

# “Bayes Affinity Fingerprints” Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept?

Andreas Bender,\* Jeremy L. Jenkins, Meir Glick, Zhan Deng, James H. Nettles, and John W. Davies

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

Received May 15, 2006

Conventional similarity searching of molecules compares single (or multiple) active query structures to each other in a *relative* framework, by means of a structural descriptor and a similarity measure. While this often works well, depending on the target, we show here that retrieval rates can be improved considerably by *incorporating an external framework describing ligand bioactivity space* for comparisons (“Bayes affinity fingerprints”). Structures are described by Bayes scores for a ligand panel comprising about 1000 activity classes extracted from the WOMBAT database. The comparison of structures is performed via the Pearson correlation coefficient of activity classes, that is, the order in which two structures are similar to the panel activity classes. Compound retrieval on a recently published data set could be improved by as much as 24% relative (9% absolute). *Knowledge about the shape of the “bioactive chemical universe” is thus beneficial to identifying similar bioactivities.* Principal component analysis was employed to further analyze activity space with the objective to define orthogonal ligand bioactive chemical space, leading to nine major (roughly orthogonal) activity axes. Employing only those nine activity classes, retrieval rates are still comparable to original Bayes affinity fingerprints; thus, the concept of *orthogonal bioactive ligand chemical space* was validated as being an *information-rich* but *low-dimensional* representation of bioactivity space. Correlations between activity classes are a major determinant to gauge whether the desired multitarget activity of drugs is (on the basis of current knowledge) a feasible concept because it measures the extent to which activities can be optimized independently, or only by strongly influencing one another.

## INTRODUCTION

Molecular similarity searching<sup>1–4</sup> attempts to identify molecules from a database which are “similar” to one or multiple query structures, in a way that needs to be defined explicitly in order to be amenable to computerized treatment. Many “descriptors” to establish molecular representations are known, and they span a wide variety of classes.<sup>5</sup> Some aspects of molecular similarity searching indeed possess analogies in textual searching<sup>6</sup> based on the perception that molecules can be described as an (per se nonlinear) arrangement of structural fragments whereas text can be seen as an (per se linear) arrangement of words.

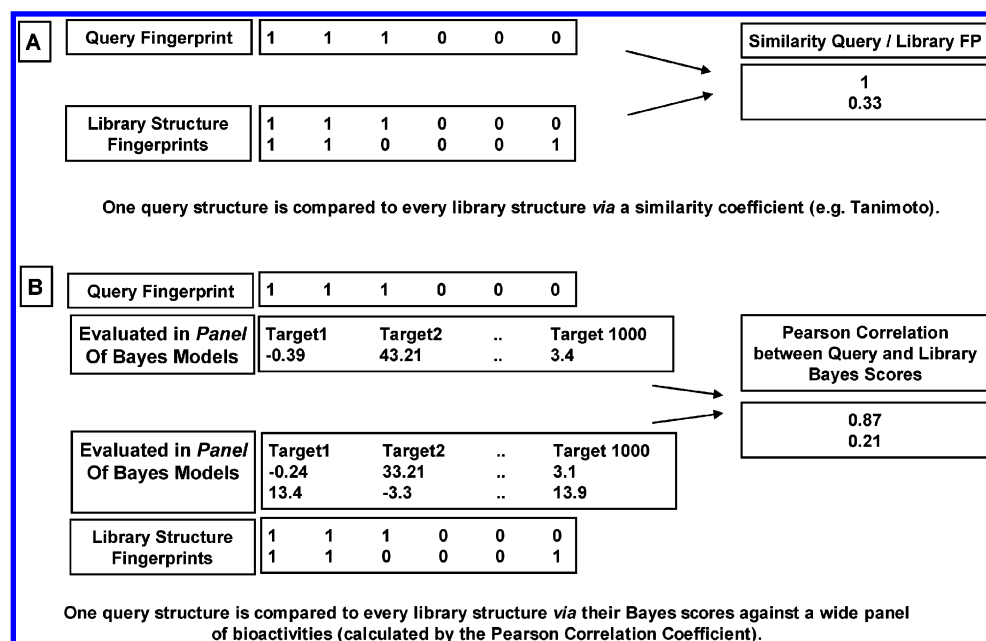
Progress has been made for example with the development of translationally and rotationally invariant descriptors<sup>7–10</sup> which allow back-projection of important (or unnecessary) features for lead optimization. Also, recent comparative studies on large data sets<sup>11–13</sup> allowed for a reproducible benchmark between similarity searching algorithms. Still, there are challenges that need to be overcome. For example, today’s descriptors are—depending on the particular data set—still quite limited in the enrichment factors they are able to achieve, sometimes not even outperforming (nonstructural) atom counts.<sup>14</sup> Addressing the data set issue, they often do not represent “even sampling of chemical space”, not even to the extent that would be possible, meaning that perform-

ance measures established on them are often not matched in prospective studies. Often incomplete data matrixes (molecules vs activities) are found in in-house as well as commercially available databases, so in effect, “false positives” are probably often true (but yet untested) positives. This fact is explored by recent extensions of conventional similarity searching algorithms such as “Turbo Similarity Searching”, which takes into account information about similar molecules (this additional information is analogous here to compressed air containing additional oxygen, thus the name “Turbo”).<sup>15</sup>

While some of the above-mentioned problems—such as incomplete data—are difficult to resolve nonexperimentally, the work presented here aims at a novel method to make predictions about novel molecules which *explicitly capitalizes upon knowledge about the currently known bioactivity universe*. More specifically, molecules are not simply compared to one another but, instead, to a panel of “prototype activities” derived from about 100 000 bioactive compounds from a large database (the WOMBAT database<sup>16</sup>). This transformation possesses two major advantages: (1) the identification of molecular structures showing desired bioactivity with novel scaffolds, at high retrieval rates, and (2) an absolute reference frame of compounds in chemical space.

The conceptual difference between conventional feature-based similarity searching and the current approach is shown in Figure 1. Conventionally (Figure 1A), descriptors are

\* Corresponding author phone: +1 (617) 871-3972; fax: +1 (617) 871-4088; e-mail: Andreas.Bender@novartis.com.



**Figure 1.** Illustration of the Bayes affinity fingerprint approach. In conventional similarity searching (A), molecular descriptors derived from the query molecule are compared to molecular descriptors derived from the library structure directly (e.g., via the Tanimoto coefficient). Bayes affinity fingerprints (B) on the other hand employ external information to compare molecules, namely, the similarity of the molecules to be compared to several sets of molecules belonging to different activity classes. This can be seen as an extension of (protein-based) affinity fingerprints to ligand bioactive chemical space. Classification is here performed via the Pearson correlation coefficient of the activity class scores.

employed directly to compare compounds, most commonly by means of a similarity coefficient. The characteristic important here is that molecules are compared by a measure of similarity based *directly* on features derived from the molecules. In the current work (Figure 1B), descriptors are initially transformed by calculating the descriptor similarity of the compound to a wide panel of bioactivity models. The library compound is evaluated in *all* class-specific Bayes models, and scores are kept. In a subsequent step, this vector, representing similarity to the bioactivity classes present in the panel data set, is employed to compare molecules. This is performed by employing a correlation measure of the list of target classes, sorted by Bayes score (see Materials and Methods section). Thus, the order in which a molecule is similar to the known bioactivity classes determines its position in chemical space and thus its similarity to other molecules.

We will now outline how this multitarget-dependent representation via Bayes affinity fingerprints is related to earlier work and where “absolute positions in chemical space” have already been explored. Affinity fingerprints characterize small molecules by their binding affinities to a set of panel proteins, and they were first established *in vitro*.<sup>17</sup> They possess interesting properties, such as that binding affinities to proteins not in the panel can often be predicted quite precisely by a linear combination of measured binding affinities to the panel proteins.<sup>18</sup> On the basis of molecular docking, affinity fingerprints were soon later translated into the *in silico* area<sup>19–21</sup> with experimentally determined binding energies being replaced by predicted ones (via scoring functions). More recently, and starting with work at the National Cancer Institute,<sup>22</sup> *in vitro* binding profiles have begun to be used on a larger scale. The BioPrint approach<sup>23</sup> attempts to relate molecules by measured binding affinities to a set of pharmacologically relevant target proteins, data

which can be applied to the prediction of side effects as well as more general molecular properties such as logP or absorption (which are also contained in the BioPrint database). Recent work by Pfizer followed a related route,<sup>24</sup> employing bioactivity spectra to predict activities for “novel” compounds, based also on the BioPrint database.

Our work is based on similar representations to the ones above, but both modifies and extends it in multiple ways. It is a statistical *in silico* method based on a large database of >100 000 bioactivity data points which are employed to characterize molecules. Also, *ligand in silico* profiles of molecules are employed and benchmarked for the purpose of *similarity searching*. Finally, the resulting “absolute chemical space” is analyzed further.

Axes may be added in the future as novel targets as well as novel bioactive compounds become known. However, when the observation from affinity fingerprints that affinity to (novel) targets can surprisingly often be represented as a linear combination of affinities to a small panel of target proteins is followed,<sup>18</sup> it can be assumed that affinities to new targets do not contain completely new information. Conversely, information contained in about 1000 bioactivity dimensions employed here can be reduced to much fewer dimensions without losing much of the information, as is shown later.

No precise measure of the size of chemical space exists; nonetheless, attempts have been made to characterize it. The ChemGPS system<sup>25</sup> spans “chemical space” in a way much larger than the desired druglike space in order to establish fixed points within which all other molecules are located. The most characteristic axes employed are related to size, hydrophobicity, and rigidity of the structures. In fact, these are also the first principal components of physicochemical space which can be derived from large data sets and common diverse descriptors,<sup>26</sup> where the first axis is related to size,

**Table 1.** Description of the Data Set Employed, Showing the Data Set Size Published Previously and the Sizes Observed in the Current Version of the Database<sup>a</sup>

activity name	MDDR activity ID	published data set size	used data set size	overlap with WOMBAT absolute (% , target name)
5HT3 antagonists	06233	752	744	79/744 (10.6%, 5HT3)
5HT1A agonists	06235	827	846	98/846 (11.6%, 5-HT1A)
5HT reuptake inhibitors	06245	359	367	30/367 (8.2%, 5-HTT)
D2 antagonists	07701	395	398	70/398 (17.6%, D2)
renin inhibitors	31420	1130	1139	55/1139 (4.8%, renin)
angiotensin II AT1 antagonists	31432	943	2103	235/2103 (11.2%, AT1)
thrombin inhibitors	37110	803	806	81/806 (10.0%, thrombin)
substance P inhibitors	42731	1246	1262	90/1262 (7.1%, NK1)
HIV protease inhibitors	71523	750	758	109/758 (14.3%, HIV-1 P)
cyclooxygenase inhibitors	78331	636	632	43/632 (6.8%, COX, COX-1, COX-2)
protein kinase C inhibitors	78374	452	444	13/444 (2.9%, PKC)

<sup>a</sup> Also shown is the overlap between compounds in the WOMBAT database and those in the MDDR benchmarking library. Note that *both* the training and test sets are derived from the MDDR and that the WOMBAT database is only employed to project compounds into chemical space. This overlap is therefore *no overlap between the training and test sets*.

the second one to lipophilicity, and the third axis to the sign of surface partial charges. While this certainly captures overall chemical properties, a definition of axes more relevant to particular bioactivities is desired.

Consequently, in the present approach, the characterization of compounds by *axes in chemical space which are related to bioactivity* is proposed. When observations from affinity fingerprinting are extended, it can be assumed (and indeed observed) that this bioactivity space is amenable to similarity searching with some interesting properties (such as scaffold hopping). The current study is in a certain way similar to previous work which employed multitarget-dependent descriptor transformation via partial least squares to generate low-dimensional molecular descriptors.<sup>27</sup> It also possesses implicit similarity to the work by Sheridan and Sphungin,<sup>28</sup> who calculated similarities between the activity indices contained in the MDL Drug Data Report (MDDR) database (which are partially found in the variable loadings of the principal component analysis of bioactivity space; see below). It still is more comprehensive in the way that the similarities in activity space are evaluated as a way to improve virtual screening protocols, and also in the facilitation of an orthogonal representation of bioactivity space.

We will, in the following, demonstrate that Bayes affinity fingerprints (1) define a chemical space amenable to similarity searching, (2) improve virtual screening results over conventional “one-feature matching” approaches by including domain knowledge about the ligand bioactive chemical space, and (3) can be transformed via principal component analysis to define low-dimensional, information-rich chemical space which can be employed to gauge the degree of ligand-based relatedness of different bioactivities.

## 2. MATERIALS AND METHODS

**2.a. Database for Spanning Chemical Space.** All compounds belonging to one of 1003 activity classes and with activity values of better than 30  $\mu$ M were extracted from the WOMBAT 2005.01 database<sup>16</sup> (see the Supporting Information for a full list of activities used). This database was employed to generate prototype models for a large number of bioactivities to span “chemical space”. While the activity classes employed in the MDDR span both activity against specific targets and rather generic activity classes (such as “Pharmacological Tool”), entries of the WOMBAT

database are always associated with a specific target. In many situations, this knowledge is invaluable. Throughout the study, compounds were pretreated using StandardizeStereo and StandardizeCharges before the calculation of ECFP<sub>4</sub> fingerprints in PipelinePilot 5.1.<sup>29</sup> ECFP<sub>4</sub> fingerprints are circular in nature, describing the atom environment around every heavy atom in the molecule, keeping counts of atom types differentiated by distance from the central atom. Multiclass Bayes models were generated on all compounds of the WOMBAT data sets, and the relative Bayes scores from each model were used as descriptors further on. This model generation step was analogous to a multiclass Bayes model employed for the prediction of small molecule targets in a previous study, and for detailed information, the reader is referred to this publication.<sup>30</sup> ECFP<sub>4</sub> circular fingerprints were used (instead of ECFP<sub>6</sub> fingerprints) because they were found to perform slightly better in retrospective virtual screening studies.<sup>12,31</sup> Laplacian correction of the standard Bayes scores (fraction of compounds showing a feature associated with activity) was employed to scale properties associated in very small samples, which would otherwise lead to overconfident predictions. The probability of encountering an active compound (class A), given feature  $F_i$ , is calculated as  $P(A|F_i) = (A_{F_i} + 1)/(T_{F_i}(A/T) + 1]$ , where  $A_{F_i}$  denotes the number of samples showing feature  $F_i$  which are active,  $T_{F_i}$  denotes the total number of samples showing feature  $F_i$ , and  $A$  and  $T$  denote the total number of active compounds and the total number of compounds, respectively. Relative scores resulting from this model were employed as descriptors for each compound from the benchmarking database.

**2.b. Benchmarking Database.** A recently published benchmarking data set<sup>11</sup> was employed to investigate the performance of the current algorithm with respect to the retrieval of active compounds. The data set was comprised of 11 active classes from the MDDR, which are shown in Table 1. Because annotations of the database change between both the data set sizes published earlier and those found in the current version (which is 2005.02), the current data set sizes are also given. To ensure comparability to established descriptors, the protocol performed previously has been repeated in the current study. Also given in Table 1 is the overlap between the database used for spanning chemical space (WOMBAT) and the benchmarking database (derived



**Table 2.** Retrieval Rates of Bayes Affinity Fingerprints, Compared to Conventional (ECFP\_4) Fingerprints in Combination with the Tanimoto Coefficient<sup>a</sup>

class	Bayes affinity fingerprints				conventional descriptors (ECFP_4)				comparison	
	recall	stdev	unique	stdev	recall	stdev	unique	stdev	$\Delta$ recall absolute	$\Delta$ recall relative
06233	37.47%	13.27%	86	26	30.66%	11.30%	35	28	6.81%	22.21%
06235	44.11%	15.49%	193	80	24.26%	9.16%	25	10	19.85%	81.80%
06245	31.09%	15.75%	35	18	24.67%	11.91%	12	7	6.42%	26.02%
07701	48.59%	17.87%	93	46	28.46%	12.11%	13	8	20.13%	70.71%
31420	90.04%	15.36%	50	76	88.36%	17.96%	33	58	1.69%	1.91%
31432	82.48%	28.53%	430	347	64.60%	27.97%	54	84	17.88%	27.68%
37110	48.51%	24.66%	121	98	40.70%	20.87%	58	23	7.81%	19.20%
42731	27.76%	23.51%	112	130	27.29%	8.07%	106	72	0.47%	1.71%
71523	60.09%	22.49%	121	137	48.87%	25.41%	34	42	11.22%	22.97%
78331	16.91%	7.81%	27	16	16.61%	6.21%	25	12	0.30%	1.81%
78374	19.59%	15.59%	32	18	15.15%	13.11%	12	8	4.45%	29.36%
<b>average</b>	<b>46.06%</b>	<b>18.21%</b>	<b>118</b>	<b>90</b>	<b>37.24%</b>	<b>14.92%</b>	<b>37</b>	<b>32</b>	<b>8.82%</b>	<b>23.69%</b>

<sup>a</sup> On average, retrieval rates are 8.8% in absolute number (24% relative) larger when Bayes affinity fingerprints are employed (same ligands used in both cases).

from the MDDR). The overall overlap is low, between 2.9% and 17.6% (average overlap 9.6%). The reader should be aware that this overlap is *not* the overlap between the training and test sets. The training and test sets are *both* derived from the MDDR database. The WOMBAT compounds are only used to project compounds into chemical space and to define the “activity axes”. The influence of the relationship between both databases is further discussed in the Results and Discussion section.

**2.c. Benchmarking Protocol.** A 10-fold random selection of single active compounds from each of the 11 activity classes was performed, and the rest of the library was ranked according to Tanimoto similarity (in the case of ECFP\_4 fingerprints) and the Pearson correlation coefficient of class scores (in the case of Bayes affinity fingerprints), in PipelinePilot 5.1.<sup>29</sup> In all cases, the fraction of active compounds in the top 5% of the sorted database was calculated, along with the overlap between conventional (ECFP\_4) descriptors and Bayes affinity fingerprints.

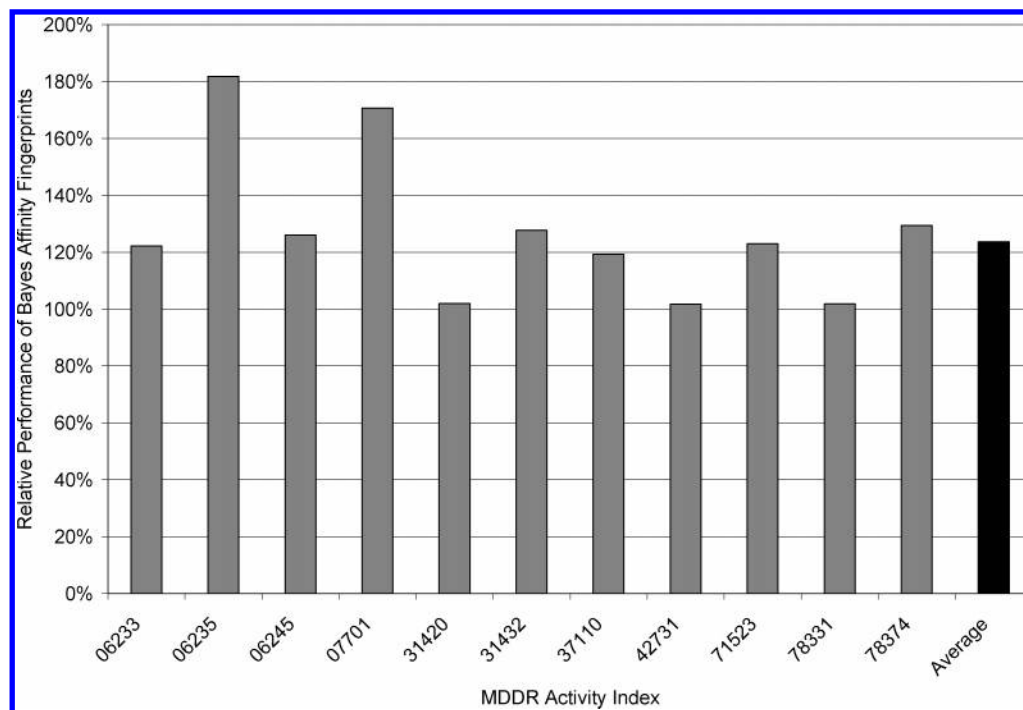
**2.d. Principal Component Analysis.** A total of 102 500 compounds<sup>11</sup> from the MDDR database were evaluated with their respective scores by the multiclass model to probe its response surface, and the 100 target classes (dimensions) possessing the largest number of ligands were retained, on one hand, to reduce the number of components before principal component analysis and, on the other hand, to focus on activity classes which seem to be of relevance. Another possibility to select a smaller number of input dimensions for the principal component analysis (PCA) would be some kind of diversity/orthogonality selection, a route which was not followed here because the 100 largest classes were found to represent targets of current interest quite well. Principal component analysis was performed in Statistica 7.0<sup>32</sup> followed by normalized Varimax rotation. All principal axes showing eigenvalues larger than 1 were kept.

### 3. RESULTS AND DISCUSSION

Retrieval rates of the conventional similarity-searching approach (ECFP\_4 in combination with the Tanimoto coefficient) compared to Bayes affinity fingerprints (in combination with the Pearson correlation coefficient) are shown in Table 2 and Figure 2, along with the overlap of compounds retrieved. Overall, Bayes affinity fingerprints

retrieve in absolute numbers 8.82% (23.69% in relative numbers) more active compounds than ECFP\_4 fingerprints. Because ECFP\_4 fingerprints were established as one of the best-performing similarity searching methods on this data set in a recent study,<sup>12</sup> this is encouraging. The best recall is seen for renin inhibitors (90.04% vs 88.36% for ECFP\_4 fingerprints) and the smallest recall for cyclooxygenase inhibitors (16.91% vs 16.61%). The largest absolute retrieval difference (+20.13%) is given for dopamine D2 antagonists (48.59% vs 28.46%), while the highest relative performance difference (improvement of +81.80%) is observed for 5HT1A agonists (44.11% vs 24.26%). The smallest absolute difference is seen for cyclooxygenase inhibitors (+0.30%), which also results in the smallest relative retrieval difference (1.81%). Retrieval rates are improved across all activity classes, leading to the conclusion that the incorporation of domain knowledge about bioactive chemical space is generally beneficial.

The relationship between the improvement in retrieval rates and the overlap between the benchmarking database and the database used to project compounds into chemical space can be derived from Tables 1 and 2. Table 1 gives the percentage overlap between the benchmarking (MDDR) and projection (WOMBAT) databases, while Table 2 gives the improvement per class. A relationship between both values can be established; for example, D2 antagonists show the largest absolute improvement in retrieval rates (20.1%) and also the largest overlap between the databases (17.6%). On the other hand, the class with the largest relative retrieval rate increase (5HT1A agonists, 81.8% relative improvement) shows only an 11.6% compound overlap, while showing a 19.9% improvement in absolute numbers. Therefore, while a relationship between both values exists, it is not clear-cut. Also, the possible misperception that an improvement in performance might simply be due to some overlap between training and test sets needs to be countered here. The WOMBAT database is employed for *describing* chemical space and for projecting compounds onto it. The MDDR data is employed for the *benchmarking* part of the study. Compounds present in both sets contribute, to a small extent (1/100, if for example 100 compounds are present in the respective WOMBAT activity class), to the activity model for that particular chemical space dimension. However, they



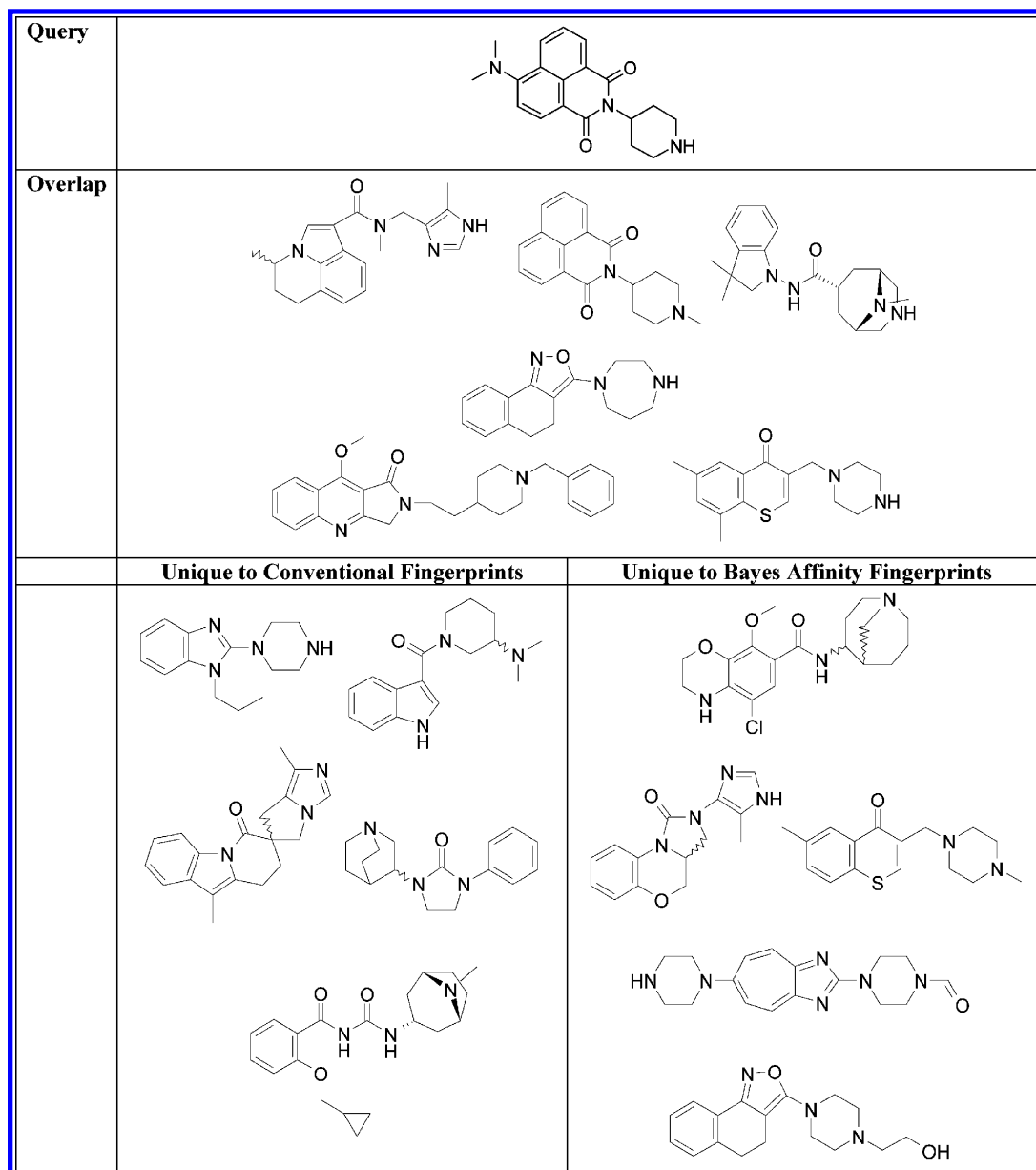
**Figure 2.** Performance of Bayes affinity fingerprints and the Pearson correlation coefficient on individual query structures, as compared to Scitegic ECFP\_4 fingerprints with the Tanimoto coefficient. On average, about 24% higher retrieval is achieved in the top 5% of the recalled compounds (between 2% and about 80%) with no activity class giving inferior results, compared to conventional similarity searching.

do *not* contribute in any way to the similarity between the training set (query) and the test set (library) structures. The similarity of the query compound to all activity dimensions is calculated first, and the order of Bayes scores of the activity dimensions is compared to the order of Bayes scores of library compounds. Only a small number (9.6% average) of compounds is overlapping. Even *if* a compound present in WOMBAT is chosen as the query structure, it is in no case present in the library (test set), therefore eliminating any possible bias. However, an influence of the overlap between the database used to span chemical space and the benchmarking library is certainly given. The authors attribute this effect to the better description of the respective activity dimension: If many compounds *similar* and *with the same activity* as the query structure are present in the database used for laying out chemical space, this space will be *more able* to pick up many different kinds of active compounds, thereby increasing retrieval rates. Thus, *overlap* between the data sets is not crucial, but rather the ability of the database used to describe chemical space to describe activity space as comprehensively as possible. In turn, this leads to the assumption that the most comprehensive database used for mapping is the most suitable one, because it provides a more detailed map of chemical space. This observation is currently still being investigated.

While the absolute number of compounds retrieved is one performance measure, additional attention needs to be paid to the type of compounds identified, for example, the different scaffolds retrieved.<sup>33</sup> As also given in Table 2, the number of compounds identified unique to each method is listed. As can be calculated from the data set sizes, the number of unique structures is larger than what could be expected from the different retrieval rates alone, leading to the conclusion that the descriptors compared are to some degree complementary. In addition to the number of unique

structures which were retrieved by each approach, also the numbers of Murcko scaffolds<sup>34</sup> identified have been calculated for both methods. Here, it was found that the ratio of unique Murcko scaffolds identified by each method divided by the total number of unique compounds gave very similar results for both approaches, slightly disfavoring Bayes affinity fingerprints. This can be explained by the larger denominator (the total number of structures) in the case of Bayes affinity fingerprints, which makes it harder to find additional novel scaffolds, given a finite data set size and diversity. Therefore, this diversity measure was not employed further, and only the total number of retrieved compounds was used as a performance measure.

Sample structures retrieved by a serotonin receptor ligand ("5HT3 antagonist", MDDR activity class 06233) are shown in Figure 3. For this query, a total of 233 structures are retrieved for Bayes affinity fingerprints and 162 for conventional ECFP\_4 fingerprints, out of which 120 (49) are unique to each of the approaches. A total of 113 compounds occur in both hit lists. Members of each data set were clustered in PipelinePilot with standard settings to give six clusters for compounds from the overlapping hit list and five compounds from each of the individual sets. Cluster centroids are displayed in Figure 3. It can be seen that the scaffold similarity is biggest for the query compound and the combined hit list from both approaches. For example, the piperidine ring is present in four of the retrieved cluster center structures, once it is replaced by a piperazine ring and, in another case, another carbon is introduced to give a seven-membered heterocycle. The scaffold similarity to both individual hit lists is smaller, which is true for both the original ECFP\_4 fingerprints and Bayes affinity fingerprints. When the fact that circular fingerprints are a topological (2D) descriptor is solely focused upon, this might be surprising. It is nonetheless in line with recent research showing that in



**Figure 3.** Structures (cluster centers) retrieved for a serotonin receptor ligand query (MDDR activity index 06233). In the cases of both ECFP\_4 fingerprints and Bayes affinity fingerprints, considerable diversity can be observed, which is to a lesser extent true for the structures retrieved by both algorithms.

addition to a large number of compounds<sup>12,13,31</sup> circular fingerprints are also able to retrieve a variety of scaffolds.<sup>12</sup> The general notion that 2D fingerprints tend to “stick” to scaffolds may thus be more attributable to Daylight or Unity fingerprints<sup>35</sup> than to circular fingerprints, despite their topological nature. To put this result into context (and as further discussed in ref 35), it should be mentioned that some of the fingerprints based on 2D structure were indeed designed to find compounds with similar scaffolds or display certain functional groups (keys), and this is the task they fulfill well.

While the retrieval of active compounds could, as shown, be improved by employing external knowledge about biologically active chemical space, the activity models derived from descriptor transformation can also be employed to investigate which biological activities are related with each other and which show orthogonal behavior. This is an important task because it addresses the question of whether activities of compounds against certain targets, leading to

off-target effects, could be *eliminated*, or, conversely, whether multitarget drugs<sup>36</sup> are *feasible* between a pair (or a group) of targets. In the following, we describe the orthogonal axes of bioactive ligand chemical space, as derived from the current Bayes multicategory model.

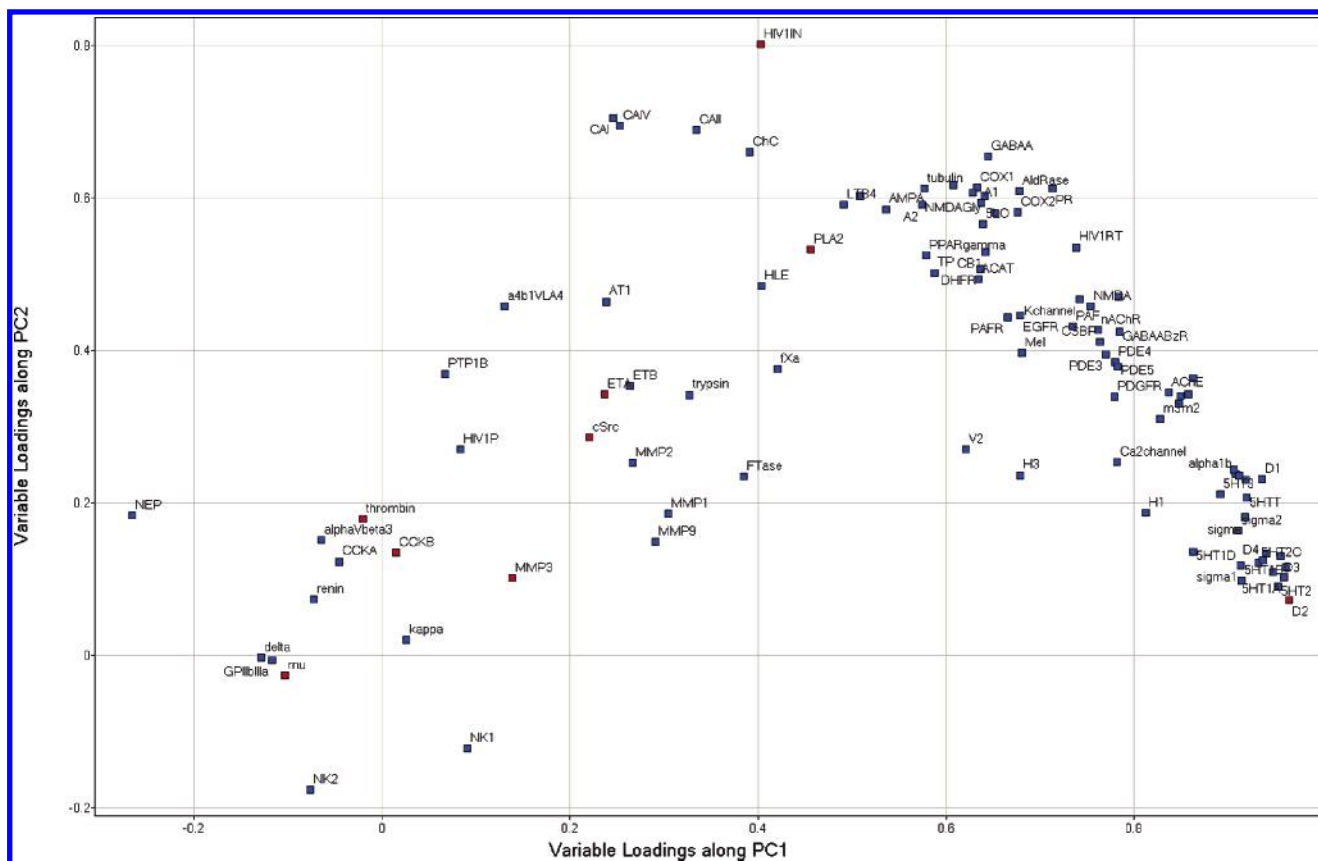
One could in principle perform a PCA of the Bayes activity models to derive orthogonal axes in bioactivity space. Because a large number of substructural features (= dimensions) can be conceived, this is not currently feasible because of computational reasons. Also, not individual features but combinations of features which result in particular prototypical bioactivities were deemed to be of more interest. Here, a large set of about 102 500 structures extracted from the MDDR<sup>11</sup> was employed to test the response surface of the different models. The 100 largest activity classes only (shown in Table 3) were employed to describe compounds in order to limit the number of dimensions for principal component analysis and also to focus on the more relevant bioactivities. Subsequently, principal components were calculated employ-

**Table 3.** First Nine Principal Axes (showing eigenvalues larger one) with Variable Loadings

target	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	target	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
PTP1B	0.056	0.368	0.087	0.507	0.148	0.317	0.274	0.070	0.408	sigma1	0.911	0.139	0.112	0.169	-0.024	-0.054	-0.029	-0.097	0.123
GABAA	0.633	0.669	0.095	0.085	-0.093	-0.058	-0.114	-0.030	0.003	PDE5	0.774	0.385	-0.206	0.060	0.096	0.192	0.026	0.233	0.046
Me1	0.671	0.408	0.067	0.233	0.052	0.198	0.164	0.046	0.175	alpha1a	0.907	0.249	-0.057	0.030	0.056	0.067	-0.020	0.044	-0.039
LTB4	0.478	0.599	-0.117	0.215	0.070	0.180	0.038	0.002	0.466	FTase	0.379	0.242	0.167	0.446	0.179	-0.012	0.263	0.134	0.069
AldRase	0.665	0.619	-0.037	0.067	0.132	0.151	0.112	0.034	0.090	cSrc	0.211	0.277	0.240	0.091	0.297	0.341	0.022	<b>0.498</b>	0.147
ETB	0.253	0.349	0.132	0.446	0.058	0.635	0.003	0.036	0.081	trypsin	0.322	0.343	0.017	0.284	0.724	-0.038	0.155	-0.030	0.055
NMDA	0.744	0.472	0.095	0.125	0.147	0.013	0.087	0.045	0.141	EGFR	0.669	0.450	-0.137	0.161	0.184	0.194	0.138	0.253	0.110
delta	-0.131	0.002	0.936	0.074	0.134	0.039	-0.006	-0.014	0.046	5HT1D	0.859	0.149	-0.086	0.144	0.088	0.147	0.105	0.051	0.011
ETA	0.225	0.337	0.111	0.406	0.093	<b>0.674</b>	0.054	0.069	0.081	CCKA	-0.052	0.130	0.409	0.456	0.179	0.033	0.669	0.014	0.058
mu	-0.105	-0.020	<b>0.944</b>	0.077	0.117	0.015	-0.016	-0.005	0.066	MMP1	0.303	0.187	0.102	0.799	0.276	0.141	-0.026	-0.050	0.032
kappa	0.024	0.031	0.934	0.050	0.114	-0.044	-0.041	-0.049	0.058	GABAAABzR	0.775	0.434	-0.068	0.148	0.123	0.129	0.157	0.195	0.074
NEP	-0.271	0.183	0.547	0.559	0.031	0.103	0.226	-0.013	-0.047	CB1	0.632	0.542	0.069	0.201	-0.009	-0.003	-0.007	0.080	0.239
5LO	0.627	0.575	-0.096	0.168	0.081	0.158	0.000	0.024	0.297	alphaVbeta3	-0.068	0.141	0.311	0.273	0.708	0.162	-0.005	0.118	-0.048
NMDAGly	0.627	0.610	-0.012	0.083	0.214	0.122	0.161	0.144	0.016	<b>D2</b>	<b>0.962</b>	0.090	0.021	0.053	0.060	0.103	0.016	-0.035	0.017
5HT2C	0.929	0.136	-0.091	0.068	0.090	0.150	0.105	0.022	0.003	m3	0.824	0.327	0.077	0.185	0.034	-0.179	-0.128	0.067	0.116
MMP9	0.289	0.151	0.045	0.827	0.225	0.198	0.141	-0.051	0.018	ChC	0.378	0.667	0.056	0.417	0.192	0.134	0.108	-0.113	-0.035
AMPA	0.524	0.590	0.061	0.149	0.210	0.046	0.106	0.247	0.029	m2	0.844	0.346	0.055	0.155	0.044	-0.156	-0.119	0.050	0.082
5HT1B	0.911	0.113	-0.113	0.081	0.048	0.137	0.081	0.019	0.001	PAF	0.726	0.441	-0.142	0.156	0.083	0.028	0.059	0.124	0.155
MMP2	0.261	0.254	0.024	0.810	0.185	0.220	0.156	-0.088	-0.006	A3	0.617	0.611	-0.196	0.126	0.049	0.009	0.044	0.314	0.023
sigma2	0.908	0.185	0.131	0.124	-0.050	-0.036	-0.014	-0.110	0.078	fXa	0.414	0.378	-0.046	0.238	0.676	0.034	0.181	0.004	0.104
COX1	0.620	0.619	-0.184	0.135	0.187	0.189	0.055	0.047	0.059	<b>MMP3</b>	0.138	0.101	0.128	<b>0.873</b>	0.219	0.153	0.067	-0.002	0.074
HIV1RT	0.727	0.544	-0.119	0.131	0.098	0.070	0.105	0.174	0.036	NK1	0.090	-0.113	0.693	0.258	0.058	0.024	0.281	-0.018	-0.152
PPARGgamma	0.569	0.531	-0.099	0.269	0.156	0.162	-0.060	0.058	0.343	A2	0.563	0.595	-0.079	0.141	-0.065	-0.026	0.014	0.402	-0.005
5HT2	0.950	0.107	-0.024	0.042	0.032	0.094	0.072	-0.034	-0.037	<b>HIV1IN</b>	0.386	<b>0.806</b>	-0.009	0.086	0.102	0.159	0.074	0.007	0.075
PDE4	0.772	0.391	-0.204	0.073	0.146	0.113	0.021	0.196	0.058	<b>thrombin</b>	-0.022	0.176	0.223	0.356	<b>0.760</b>	-0.098	0.155	-0.006	0.064
D2L	0.936	0.151	-0.012	-0.008	0.042	0.074	0.097	0.010	0.021	A1	0.626	0.597	-0.164	0.122	0.030	0.001	0.048	0.343	0.025
5HT3	0.886	0.225	-0.055	-0.008	0.109	0.078	0.142	0.143	-0.016	GPIIbIIIa	-0.116	-0.014	0.492	0.230	0.651	0.115	-0.160	0.082	0.052
TP	0.578	0.509	-0.149	0.272	0.157	0.082	-0.045	0.007	0.277	PAFR	0.658	0.453	-0.145	0.218	0.061	-0.041	-0.109	0.097	0.306
a4b1VLA4	0.119	0.452	0.370	0.253	0.446	0.336	-0.063	0.071	-0.048	DHFR	0.625	0.513	-0.088	0.234	0.223	0.211	0.085	0.101	0.243
D4	0.933	0.142	-0.098	0.000	0.074	0.066	0.068	-0.028	0.065	PDE3	0.762	0.402	-0.188	0.084	0.102	0.080	-0.023	0.216	0.133
D1	0.930	0.250	0.044	0.010	0.014	0.056	0.035	-0.037	-0.044	alpha1	0.959	0.132	-0.053	0.061	0.061	0.062	0.012	0.001	-0.035
5HT2A	0.944	0.124	-0.062	0.064	0.077	0.141	0.095	0.009	-0.021	H1	0.810	0.199	-0.094	0.237	0.176	-0.019	-0.091	0.063	0.211
5HT1A	0.957	0.119	-0.037	0.073	0.067	0.086	0.035	-0.003	0.004	V2	0.616	0.282	0.213	0.104	0.337	-0.017	0.025	-0.078	-0.149
TSase	0.495	0.606	0.023	0.200	0.256	0.149	0.130	0.111	-0.035	NK2	-0.075	-0.172	0.776	0.208	0.023	0.042	0.170	0.086	-0.161
HLE	0.396	0.491	0.082	0.394	0.334	-0.061	0.024	-0.028	0.068	CSBP	0.754	0.419	-0.125	0.102	0.168	0.095	0.118	0.207	-0.007
D3	0.952	0.147	-0.049	0.051	0.064	0.083	0.065	-0.010	0.036	AT1	0.228	0.463	-0.134	0.054	0.359	0.133	0.246	0.158	0.282
alpha1b	0.903	0.253	-0.041	0.024	0.054	0.049	-0.019	0.042	-0.055	5HTT	0.916	0.226	0.045	0.160	0.002	0.025	0.022	-0.051	0.074
alpha2	0.915	0.245	-0.030	0.046	0.057	0.048	-0.077	0.004	-0.037	PDGFR	0.771	0.344	-0.181	0.118	0.180	0.199	0.063	0.223	0.085
sigma	0.914	0.200	0.059	0.100	-0.003	0.012	-0.041	-0.054	0.165	CAII	0.322	0.693	0.099	0.396	0.290	0.025	-0.054	-0.090	-0.009
ACAT	0.626	0.502	-0.085	0.298	0.084	0.064	-0.020	0.073	0.268	NET	0.857	0.381	0.073	0.148	0.035	-0.054	-0.028	-0.017	0.082
Ca2channel	0.778	0.266	-0.047	0.236	0.038	-0.012	-0.132	0.043	0.146	A2A	0.596	0.620	-0.180	0.115	0.055	0.004	0.022	0.347	0.030
AChE	0.830	0.359	0.007	0.157	0.070	0.015	-0.045	0.015	0.170	H3	0.677	0.243	-0.078	0.383	0.214	-0.037	-0.099	0.198	0.251
nAChR	0.754	0.440	0.156	0.166	0.032	-0.050	-0.104	0.119	0.036	m1	0.845	0.356	0.053	0.152	0.030	-0.142	-0.120	0.084	0.062
tubulin	0.566	0.624	-0.027	0.048	-0.127	0.034	-0.087	-0.085	0.069	CAIV	0.241	0.698	0.144	0.388	0.337	0.050	0.031	-0.096	-0.108
alpha1d	0.901	0.258	-0.058	0.020	0.055	0.061	-0.015	0.058	-0.054	renin	-0.070	0.071	0.307	0.689	0.019	-0.281	-0.042	0.348	-0.002
<b>CCKB</b>	0.008	0.143	0.412	0.433	0.175	0.031	<b>0.679</b>	0.001	0.021	CAI	0.234	0.708	0.115	0.393	0.330	0.025	0.018	-0.108	-0.042
COX2	0.664	0.588	-0.163	0.147	0.172	0.179	0.055	0.059	0.074	PR	0.701	0.626	0.102	0.051	-0.010	0.044	-0.045	-0.028	-0.044
Kchannel	0.734	0.478	-0.057	0.213	0.137	0.018	0.036	0.020	-0.003	muscarinic	0.775	0.483	0.052	0.135	0.013	-0.092	-0.067	0.173	0.040
<b>PLA2</b>	0.444	0.538	-0.082	0.228	0.157	0.133	0.092	0.119	<b>0.498</b>	HIV1P	0.080	0.271	0.220	0.727	0.072	-0.083	-0.024	0.124	0.064
KAIN	0.639	0.589	-0.022	0.111	0.185	0.083	0.085	0.141	0.126	DAT	0.852	0.361	0.101	0.202	0.023	-0.052	-0.020	-0.012	0.119

<sup>a</sup> Selected are the activity classes which show the highest correlation with each principal component. Single activities were chosen over linear combination for better interpretability.





**Figure 4.** Variable loadings of the two first principal components of ligand bioactivity space. On the basis of the Bayes activity models, inhibitors of HIV-1 integrase and ligands of the dopamine D2 receptor define the first two axes. The activity classes most highly correlated with the first nine principal components (eigenvalues larger than 1) are marked in red.

ing Varimax rotation. Surprisingly, only the first nine eigenvalues resulting from 100-dimensional activity space were larger than 1, indicating “low” dimensionality of (this) bioactive ligand chemical space. Variable loadings are shown in Table 3 and plotted for the first two principal components in Figure 4. To give interpretable orthogonal axes, no linear combination of activity classes was employed to derive the prototypical activity axes of bioactive chemical space. Instead, those activity classes possessing the highest variable contributions to each of the axes were chosen to represent that particular dimension in bioactivity space. This results in biological activity dimensions which behave (approximately) orthogonal to each other. Probably the most important property of this space is its ability to derive which bioactivities are related or dependent on each other and which are not, and the rational design of activity profiles of compounds will be one of its most important measures of success.

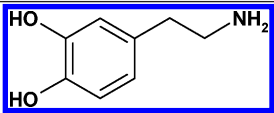
One should note that the axes described here are resulting from the particular settings chosen, including the data set used (which is limited both to known targets and known ligands), the particular descriptor we are using (circular ECFP<sub>4</sub> fingerprints), the modeling procedure employed (the Bayes classifier), and the PCA and selection of particular “orthogonal” activities. If only one of those steps is modified, the set of “orthogonal” activities will be different. Nonetheless, if only gradual changes are introduced in any of the steps, it can be expected that considerable collinearity of the novel principal components with the ones derived here should be observed.

The first two dimensions of bioactivity space (Figure 4 and Table 4) are spanned by dopamine D2 ligands (PC1) and HIV-1 integrase inhibitors (PC2). (More precisely, this should read “by similarity to known binders to the targets ...”, which is omitted in the following for easier readability.) Higher dimensions are formed by the  $\mu$  opioid receptor (PC3) and inhibitors of the MMP3 (PC4) and thrombin (PC5) proteases. The endothelin A and cholecystokinin B receptor form PC 6 and PC7, while the oncogene-related SRC kinase (PC8) and PAF2-acylhydrolase (PC9) form the remaining dimensions of ligand bioactivity space. The fraction of explained variance decreases rapidly, from 58.5% in dimension one to 1.1% in dimension nine. Overall, 85.4% of the variance is explained by the first nine principle components possessing eigenvalues larger than 1.

The first principal component was found to explain a high value of about 59% of the total variance. This can be explained by looking deeper into Table 3. In fact, many activity data sets possess large variable loadings in PC1, besides the D2 class, typically belonging to the group of monoamine receptor ligands. For example, various serotonin receptor subtypes (e.g. 5HT<sub>2C</sub>, 5HT<sub>1B</sub>, 5HT<sub>2A</sub>, and 5HT<sub>1A</sub>) possess variable contributions of >0.9. In addition, less phylogenetically related targets such as  $\sigma$  receptor binders belong to this group, as well as  $\alpha$  and  $\beta$  adrenergic receptor ligands. Therefore, similarity to D2 receptor ligands as the prototypical class of bioactivity PC1 can be interpreted as monoamine ligand-likeness. The close proximity of D2 and 5HT ligands also might give an idea as to why circular fingerprints are capable of performing scaffold hopping:



**Table 4.** Analysis of the First Nine Principal Components of Ligand Bioactivity Space, Listing Targets with Their Classes and Endogenous Ligands or Substrates

principal component (PC)	target with highest variable score ("most characteristic"), (score)	target characterization	percent explained variance (cumul.)	endogenous ligand/ endogenous substrate
PC1	D2: dopamine D2 receptor (0.962)	GPCR/class A/amine binding/dopamine	58.5 (58.5)	
PC2	HIV1IN: HIV 1 integrase (0.806)	enzyme/transferase/transferring P-containing groups/nucleotidyl transferase	11.2 (69.7)	DNA
PC3	$\mu$ : $\mu$ opioid receptor (0.944)	GPCR/class A/peptide binding/opioid	5.2 (74.9)	endomorphin 1, Tyr-Pro-Trp-Phe-NH <sub>2</sub> ; endomorphin 2, Tyr-Pro-Phe-Phe-NH <sub>2</sub>
PC4	MMP3: matrix metalloprotease 3 (stromelysin 1) (0.873)	enzyme/hydrolase/peptidase/metalloendopeptidase	2.7 (77.6)	proteoglycans; fibronectin; collagens III, IV, V, IX; procollagenase
PC5	thrombin:thrombin (0.760)	enzyme/hydrolase/peptidase/serine-endopeptidase	2.2 (79.8)	fibrinogen
PC6	ETA: endothelin A receptor (0.674)	GPCR/class A/amine binding/peptide binding	1.8 (81.6)	endothelin A
PC7	CCKB: cholecystokinin B receptor (0.679)	GPCR/class A/peptide binding/CCK	1.4 (83.0)	cholecystokinin
PC8	cSRC: SRC tyrosine-protein kinase	enzyme/transferase/transferring P-containing groups/protein tyrosine kinase	1.3 (84.3)	various
PC9	PLA2: PAF2-acylhydrolase	enzyme/hydrolase/esterase/carboxylic ester hydrolase	1.1 (85.4)	phospholipids

while structurally D2 and 5HT are not significantly similar, their activity models (at least in the first two dimensions) comprise very similar features. On the basis of empirical evidence found here, circular substructures seem to be a very useful representation of molecules which enables both high retrieval rates and the identification of new chemotypes. Component 2 in activity space is spanned with HIV-1 integrase inhibitors, with carbonic anhydrase inhibitors showing the highest collinearity. PC3 is most closely correlated with  $\mu$  opioid receptor ligands, which partly show the classical morphine scaffold (center structure of Table 5a, row PC3). The  $\delta$  and  $\kappa$  opioid receptors show the highest collinearity. PC4 is principally explained by MMP3 affinity, with other matrix metalloproteases (MMPs 1, 2, and 9) showing similar (high) variable loadings along this axis. In addition, renin has a variable loading close to 0.7, indicating that those activities are transferred by similar ligand features. Thrombin inhibitors define PC5, closely related with  $\alpha$ II/ $\beta$ III integrin binders and factor Xa inhibitors. Because thrombin as well as  $\alpha$ II/ $\beta$ III integrin are known to bind fibrinogen, this link is plausible; thrombin and factor Xa are both serine proteases, and ligand similarity is also established. The highest correlation on PC6 is that between endothelin A and endothelin B ligands. PC7 is defined by cholecystokinin B receptor ligands with the highest correlations to the A type CCK receptor. PC8 is spanned by inhibitors of the cSrc kinase and correlates to a certain extent (variable loadings around 0.3) with adenosine receptor binders. PAF2-acylhydrolase describes PC9, with the highest correlation to protein tyrosine phosphatase B (PTP1B) and leukotriene receptor B4 (LTB4) binders. Overall, in the higher principle components, it can be observed that no single activity class has a (very) high correlation with a particular principle component, but that several classes are present with medium correlation (see Table 3).

Tables 4 and 5 show the "diversity" of the dimensions covered from the target class, as well as the endogenous ligand/endogenous substrate kind of view (Table 4). They also show representative members (cluster centroids) of each activity dimension (Table 5), derived from the bioactive

ligand chemical space side. It may initially be surprising that only G-protein-coupled receptors (GPCRs) and enzymes are necessary to describe orthogonal ligand bioactive chemical space, and that no nuclear hormone receptors and ion channels seem to be necessary for this purpose. It should be mentioned that a similar representation of bioactivity space, with high collinearity to the given axes, can probably also be constructed solely, for example, on the basis of nuclear hormone receptor ligands. Here, we restricted ourselves to the 100 largest activity classes from the WOMBAT database. Also, our analysis is based on the ligand structures only, and (formal) attribution to one target class must not necessarily be an indicator of ligand similarity. Accordingly, if one recognizes the variety of endogenous ligands or substrates listed in Table 4, considerable diversity on the ligand site of the targets can be observed.

Next, we validated the information content of this nine-dimensional space by performing similarity searching using solely similarities to those prototypical activity classes. Table 6 shows the retrieval rates when performing similarity searching in nine-dimensional (approximately) orthogonal ligand bioactivity space. Similarities of the query compound to each of the nine Bayes models are calculated, analogous to the library compounds. Structures are then compared, as above, by calculating the Pearson correlation coefficient of the nine axes. Overall, it can be seen that average retrieval rates are slightly inferior by only 2.89% in absolute terms (7.05% in relative terms), with performance differing greatly between the different activity classes. This way, an information-optimal panel of small molecule models has for the first time successfully been employed for similarity searching which possesses a sound theoretical basis and shows high retrieval rates and is as well computationally inexpensive (because of the reduced number of dimensions).

This kind of analysis can now be applied to guide the design of multitarget drugs, which is currently hotly debated,<sup>37</sup> as well as the elimination of undesired off-target effects of ligands. If activity classes show similar variable loading along one of the axes, this reflects a similarity of the particular activities, indicating that off-target effects

**Table 5.** (a) Cluster Analysis of the Principal Components of Ligand Bioactive Chemical Space PC1–PC3, Showing Dopamine D2 and  $\mu$  Opioid Receptor Ligands as well as HIV-1 Integrase Inhibitors,<sup>a</sup> (b) Cluster Analysis of the Principal Components of Ligand Bioactive Chemical Space PC4–PC6, Showing MMP3 and Thrombin Inhibitors as well as Ligands of the Endothelin A Receptor,<sup>a</sup> and (c) Cluster Analysis of the Principal Components of Ligand Bioactive Chemical Space PC7–PC9, Showing SRC Kinase and PAF2 Acylhydrolase Inhibitors as well as Ligands of the Cholecystokinin B Receptor

Principal Component (PC)	Target	Ligand Cluster Center 1	Ligand Cluster Center 2	Ligand Cluster Center 3
PC1	D2: Dopamine D2 Receptor (0.962)			
PC2	HIV1N: HIV 1 Integrase (0.806)			
PC3	Mu: $\mu$ Opioid Receptor (0.944)			
PC4	MMP3: Matrix Metalloprotease 3 (Stromelysin-1) (0.873)			
PC5	Thrombin: Thrombin (0.760)			
PC6	ETA: Endothelin A Receptor (0.674)			
PC7	CCKB: Cholecystokinin B Receptor (0.679)			
PC8	cSRC: SRC tyrosine-protein kinase			
PC9	PLA2: PAF2-acylhydrolase			

<sup>a</sup> Fragments indicating those bioactivity classes behave orthogonal to each other.

**Table 6.** Retrieval Rates of the Nine First Components of Bayes Affinity Fingerprints, Compared to Conventional (ECFP\_4) Fingerprints in Combination with the Tanimoto Coefficient<sup>a</sup>

class	Bayes affinity fingerprints				conventional descriptors (ECFP_4)				comparison	
	recall	stdev	unique	stdev	recall	stdev	unique	stdev	$\Delta$ recall absolute	$\Delta$ recall relative
06233	20.00%	8.00%	73	40	31.97%	10.53%	162	50	-11.97%	-37.43%
06235	46.11%	13.20%	240	83	25.18%	10.11%	63	27	20.92%	83.08%
06245	29.78%	17.32%	57	49	20.79%	13.37%	24	12	8.99%	43.23%
07701	57.78%	10.97%	160	45	23.93%	9.16%	26	11	33.85%	141.47%
31420	79.58%	16.42%	30	35	90.03%	7.72%	149	83	-10.45%	-11.61%
31432	53.15%	19.62%	145	99	70.51%	24.07%	510	242	-17.36%	-24.63%
37110	66.17%	24.77%	135	88	55.52%	11.16%	49	71	10.66%	19.20%
42731	7.64%	14.28%	39	26	32.16%	10.49%	348	139	-24.51%	-76.23%
71523	37.32%	10.79%	21	12	65.25%	9.71%	238	55	-27.93%	-42.80%
78331	9.33%	3.26%	29	12	13.31%	4.80%	54	21	-3.98%	-29.88%
78374	12.33%	10.11%	24	29	22.35%	17.68%	67	59	-10.02%	-44.85%
Average	38.11%	13.52%	87	47	41.00%	11.71%	154	70	-2.89%	-7.05%

<sup>a</sup> On average, retrieval rates are slightly inferior, with performance differing greatly between the different activity classes (same ligands used in both cases).

between the classes are more likely. Also, the likelihood is increased that performed structural modifications will influence activity against both (or a set of) targets, due to the fact that similar features in the models confer activity to both classes at the same time. Independent optimization of the activities is more difficult. The situation is different if activities are found to be orthogonally related to each other; in other words, orthogonal activities translate to structural features which confer activity against one target but are not related to activity against another. In those cases, activities are independent of each other, and affinity optimization should be easier.

Here, it needs to be noted that our analysis is based on current knowledge about the bioactivity of small molecules. It does not make statements about individual structures, and new ligand chemistry may well be able to guide drug discovery to a desired multitarget drug.

#### 4. CONCLUSIONS

We proposed a novel method that employs descriptor transformation for similarity searching which improves retrieval rates by about 24% compared to today's state-of-the-art methods. It capitalizes upon domain knowledge about current ligand bioactive chemical space to transform descriptors from *feature space* into *bioactivity space*. Knowledge about the shape of the "bioactive chemical universe" is thus beneficial to identify similar bioactivities. Via principal component analysis, a nine-dimensional orthogonal and information-optimal small-molecule bioactivity space was proposed and its information content relevant to bioactivity validated by similarity searching. By analyzing activity collinearities within the model, it can be employed to gauge whether activity classes can be independently optimized or not. The purpose of this analysis is 2-fold. First, it addresses the *elimination of off-target effects of drugs* ("can I modulate off-target activity B independently of on-target activity A?"). Second, it gauges whether two or more *desired on-target activities of drugs* would—on the basis of current chemical activity knowledge—be achievable or not.

#### ACKNOWLEDGMENT

The NIBR Program Office is thanked for supporting A.B.

**Supporting Information Available:** List of 100 target classes used for PCA, along with class sizes. Also correlations between the models of the nine target classes most highly correlated with the nine principle components of the whole bioactivity space. This information is available free of charge via the Internet at <http://pubs.acs.org>.

#### REFERENCES AND NOTES

- (1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (2) 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*, 3204–3218.
- (4) Bender, A.; Jenkins, J. L.; Li, Q.; Adams, S. E.; Cannon, E. O.; Glen, R. C. Molecular Similarity: Advances in Methods, Applications and Validations in Virtual Screening and QSAR. *Annu. Rep. Comput. Chem.* **2006**, in press.
- (5) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
- (6) Willett, P. Textual and Chemical Information Retrieval: Different Applications but Similar Algorithms. *Inf. Res.* **1999**, *5*, pp–pp.
- (7) Stiefl, N.; Baumann, K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure–Activity Relationship Technique. *J. Med. Chem.* **2003**, *46*, 1390–1407.
- (8) 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*, 6144–6159.
- (9) Baumann, K. Distance Profiles (DiP): A Translationally and Rotationally Invariant 3D Structure Descriptor Capturing Steric Properties of Molecules. *Quant. Struct.-Act. Relat.* **2002**, *21*, 507–519.
- (10) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569–6583.
- (11) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- (12) 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*, 3256–3266.
- (13) 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*, 1708–1718.

- (14) 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*, 1369–1375.
- (15) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.
- (16) WOMBAT World of Molecular BioAcTivity (WOMBAT); Sunset Molecular Discovery LLC: Santa Fe, NM. <http://www.sunsetmolecular.com/> (accessed Jul 2006).
- (17) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.
- (18) Beroza, P.; Villar, H. O.; Wick, M. M.; Martin, G. R. Chemoproteomics as a Basis for Post-Genomic Drug Discovery. *Drug Discovery Today* **2002**, *7*, 807–814.
- (19) Briem, H.; Lessel, U. In Vitro and in Silico Affinity Fingerprints: Finding Similarities beyond Structural Classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231–244.
- (20) Lessel, U. F.; Briem, H. Flexsim-X: A Method for the Detection of Molecules with Similar Biological Activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246–253.
- (21) Weber, A.; Teckentrup, A.; Briem, H. Flexsim-R: A Virtual Affinity Fingerprint Descriptor To Calculate Similarities of Functional Groups. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 903–916.
- (22) Weinstein, J. N.; Myers, T. G.; Oconnor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; vanOsdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **1997**, *275*, 343–349.
- (23) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470–480.
- (24) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra Analysis: Model Proteome Characterizations for Linking Molecular Structure and Biological Response. *J. Med. Chem.* **2005**, *48*, 6918–6925.
- (25) Oprea, T. I. Chemical Space Navigation in Lead Discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384–389.
- (26) Karthikeyan, M.; Glen, R. C.; Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *J. Chem. Inf. Model.* **2005**, *45*, 581–590.
- (27) Givchchi, A.; Bender, A.; Glen, R. C. Analysis of Activity Space by Fragment Fingerprints, 2D Descriptors, and Multitarget Dependent Transformation of 2D Descriptors. *J. Chem. Inf. Model.* **2006**, *46* (3), 1078–1083.
- (28) Sheridan, R. P.; Shpungin, J. Calculating Similarities between Biological Activities in the MDL Drug Data Report Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 727–740.
- (29) PipelinePilot 5.1; Scitegic: San Diego, CA. <http://www.scitegic.com/> (accessed Jul 2006).
- (30) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Comput. Sci.* **2006**, *46* (3), 1124–1133.
- (31) 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*, 199–204.
- (32) Statistica 7.0; Statsoft: Tulsa, OK. <http://www.statsoft.com> (accessed Sept 2006).
- (33) 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*, 529–536.
- (34) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (35) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.
- (36) Morphy, R.; Kay, C.; Rankovic, Z. From Magic Bullets to Designed Multiple Ligands. *Drug Discovery Today* **2004**, *9*, 641–651.
- (37) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can We Rationally Design Promiscuous Drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.

CI600197Y