: Topomer similarity searching of conventional structure databases. *J. Mol. Graph. Model.* **2002**, *20*, 447−462.
(6) Lloyd, D. G.; Buenemann, C. L.; Todorov, N. P.; Manallack, D. T.; Dean, P. M. Scaffold Hopping in De Novo Design. Ligand Generation in the Absence of Receptor Information. *J. Med. Chem.* **2004**, *47*, 493−496.
(7) Bohl, M.; Dunbar, J.; Gifford, E. M.; Heritage, T.; Wild, D. J.; Willett, P.; Wilton, D. J. Scaffold searching: Automated identification of similar ring systems for the design of combinatorial libraries. *Quant. Struct.−Act. Relat.* **2002**, *21*, 590−597.
(8) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: A program for rapidly producing pharmacophorically relevent molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505−1514.

(9) Naerum, L.; Norskov-Lauritsen, L.; Olesen, P. H. Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525−1528.
(10) Abrahamian, E.; Fox, P. C.; Naerum, L.; Christensen, I. T.; Thogersen, H.; Clark, R. D. Efficient generation, storage, and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 458−468.
(11) 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.
(12) 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.
(13) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497−520.
(14) Paris, C.G. Chemical Structure Handling by Computer. In *Annual Review of Information Science and Technology*; Williams, M. E., Ed.; ASIS&T: Silver Spring, MD, 1997.
(15) Gasteiger, J.; Engel, T. *Chemoinformatics*; Wiley-VCH GmbH & Co. KGaA: Weinheim, 2003.
(16) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1999**, *39*, 2887−2893.
(17) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *45*, 5311−5320.
(18) Lipkus, A. H. Mining a large database for peptidomimetic ring structures using a topological index. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 582−586.
(19) Nilakantan, R. A ring-based chemical structural query system: Use of a novel ring-complexity heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 65−68.
(20) Feher, M.; Schmidt, J. M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(21) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.
(22) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46*, 3257−3274.
(23) Lipkus, A. H. Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 430−438.
(24) Maggiora, G. M.; Johnson, M. A. *Concepts and applications of molecular similarity*; John Wiley & Sons: New York, 1990; pp 99−117.
(25) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(26) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching [Review]. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(27) Makara, G. M. Measuring molecular similarity and diversity: Total pharmacophore diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.
(28) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(29) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? [Review]. *Drug Discovery Today* **2002**, *7*, 903−911.
(30) Lemmen, C.; Zimmermann, M.; Lengauer, T. Multiple molecular superpositioning as an effective tool for virtual database screening [Review]. *Perspectives Drug Discovery Des.* **2000**, *20*, 43−62.
(31) Raymond, J. W.; Willett, P. Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrixes. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 908−916.
(32) Rhodes, N.; Willett, P.; Calvet, A.; Dunbar, J. B.; Humblet, C. CLIP: Similarity searching of 3D databases using clique detection. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 443−448.
(33) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(34) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity searching in files of three-dimensional chemical structures − evaluation of the eva descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.
(35) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. Mimic—A molecular-field matching program—Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **1997**, *18*, 934−954.
(36) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Similarity searching in files of three-dimensional chemical structures—Flexible field-based searching of molecular electrostatic potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.
(37) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(38) Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.
(39) Pirard, B.; Pickett, S. D. Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(40) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: A flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60−66.
(41) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion [Review]. *Perspectives Drug Discovery Des.* **2000**, *20*, 1−16.
(42) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (Comma)—3D-qsar without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.
(43) Silverman, B. D. Three-dimensional moments of molecular property fields. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1470−1476.
(44) Xu, Y. J.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181−185.
(45) Xu, Y. J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.
(46) Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos. Trans. Royal Soc. London A Math. Phys. Eng. Sci.* **1896**, *187*, 253−318.
(47) Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.
(48) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativities—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(49) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
(50) Glick, M., Richards, W. G., and Robinson, D. D. (Isis Innovation) Method for binding site identification using a multiscale approach. WO2003GB00241 20030120, 2003. Great Britain. 7−31−2003. Patent
(51) Glick, M.; Grant, G. H.; Richards, W. G. Docking of flexible molecules using multiscale ligand representations. *J. Med. Chem.* **2002**, *45*, 4639−4646.
(52) Mason, J. S.; Cheney, D. L. Ligand−receptor 3-D similarity studies using multiple 4-point pharmacophores. In *Pacific Symposium on Biocomputing*, **1999**, *4*, 456−467.
(53) Pickett, S. D.; Luttman, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB—Strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.
(54) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning*; Springer: New York, 2001.
(55) Glick, M.; Robinson, D. D.; Grant, G. H.; Richards, W. G. Identification of ligand binding sites on proteins using a multiscale approach. *J. Am. Chem. Soc.* **2002**, *124*, 2337−2344.
(56) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(57) 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.
(58) Kier, L. B. *Molecular orbital theory in drug research*; Academic Press: New York, 1971.
(59) Apaya, R. P.; Lucchese, B.; Price, S. L.; Vinter, J. G. The matching of electrostatic extrema: A useful method in drug design? A study of phosphodiesterase III inhibitors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 33−43.
(60) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.
(61) Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810−818.
(62) Kaufman, L.; Rousseew, P. J. *Finding groups in data—An introduction to cluster analysis*; John Wiley & Sons: New York, 1990.

(63) R Development Core Team. R: A language and environment for statistical computing (computer program); R Foundation for Statistical Computing: Vienna, Austria, 2003.
(64) Morgan, H. L. The generation of a unique machine description for chemical structures—A technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(65) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J. J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947−955.
(66) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.
(67) Azam, M.; Latek, R. R.; Daley, G. Q. Mechanisms of autoinhibition and STI-51/Imatinib resistance revealed by mutagenesis of *BCR-ABL*. *Cell* **2003**, *112*, 831−843.
(68) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature* **1996**, *384*, 644−648.
(69) Emilian, G.; Maloteaux, J. M.; Geurts, M.; Hoogenberg, K.; Cragg, S. Dopamine receptors—Physiological understanding to therapeutic intervention potential. *Pharmacol. Ther.* **1999**, *84*, 133−156.
(70) Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. In silico discovery of novel retinoic acid receptor agonist structures. *BMC Struct. Biol.* **2001**, *1*, 1.
(71) Natesh, R.; Schwager, S. L. U.; Sturrock, E. D.; Acharya, K. R. Crystal structure of the human angiotensin-converting enzyme−lisinopril complex. *Nature* **2003**, *421*, 551−554.
(72) Jenkins, J. L.; Glick, M.; Davies, J. W. FEPOPS (feature point pharmacophores): a novel three-dimensional (3D) similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. U.S. Provisional Patent 60/546,835, February 19, 2004.
(73) Jenkins, J. L.; Glick, M.; Davies, J. W. Use of feature point pharmacophores (FEPOPS). U.S. Provisional Patent 60/559,602, April 5, 2004.

JM049654Z