How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space

Andreas Bender,*,† Jeremy L. Jenkins, Josef Scheiber, Sai Chetan K. Sukuru, Meir Glick, and John W. Davies

Center for Proteomic Chemistry, Lead Discovery Informatics, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received July 25, 2008

Different molecular descriptors capture different aspects of molecular structures, but this effect has not yet been quantified systematically on a large scale. In this work, we calculate the similarity of 37 descriptors by repeatedly selecting query compounds and ranking the rest of the database. Euclidean distances between the rank-ordering of different descriptors are calculated to determine *descriptor* (as opposed to *compound*) similarity, followed by PCA for visualization. Four broad descriptor classes are identified, which are circular fingerprints; circular fingerprints considering counts; path-based and keyed fingerprints; and pharmacophoric descriptors. Descriptor behavior is much more defined by those four classes than the particular parametrization. Using counts instead of the presence/absence of fingerprints significantly changes descriptor behavior, which is crucial for performance of topological autocorrelation vectors, but not circular fingerprints. Four-point pharmacophores (piDAPH4) surprisingly lead to much higher retrieval rates than three-point pharmacophores (28.21% vs 19.15%) but still similar rank-ordering of compounds (retrieval of similar actives). Looking into individual rankings, circular fingerprints seem more appropriate than path-based fingerprints if complex ring systems or branching patterns are present; count-based fingerprints could be more suitable in databases with a large number of repeated subunits (amide bonds, sugar rings, terpenes). Information-based selection of diverse fingerprints for consensus scoring (ECFP4/TGD fingerprints) led only to marginal improvement over single fingerprint results. While it seems to be nontrivial to exploit orthogonal descriptor behavior to improve retrieval rates in consensus virtual screening, those descriptors still each retrieve different actives which corroborates the strategy of employing diverse descriptors individually in prospective virtual screening settings.

INTRODUCTION

The 'molecular similarity principle' states that similar molecules are more likely to have similar properties than dissimilar ones.[1−5] Despite recent examples that one cannot conclude property similarity from structural similarity in every case[6] this is still the underlying assumption of current drug design efforts. Lots of descriptors exist, but little effort has until now been made to understand the 'behavior' of those descriptors relative to each other and in an absolute sense. Correlating how similar descriptors behave, such as not only in the current work, is fundamental for applications such as ligand-based virtual screening, but also in areas such as chemogenomics,[7−13] where one attempts to relate ligands classes on a larger scale, it is crucial to know about the behavior of the descriptors one employs to construct those relationships in the first place.

Both variables we are trying to relate have their weaknesses, the input variable (structural descriptors) as well as the output variables (experimental measurements). The weakness on the side of structural descriptors is that a plethora of descriptors exists, but little effort has until now been made to understand the information content of those descriptors. The weakness on the side of experimental measurements lies in the fact that many experiments are notoriously noisy, namely contaminated with false positive and false negative readouts. Also when it comes to reproducibility HTS results differ hugely between screening runs[14] due to differences in the experimental conditions and kinetic solubility measurements are hard to reproduce. Even given 'optimal' molecular descriptors it is not sure that one would be able to devise those, given the noise in the output variable, leading to predictive models which can even *in principle* only be as good as the underlying data.

While the problems with experimental data are difficult to resolve by modelers, the modeler can contribute an answer to the question which descriptors are more suitable for a given task than others. E.g., for modeling the permeability of the cell membrane one would likely choose descriptors related to logP (or an experimentally determined logP value itself) and molecular weight which capture much of the passive transport abilities of a membrane system. For virtual screening for bioactive compounds[15] structural descriptors are more likely to succeed—since ligand-target binding is not (only) determined by the size of the molecule and its lipophilicity but also by the spatial arrangement of possible interaction points in space. Accordingly, structural descriptors

* Corresponding author phone: (617)671-0667; e-mail: Andreas.Bender@cantab.net.

† Current address: Leiden/Amsterdam Center for Drug Research, Medicinal Chemistry Division, Einsteinweg 55, 2333 CC Leiden, The Netherlands.

10.1021/ci800249s CCC: $40.75
Published on Web 01/06/2009

Table 1. Description of the Data Set Used, Derived from Previously Published Work by Hert et al. and Using All Compounds from Each Previously Described Activity Class for Which *All* the Descriptors Used in the Current Study Could Be Calculated

MDDR activity index	activity class	class size
06233	5HT3 antagonists	727
06235	5HT1A agonists	805
06245	5HT reuptake inhibitors	275
07701	D2 antagonists	375
31420	renin inhibitors	1065
31432	angiotensin II AT1 antagonists	2087
37110	thrombin inhibitors	503
42731	substance P inhibitors	1023
71523	HIV protease inhibitors	652
78331	cyclooxygenase inhibitors	620
78374	protein kinase C inhibitors	336
total data set size (including negative compounds)		83793

such as circular fingerprints seem to be more appropriate for virtual screenings studies.[16]

Given the large number of molecular descriptors at hand, which is clearly at least in the thousands,[17] the questions arise as to which descriptors contain independent (orthogonal) information and which descriptors are redundant since correlated with each other. Knowledge about this descriptor behavior would in practical terms eliminate the need to test every descriptor for its suitability to model a given phenomenon—and only a set of orthogonal 'reference descriptors' could be relied upon for a given study. In virtual screening studies sometimes the overlap between hit lists retrieved by methods is presented,[18] which is a measure of similar rank-ordering of two methods or descriptors. In the most extreme case of very similar methods the rank-ordering of compounds is identical, and none of the methods offers any advantage over another. In the other extreme, the rank-orderings are completely different, and under the assumption that both methods retrieve active molecules, they are complementary to each other. In practice, most descriptors will be in between those extremes—they capture different aspects of molecular structures (which leads to deviations between the rank orderings), but they are still aiming at modeling certain physicochemical or biological properties of molecules (which leads to agreements between them).

In the current study we are examining not the question how similar *molecules* are to one another but how similar *descriptors* behave when retrieving lists of molecules. The general idea that different descriptors consider different aspects of structures has been reported before,[19−21] but it has not yet been numerically evaluated on a large scale such as in the present work. Regarding the comparison of different descriptors usually individual cases are discussed, such as when investigating the behavior of 2D vs 3D descriptors for scaffold hopping.[22] A recent publication by Zhang and Muegge[23] found that both atom pair descriptors and 3D pharmacophore fingerprints performed well in finding novel bioactive scaffolds when only unique scaffolds were used in a retrospective virtual screening experiment.

Conceptually, we randomly and repeatedly select single molecules from a database in turn as queries, which are used to rank a database of library molecules. The difference in compound ranks between the methods is used to establish a similarity measure between the *methods* (more precisely, this is the run-averaged Euclidean distance of the ranks of identical compounds to the same query, as assigned by two different descriptors). These Euclidean distances are further analyzed by principal component analysis which enables us to visualize the *differences in descriptor behavior* in a three-dimensional coordinate system.

METHODS

a) Data Set for Principal Component Analysis. 1000 compounds have been randomly selected in the weight range between 200 and 1000 Dalton from the World Drug Index 2005 (WDI)[24] using PipelinePilot 6.1.[25] However, most compounds were significantly smaller and 'druglike', with molecular weight peaking at 350 and 733 of the molecules passing the Lipinski filter in PipelinePilot. Structures were standardized (NormalizeStereo, NormalizeCharges) using the same software. To generate 3D conformations the exported SD files were imported into MOE2006.1, washed, and protonated, and 3D coordinates were generated from scratch with the 'keep stereo' option being selected. Single conformers were used for further processing (i.e., no conformational sampling was performed).

b) Data Set for Retrieval Evaluation. The second data set used to evaluate descriptor performance was based on the one published previously by Hert et al.[26,27] which was also used in other follow-up studies.[28] It contains ligands from 11 activity classes from the MDL Drug Data Report (MDDR)[29] database as described in Table 1. In our case, version 2005.3 of the database was used. Note that the previously reported size of a total of 102,535 active compounds was trimmed down in the current study to 83,793 molecules since only those compounds for which descriptors could be calculated in *all* virtual screening tools were considered (where the majority of compounds failed in the calculation of FEPOPS descriptors). In order to have a common baseline for comparison, in the current study the work performed by Hert et al. was repeated here on the various PipelinePilot fingerprints in combination with the Tanimoto Coefficient. While recently a large number of publications appeared on how to benchmark virtual screening methods to estimate their prospective value such as 'Maximum Unbiased Validation' (MUV),[30] in this work we confine ourselves to this previously used data set derived from the MDDR since it was used in a large number of studies until today.[26,28]

c) Fingerprints. 37 different fingerprints implemented in different external and internal software tools have been calculated for each structure from the above data sets. Fingerprints used are listed in Table 2 and briefly described in the following.

1. MDL/MACCS (PipelinePilot, MOE). 166 predefined keys[31] in their PipelinePilot and MOE implementations (abbreviated MDL and MACCS, respectively). Both implementations were used on the one hand to check whether implementations agree between the programs and also to have a 'negative control' in the protocols.

2. PipelinePilot xyFz_n Fingerprints (23 in Total) (PipelinePilot). Topological fingerprints as implemented in PipelinePilot 6.1. The first letter, x, represents the atom typing: F = Functional Class; E = Atom type; A = AlogP Code; S = Sybyl (mol2) atom type. The second letter, y, represents the fingerprint type: C = Extended Connectivity

Table 2. Fingerprints Used in the Comparative Study[a]

fingerprint name	implemented in (PP = PipelinePilot)	description
MDL	PP	166 predefined MDL keys (public set)
ECFP4	PP	atom type based, extended connectivity fingerprint, MaxDistance = 4
ECFP2	PP	atom type based, extended connectivity fingerprint, MaxDistance = 2
ECFP6	PP	atom type based, extended connectivity fingerprint, MaxDistance = 6
FCFP2	PP	functional class based, extended connectivity fingerprint, MaxDistance = 2
FCFP4	PP	functional class based, extended connectivity fingerprint, MaxDistance = 4
FCFP6	PP	functional class based, extended connectivity fingerprint, MaxDistance = 6
FCFC2	PP	functional class based, extended connectivity counts, MaxDistance = 2
FCFC4	PP	functional class based, extended connectivity counts, MaxDistance = 4
FCFC6	PP	functional class based, extended connectivity counts, MaxDistance = 6
ECFC2	PP	atom type based, extended connectivity counts, MaxDistance = 2
ECFC4	PP	atom type based, extended connectivity counts, MaxDistance = 4
ECFC6	PP	atom type based, extended connectivity counts, MaxDistance = 6
ACFP4	PP	AlogP type based, extended connectivity fingerprint, MaxDistance = 4
APFP4	PP	AlogP type based, path fingerprint, MaxDistance = 4
ACFC4	PP	AlogP type based, extended connectivity counts, MaxDistance = 4
APFC4	PP	AlogP type based, path counts, MaxDistance = 4
FPFP4	PP	functional class based, path fingerprint, MaxDistance = 4
FPFP6	PP	functional class based, path fingerprint, MaxDistance = 6
SPFP4	PP	Sybyl type based, path fingerprint, MaxDistance = 4
SPFP6	PP	Sybyl type based, path fingerprint, MaxDistance = 4
EPFP4	PP	atom type based, path fingerprints, MaxDistance = 4
EPFP6	PP	atom type based, path fingerprints, MaxDistance = 6
SEFP4	PP	Sybyl type based path fingerprints, MaxDistance = 4 (PP implementation of atom environments[28,47])
ESshape3D[b]	MOE	EigenSpectrum shape fingerprint
ESshape3D_HYD[b]	MOE	EigenSpectrum shape fingerprint considering hydrogen atoms
GpiDAPH3	MOE	graph-based 3-point pharmacophores, atom typing (in pi system, is donor, is acceptor)
MACCS	MOE	166 predefined MDL keys (public set)
TAD	MOE	typed atom distances, atom typing (Donor, Acceptor Polar, Anion, Cation, Hydrophobe)
TAT	MOE	typed atom triangles, atom typing (Donor, Acceptor Polar, Anion, Cation, Hydrophobe)
TGD	MOE	typed graph distances, atom typing (Donor, Acceptor Polar, Anion, Cation, Hydrophobe)
TGT	MOE	typed graph triangles, atom typing (Donor, Acceptor Polar, Anion, Cation, Hydrophobe)
piDAPH3	MOE	(spatial) 3-point pharmacophores, atom typing (in pi system, is donor, is acceptor)
piDAPH4	MOE	(spatial) 4-point pharmacophores, atom typing (in pi system, is donor, is acceptor)
Unity	Sybyl	Unity fingerprints (various contributions)
Atom Counts[18]	PP (Script)	12-dimensional vector counting total number of atoms and number of atoms per elemental type

[a] A total of 37 different fingerprints were employed derived using a variety of different underlying concepts. [b] No reasonable retrieval rates were obtained from ESshape3D fingerprints which is likely due to not identifiable errors in our data processing procedures. While listed here, ESshape3D fingerprints were thus left out of the further analysis.

Fingerprints; P = Path Fingerprint; E = Atom Environment. The third letter, z, represents either the presence/absence of a feature (P = Fingerprint) or feature counts (C = Count). The parameter n describes the extension of the fingerprints, which is the diameter of the feature in case of Extended Connectivity Fingerprints and the largest path length in case of path-based fingerprints. (See Table 2 for complete list.)

3. ESshape3D, ESshape3D_HYD (MOE). EigenSpectrum shape fingerprints as implemented in MOE. (Method has been excluded from further analysis due to insignificant enrichment achieved with the Euclidean Distance measure employed here, which was overall only marginally better than random retrieval.)

4. GpiDAPH3 (MOE). Graph-based three-point pharmacophores employing any set of three possible atom types (in pi system, is donor, is acceptor).

5. TAD, TAT, TGD, TGT (MOE). Typed atom distances (TAD; similar to atom pairs), typed atom triangles (TAT), typed graph distances (TGD), and typed graph triangles (TGT). Six different atom types are possible which are Donor, Acceptor Polar, Anion, Cation, and Hydrophobe.

6. piDAPH3, piDAPH4 (MOE). Spatial three-point (piDAPH3) and four-point pharmacophores (piDAPH4). Three different features are possible per atom, which are in pi system, is donor, is acceptor.

7. Unity (Sybyl). Fingerprint of 992 bits length containing both predefined features and hashed paths.

8. Atom Counts (MOE Counts of Individual Atom Types). 12-Dimensional vectors containing for each molecule the total number of atoms; the number of heavy atoms; and the numbers of boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus, and sulfur atoms. This 'descriptor' has been published previously,[18] and it has been shown that on some standard virtual screening data sets its performance was, on average, about half as good as that of established methods.

9. Similog Keys (Novartis Implementation). Similog keys are graph-based three-point pharmacophores considering

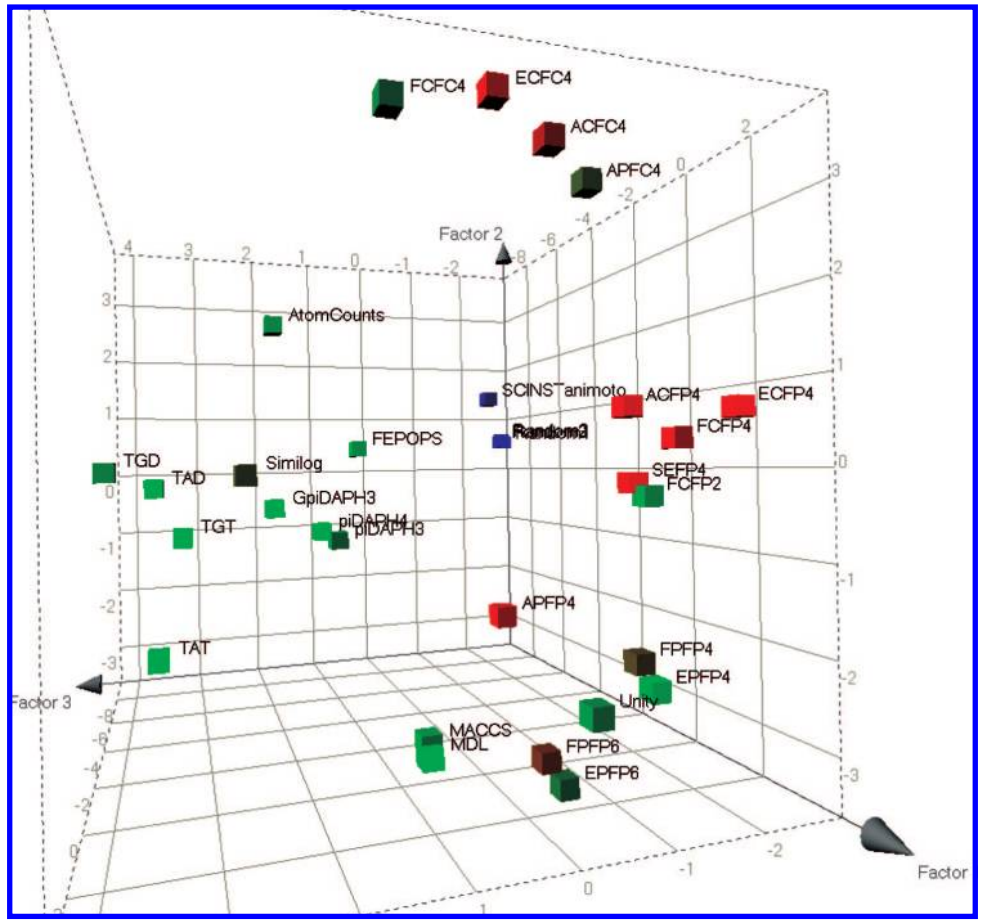

Figure 1. Principal component analysis of the rank-ordering of 37 descriptor spaces plus random selection. The size of each data points is determined by the performance (retrieval rate) of the method (red = high, green = medium, blue = low). Nearly half of the cumulative variance is captured in the first three principal components shown here, thus also descriptors located close to each other can show quite some difference in rank-ordering compounds.

count information.[7] Each triplet of atoms is characterized by the graph distances as the constituent atom types. Atoms are typed according to hydrogen bonding capabilities, electronegativity, and steric requirements. This descriptor has been found to be able to perform 'chemogenomics'-type virtual screening, when ligands from one member of a target family are employed to find ligands for another member of the same target family.

10. SCINS (Novartis Implementation). SCINS ('Scaffold Class Identification and Naming System') describes a reduced graph of a scaffold by a string of numbers in the format ABCDE-FGHI-JKLM[32] which considers both ring characteristics and binned chain lengths. These numbers stand for the following: A. number of chain assemblies (= linkers between ring assemblies). B. number of all chains. C. number of rings. D. number of ring assemblies. E. number of bridge bonds. F. number of ring assemblies consisting of exactly one ring. G. number of ring assemblies consisting of exactly two rings. H. number of ring assemblies consisting of three or more rings. I. number of macrocycles. J.−M. binned length of n-th shortest chain ($n = 1-4$). If the binned length of the n-th shortest chain exists, it is used; otherwise, it is zero. Four bins are used which are 1−2 bonds, 3−4 bonds, 5−7 bonds, or more than 7 bonds.

The purpose of this 'descriptor' was both not only to get a feeling for how target-specific a description of the underlying molecular frameworks already is but also to get a performance evaluation of each virtual screening method with regard to the different framework/SCINS retrieved.

11. FEPOPS (Novartis Implementation). The FEPOPS descriptor is a three-dimensional descriptor developed by Jenkins et al.[33] that is constructed as follows (here the standard protocol described in the original presentation of the descriptor was followed). First, tautomers as well as conformers are generated for each molecule. A set of four putative pharmacophoric interaction points is assigned to capture a fuzzy definition of the interaction capabilities of a structure, followed by clustering of the tautomers and conformers based on pharmacophore centers. A set of cluster centers is chosen to describe each compound, and comparison between compounds is based on the comparison of the distances and charges of the pharmacophoric interaction points which are sorted according to charge.

12. Random 1, 2, 3. Negative control experiment, selecting a random set of compound in each run. Expected are a retrieval of actives in agreement with random selection. Additionally, random selection provided a reference point to compare the similarity and differences in retrieval behavior between the other descriptors used (see Figures 1 and 2).

13. Cosine Coefficient with MDL, Unity, and ECFP4 Fingerprints. The Tanimoto Coefficient has been replaced by the Cosine Coefficient to compare both retrieval of actives as well as rank-ordering of compounds (see Computational Details below for a description of those values).

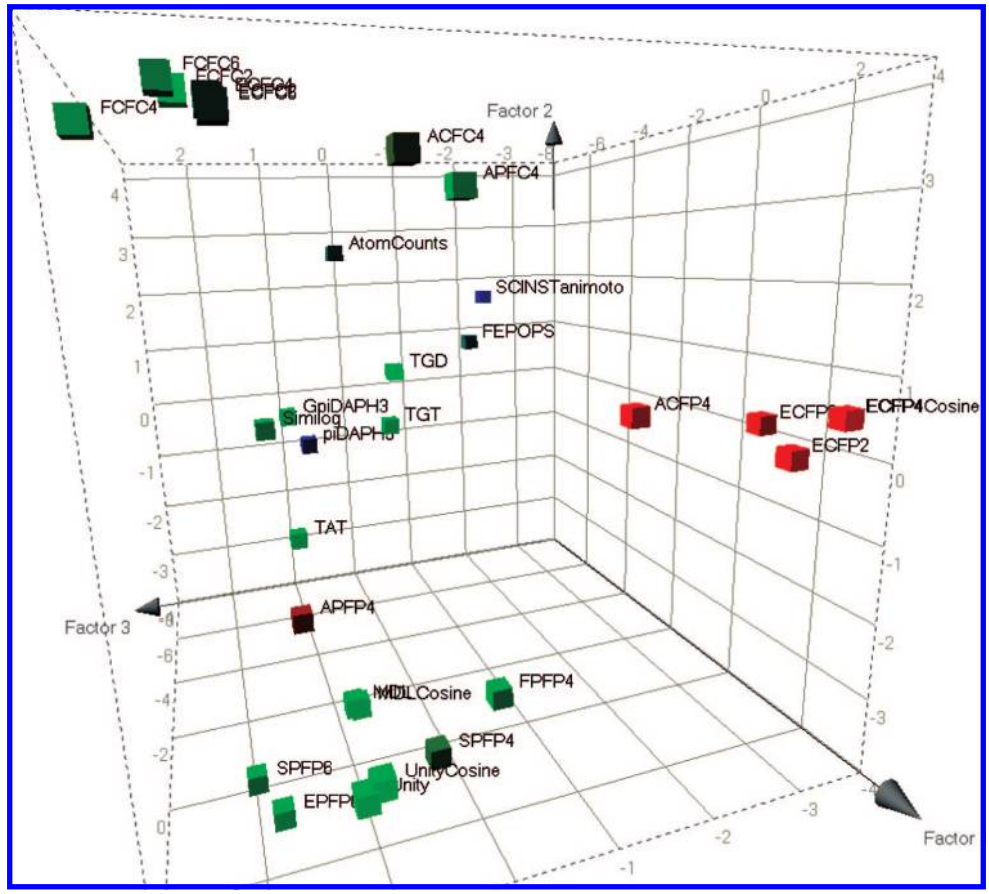

Figure 2. Principal component analysis of 37 diverse descriptor spaces without random compound selection, based on the similarity of their compound rank-ordering. The color of each data point is determined by the performance (retrieval rate) of each method (red = high, green = medium, blue = low). Nearly half (47.68%) of the cumulative variance is captured in the first three principal components shown here, thus also descriptors located close to each other can show quite some difference in rank-ordering compounds.

d) Computational Details. *1. Analysis of Descriptor Rank-Ordering.* 100 times a random structure out of the 1000 compounds has been picked as a query, and the remaining 999 compounds were ranked according to similarity to the query using the Tanimoto coefficient as implemented in PipelinePilot 6.1. Rank positions were assigned to every ligand in every descriptor space, and the original sort order has been established, thus annotating every compound with its ranking position using the different similarity metrics. Identical similarity values of two molecules were assigned random ranks within the given range. The Euclidean distance between the rank vectors of two methods was used to define the distance between two methods for a particular compound. The final distance between two methods was the mean of all Euclidean distances obtained from the rank position assignments for 100 query/database splits.

In addition to using the 37 well-established descriptors as well as simple 'atom counts', three random rankings have been performed in parallel and also been repeated 100 times. The hypothesis was that, averaged over all runs, the random rankings should occupy a distinct area in 'methods space' than the actual descriptors, since ranking with reasonable descriptors should precisely introduce nonrandom ranking results.

2. Retrieval Assessment. For the MDDR data set shown in Table 1, random query compounds were selected 10 times from each of the 11 active sets, and the number of compounds annotated with the same activity class in the top 5% of the ranked database were counted. (This is a higher fraction of the database than would be used in practice; however, this is the protocol published previously using this particular data set[26] and on which also other studies have been performed.[28]) Analogously, the number of SCINS was calculated as a measure of chemotype retrieval.

e) Principal Component Analysis. In order to obtain a visualization principal component analysis of the distance matrix was performed in Statistica 6.0 and visualization in Spotfire.[34] In addition the datapoints representing methods were colored according to the retrieval of the method (red = high, green = medium, blue = low; for full data see Table 3 and the Supporting Information).

RESULTS

The first part of the current study was to investigate the correlation of molecular descriptors in repeatedly rank-ordering a database, using identical queries in every ranking. Figure 1 visualizes the location of each of the 37 descriptors in '3D chemical descriptor space'. It can be seen that overall all compound rankings are significantly apart from random ranking, for which all three random selections of compounds are similarly unrelated to any of the descriptors considered. What is also apparent is that 'good' descriptors (those showing high retrieval, visualized by large data points) are spread throughout the plot, corresponding to different actives recovered. Figure 2 contains the same set of descriptors but excluding the random selection of compounds as a fix point. (Visualization of this large number of datapoints is not trivial,

Table 3. Factor Coordinates of Descriptors after Principal Component Analysis[a]

descriptor	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	retrieval
MDL	0.1794	−3.0336	0.9608	−0.8398	2.1496	31.15%
ECFP4	2.7909	0.4946	−3.4581	−0.4072	−2.4427	40.66%
ECFP2	2.8193	0.1252	−2.8665	−0.5951	−1.6525	40.00%
ECFP6	1.7357	0.3276	−3.0789	−0.0787	−2.3537	39.51%
FCFP2	1.7772	−0.4283	−2.7161	−0.0031	2.9779	33.00%
FCFP4	2.1951	0.2027	−2.9583	0.0265	2.9273	37.96%
FCFP6	1.6752	0.2548	−2.8260	0.1790	2.5988	38.91%
FCFC2	1.3915	4.0345	2.0698	0.3334	1.4542	33.00%
FCFC4	2.1678	3.6668	2.6779	−0.0196	2.3292	34.24%
FCFC6	1.9205	4.0508	2.1535	0.2969	1.3475	34.64%
ECFC2	2.2756	3.7953	1.8214	0.1814	−1.6325	36.60%
ECFC4	2.4628	3.8117	1.8512	0.1843	−1.7442	37.91%
ECFC6	2.4690	3.7529	1.8349	0.1734	−1.7533	37.61%
ACFP4	1.8358	0.6365	−1.6123	0.1042	0.6946	39.80%
APFP4	0.1020	−1.7669	1.4059	0.8340	−1.8444	37.93%
ACFC4	2.5386	3.4283	0.5736	0.1569	0.1863	36.92%
APFC4	2.1258	3.1431	0.0654	0.5132	−0.3027	35.12%
FPFP4	1.9058	−2.4652	−0.2112	0.7100	−0.3772	35.47%
FPFP6	0.9194	−3.4787	1.6065	0.9511	−1.4903	36.02%
SPFP4	1.9495	−3.0135	0.4055	0.6364	−1.0960	36.69%
SPFP6	0.1079	−3.8288	1.9803	0.9756	−2.0052	35.21%
EPFP4	2.2032	−2.8010	0.2501	0.8506	−0.6835	32.35%
EPFP6	1.0816	−3.8214	1.7816	1.0788	−1.7528	34.63%
SEFP4	1.5610	−0.2693	−2.2655	0.0438	1.6452	40.22%
GpiDAPH3	−4.4533	0.0609	1.1086	2.2538	0.5485	31.07%
MACCS	−0.0362	−2.9271	0.8578	−0.7673	2.0926	33.65%
TAD	−4.8239	0.4824	−0.1615	−3.7714	−0.6913	27.93%
TAT	−3.2139	−1.6736	1.1743	−3.2415	−0.7989	30.18%
TGD	−3.8734	0.7068	−0.2434	−4.6140	−0.2210	32.73%
TGT	−4.1363	−0.2097	−0.1900	−3.4339	−0.0374	29.90%
piDAPH3	−4.7108	−0.4852	0.8464	3.2935	0.5030	19.15%
piDAPH4	−5.0630	−0.3054	−0.0701	1.9049	0.5240	28.21%
Unity	3.0193	−2.6955	1.1951	0.0605	1.3261	33.95%
Similog	−1.6217	0.3679	1.5218	−2.0761	−1.1712	35.22%
AtomCounts	−5.0438	2.5921	0.3046	0.0358	0.6589	22.48%
UnityCosine	2.9797	−2.6002	1.0518	0.1830	1.3604	33.80%
ECFP4Cosine	2.6367	0.4610	−3.5825	−0.2060	−2.4957	40.32%
MDLCosine	0.1353	−3.0926	0.9041	−0.7237	2.1020	31.42%
FEPOPS	−6.2369	0.8726	−1.7836	0.3585	−0.2521	22.48%
SCINSTanimoto	−7.7484	1.6276	−2.3789	4.4579	−0.6272	12.57%

[a] Excluding Random selection; the first four principal components are visualized in Figure 2.

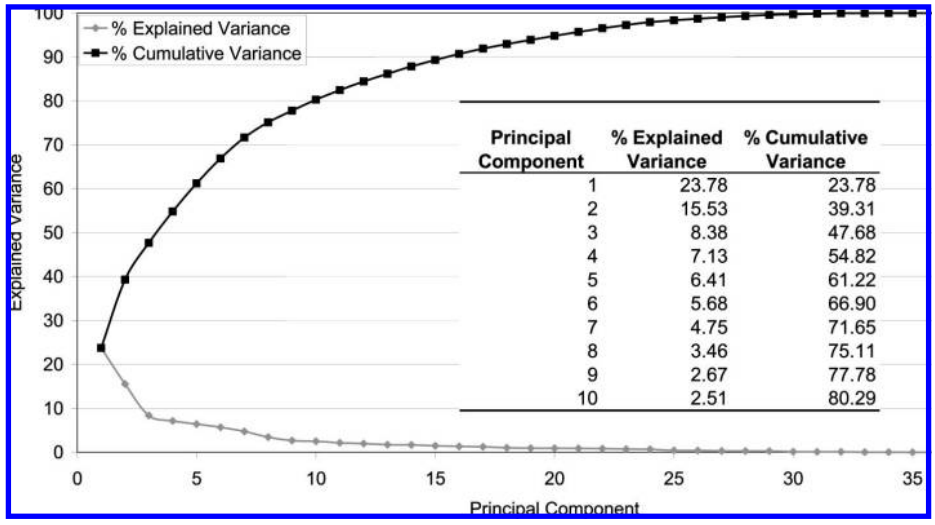

Figure 3. Scree plot of the principal component analysis of the 37-dimensional descriptor space. The first three principal components capture nearly 50% of the total variance, while for 80% of the total variance ten principal components are required.

and in both Figures 1 and 2 a small set of datapoints has been deleted for visualization reasons; the full coordinate set corresponding to Figure 2 is given in Table 3 for the first 5 principal components as well as the enrichment values, full data in the Supporting Information). Figure 3 shows the scree plot of the principal component analysis visualized in Figure 2. The first three principal components capture nearly 50% of the total variance, while for 80% of the total variance

Table 4. Retrieval Rates for the 37 Different Descriptors Evaluated in the Current Study Across 11 Activity Classes[a]

activity class	MDL	ECFP2	ECFP4	ECFP6	FCFP2	FCFP4	FCFP6	FCFC2	FCFC4	FCFC6	ECFC2	ECFC2	ECFC6	LCFC4	LPFP4
06233	27.76%	31.03%	30.56%	29.06%	22.59%	24.33%	26.57%	22.59%	27.63%	25.64%	31.77%	29.67%	27.96%	23.80%	22.61%
06235	24.19%	28.98%	28.88%	26.94%	21.37%	22.72%	23.12%	21.37%	18.96%	18.09%	27.76%	29.73%	29.34%	23.47%	25.78%
06245	20.98%	22.87%	22.22%	19.60%	26.15%	26.40%	25.64%	26.15%	25.93%	24.76%	25.82%	24.91%	23.16%	28.91%	24.58%
07701	26.43%	29.49%	27.12%	24.32%	28.56%	32.27%	29.20%	28.56%	17.52%	16.67%	23.79%	23.92%	23.15%	19.89%	22.13%
31420	64.10%	92.57%	93.83%	92.90%	71.01%	91.74%	92.17%	71.01%	89.16%	90.07%	78.85%	80.80%	81.10%	81.56%	96.64%
31432	40.57%	57.82%	62.27%	61.61%	47.78%	54.04%	55.13%	47.78%	39.44%	41.45%	49.20%	56.10%	57.39%	57.15%	59.43%
37110	46.52%	53.38%	52.58%	50.50%	46.64%	51.45%	51.75%	46.64%	38.43%	39.07%	36.64%	37.04%	36.04%	43.98%	48.61%
42731	25.05%	26.01%	28.19%	29.58%	17.96%	21.32%	23.77%	17.96%	27.63%	29.32%	29.01%	30.33%	31.00%	29.14%	26.00%
71523	38.05%	59.36%	62.90%	63.11%	45.80%	56.18%	63.02%	45.80%	49.57%	52.13%	51.23%	55.74%	56.84%	53.60%	58.07%
78331	13.16%	14.23%	13.68%	12.95%	13.39%	13.69%	14.37%	13.39%	18.56%	19.34%	19.60%	19.94%	19.44%	19.69%	10.11%
78374	15.83%	24.26%	25.03%	24.05%	21.76%	23.45%	23.30%	21.76%	23.81%	24.46%	28.90%	28.84%	28.27%	24.94%	23.27%
average	31.15%	40.00%	40.66%	39.51%	33.00%	37.96%	38.91%	33.00%	34.24%	34.64%	36.60%	37.91%	37.61%	36.92%	37.93%

activity class	LCFP4	LPFC4	EPFP4	EPFP6	SPFP4	SPFP6	FPFP4	FPFP6	SEFP4	MACCS	TAD	TAT	TGD	TGT
06233	28.05%	23.00%	20.83%	19.63%	22.81%	18.90%	17.98%	16.57%	30.25%	30.88%	25.42%	21.57%	29.89%	22.50%
06235	25.37%	24.62%	23.64%	23.37%	30.26%	27.33%	23.95%	22.52%	30.53%	30.21%	20.73%	24.62%	23.29%	19.38%
06245	24.18%	21.93%	22.18%	21.71%	23.31%	19.71%	24.15%	23.42%	25.67%	26.76%	18.44%	16.98%	22.33%	21.67%
07701	27.52%	16.56%	22.61%	24.13%	31.79%	26.35%	32.88%	28.85%	30.29%	29.81%	28.24%	23.41%	27.09%	21.71%
31420	92.81%	78.70%	78.94%	91.62%	91.00%	92.15%	84.08%	91.39%	96.38%	64.52%	63.69%	63.78%	82.14%	88.57%
31432	58.01%	56.46%	42.47%	51.70%	46.02%	53.20%	50.24%	54.10%	52.98%	47.69%	33.32%	64.08%	51.31%	32.38%
37110	48.65%	39.40%	49.05%	48.41%	51.47%	46.30%	56.42%	54.37%	51.99%	46.16%	25.98%	23.68%	38.55%	35.29%
42731	26.47%	28.89%	32.75%	32.28%	31.04%	31.97%	24.78%	27.01%	24.71%	25.12%	20.01%	18.50%	15.83%	11.64%
71523	67.24%	57.01%	33.17%	39.00%	46.35%	46.99%	41.01%	46.70%	62.01%	38.83%	44.29%	41.52%	40.78%	50.92%
78331	15.16%	15.56%	10.66%	9.27%	9.76%	7.68%	12.00%	9.40%	14.79%	14.94%	12.58%	12.68%	13.82%	9.34%
78374	24.38%	24.23%	19.58%	19.76%	19.82%	16.76%	22.65%	21.88%	22.80%	15.18%	14.52%	21.16%	15.03%	15.51%
average	39.80%	35.12%	32.35%	34.63%	36.69%	35.21%	35.47%	36.02%	40.22%	33.65%	27.93%	30.18%	32.73%	29.90%

activity class	GpiDAPH3	piDAPH3	piDAPH4	SIMILOG	Unity	SCINS	FEPOPS	AtomCounts	Random1	Random2	Random3	ECFP4_ Cos	MDL_ Cos	Unity_ Cos
06233	28.42%	10.50%	18.65%	29.35%	20.04%	14.55%	34.40%	26.31%	4.37%	5.19%	4.87%	30.03%	27.41%	19.68%
06235	19.24%	11.71%	21.73%	24.20%	21.38%	9.14%	16.57%	12.61%	5.14%	5.24%	4.89%	28.53%	23.71%	20.94%
06245	20.44%	8.22%	21.42%	27.05%	24.69%	9.67%	23.09%	14.58%	4.73%	4.69%	5.31%	21.85%	20.69%	24.29%
07701	16.03%	7.60%	22.00%	25.41%	20.11%	7.36%	17.84%	15.28%	5.31%	4.99%	4.93%	26.61%	26.67%	20.05%
31420	85.58%	54.31%	73.41%	85.73%	87.50%	25.12%	31.62%	67.54%	5.13%	4.94%	5.10%	93.26%	64.76%	87.16%
31432	53.87%	26.89%	34.17%	50.87%	49.27%	13.22%	35.52%	15.15%	4.93%	4.99%	5.26%	61.92%	41.28%	49.30%
37110	26.30%	18.47%	26.16%	35.84%	42.72%	8.57%	31.25%	27.32%	5.25%	4.55%	5.01%	52.01%	47.81%	42.27%
42731	15.43%	8.64%	19.41%	21.10%	28.94%	10.79%	17.53%	11.82%	5.11%	4.77%	4.83%	28.50%	25.31%	29.05%
71523	41.84%	39.49%	47.73%	50.31%	48.47%	18.63%	15.97%	33.13%	4.69%	4.91%	4.46%	62.35%	39.02%	48.94%
78331	13.69%	10.05%	11.53%	15.71%	9.68%	8.35%	10.24%	13.23%	4.98%	5.29%	5.21%	13.44%	13.08%	9.60%
78374	20.92%	14.73%	14.11%	21.88%	20.63%	12.80%	13.27%	10.27%	5.15%	4.97%	5.48%	25.03%	15.92%	20.48%
average	31.07%	19.15%	28.21%	35.22%	33.95%	12.57%	22.48%	22.48%	4.98%	4.96%	5.03%	40.32%	31.42%	33.80%

[a] In addition to the actual descriptor performances also 3 random runs (as a 'negative control') and the Cosine coefficient instead of the Tanimoto Coefficient were used.

ten principal components are required. It should thus be kept in mind that the principal component analysis, while reducing dimensionality of descriptor space, also needs to be taken with care since only about half of the variance between descriptors is displayed in Figures 1 and 2.

From these Figures (1 and 2) the following observations can be made:

(a) The extension of PipelinePilot descriptors (of 2, 4, and 6 bonds) does not play a major role in the ranking of compounds. Otherwise identical descriptors cluster in very similar areas of 'descriptor space', mainly in two large clusters around the top right corner of the 2D representation of the plot and in the middle-right position (Figure 2). This is most profound for ECFC2, ECFC4, and ECFC6 descriptors with virtually identical behavior, slightly less so for ECFP2, ECFP4, and FCFP6 fingerprints, and the least (in relative terms) for the analogous series of (path-based) SPFP descriptors.

(b) Using the Cosine instead of the Tanimoto Coefficient leads to very minor variations in both retrieval rates and rank ordering of compounds. Data points for ECFP4 fingerprints and MDL keys as well as Unity fingerprints using either of the above similarity coefficients are at very similar positions in descriptor space (Figure 2).

(c) For MOE descriptors, three-point and four-point pharmacophores show overall similar rank-ordering behavior (piDAPH3, piDAPH4; Figure 1), while piDAPH4 pharmacophores show significantly higher retrieval rates (Table 4).

(d) Typed atom triangles (TAT), typed graph triangles (TGT), typed graph distances (TGD), and typed atom distances (TAD) show overall similar behavior, with TAT being most dissimilar from the other descriptors (Figure 1). Similog keys retrieve similar compounds to GpiDAPH3 (and to a lesser extent TGT) descriptors, probably since they are all based on typed, graph-based sets of three atoms each. The proximity of FEPOPS descriptors as well as graph-based and spatial pharmacophores in this region of descriptor space might be surprising (Figure 1).

(e) Atom Counts, SCINS, and count fingerprints (ECFC series) show similar behavior in the sense that they all, as size-dependent descriptors, are characteristic for factor 2 of the principal component analysis. Hence, this axis could be described as the 'size axis' of descriptor space.

(f) Using counts instead of fingerprints makes a major difference in retrieving compounds. This can be seen from the two clearly distinct clusters of PipelinePilot descriptors, one of which contains the fingerprint descriptors (e.g., ECFP4), while the other contains the equivalent count descriptors (e.g., ECFC4). This is in line with earlier research[35] and could enable the use of both binary and count vectors in data fusion studies. The influence of using descriptors that consider size vs their size-oblivious counterparts has been discussed in detail recently[36] for conventional QSAR models where size-intensive descriptors (which are normalized with respect to the size of the molecule) were often shown to lead to better models than their size-extensive counterparts.

(g) Using path-based fingerprints instead of extended connectivity fingerprints makes a major difference to the compounds retrieved (e.g., ECFP4 vs EPFP4 fingerprints and so on). This is true for all path-based fingerprints considered, which are APFP4, FPFP4, EPFP4, and SPFP4. Overall, going from extended connectivity fingerprints to path-based fingerprints makes more difference to the rank-ordering of compounds than the atom typing used.

(h) Changing atom typing from Sybyl (S) to atom type (E) does not have a major influence in case of PipelinePilot fingerprints, while going from either to AlogP atom types (A) or functional class fingerprints (F) makes some difference.

(i) The MACCS/MDL implementations in MOE and PipelinePilot give very similar, however not identical rankings.

Previous work examined the influence of using counts (instead of binary presence/absence values) as well as overlap between atom type definitions (fuzzy atom types).[37] In this previous work, the correlations between the Tanimoto similarities of all molecule pairs in the set used were found to be as follows: MACCS key vs MACCS counts had an $r^2 = 0.70$; TGT vs TGT counts an $r^2 = 0.81$, TGT vs fuzzy TGT an $r^2 = 0.41$, and TGT vs fuzzy count TGT an $r^2 = 0.33$. As confirmed in this (and previous[38]) work, considering counts of features and not only their presence or absence makes a large difference with respect to the ranked list of compound databases. Interestingly, when comparing ECFP4 and ECFC4 fingerprints over a wide variety of activity classes with respect to their hit rates (as done below), using only presence/absence values performed superior on this data set. This, in case it holds generally true, would mean that molecular size (corresponding to using feature counts) is of less importance in case of circular fingerprints, than in case of e.g. CATS autocorrelation descriptors[35] where 'holographic' (count-based) descriptors perform superior to presence/absence of features. This may be due to the very different dimensionality of the descriptor space, which is only 150 for CATS descriptors, while it is virtually unlimited for circular fingerprints (so 'descriptor saturation' is less likely to occur).

We can conclude that four large groups of descriptors can be distinguished. These are

1. Extended connectivity fingerprints using the presence/absence of features (with any atom typing, any diameter; also SEFP4/Molprint2D fingerprints belong to this category; Figure 1),

2. Extended connectivity fingerprints using counts (with any atom typing, any diameters; contributing to principle factor 2; Figures 1 and 2),

3. Spatial pharmacophores (all MOE pharmacophoric descriptors used here, Similog keys, FEPOPS; also to a lesser degree atom counts; cluster in the center/back in Figures 1 and 2),

4. Path-based fingerprints and predefined keys (MACCS, all xPFP4 descriptors from Pipelinepilot, Unity fingerprints; bottom of the plots in Figures 1 and 2),

In other words, the membership of a descriptor to one of the above classes—circular fingerprints, circular fingerprints considering counts, path-based/keyed fingerprints, or pharmacophore-based fingerprints—is much more defining the behavior of the ranked list of compounds than the particular parametrization, such as atom typing, whether 3-point or 4-point pharmacophores are used, or the particular diameter of a circular fingerprint.

Overall, the above plot can be used if it should be established where similarity searching methods are likely to behave similar or different. This is relevant not only for data fusion methods (as used later in this study) but also simply to the question of how to choose parameters for a similarity searching methods. In the next part of the study, we were examining the retrieval rates of descriptors to be able to select a set of descriptors that behave both orthogonally and also show high retrieval rates for later consensus scoring calculations.

Figure 4 (with numerical values presented in Table 4) shows the retrieval rates for the 37 similarity searching methods. It can be seen that circular fingerprints generally show good performance on this data set, with graph-based methods generally outperforming their 3D counterparts (e.g., TAD vs TGD, TAT vs TGT). This is only the case by a larger margin though in case of GpiDAPH3 vs piDAPH3 fingerprints. Four-point pharmacophores (piDAPH4) in turn perform very well on this data set and outperform their 3-point counterparts (piDAPH3) significantly. While the performance of four-point pharmacophores has been described previously in discriminating actives against closely related targets,[39] the better performance compared to three-point pharmacophores on virtual screening sets has not been reported before. One reason might be that 3-point pharmacophores show better performance with smaller bin sizes (more bins) than 4-point pharmacophores—a variable we did not examine in the current work. Exchanging the Tanimoto Coefficient for the Cosine Coefficient made only very minor differences for the retrieval, in agreement with the high correlation of ranked compound lists (Figure 2).

While retrieval of actives is a necessary requirement for a virtual screening method, retrieval of different chemotypes is another crucial benchmark in this area.[40] In the current study, we employed SCINS as described in the methods section to evaluate chemotype enrichment in the ranked database which is shown in Figure 5. Contrary to what one might have assumed (but in line with the experience of the authors), circular fingerprints in addition to retrieving a large number of actives also show good scaffold retrieval (examples of which are the ECFP4, LCFP4, and SEFP4/Molprint2D fingerprints). It is interesting to see that while random retrieval of compounds only finds, as expected, about 5% of the active compounds in a random set the size of 5%

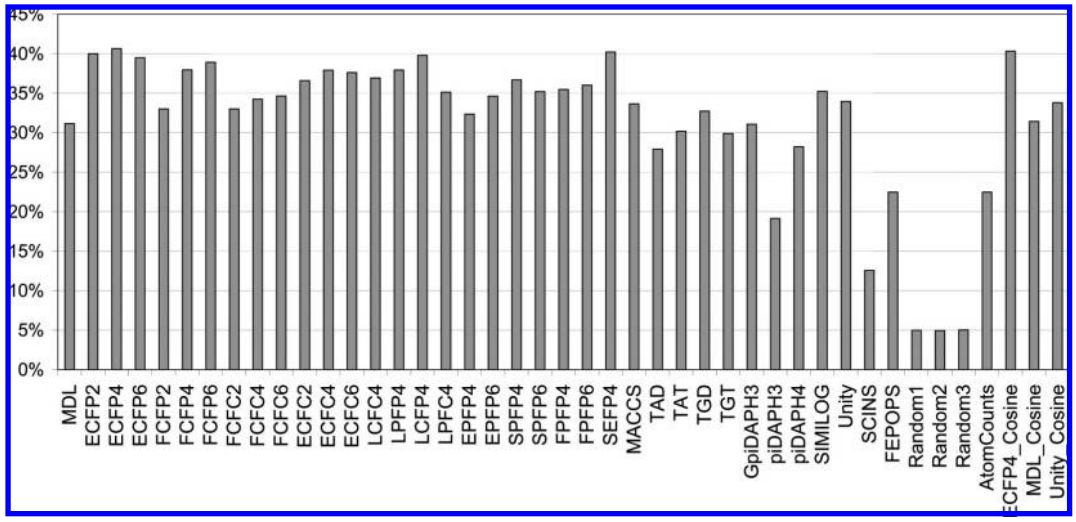

Figure 4. Retrieval rates for 37 similarity searching methods, random selection, and for the Cosine Coefficient. It can be seen that circular fingerprints generally show good performance on this data set, with graph-based methods generally outperforming their 3D counterparts (e.g., TAD vs TGD, TAT vs TGT), though only in the case of GpiDAPH3 vs piDAPH3 by a larger margin. Four-point pharmacophores (piDAPH4) in turn perform very well on this data set and outperform their 3D counterparts (piDAPH3) significantly.

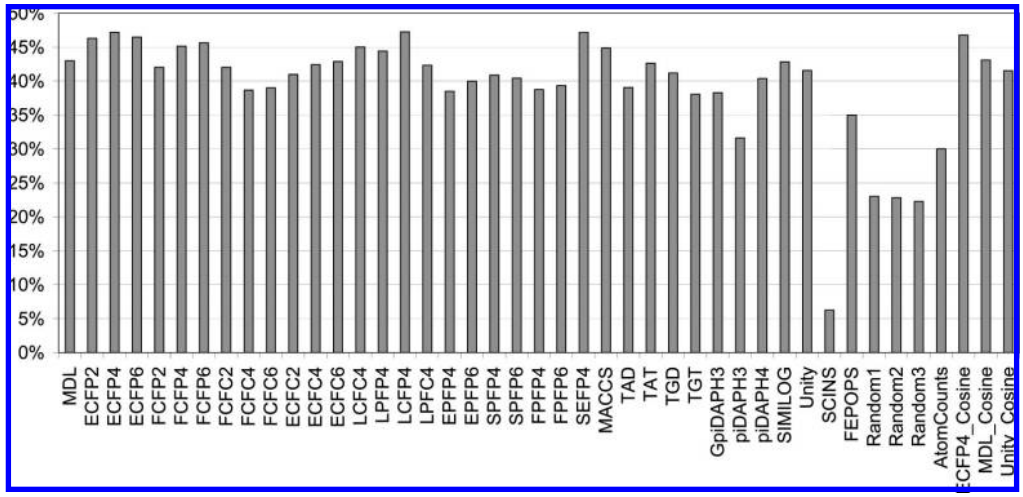

Figure 5. Retrieval rates of chemotypes (defined as distinct SCINS) for 37 similarity searching methods. It can be seen that circular fingerprints also show good scaffold retrieval (e.g., ECFP4, LCFP4, SEFP4/Molprint2D fingerprints), while random retrieval already gives a remarkable fraction of active scaffold in the top of the sorted database.

of the total database, the diversity of actives with respect to different SCINS retrieved is much higher, at around 22%−23% in every run. Of course, this depends strongly on the database used, but since the MDDR employed here contains even more close analogues than most screening libraries, the effect may actually be even more profound in other environments.

Given the high chemotype retrieval even using random selection of compounds, one might reconsider the value of judging the performance of virtual screening methods by only considering the number or fraction of different active scaffolds retrieved. For one, even random retrieval is considerable as shown above; but even more importantly, in the hypothetical case that a descriptor achieves only random chemotype retrieval but high absolute retrieval of actives, this would be a descriptor with quite a good performance overall (in particular when following up HTS hits, when one wants to explore the SAR around a particular scaffold). One option would be to use both contributions to descriptor performance, absolute retrieval, and scaffold retrieval, for example in a multiplicative manner to judge descriptor performance, which covers both of the above aspects in a single number.

Next, based on the principal component analysis of molecular descriptors, we attempted to perform an 'informed' consensus scoring protocol, based on both the retrieval rates of descriptors (Table 4) as well as their different behavior visualized in Figures 1 and 2. Consensus scoring has a long history in the chemical similarity searching arena with varying results, and also theoretical models of its performance have recently been developed for both ligand-based and target-based virtual screening.[41−45] The aim of consensus scoring methods is to retain a high rate of true positives (which are captured by two or more methods) while decreasing the rate of false positives (since different ranking methods will give partly random false-positives, whose removal is the aim of the consensus scoring step.) In our case, the combination of ECFP4 and FCFP6 descriptors has been selected as a descriptor pair with the highest retrieval rates (but similar ranking behavior), and the pair of ECFP4 and TGD descriptors as a pair with high retrieval rates, but dissimilar ranking behavior (according to Figure 2). MAX

Table 5. Consensus Scoring Using ECFP4/FCFP6 Fingerprints (Two Fingerprints with the Highest Retrieval Rates in Individual Scorings) and ECFP4/TAD Fingerprints (Two Fingerprints with Very Different Behavior, As Seen from the Descriptor PCA Plot)[a]

fingerprint	ECFP4	FCFP6	TGD	ECFP4_FCFP6_Max	ECFP4_FCFP6_Ave	ECFP4_TGD_Max	ECFP4_TGD_Ave
average retrieval	40.66%	38.91%	32.73%	40.89%	42.07%	41.38%	41.51%

[a] While overall slight improvement in retrieval can be found, the difference was found to be marginal on this data set.

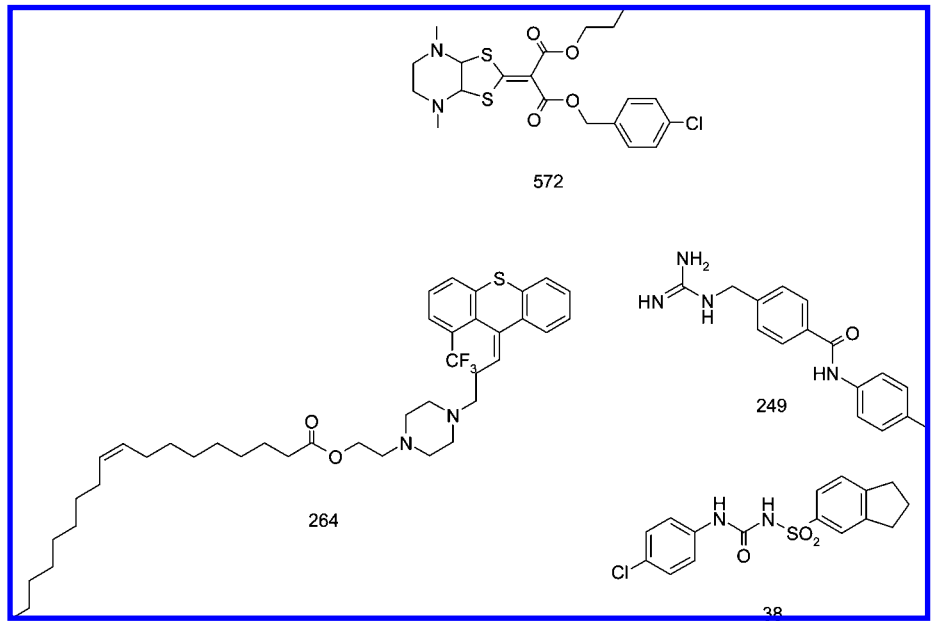

Figure 6. From the similarities obtained from this example, combined with the retrieval rates reported in this work, one hypothesis could be that while considering counts of features is crucial in the case of graph-based correlation descriptors such as CATS, this is less so the case for circular fingerprints. The size of each respective feature set might be responsible for this behavior.

and AVERAGE fusion schemes based on ranks were employed in the current study due to the different parameter ranges of circular fingerprints and TGD descriptors. In the MAX scheme, a compound is assigned the highest rank (smallest number) assigned by any of the descriptors employed. In the AVERAGE scheme, a compound is assigned the average rank assigned by any of the descriptors used. Previously it has been found that the MAX and AVERAGE (equivalent to SUM) consensus scoring improved results on some data sets,[46] although this effect is not in all cases as profound as reported in this previous work.[44]

Table 5 lists the results of the consensus scoring method, based on the current data set. Overall, small but consistent improvements in retrieval could be obtained, with the maximum retrieval being the AVERAGE rank position of ECFP4 and FCFP6 descriptors, followed by the AVERAGE rank of ECFP4 and TGD descriptors. In particular the latter result is surprising, since the inferior retrieval of the TGD descriptor (nearly 10% difference to the ECFP4 descriptors) was completely eliminated in this example. It can be concluded that in the current study descriptor fusion does not lead to greatly improved results, but it seems to make retrieval performance more stable which can be an advantage in cases of novel activity classes where no clearly superior descriptor has been established before.

Following both the results from the correlation analysis of rank-ordering compounds as well as the performance evaluation on retrospective virtual screening data sets the practical application for tasks such as following up hit lists would be the selection of *diverse* descriptors (i.e., those far away from each other in the above correlation analysis) but which still show *high retrieval rates*. For example, each pair of ECFC, ECFP, and FPFP descriptors would be a suitable choice, optimizing diversity of (virtual) hit lists, while keeping the likelihood of good retrieval rates as high as possible.

While the behavior of different descriptors in a statistical sense has been described in detail above, we will now also illustrate differences in behavior in a set of examples. These examples are intentionally selected to display the differences in behavior between descriptors, and they have been selected based on the highest variance of compound ranks between descriptors.

When compound 572 displayed in Figure 6 (numbering according to the position among the 1000 randomly selected compounds from the World Drug Index) was used as a query, differences in ranking between count-based (here e.g. ECFC4) and presence/absence extended connectivity fingerprint (e.g., ECFP4) were profound in case of e.g. compound 264. This structure was ranked 15th with MDL keys (likely due to the presence of similar features such as the piperazine ring, the ester and the benzene ring) and 66th with ECFP4 fingerprints but only 810th with ECFC4 fingerprints due to the widely different size. Compound 249 is the structure showing overall most variance with regarding to rankings by the different descriptors—e.g., ECFP4 fingerprints rank it at 73rd place compared to the query (similar to ECFC4 descriptors, which rank it 98th), while all functional atom type descriptors rank it at a considerably lower rank (e.g., 793rd in case of FCFP4 and 865th in case of FCFC4 descriptors). Compound 38 illustrates also a

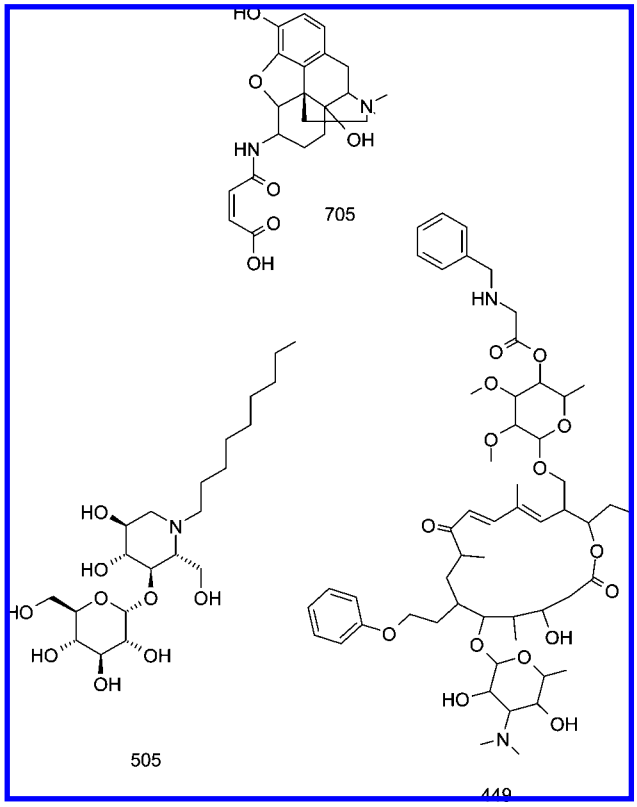

Figure 7. Sample compounds with significantly different rank-orderings between descriptors. Conclusions that might be generalizable are that extended connectivity fingerprints should be preferred over path-based fingerprints in case of more branched structures and that count-based fingerprints could be preferable over binary presence/absence fingerprints in case of repetitive substructures such as amide bonds, terpene subunits, or sugars of natural products.

shortcoming of our SCINS definition with respect to retrieving chemotypes—both the query and compound 38 are found to have the same SCINS while scaffold obviously differ. Still, we found SCINS useful in evaluating chemotypes as a fuzzier definition of Murcko scaffolds which are sometimes too restrictive and suggest the presence of a novel chemotype even in the case of only minor changes to a scaffold.

When query molecule 705, a morphine analogue, is used as a query (shown in Figure 7), interesting ranking behavior can be observed in the case of compound 505, which from the scaffold point of view bears little resemblance to the query structure. MDL keys assign a rank of 88 to the pair when using predefined keys, probably due to the lack of resolution of molecular features present. While ECFP4 descriptors—using extended connectivity units and atom types— assign a rank of 898 to the pair, using functional class fingerprints (FCFP4 descriptors) lift the pair already to rank 413. Replacing extended connectivity information by path-based fingerprints in the FPFP4 descriptor assigns both compounds a pair rank of 99, which is already in the top 10% of the database. This is likely due to the fact that the fused ring system characteristic for morphine analogues is not captured well by path-based fingerprints. The general conclusion is that extended connectivity fingerprints may well be more suitable for more molecules containing more branches or more complex scaffolds, such as in the case of fused ring systems, but this has not been investigated thoroughly in the current study. Compound 449 and its analogues also show remarkable differences in their similarity to query 705 even measured by descriptors from the same class. Typed Graph Distances (TGD) and typed Graph Triangles (TGT) as implemented in MOE rank the pair at position 644 for TGD and position 23 (!) for TGT descriptors. Similarity, their 3-D counterparts TAD and TAT rank the pair at 710th and seventh position, respectively. This is likely due to the presence of similar functional groups at a similar bond distance in both molecules which are still relatively common in the whole database when two points are taken into account but which become very significant if three atoms are considered simultaneously. Count and binary fingerprint-based PipelinePilot descriptors also widely differ on judging this compound pair, with FCFP2 descriptors ranking them 38th and FCFC2 descriptors ranking them 849th. A likely reason is the presence of repeated subunits, here in the case of sugar rings but probably also present in the case of peptide backbones or terpene structures, which influences the count similarities (of FCFC2 descriptors), which is not considered in the case of presence/absence of features (FCFP2 descriptors).

The examples above have been listed as cases where significant deviations of similarities both from expectations and the statistical analysis in the first part have been observed. Possibly the observations can be generalized, such as preferring extended connectivity fingerprints over path-based fingerprints in the case of more branched structures and using count fingerprints over binary presence/absence fingerprints for molecules which contain a large number of repetitive subunits, but a detailed analysis of this part of fingerprint behavior is beyond the scope of the current study.

CONCLUSIONS

In this work, employing a diverse subset of the MDDR database, we present the first large-scale investigation into how similar different molecular descriptors rank-order compounds. This not only is of relevance for virtual screening studies (where orthogonal descriptors are usually chosen for diverse virtual screening hit lists) but also to understand descriptor behavior better for studies such as those in the chemogenomics field. Four broad descriptor classes are identified in the current work, which are binary (presence/absence) circular fingerprints; circular fingerprints considering counts; path-based and keyed fingerprints; and pharmacophore-based fingerprints. Descriptor membership to one of those classes is much more defining the behavior of the ranked list of compounds than the particular parametrization. While consensus scoring was only able to achieve marginal improvement over the individually best descriptors, it was able to eliminate very inferior behavior at the same time. It is hoped that the descriptor analysis presented here can be used to make an informed descriptor selection for virtual screening and chemogenomics studies in the future.

ACKNOWLEDGMENT

A.B., J.S., and S.C.K.S. thank the Education Office of NIBR for a Postdoctoral Fellowship. The authors are also grateful to the referees for very valuable comments on the initial version of the manuscript.

Supporting Information Available: Descriptors, Factors 1−39, and retrievals. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - A review. *QSAR Comb. Sci.* **2004**, *22* (9−10), 1006–1026.
(2) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(3) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.
(4) 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* (16), 3049–3059.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.
(6) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45* (19), 4350–4358.
(7) 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* (2), 391–405.
(8) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.
(9) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding False Positives in Reporter Gene Assays: in silico Chemogenomics Approaches to Prioritize Cell-based HTS Data. *J. Chem. Inf. Model.* **2007**, *47* (4), 1319–1327.
(10) Jacoby, E. A novel chemogenomics knowledge-based ligand design strategy - Application to G protein-coupled receptors. *Quant. Struct.-Act. Relat.* **2001**, *20* (2), 115–123.
(11) Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M.; Mikhailov, D.; Clemons, P. A.; Davies, J. W. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High-Throughput Screening* **2007**, *10* (8), 719–731.
(12) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697* (1−2), 243–257.
(13) Spring, D. R. Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev.* **2005**, *34* (6), 472–482.
(14) Bender, A.; Bojanic, D.; Davies, J. W.; Crisman, T. J.; Mikhailov, D.; Scheiber, J.; Jenkins, J. L.; Deng, Z.; Hill, W. A.; Popov, M.; Jacoby, E.; Glick, M. Which aspects of HTS are empirically correlated with downstream success. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (3), 327–337.
(15) Bohm, H.-J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim: 2000.
(16) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9* (3), 199–204.
(17) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
(18) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45* (5), 1369–1375.
(19) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 233–245.
(20) Basak, S. C.; Gute, B. D. Characterization of Molecular Structures Using Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1–21.
(21) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. Four associative constants for relating molecular similarity measures. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (4), 909–915.
(22) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How far can you jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
(23) 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* (5), 1536–1548.
(24) *WDI (World Drug Index), version 2005*; Thomson: New York, 2005.
(25) *PipelinePilot, version 6.1*, Accelrys: San Diego, 2007.
(26) 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* (3), 1177–1185.
(27) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2* (22), 3256–3266.
(28) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.
(29) *MDL Drug Data Report (MDDR) version 2005.3*; MDL: San Ramon, CA, 2005.
(30) Rohrer, S. G.; Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **2008**, *48* (4), 704–718.
(31) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.
(32) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.* **2007**, *47* (2), 325–336.
(33) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from the known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004**, *47* (25), 6144–6159.
(34) *Spotfire, version 8.0*; TIBCO: Palo Alto, CA, 2007.
(35) Fechner, U.; Paetz, J.; Schneider, G. Comparison of Three Holographic Fingerprint Descriptors and their Binary Counterparts. *QSAR Comb. Sci.* **2005**, *24* (8), 961–967.
(36) Purvis, G. D. 3rd Size-intensive descriptors. *J. Comput.-Aided Mol. Des.* **2008**, *22* (6−7), 461–468.
(37) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2423–2431.
(38) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput.-Aided Mol. Des.* **2003**, *17* (10), 687–698.
(39) 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* (17), 3251–3264.
(40) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18* (7), 529–536.
(41) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42* (25), 5100–5109.
(42) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2005**, *46* (1), 277–288.
(43) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: theoretical model. *J. Chem. Inf. Model.* **2006**, *46* (6), 2193–2205.
(44) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *46* (6), 2206–2219.
(45) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1422–6.
(46) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20* (1), 1–16.
(47) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170–178.

CI800249S