Molecular Diversity (2006) 10: 311–332
DOI: 10.1007/s11030-006-9039-z

Full-length paper

Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance

Chris Williams
Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada, H3A 2R7
E-mail: cw@chemcomp.com, Tel.: +(514)-393-1055 ext 50, Fax: +(514)-874-9538

Received 28 October 2005; Accepted 25 January 2006

Key words: data fusion, fragment scoring, reverse fingerprints, similarity searching, scaffold hopping

Summary

Recent research has shown that using *data fusion* rules in fingerprint-based similarity searching can improve results over traditional searches. *Group* fusion scores, which use multiple reference compounds, have in particular been shown to be quite effective in increasing enrichment rates over single reference structure based searches. In this paper, the effectiveness of using data fusion with multiple reference compounds to increase similarity search recall rates was investigated using 44 biological targets and four different 2D fingerprinting systems, including a new 2D typed triangle fingerprinting system introduced here. Scaffold-hopping abilities using data fusion rules were investigated using eight (8) different classes of scaffolds active against cGMP phosphodiesterase isoform 5 (PDE5). An approach to using the reference group for ranking and visualizing important fingerprints bits, or *reverse fingerprinting,* was presented, and used to score and visualize important pharmacophore features within sample active molecules. Finally, similarity statistics within the reference groups were investigated and compared to recall rates.

Abbreviations: GpiDAPH, graph pi-donor-acceptor-polar-hydrophobe fingerprints; TGT, typed graph triangle fingerprints; PCH polar-charged-hydrophobe fingerprints; MACCS, 166 public MACCS keys; n, group count (# of reference structures); ROC, receiver operator characteristic curve; AVE, average fusion rule; MAX, maximum fusion rule; vHTS, virtual high-throughput screening; C_k, bit coverage in the training group; T_k, bit importance; w_L, pharmacophore fragment score; f_k^i, bit position

Introduction

Similarity searching with molecular fingerprints is a common virtual screening method used to mine compound libraries for new drug-like compounds [1–4]. Similarity searching is a relatively simple technique whose parameters include the reference structure (typically a known active compound), the *similarity metric* used to measure similarity between molecules (e.g., Tanimoto coefficient), and the fingerprint system (or *molecular representation*) used to describe the molecules. Many similarity metrics have been proposed, the most common being the *Tanimoto coefficient* $S(i,j)$ [1], which for two compounds i and j with fingerprints of length a and b, respectively is given by

$$\text{Tanimoto coefficient} : S(i, j) = c/[a + b - c].$$

Here a is the number of bits in molecule i, b is the number of bits in molecule j, and c is the number of bits in common between a and b. The similarity between the reference structure and each molecule in the compound library is computed, and the library is sorted and filtered based on decreasing similarity rank. Similarity searching relies on the similarity property principle [5], which states that similar molecules in general exhibit similar biological behavior. Although often violated [6, 7], the similarity property principle holds true in many cases, with many successful results [8]. Similarity searching is especially useful when little information is known about the system, because it requires no crystal structure of the biological target and a minimum one active molecule [9].

Data fusion similarity searching

Traditional similarity searching is performed with only one compound as the reference structure. The *data fusion* approach outlined by Willett [10] and Ginn [11] is a recent extension which merges multiple similarity scores into a consensus score in order to improve performance. Consensus scoring has become increasingly important in computational chemistry, having been applied to docking scoring functions [12–15] and QSAR predictions [16–20], as well as

similarity searching. A theoretical basis for why consensus methods work has been advanced by Wang and Wang [21], and extended by Feher [22]. Observations to date suggest that consensus approaches perform well because they emphasize the common wisdom of the combined methods, and suppress the weaknesses of each individual method. Consensus models can, in principle, be made using results from any type of prediction, and examples of increasing virtual high-throughput screening (vHTS) enrichment rates by combining dissimilar prediction types such as docking scores, QSAR predictions, and similarity search results have been reported [23].

There exist a number of approaches to performing data fusion on similarity search results to produce a *fusion* similarity score $\boldsymbol{S}$. These include [10]:

(a) combining similarities produced using different molecular representations,
(b) combining similarities produced using different similarity metrics,
(c) combining similarities produced using different reference compounds, and
(d) any combination of the above.

A recent study by Willett et al. [24] used collections of ten active compounds to perform similarity searches over 11 different biological targets to highlight the effectiveness of approach (c), i.e., using multiple reference structures. The multiple reference structure, or *group fusion* [25] approach, computes the score $\boldsymbol{S}$ by applying one of a number of data fusion rules to the similarities between the test molecule t and all of the reference molecules $\{Q_i\}$. Examples of the SUM, AVERAGE, MINIMUM and MAX fusion rules are given below.

SUM – fusion score is a sum of the similarities
$\boldsymbol{S} = \text{add}\,\{S(Q_i, t)\}$

AVE – fusion score is an average of the similarities
$\boldsymbol{S} = \text{ave}\,\{S(Q_i, t)\}$

MIN – fusion score is the *minimum* of the similarities
$\boldsymbol{S} = \min\{S(Q_i, t)\}$

MAX – fusion score is the *maximum* of the similarities
$\boldsymbol{S} = \max\{S(Q_i, t)\}$

Here $\{S(Q_i,t)\}$ is the set of similarities between all of the reference group molecules Q_i and the test molecule t. To facilitate discussion the set of query reference molecules $\{Q_i\}$ will be referred to as the *reference group* (or *group*) and the number of compounds (n) in the group will be referred to as the *group count,* or *count, n*. Studies to date suggest the MAX fusion rule is the most effective in fingerprint based similarity searches [24, 25].

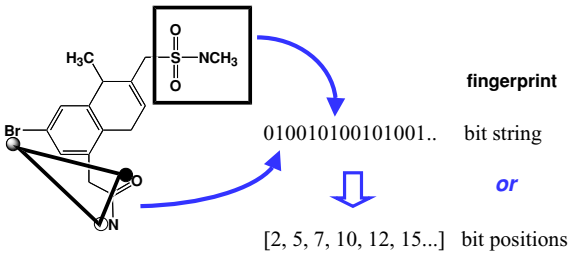


Figure 1. Molecular fingerprints: Each bit indicates the presence/absence (1/0) of a structural motif. The fingerprint may be a *bit string* indicating the on/off state of all the bits, or a *bit position vector* indicating the positions of the bits which are turned on.

Reverse fingerprinting and pharmacophore fragment scoring

In a fingerprint each molecule is typically described as a collection of *bits*, indicating the presence or absence of a structural motif [26]. The fingerprint may be represented as a bit string of 1's and 0's indicating the on/off state of each bit, or as a *bit position vector* which records only the indices of bits which are in the on state. The fingerprint bit position vector, FP_Q, of a molecule Q can be written as

$$FP_Q = [fx_1,\ fx_2,\ fx_3, \ldots fx_n],$$

where fx_i is the position of the ith bit which is turned on in the fingerprint bit string. The number of occurrences of each bit may be encoded by replacing the binary indicators in the FP vector with a frequency count for each bit [27]. There exist many types of fingerprints [28–31], which can be roughly classified based on the dimension of the molecular structure used to derive the fingerprint (2D or 3D), and the type of molecular fragments used to construct the bits – functional groups [32], extended graph connectivities [33], typed polygons [34] or descriptor value distributions [35] (see Figure 1).

2D fingerprints are derived from the chemical graph and may use graph distances, while 3D fingerprints usually imply the use of atomic coordinates and Euclidian distances. 2D fingerprints do not require conformations, and are hence fast and independent of force field accuracy, but they cannot be expected to capture 3D effects. In principle 3D fingerprints capture more information than 2D, but these methods are slower (requiring conformations) and can be sensitive to the conformation generation method used. In addition, the number of bits potentially produced in 3D can be large and may swamp the activity signal with noise bits – there is some indication in recent literature that 3D fingerprints may not substantially improve performance over than 2D fingerprints, and work continues on this issue [36, 37].

One attractive feature of fingerprints is the possible extraction of important fingerprint bits for substructure analysis, or *reverse fingerprinting*. One fingerprint type well suited for

this approach is the *typed pharmacophore polygon*, whose bits consist of polygons constructed using phamacophore-typed atoms or centroids as vertices. The polygons may be distances, triangles or tetrahedra, and the polygon edge distances can be measured in either 2D (bond graph distance) or 3D (Euclidian distance).

Given that the vertices in the typed polygon fingerprints can be the same as the feature centers used in 3D pharmacophore searches, the 2D and 3D bits may be used to help elucidate a 3D pharmacophore. 3D fingerprint bits that correspond to rigid portions of molecules may be sufficiently represented in 2D.

$$\text{2D } FP \leftrightarrow \text{3D } FP$$

As well, in cases where the 3D pharmacophore comprises a conformational restricted region of the active compounds, 2D fingerprints may be sufficient to elucidate a meaningful 3D pharmacophore.

$$\text{2D } FP \leftrightarrow \text{3D Pharmacophore query}$$

The advantage of being able to use the same vertices in 2D fingerprints, 3D fingerprints, and 3D pharmacophore searches motivated the creation of the PCH fingerprint, a typed atom triangle fingerprint based on the PCH (Polar Charged Hydrophobe) pharmacophore annotation scheme in the MOE software [38]. Details of the implementation are given in the Methods section. Typed-polygon fingerprints are not new [39–41] and other fingerprints could have been used in the bit importance and fragment visualization portion of this study. The main reason for implementing the PCH fingerprint here is simply the convenience of having the same vertices in 2D and 3D, which will help in subsequent analysis and visualization.

Fingerprint bit ranking and fragment scoring

In the context of a reference group, a *bit coverage*, C_k^Q, is defined as the fraction of the reference group molecules which contain the kth structural bit:

$$C_k^Q = \frac{1}{n}\sum_{i=1}^{n} f_k^i$$

In the above equation f_k^i is either 1 or 0, indicating the presence or absence of the kth bit in the ith molecule and n is the group count. The coverage is a partial indication of the importance of each structural bit in the reference group and can be used to help isolate bits which may be significant for activity. Note that the current definition of coverage does not take into account the possibility of multiple occurrences of bit k in the ith molecule. Although the effect of incorporating bit frequency should be studied further (especially in the context of bit-coverage), a recent study [24] using the CATS [41] and Semilog [27] fingerprints showed little difference in the recall rates produced by the binary and frequency count versions of these fingerprints. Thus, for simplicity, the current study was restricted to binary fingerprints only.

With conformationally-dependant fingerprints, f_k may include a Boltzmann weighting factor for the bit:

$$f_k^i = \frac{\sum_{\text{conf}} f_k^{\text{conf}} * e^{-\Delta E\text{conf}/kT}}{\sum_{\text{conf}} e^{-\Delta E\text{conf}/kT}}$$

Here the summations are over all the conformations of the ith molecule, and $f_k{}^{\text{conf}}$ is a 1 or 0 indicating the presence or absence of the kth bit in the conformation.

In addition to coverage within the reference group, a *bit importance* T_k, inspired by Bayesian likelihood ratios [42] and the Binary-QSAR method [43], is defined. The bit importance T_k is computed by taking a ratio of the frequency of the kth bit in the reference group with the frequency of the kth bit in a larger chemical space, typically a collection of random, inactive and/or diverse drug-like molecules.

$$T_k = \left[\frac{\frac{1}{n+1}\left(\left(\sum_{i=1}^{n} f_k^i\right) + 1\right)}{\frac{1}{m_0+1}\left(\left(\sum_{j=1}^{m_0} f_k^j\right) + 1\right)} \right]$$

In the above equation m_0 is the number of random or inactive compounds included in the analysis, and the 1's are introduced in the denominators to avoid division by zero.

The bit coverage and importance values can be used in combination to score the molecular fragments from which the fingerprint bits were made. For each pharmacophore fragment L used to construct bits (and fragment here can mean a molecular substructure, a pharmacophore triangle vertex, an atom in an extended connectivity, etc.), a score, w_L, is computed using C_k and T_k;

$$w_L = \sum_{k}^{K} (C_k * T_k)$$

Here, w_L is the score of the Lth pharmacophore fragment. The summation is over all K fingerprint bits which include the fragment L in its specification. The relative scores of each of the fragments can be visualized by drawing weighted spheres around the fragments, where the radius r_L for each fragment sphere is the fragment score scaled by the maximum fragment score as seen in the molecule:

$$r_L = w_L / \max\{w_L\}$$

Scaled this way, points with large relative radii will indicate fragments that have large relative scores.

The above definitions outline a reverse fingerprinting approach to scoring pharmacophore fragments in a molecule, as the scores and weighted radii can be used to detect the regions of the molecule which are most important for activity. The basic procedure for scoring molecular fragments with this method is as follows:

(1) Select a reference group of active compounds.
(2) Train T_k and C_k statistics on a test database that contains examples of active and inactive compounds. The exact

composition of the test database can vary depending on the intended application (activity, selectivity, etc).

(3) For a given query molecule, compute the fingerprint bits and use the C_k and T_k values to score the fragments used to construct each bit.

(4) Display the absolute fragment scores and weighted spheres on the query molecule.

The approach of using fingerprint bit frequencies to isolate important features in compounds is not new. The Stigmata approach [44] used a form of bit coverage to determine structural commonalities within compound datasets, and Xue et al. [29] have used statistical distributions to isolate bits which perform optimally for isolating active compounds. The above definitions for fingerprint bit importance and fragment scoring represent an approach to ranking molecular fragment contributions to biological activity.

Reference group similarity statistics

To further characterize reference groups, the average pair-wise similarity, S^Q_{AB}, the maximum pair-wise similarity, S^Q_{MAX}, and the minimum pair-wise similarity, S^Q_{MIN} are defined;

$$S^Q_{\mathrm{AB}} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^{n} S(i,j)$$

$$S^Q_{\mathrm{MAX}} = \max\{S(i,j)\}$$

$$S^Q_{\mathrm{MIN}} = \min\{S(i,j)\}$$

These statistics will be related to recall rates in order to determine if it is possible, *a priori*, to predict the recall performance of a given reference group. The above definitions hold regardless of the similarity metric; in this study, we limit ourselves to the square root of the Tanimoto coefficient, $\sqrt{[c/(a+b-c)]}$, the default coefficient in the MOE software. This coefficient was chosen for clarity in plotting, since we wish to visually examine correlations between recall and pairwise similarity statistics. The performance is identical to the Tanimoto coefficient. The advantage of the square root is that at low similarity values, the absolute similarities produced have more spread in the plots than with the regular Tanimoto coefficient. Other coefficients could have been used and may be the subject of future studies.

Outline of current study

The effectiveness of using multiple structures as opposed to a single reference structure in similarity searching prompted the following studies into the nature of the reference group $\{Q_i\}$, its effect on the quality of the similarity search, and its potential to help elucidate important phamacophores. This study is divided into the following four sections:

(1) *Group fusion: General behavior over multiple biological targets:* To assess the general performance of group data fusion over many biological targets, 44 biological targets (Table 1) were chosen for study. Fingerprint models were made using compounds active against each of the biological targets. Four different 2D fingerprints – the 166 public MACCS keys [45], the MOE TGT (typed graph triangles) scheme, the MOE GpiDAPH (graph pi-donor-acceptor-polar-hydrophobe) fingerprint, and the newly implemented PCH fingerprints – were used. Reference groups, varying in size from n = 1–40, were constructed using compounds active against each target. Fingerprint models were then made from the reference groups and used to virtually screen a 10000 compound in-house dataset of drug-like molecules. Both the AVE and MAX fusion results were tested. The recall rates of each fingerprint scheme and fusion rule were measured using the area under the receiver operator characteristic curve (ROC) [46], and reported in percent units (ROC area × 100). On this scale, 100 is a perfect model and 50 is a random model; areas of 90 and greater are considered excellent, while areas of 60 or less indicate no significant model. Gains in ROC areas as a function of group count were examined. The performance of the newly implemented PCH fingerprint was compared to that of the MACCS, TGT and GpiDAPH fingerprints, and the performances of the AVE and MAX fusion rules were discussed.

(2) *Group fusion: PDE5 scaffold hopping:* To test the ability of the data fusion similarity searches to 'scaffold-hop' – i.e., retrieve active compounds which do not belong to the same chemical class as the reference group – an experiment focusing only on the PDE5 inhibitors was performed. Eight (8) PDE5 inhibitor scaffold classes (Figure 2) were identified and used to make 8 different *single class* models consisting of scaffolds from one class only. The AVE and MAX fusion rules were tested using group counts of $n = 1$, 4, 7, and 10. Each single-class model was used to virtually screen a test database containing molecules belonging to other scaffold classes, but no active molecules belonging to the training set class. To test how effectively new scaffolds were hit by single class models, the number of new scaffolds and the total number of compounds retrieved in the top 100 ranked compounds were examined.

(3) *Reverse fingerprinting – Fragment scoring and visualization:* The potential of the fragment scoring method was briefly surveyed by computing and visualizing the fragment scores in example compounds (Figure 3) active against biological targets 5-HT2A, CDK2/Cyclin-A, FXa, PDE4 and PDE5. The fragment scores were compared with known pharmacophores for the biological target. The sensitivity of the fragment scoring was examined by comparing scores produced by fingerprint models trained using compounds active against different biological targets.

(4) *Reference group similarity statistics:* Reference group similarity statistics were gathered from the models created in Sections (1) and (2) and compared to recall rates. The ability of these measures to predict recall rates *a priori* was investigated.

Table 1. Biological targets codes: The names and abbreviation codes of the 44 biological targets considered in this study

Biological target	Code	Biological target	Code
5-HT1A serotonin	5-HT1A	Faresyl tranferase	FTase
5-HT1B serotonin	5-HT1B	Factor X alpha	FXa
5-HT2A serotonin	5-HT2A	HIV protease (type-1)	HIV-1-PR
5-HT2C serotonin	5-HT2C	HIV reverse transcriptase	HIV-1-RT
5-HT4 serotonin	5-HT4	HIV protease	HIVPR
Acetylcholinesterase	AChE	Hydrolase (multiple targets)	HLase
Adenosine A1 receptor	ADORA1	Histamine Receptor H3	HRH3
Alpha-1 adreneric receptor	ADRA1	K-opioid receptor	KOR
Adenosine kinase	AK	Matrix metalloprotease-2	MMP-2
Butylcholinesterase	BChE	Matrix metalloprotease-3	MMP-3
Carbonic anhydrase 1	CA-1	Matrix metalloprotease-8	MMP-8
Carbonic anhydrase 1	CA-2	M-opioid receptor	MOR
Cannabinoid CB1 receptor	CCB1	Melatonin MT1 receptor	MT1
Cholecystokinin B	CCKB	Neuropeptide Y receptor 5	NPY5
CDK2/cyclin-a-dependant kinase	CDK2-Cyclin_A	cAMP phosphodiesterase 4	PDE4
Cyclooxygenase-2	COX-2	cGMP phosphodiesterarse 5	PDE5
Dopamine D2	D2	Platelet derived growth factor b receptor	PDGFRB
Dihyrofolate reductase	DHFR	Pyruvate dehydrogenase kinase	PDHK
Dimerization partner 2	DP2	Protein tyrosine phosphatase 1B	PTP-1B
Epidermal growth factor receptor	EGFR	Streoid receptor coactivator	SRC
Endothelin A	ETA	Tumor necrosis factor-alpha converting enzyme	TACE
Factor II alpha	FIIa	Trypsin	Trypsin

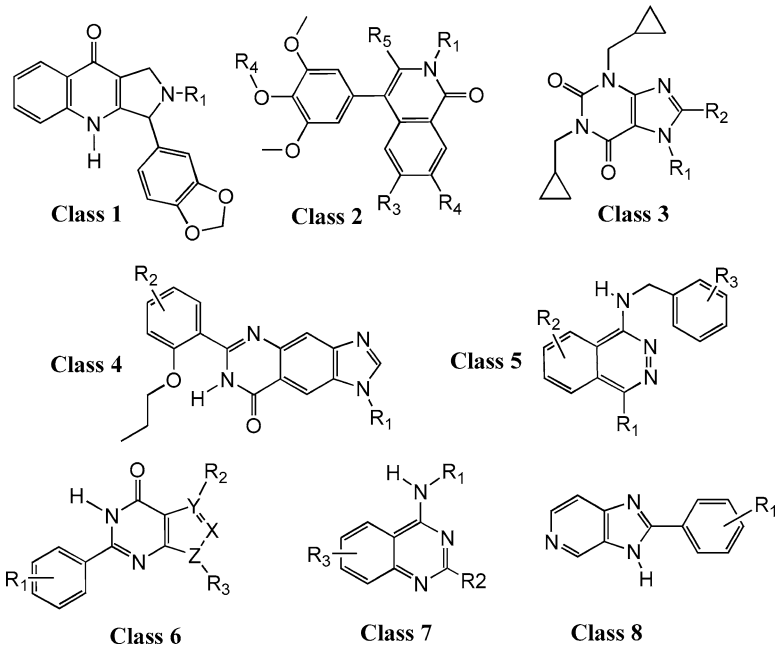


Figure 2. PDE5 inhibitor scaffold classes: Eight (8) scaffold classes identified from PDE5 active compounds in the entire dataset.

Methods

PCH typed polygon fingerprint implementation

The MOE software pharmacophore annotation points [38] were chosen as vertices for the typed graph triangles. In MOE, an annotation point scheme derives pharmacophore points from a molecule based on the rules of the scheme. These points are the same ones used by MOE 3D pharmacophore searching routines. Currently there are four (4) annotation schemes in MOE, some of which have projected features; a detailed explanation of the annotation schemes can be found in the MOE manual. For simplicity, only the PCH annotation scheme was considered in the current work.

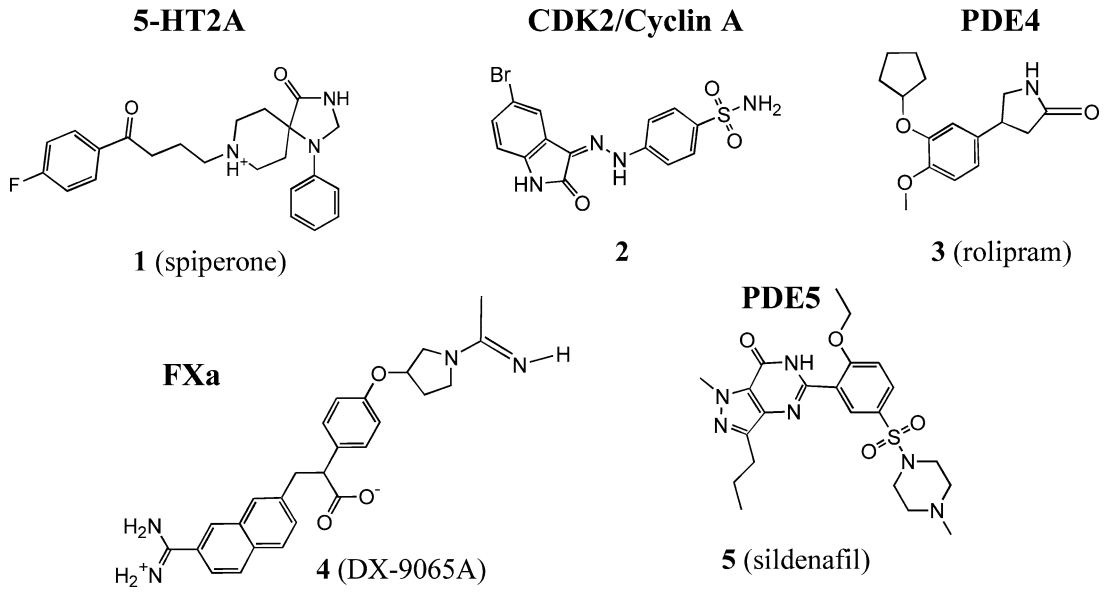


Figure 3. Sample active compounds for bit importance and visualization: Compounds active against 5-HT2A (**1**, spiperone), CDK2/Cyclin A (**2**), PDE4 (**3**, rolipram), FXa (**4**, DX-9065A) and PDE5 (**5**, sildenafil) used in bit importance and visualization examples.

The PCH scheme produces atom-centered annotation points for all features (donor, acceptors, etc.) except for aromatic rings, (where ring center annotation points are produced), and for large hydrophobic features, where annotation points are placed near the hydrophobic bulk centroids. The PCH scheme does not contain any annotation points projected from polar atoms. For the 2D version of the fingerprints, the shortest path graph distance is used as the distance between points. Annotation points not centered on atoms are assigned to be one bond distance from the atoms that derived the point. A final requirement for bit specification is the setting of distance bins for the triangle edges. All triangle edges whose distances fall within one bin range are considered to be equal. No attempt was made here to fine-tune the binning for better results. For now, the bin distances are used as is, but this does not preclude future efforts to optimize the binning for better 2D/3D correspondence. The following bin distances were used for the 2D PCH fingerprint.

1, 2, 3, 4, 5, 6, 7, 10, 12, 15

These bin distances are a compromise between the graph distance bins used in the GpiDAPH and the TGT fingerprints. For the 3D PCH fingerprints, the distance between the vertices is simply the Euclidian distance; the same distance bins used by the 3D TAT (typed-atom-triangle) fingerprints in MOE were chosen (in Å).

1.0, 2.19, 2.64, 3.05, 3.47, 3.92, 5.15, 8.26, 14.72

Although the 3D PCH fingerprints were implemented, they were not considered for this work, which focused entirely on 2D fingerprints.

Test and training datasets

The *entire dataset* used in this study consisted of 10106 unique compounds with biological measurements spanning 438 biological targets, compiled from *Journal of Medicinal Chemistry* articles spanning the years 1994 to 2004. Details of the dataset preparation are available upon request. Forty-four (44) of the biological targets (listed in Table 1) were chosen for this study to examine the effects of group fusion on similarity search recall rates. It is important to note that the data contains many congeneric series for each target, and there are many examples of active and inactive compounds for each target/chemical class combination. Since most of the measurements in the dataset are continuous, consisting mainly of pIC50 and IC50 values in different units, they were all normalized to pIC50 (M) activity values, and a threshold of pIC50 (M) >6 was set to separate the 'active' compounds for the study. The normalization procedure was less than perfect (due to variation in activity experiments among different journal articles), so this procedure introduced some noise into the activity data. The pIC50 threshold of 6 is also somewhat arbitrary, and other thresholds could have been chosen. By computing the number of active compounds as a function of different activity thresholds (Table 2), one can see that compounds for each of the forty-four targets span a number of activity ranges.

Choosing an activity threshold of pIC50 > 6 means the resulting data set contains a substantial number of inactive compounds which belong to the same chemical classes as actives, and lie on the interface region of pIC50s between 4 and 6. At an activity threshold of pIC50 > 4, the sum of all the active compounds is *greater* than the number of compounds in the dataset, because many compounds are active against

Table 2. Number of active compound for each biological target at various pIC50 Thresholds: Number of active compounds in the entire database at three different activity thresholds (pIC50 = 4, 6, 8) for each of the 44 biological targets considered in this study. The total number of actives for all targets is greater than the total number of compounds in the entire dataset, because many compounds are active against multiple targets

	Activity threshold				Activity threshold		
Code	pIC50 > 4	pIC50 > 6	pIC50 > 8	Code	pIC50 > 4	pIC50 > 6	pIC50 > 8
5-HT1A	505	480	165	FTase	327	216	55
5-HT1B	161	118	9	FXa	330	221	63
5-HT2A	214	193	58	HIV-1-PR	336	319	233
5-HT2C	145	122	23	HIV-1-RT	198	121	4
5-HT4	91	89	15	HIVPR	226	219	186
AChE	186	87	30	HLase	304	211	16
ADORA1	457	355	55	HRH3	92	88	34
ADRA1	573	522	161	KOR	320	248	106
AK	107	94	36	MMP-2	446	412	223
BChE	109	70	9	MMP-3	477	418	117
CA-1	696	483	30	MMP-8	313	279	148
CA-2	723	701	290	MOR	381	301	153
CCB1	107	98	27	MT1	170	164	102
CCKB	81	73	21	NPY5	188	158	61
CDK2-Cyclin_A	123	97	42	PDE4	221	162	44
COX-2	244	112	6	PDE5	549	400	234
D2	616	535	116	PDGFRB	474	184	12
DHFR	175	132	47	PDHK	124	76	7
DP2	85	84	5	PTP-1B	210	102	0
EGFR	359	184	61	SRC	344	225	48
ETA	417	375	207	TACE	144	141	34
FIIa	376	158	58	Trypsin	281	107	1
Total (all Targets)	13005	9934	3352				

a number of the targets at this threshold, and thus count for more than one active compound when computing the total number of active compounds in the dataset.

A *training set* of 1760 unique compounds was constructed by randomly extracting from the 10106 compound entire dataset 40 active compounds for each of the 44 targets in Table 1. All of the training compounds were removed from the entire dataset to create an 8346 compound *external test set* which was devoid of any of the training compounds. The external test set was used for determining the recall rates of the fingerprint models, and for the reverse fingerprint fragment visualization studies.

Group fusion: general behavior over multiple biological targets

To assess the general behavior of group fusion over many biological targets, a fingerprint model for biological activity against each target in Table 1 was made by using reference groups consisting of compounds active against the target being modeled. Four different 2D fingerprinting schemes - the MACCS, TGT, GpiDAPH and the new PCH fingerprints – were used. Reference groups of the following sizes – 1, 2, 4, 7, 10, 15, 25, 40 – were constructed for each target. The MOE 2005.06 Fingerprint-Model application was used to create fingerprint models using the active compounds. The fingerprint models were then used to virtually screen the external test dataset for active compounds. For reference group counts of $n = 1$–25, the reference groups were chosen by randomly selecting the appropriate size subset from the 40 training compounds for each target; this was repeated five (5) times for each value of n, and the ROC areas averaged over all repetitions to give the ROC area for the group count. For the group count $n = 40$, all the training compounds for each target were used to construct the model. Both the AVE and MAX fusion rules were used.

Group fusion: PDE5 scaffold hopping

To create the training and test sets for the PDE5 scaffold hopping study, eight (8) scaffold classes shown in Figure 2 were separated from the 549 compounds with PDE5 pIC50 values greater than 4. PDE5 active compounds which did not belong to one of the eight classes were removed from the dataset, leaving a total of 325 PDE5 active compounds in the 'pruned' entire dataset. For each scaffold class, a *single-class* training set containing only compounds belonging to that class was created, along with a corresponding test set that consisted of the pruned entire test set minus the PDE5 inhibitors used to make the single class model. The ability of a single-class model to hit itself was not considered. Fingerprint models constructed with each single-class training set were used to virtually screen the corresponding test set. The AVE and MAX fusion rules were tested. The reported ROC was averaged over five (5) repetitions of each of the tested group counts, $n = 1, 4, 7, 10$. The groups were

randomly chosen from the single class training sets at each repetition.

Reverse fingerprinting fragment scoring and visualization

An SVL program was written in the MOE to train the C_k and T_k statistics, to calculate the fragment scores w_L, and to display the scaled spheres and scores on a sample structure. For each example target – 5-HT2A, CDK2/Cyclin-A, PDE4, FXa, and PDE5 – a sample active compound (Figure 3) was chosen. The training and test databases used in this portion of the study are the same as those used in the multiple target study. For each target, the reference group of size 10 was randomly chosen from the 40 active structures in the 1760 compound training dataset. The external test dataset was used to train the C_k and T_k statistics.

Reference group similarity statistics

The reference group similarity statistics for all of the models created in the testing of fusion rules on multiple targets were used to accumulate the reference group pair-wise similarity statistics. The S^{Q}_{AVE}, S^{Q}_{MIN} and $S^{Q}{}_{\mathrm{MAX}}$ values for each model were computed and compared with model ROC areas to correlate these statistics with recall rates.

Results and discussion

Group fusion: General behavior over multiple biological targets

Over multiple biological targets both the AVE and MAX fusion rules on average increased recall rates over single reference compound similarity searches, with the MAX rule and a group count $n = 40$ producing the best results overall. The ROC area averaged over all 44 targets using the AVE and MAX fusion rules are plotted as a function of group count n in Figures 4 and 5.

It was somewhat surprising to note that all fingerprint systems using the AVE fusion rule showed a leveling off in the ROC area gain when n exceeds 10. At the maximum group count of 40, the MAX fusion rule ROC gains seem to still be increasing, albeit slowly, suggesting they could continue to improve as the group count increases beyond 40. The largest improvements were exhibited by the MAX rule, with ROC area gains of 15–20 produced at $n = 40$. The average ROC area gains using the AVE fusion rule ranged from ~5 for the GpiDAPH fingerprint to 10–12 for MACCS and PCH fingerprints. The newly implemented PCH fingerprints performed at least as well as the other fingerprints, indicating they are reasonable for use in the fingerprint bit analysis and fragment visualization sections of the study.

To better compare the performance of data fusion across targets, a plot of the average ROC area for each target using the MAX and AVE fusion rules and group count $n = 10$ was

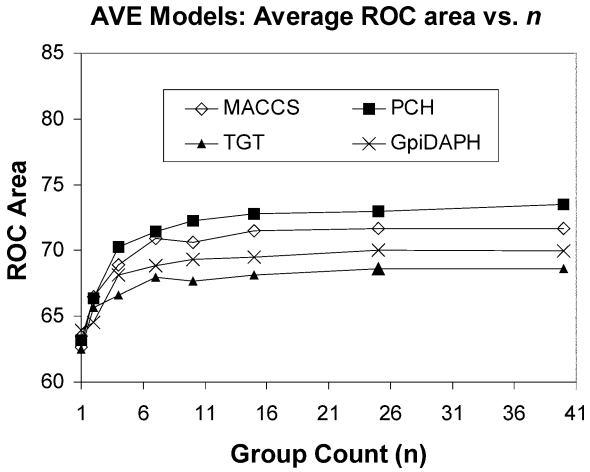


Figure 4. AVE model ROC area vs. group count: The ROC area using the AVE rule as a function of group count n as averaged over all the biological targets.

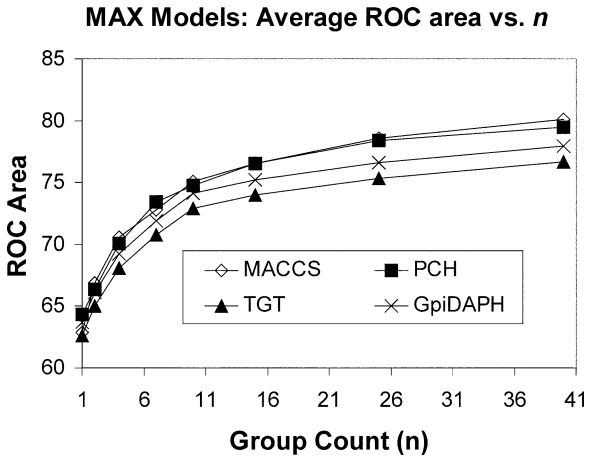


Figure 5. MAX model ROC area vs. group count: The ROC area using the MAX rule as a function of group count n as averaged over all the biological targets.

constructed. In both cases the plots are sorted in descending order based on the PCH ROC model value for $n = 10$. The AVE fusion rule ROC area plot in Figure 6 shows that for targets where the PCH ROC area at is greater than 85, the MACCS and GpiDAPH fingerprints ROC area values are also quite good, lying in the 80–100 range. Except for the AK target, the TGT fingerprint also performs well for these targets, with ROC areas in the 80–100 range. As the PCH $n = 10$ model ROC area decreases, the variation in ROC area between fingerprinting systems becomes more pronounced. Similar trends are observed for the MAX fusion rule (Figure 7).

For all of the 44 targets studied, a group count of 10 and either the MAX or AVE fusion rules produced an ROC area of greater than 60 with at least one of the fingerprints. All fingerprint systems produced an ROC area of 50 or less for at least one target, suggesting a worse than random model.

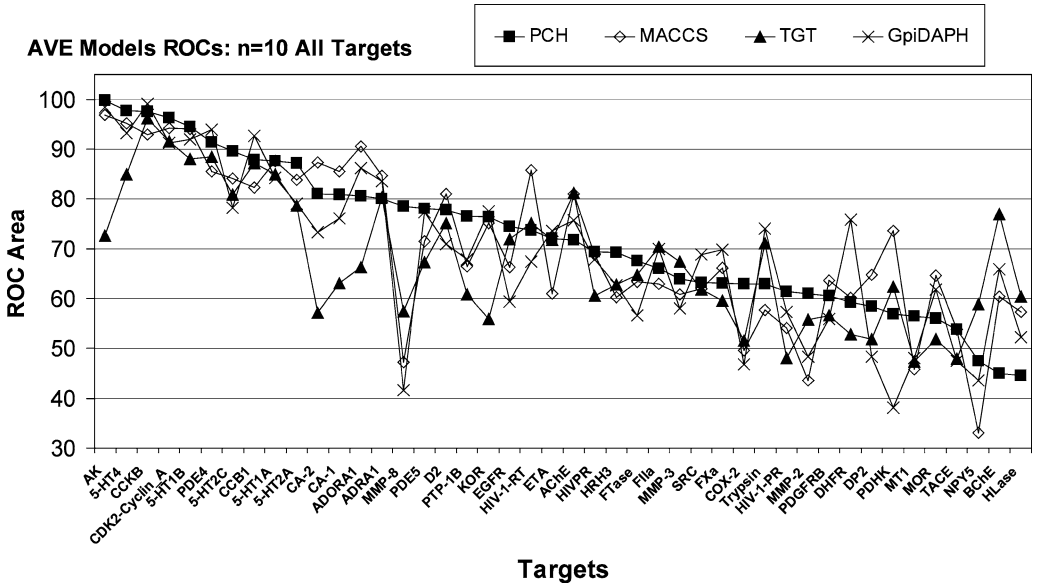


Figure 6. AVE models ($n = 10$) over biological targets: The ROC area for each biological target using a group count of 10. Plotted by descending order of the PCH fingerprint ROC area. Results for the PCH, MACCS, TGT and GpiDAPH fingerprints are shown.

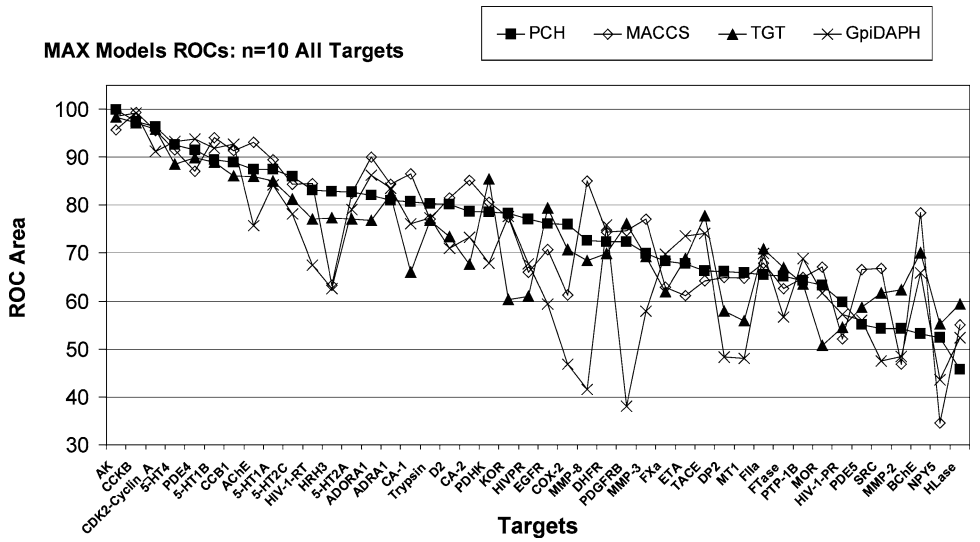


Figure 7. MAX models ($n = 10$) over biological targets: The ROC area for each biological target using a group count of 10. Plotted by descending PCH fingerprint ROC area. Results for the PCH, MACCS, TGT and GpiDAPH fingerprints are shown.

Some targets (such as CDK2/Cyclin-A and CCKB) were well modeled by all fingerprints, possibly indicating chemotype biases in the training sets and/or an activity class amiable to this type of analysis. Other targets such as the HLase and NPY5 were poorly modeled by all fingerprints and fusion rules, suggestion either the fusion rule, the fingerprints, or both, were inappropriate for modeling these activities. Some targets (notably COX-2 and MMP-8) were well modeled by some fingerprints and poorly by others, possibly highlighting the strengths and weakness of the different fingerprint systems.

In many cases when an ROC area of 50 or less was produced for a target using a particular fingerprint, a higher ROC area was obtained for the same target using another fingerprint, suggesting that combining similarity fusion scores from different fingerprinting systems could smooth out inconsistencies and improve overall performance.

For many of the biological targets, the results obtained using the PCH fingerprints and group counts of 10 and 40 were quite acceptable, using either the AVE or MAX fusion rules. It should be noted that the averaged ROC area gains do not reflect the variation between targets. Some systems experienced no ROC area gains, while others experienced dramatic gains in ROC area. Some targets even experienced a *loss* of predictive ability upon increasing n.

The MAX and AVE fusion rules at $n = 10$ and $n = 40$ were compared directly using the PCH fingerprints. In Figure 8 the PCH ROC areas over all targets using the MAX and AVE fusion rules and group counts of $n = 10$ and $n = 40$ are plotted, sorted by descending ROC area as obtained using the AVE rule and $n = 10$.

The AVE rule outperforms the MAX rule in some cases when $n = 10$, but the MAX rule is always either comparable or better than the AVE rule at $n = 40$. There are many targets where the MAX and AVE rules give similar results for both group counts. The plot in Figure 8 also shows how performance improves for many targets with the MAX rule, while performance remains relatively unchanged with the AVE rule. The MAX fusion rule ROC area gains from $n = 10$ to $n = 40$ are dramatic for a few targets, but there is little change with other targets. The MAX ROC area surprisingly *decreases* for a few targets as n increases from 10 to 40 – these cases should be the subject of future studies.

Group fusion: PDE5 scaffold hopping

The recall rates achieved when filtering for other scaffold classes using single class models are plotted in Figure 9. The percents of the database filtered before all of the other classes are hit were also computed. Interestingly, the AVE models performed somewhat better than the MAX models, but not by much. With the AVE rule, increasing the group count from 1 to 10 resulted in ROC gains for 4 of the models, with a decrease in ROC gain exhibited by one model. In contrast, the MAX fusion rule resulted in increased ROC gains for only 3 models, with 3 other models exhibiting a net decrease in ROC with increasing n.

To more deeply examine new scaffold retrieval, the percentage of the test database filtered when each single class model hits another scaffolds for the first time was examined (Table 3). This percentage gives some indication of how quickly a given single class models can retrieve other scaffolds, and how much better the model is than random selection. A group count of $n = 10$ and the AVE fusion rule was used in all cases. The ROC area of the model is listed in Table 3, along with the total number of compounds within each training scaffold class. The total number of hits and the number of new scaffolds hit in the top 100 ranked compounds was also recorded for each class model (Table 4). Since 100 compounds correspond to roughly 1% of each class test database, a 'random' model would be expected to hit 1% of the PDE5 active compounds, or 3 hits. Models which hit more than 3 compounds are better than random.

The drawings in Figures 10a–10c show the first examples of new scaffolds hit in the top 100 ranked compounds, along with the rank in hit list (1 = first hit, 100 = last hit), the class membership, and the PDE5 activity (pIC50 M) of each compound. The number of scaffolds hit, the total number of compounds hits, and the relative rankings of the compounds all vary greatly between class models.

Table 3. PDE5 single scaffold class models: For scaffold classes 1–8, the percentage (%) of the test database filtered at which the first and last instances of hitting each of the other scaffold classes is listed. The AVE fusion rule was used, with a group count of 10 in all cases. The ROC area of the model is also listed, along with the total number of compounds in each scaffold class

	Class 1 model		Class 2 model		Class 3 model		Class 4 model		Class 5 model		Class 6 model		Class 7 model		Class 8 model	
Total # of compounds in scaffold class	46		28		36		14		36		47		92		26	
ROC area: (%)	76.50		71.20		73.10		74.20		86.40		73.20		85.60		81.80	
	% Test DB filtered		% Test DB filtered		% Test DB filtered		% Test DB filtered		% Test DB filtered		% Test DB filtered		% Test DB filtered		% Test DB Filtered	
Classes hit	First hit	All hit	First hit	All hit	First hit	All hit	First hit	All hit	First hit	All hit	First hit	All hit	First hit	All hit	First hit	All hit
Class 1 hits	–	–	2.40	32.94	2.06	52.75	5.85	56.52	1.97	25.76	6.29	55.82	3.89	42.86	15.93	59.68
Class2 hits	0.12	5.16	–	–	0.54	11.07	5.31	28.93	2.97	27.87	1.28	19.45	11.58	37.85	27.44	59.00
Class 3 hits	0.14	76.52	0.01	79.31	–	–	0.29	83.67	0.38	86.86	0.19	69.48	3.25	67.13	3.66	55.52
Class 4 hits	0.52	36.28	3.73	32.76	0.49	21.86	–	–	1.58	11.64	0.01	0.17	7.80	15.38	0.22	7.89
Class 5 hits	0.41	19.13	8.61	29.68	6.57	44.55	4.17	44.27	–	–	5.41	30.91	0.01	3.42	10.70	38.39
Class 6 hits	2.27	60.56	1.50	67.75	0.82	56.80	0.01	12.61	1.06	63.93	–	–	8.37	43.88	0.08	10.83
Class 7 hits	0.81	98.80	2.32	98.23	2.26	98.34	6.40	96.47	0.01	91.26	7.30	97.29	–	–	0.38	26.93
Class 8 hits	24.46	96.44	26.11	98.22	35.38	97.60	11.83	94.14	26.01	89.81	0.15	89.54	5.77	35.03	–	–

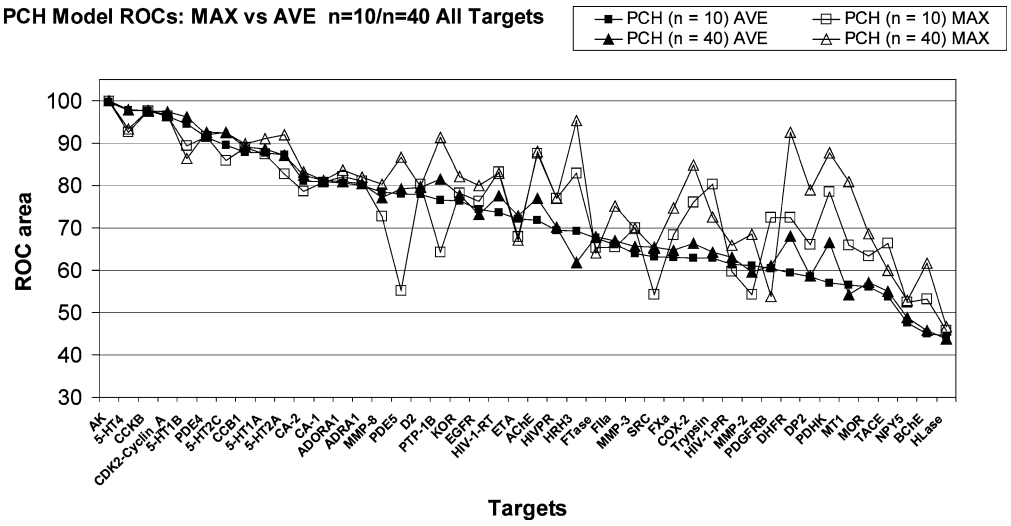


Figure 8. AVE vs MAX models ($n = 10/n = 40$) over all biological targets: PCH fingerprint ROC area for all biological targets using the MAX and AVE fusion rules. Group counts of $n = 10$ and $n = 40$ are shown. Data is plotted by descending $n = 10$ PCH AVE ROC area.

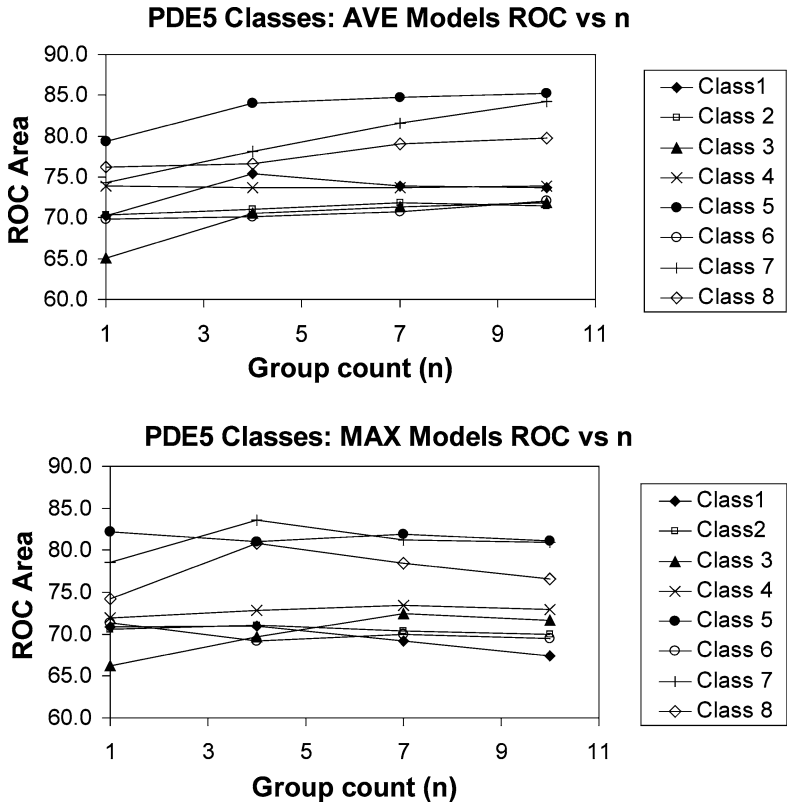


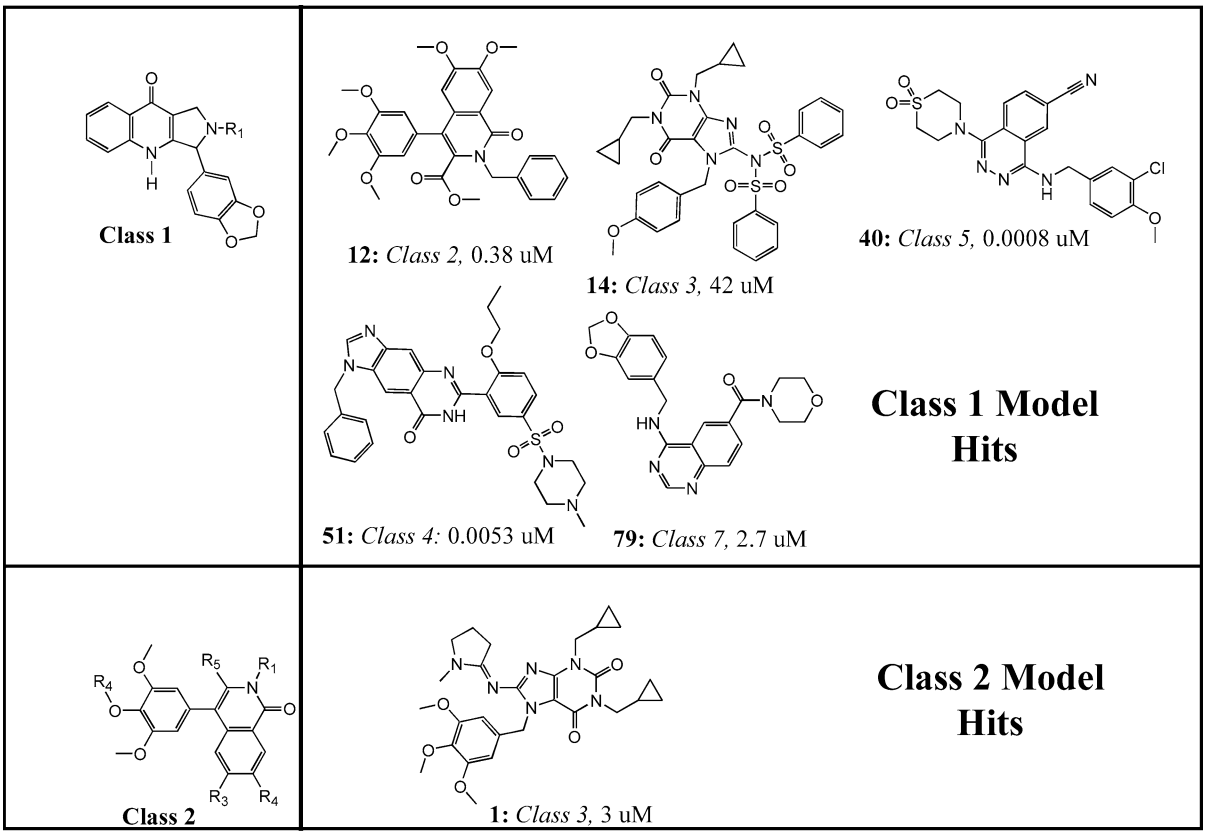
Figure 9. ROC Area vs. Group Count for PDE5 Single Class Models: The AVE and MAX fusion rule ROC areas obtained with PCH fingerprint single scaffold-class models (classes 1–8) plotted vs. group count n.

At one extreme, the Class 1 and Class 3 models both hit a number of different scaffolds (5 and 4 respectively – Table 4), but overall picked up a relatively small number of hits (15 and 12) within the top 100 compounds. This hit rate is still much better than random. The Class 2 and 7 models were the poorest at retrieving new scaffolds, each hitting only a small number of compounds from one scaffold class in the top 100 compounds. The class 4, 5 and 6 models all retrieved a substantial amount of compounds (>10) from one scaffold in particular.

The relative ranking of hits also varies between class models. Most of the 1, 2, 3, 7, and 8 class model hits spread out

Table 4. Single-class model hits in the top 100 compounds: Number of hits (by class) retrieved in the top 100 ranked compounds by each single class model

	Class Models							
	Class 1 model	Class 2 model	Class 3 model	Class 4 model	Class 5 model	Class 6 model	Class 7 model	Class 8 model
Class 1 hits	–	0	0	0	0	0	0	0
Class 2 hit	6	–	7	0	0	0	0	0
Class 3 hits	4	8	–	3	4	4	0	0
Class 4 hits	2	0	2	–	0	14	0	1
Class 5 hits	2	0	0	0	–	0	11	0
Class 6 hits	0	0	3	43	0	–	0	20
Class 7 hits	1	0	0	0	26	0	–	0
Class 8 hits	0	0	0	0	0	10	0	–
Total hits in top 100 ranked compounds	15	8	12	46	30	28	11	21

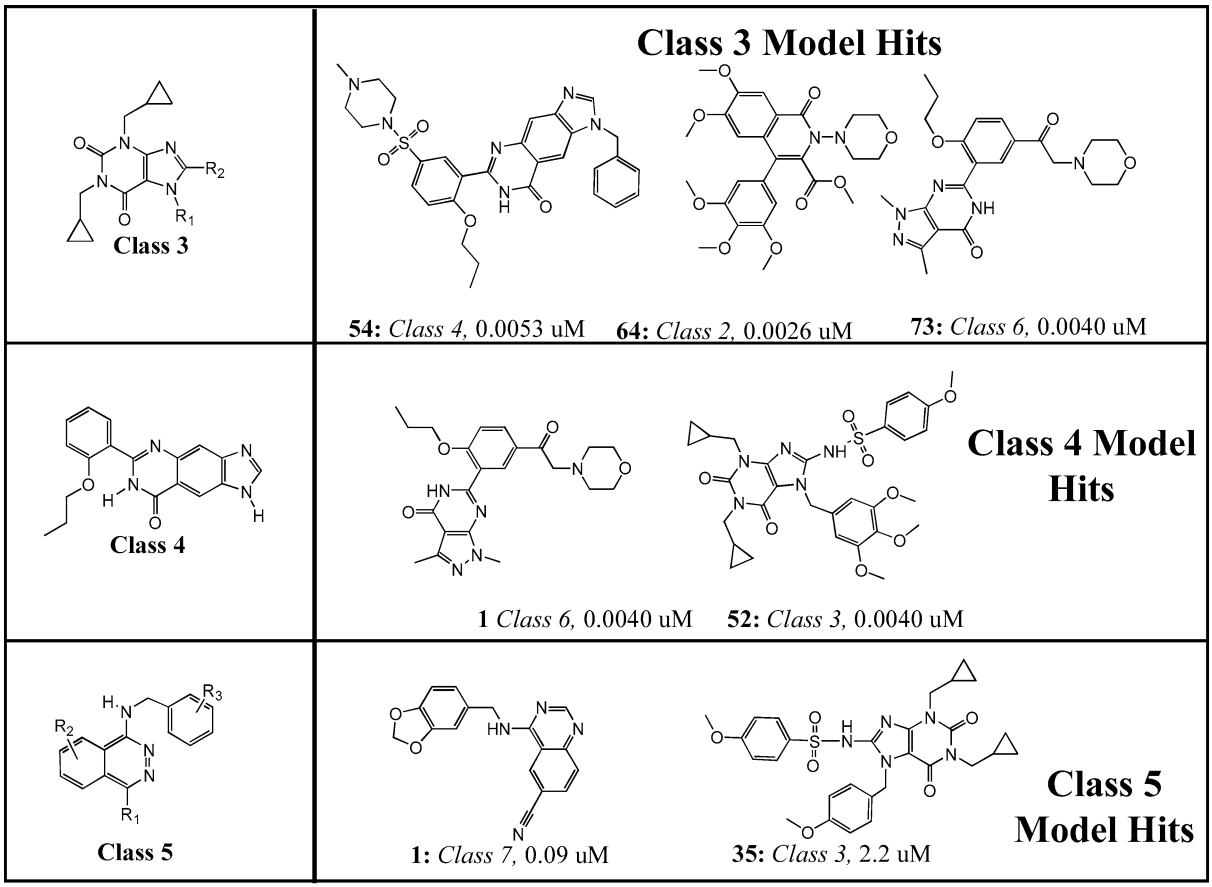


(a)

Figure 10.a. Sample hits in Top 100 ranked compounds using single class models – class models 1 and 2: The first hit compound from other scaffold classes using single class models 1 and 2. The list position (in **bold,** from 1–100)), class (in italics) and PDE5 IC50 (in *u*M) are given.

across the top 100 compounds (as demonstrated by the ranks of the 5 different scaffold hits of the class 1 model), while hits obtained with class models 4, 5 and 6 clustered at the top of the list. The class 4, 5 and 6 models all retrieve 18 or more hits in the top 25 compounds; the class models 4 and 5 retrieved only one new scaffold in the top 25 compounds, while the class 6 model retrieved three new scaffolds in the top 25. The class 1 and class 8 scaffolds seemed to be the most difficult to detect by the other classes – no class detected a class 1 hit in the top 1% of compounds, while only the class 6 model detected a class 8 compound in the top 100.

Overlap between classes and variation in class retrieval was examined using the percentage of the database filtered at which each single class model retrieved *all* of the compounds



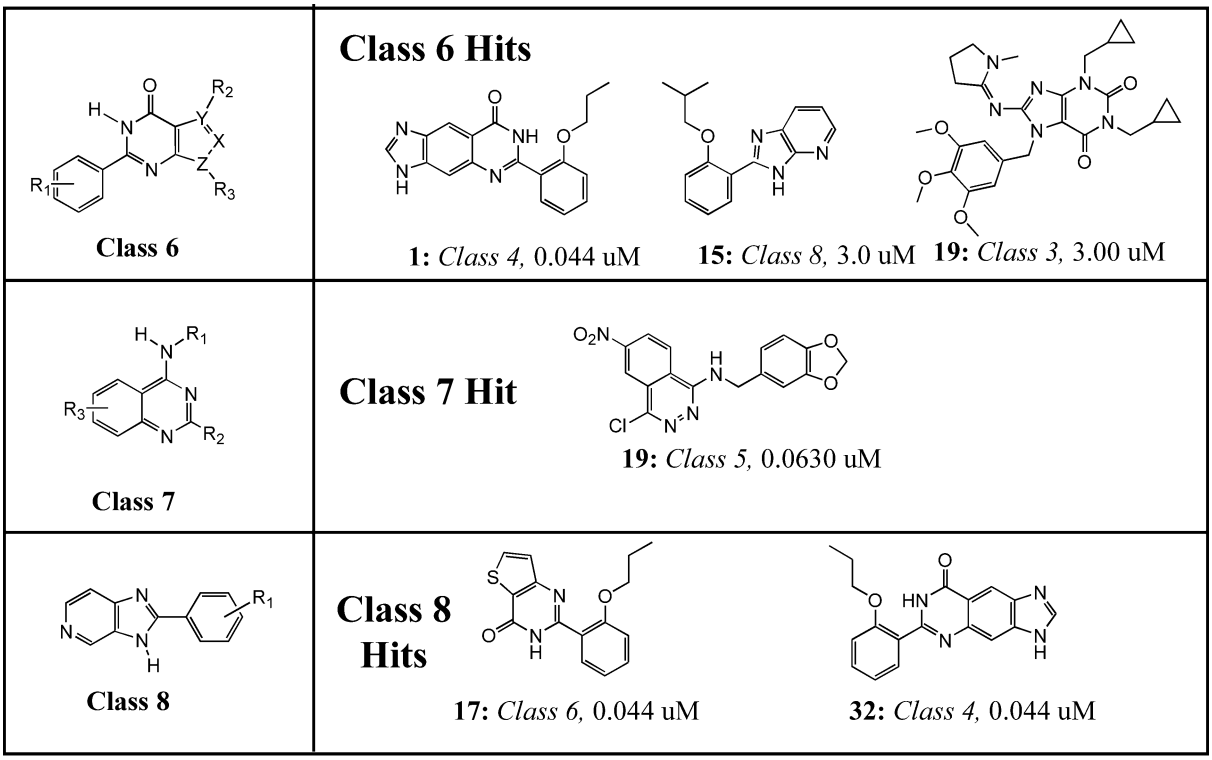
(b)

Figure 10b. Sample hits in top 100 ranked compounds using single class models – class models 3, 4, and 5: The first hit compound from other scaffold classes using single class models 3,4 and 5. The list position (in **bold,** from 1–100)), class (in italics) and PDE5 IC50 (in *u*M) are given.

from each of the other classes (Table 3). This percentage gives an idea of how spread out one class is relative to another. For the purpose of discussion, class hopping from class *A* to class *B* was considered to be 'solid' only if the class *A* model retrieved *all* of the class *B* compounds in the top 20% of the database filtered. The scaffolds were arranged in a diagram (see Figure 11) to visualize the data.

The solid arrows in Figure 11 are drawn from one class to another to indicate that a single class models retrieved all compounds from the other class in the top 20% of the test database. Dashed arrows show selected weaker hops, where more than 20% of the test database was filtered before all compounds of the other class were retrieved. Some interesting results become apparent when the data is arranged in this fashion. In all cases the most successful PDE5 inhibitor class hops occurred between scaffolds with minor differences. This result is to be expected from similarity search based methods. Classes devoid of exocyclic carbonyl groups (Classes 5, 7 and 8) successfully hopped only to scaffolds classes with one exocyclic carbonyl group or less. Class 7 picked up all of class 5 in the top 3.4% of filtered compounds, which is not surprising because this scaffold change involves only the shift in position of a heterocyclic nitrogen atom. Classes 5 and 7 both hopped with moderate success to class 4. Class 4 and Class 6 are the only pair of classes that mutually retrieved each other in the top 20% of filtered compounds. In addition, three (3) other classes hop to class 4 (5, 7 and 8) and class 8 hops to class 6. The results indicate that, at least for this dataset, classes 4 and 6 are quite close to each other, and lie in a central location of the chemical activity space defined by this dataset. Not one class hopped to class 3 very convincingly, the best result being class 8, which required filtering >50% of the database to hit all the class 3 compounds. Despite the low overall recall rates, five classes (1, 2, 4, 5 and 6) managed to retrieve examples of class 3 compounds in the top 100 scoring compounds (Figure 10), suggesting that class 3 compounds occupy a region of chemical space located somewhere in-between these classes.

Class 1 is interesting case that merits further discussion; single class models made with class 1 compounds show mediocre recall rates for all other classes, requiring filtering of 25%–60% of the test database in order to completely retrieve all other classes. In addition, no single class model retrieves a class 1 molecule in the top 100 ranked compounds.



(c)

Figure 10.c. Sample hits in Top 100 ranked compounds using single class models - class models 6, 7, and 8: The first hit compound from other scaffold classes using single class models 6, 7 and 8. The list position (in **bold,** from 1–100)), class (in italics) and PDE5 IC50 (in *u*M) are given.

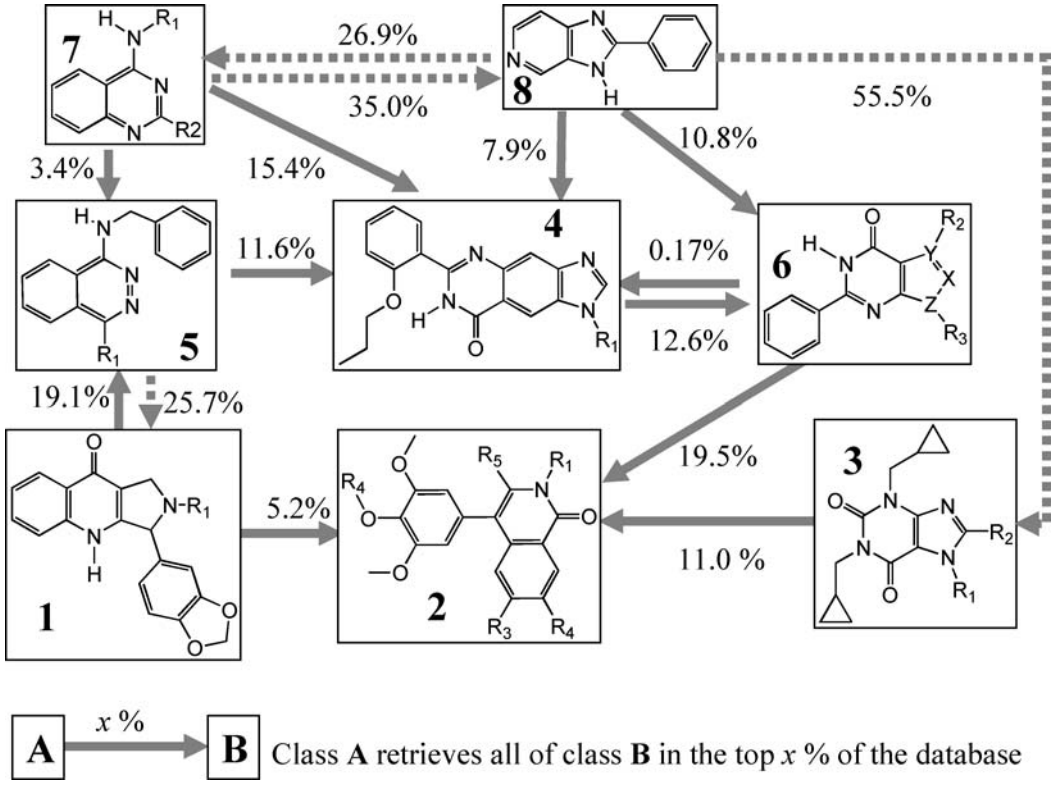


Figure 11. Complete class retrieval by single class models: Arrows drawn from one class to another indicate the percentage (%) of the database filtered when all members of the other class are retrieved. The arrow is drawn *from* the class used in the model *to* the class retrieved by the model. Solid arrows are drawn when the percentage of the database filtered is below 20%. Dashed arrows show selected weaker relationships.

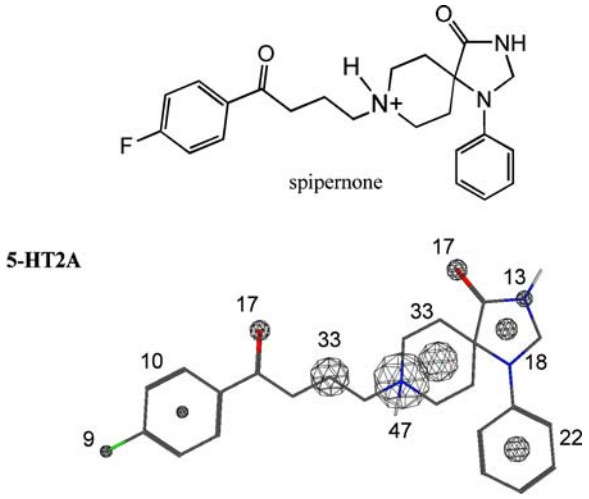


Figure 12. Fragment Scores for 5-HT2A active compound spipernone: The line drawing of spipernone along with the fragments scores and weighted spheres produced by the 5-HT2A model.

However, despite the low overall recall rates produced with the class 1 group, the class 1 model does find five (5) new classes (classes 2, 3, 4, 5, 7 – Figure 10a) in the top 100 ranked compounds, making it the most successful single class model by this measure. A possible explanation for this behavior is that class 1 compounds occupy a distinct region of chemical space somewhere in-between and equidistant to the other classes. Most of the other classes are sufficiently close to other classes that the majority of hits retrieved come from proximal classes, and do not include class 1 compounds in the top ranks. Class 1 compounds, on the other hand, are not close to any one class in particular, but are close enough to a number of other classes that active compounds can be retrieved from each. As a result, class 1 shows little preference for one class, and instead retrieves examples (albeit few) from a number of the other classes.

Overall the results of the scaffold-hopping experiment suggest that successful class hops can occur between dissimilar compounds by going through intermediate classes – for example, class 7 could eventually lead to class 2 via classes 4 and 6.

Reverse fingerprinting – fragment scoring and visualization

The bit importance and fragment visualization portion of the study was restricted to the PCH fingerprints only. A sample compound was chosen for each of five biological targets – spiperone for 5-HT2A (1) [47], an oxindole-based inhibitor for CDK2/CyclinA (2) [48], rolipram for PDE4 (3) [49], compound DX-9065A (from the 1FAX x-ray PDB structure) (4) [50] for FXa and sildenafil [51] (5) for PDE5 (Figure 2). For each target, a reference group of size $n = 10$ was used to train the C_k and T_k data on the external dataset. Examination of the resulting C_k and T_k values showed that typically only a few bits have high coverage values, but these bits also tended to have high T_k values. In the 5-HT2A, CDK2/Cyclin-A and PDE4 cases, all bits with coverage values above 0.3 had T_k values greater than 1. As the bit coverage values fall below 0.3, the number of bits with $T_k < 1$ increases. The T_k values split evenly between values greater than 1 and less than 1 when the bit coverage approaches 0.1.

5-HT2A fragment scoring

The $n = 10$ 5-HT2A model fragment scores are displayed on spipernone in Figure 12 The line drawing of spipernone is shown along with the weights of the fragments (bold score numbers) and fragment spheres of radius r_L. The point scores and weighted spheres show that the fragment scoring clearly picked a basic nitrogen center as the most important point, with proximal aliphatic groups having nearly the same importance. An aromatic center was also identified in the four top scoring points. These features correspond well with experimental 5-HT2A structure-activity relationships, which suggest the main 5-HT2A pharmacophore consists of two planar aromatic or heterocyclic systems, connected through an aliphatic group containing a basic nitrogen [52, 53]. Although this is not identical to the complete 3D pharmacophore for 5-HT2A reported in reference [47], the method convincingly located two of the important phamacophore centers.

CDK2 inhibitor fragment scoring

The $n = 10$ CDK2/Cyclin-A model fragment scores and weighted spheres are displayed on compound **2** (compound **16** from reference [48]) in Figure 13. The fragment scoring clearly picked out the oxindole ring system central to the CDK2 activity of this chemotype. Three of the top five scoring points (scores of 930, 603 and 479) corresponded to atoms from x-ray structures [48] known to form hydrogen bonds with the backbone residues in the CDK2 receptor. The hydrogen bond locations are indicated schematically in Figure 13. In contrast with the spipernone fragment scores, all of which were below 100, the fragment scores in compound 2 were all quite high, with many points having scores >200. This may have been anticipated from the C_k/log10 T_k plots, which for 5-HT2A show many reference group bits with $T_k < 1$, while CDK2 C_k / log10 T_k plots show only few group bits with $T_k < 1$.

PDE4 inhibitor fragment scoring

The $n = 10$ PDE4 model fragment scores and weighted spheres are displayed on the PDE4 inhibitor rolipram in Figure 14. The catechol fragment, by far the most important fragment in many known PDE4 inhibitors [49], is isolated well by the scores. The catechol unit is known from x-ray crystal results to form hydrogen-bond interactions with the 'Q-switch' residue GLN 443 [54] in PDE4 – the location of which are given schematically in Figure 14.

FXa inhibitor fragment scoring

The $n = 10$ FXa model fragment scores and weighted spheres for the FXa inhibitor DX-9065A are shown in Figure 15. Important H-bond interactions as determined from the 1FAX

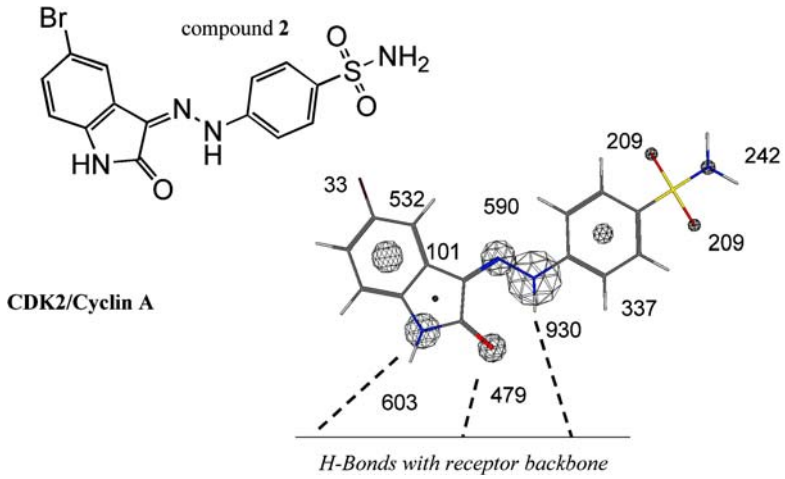


Figure 13. Fragment scores for CDK2/cyclin a active compound **2***:* The line drawing of **2** along with fragments scores and weighted spheres produced by the CDK2/Cyclin A model. H-bond interactions with the receptor known from x-ray structures are shown schematically.

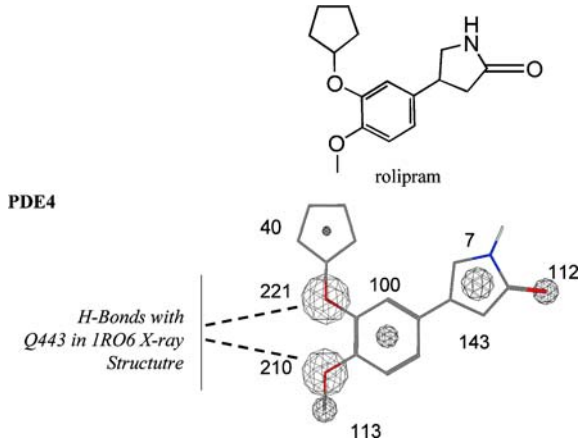


Figure 14. Fragment scores for PDE4 active compound rolipram: The line drawing of rolipram along with the fragments scores and weighted spheres produced by the PDE4 model. H-bond interactions with the receptor known from x-ray structures are shown schematically.

crystal structure [50] are also indicated. The scores highlighted points where the DX-9065A inhibitor forms H-bonds with the receptor. The highest scores were assigned to the cationic centers in the benzamidine ring, reflecting the importance of binding to the ASP residues in the *S* pocket of many proteases [55], including FXa. Surprisingly, the next highest scored point is the imine nitrogen, know from the crystal structure to form an H bond with residue E97 [50].

FXa inhibitor: Scoring fragments with other activity models

To test the effect of using inappropriate, or 'random', models on a query compound, the fingerprint models of 5-HT2A, PDE4 and PDE5 activity were used to score the FXa inhibitor DX-9065A. The results are given in Figure 16 along with the scores from the 'real' FXa fingerprint model for comparison.

The results of using incorrect activity models on DX-9065A suggest that a combination of the score magnitudes and the relative score spreads may help distinguish active from inactive compounds. The first striking difference between the scores produced by the real and random models is overall magnitude. None of the random models produced a fragment score >100, while the real FXa model produced scores >200. The 5-HT2A model produced low absolute scores which range from 2–9, much smaller than the 30+ scores produced for a true 5-HT2A active compound like spipernone.

The spread of point scores was often less pronounced with random models than with real models. With 5-HT2A and PDE5 'random' models, ratios of the largest and smallest point scores in DX-9065A were $\sim$2.5, and $\sim$3 respectively. For the real model, this ratio was >10. A smaller point score ratio is not always produced by random models on DX-9065A, as is shown by the PDE4 example. The PDE4 model produced low overall scores for DX-9064A, but the score ratio is comparable to that produced by the real FXa model. Interestingly, the portion of the DX-9065A molecule which was best scored by the PDE4 activity model is the portion that most looks like rolipram.

PDE5 inhibitor (sildenafil)- fragment scoring with PDE5, 5-HT2A and PDE4 models

As a final comparison of 'real' and 'random' models, three activity models – PDE5, 5HT2A and PDE4 – were used to score the fragments in sildenafil. The PDE5 model, or 'correct' activity model fragment scores are given for sildenafil in Figure 17. When the 'real' PDE5 activity model was applied to sildenafil, the resulting score magnitudes were relatively large ($>$ 100), with the largest scores assigned to the guanidine mimic central to the active core of the molecule. The carbonyl oxygen, which forms a hydrogen bond with GLN-817 in the 1TBF [56] x-ray structure, received the largest score.

Unlike the clear separation between the high and low scoring fragments in the FXa fragment scoring example, many

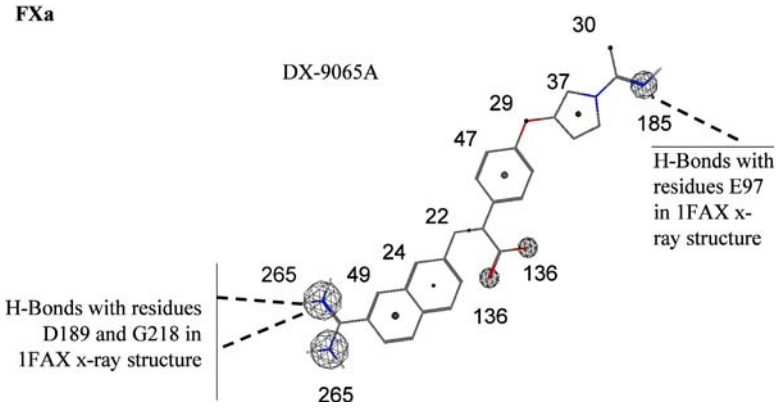


Figure 15. Fragment scores for FXa active compound DX-9065A: DX-9065A is shown along with the fragments scores and weighted spheres produced by the FXa model. H-bond interactions with the receptor known from x-ray structures are shown schematically.

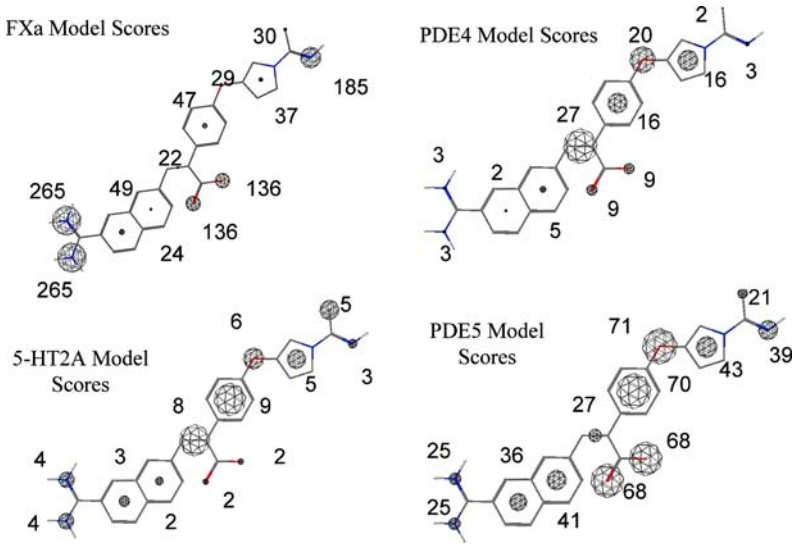


Figure 16. DX-9065A Fragment scoring using FXa, 5-HT2A, PDE4 and PDE5 Activity Models: Only the 'correct' FXa model produces both high scores (>100) and significant spread in point scores.

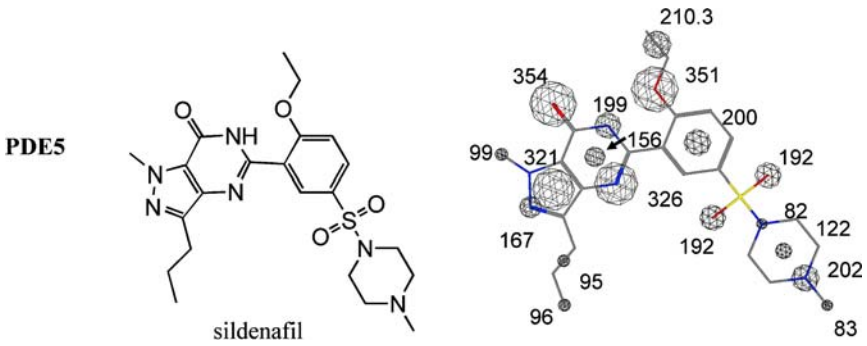


Figure 17. Fragment scores for PDE5 active compound sildenafil: The line drawing sildenafil along with fragment scores and weighted fragment spheres produced with the PDE5 activity model.

peripheral fragments in sildenafil not directly responsible for core activity also received high scores, notably the sulphone oxygen atoms (192) and one of the pipyridyl ring nitrogens. These high scores may reflect genuine contributions to activity or biases within the substituent sets used to functionalize the core scaffold. Despite the lack of unequivocal

fragment separation in the sildenafil scores, the high scores were not randomly distributed across the molecule, and for the most part centered around important regions of the molecule.

The 5-HT2A and PDE4 scores for sildenafil are drawn in Figure 18. The 5-HT2A model scores for sildenafil display a similar trend as was observed for 'incorrect' 5-HT2A, PDE4 and PDE5 model scores for DX-9065A – 'incorrect' activity models produced small absolute point scores and a small range of values within the scores. Sildenafil is not active against 5-HT2A, so the 5-HT2A model scores were expected to be small. The largest 5-HT2A model fragment score for sildenafil was 35, compared with a maximum score of 354 produced by the PDE5 model.

The PDE4 activity model is not exactly an incorrect activity model for sildenafil, which has moderate PDE4 activity (pIC50 = 4.6 vs pIC50 > 8 for PDE5). The moderate overlap in chemical activity space between sildenafil and the PDE4 active compounds used to train the model was reflected in the absolute PDE4 model fragment scores for sildenafil. The scores were moderate in magnitude - the top five PDE4 scores for sildenafil ranged from 104 to 148 and centered on one half of the recognition portion of the structure. The absolute PDE4 fragment scores for sildenafil were all generally lower than the PDE5 scores, but higher than the 5-HT2A scores. As was observed in the PDE4 model scores produced for DX-9065A, the sildenafil fragments assigned the highest scores by the PDE4 model (the ethoxy-phenyl ring) were those which resemble rolipram. This may be a reflection of the high bias towards rolipram-like molecules in the PDE4 activity training set.

Reference group similarity statistics

For all fingerprinting systems studied here, over all targets, there is little correlation between the ROC area and the S_{ab}, S_{max}, and S_{min} values of the fusion group. Sample plots for the PCH fingerprint system (Figure 19) show no correlation between the ROC area and any of the group similarity statistics, S_{ab}, S_{max}, and S_{min}. Similar lack of correlation was observed for all the other fingerprints as well (results not shown).

For each group count value n the S_{ab}, S_{min}, and S_{max} values were averaged over all models and the results plotted vs. n in Figure 20. The S_{ab}, S_{min}, and S_{max} plots reveal interesting trends as a function of group count n. For all fingerprinting systems the average S_{ab} quickly levels off to a constant value; the constant is dependent on the fingerprinting system and reflects the probability of coincident bits in random chemical space. Fingerprinting systems which define only a small number of possible bits will in many cases, by chance, have a number of bits in common between any two molecules. As a result, the similarity between any two random molecules will on average be high. This trend can be seen in the MACCS and TGT S_{ab} plots, which level off to similarity values higher than PCH of GpiDAPH. The MACCS keys are restricted 166 bits and level off to the highest S_{ab} value. The TGT fingerprints level off at S_{ab} values higher than PCH or GpiDAPH because the TGT system has fewer possible bits. The PCH and GpiDAPH fingerprints both produce relatively large numbers of bits for any given molecule, only a small percentage of which are turned on in any given molecule. This results in low similarity scores between random pairs of molecules. This is reflected further in the plot of S_{min} vs. n, which shows the MACCS key and TGT S_{min} values leveling off at values much higher than PCH or GpiDAPH.

The plots of S_{max} vs group count n show that by a count of $n = 40$, all systems have on average a group with at least one pair of molecules having similarity of $\sim$1. This is probably a result of having many large congeneric series in the dataset. It is interesting to note that S_{max} approaches 1 more rapidly with small bit systems such as MACCS and TGT, again reflecting the higher probability of having bits in common between any pair of molecules.

Conclusions

Group fusion: General behavior over multiple biological targets

The results of the group fusion experiments over multiple targets show that on average, using multiple reference structures with both the AVE and MAX fusion rules increases recall performance over single reference structure searches. Overall the MAX rule produces better results than the AVE rule, confirming results of others which suggest the MAX rule is superior [25]. Increases in recall rates level off at a group count of 10 with the AVE rule, while recall rate gains using the MAX rule continue to increase as n approaches 40, suggesting they could continue to increase beyond $n = 40$, the maximum group count considered in this study. Recall rates with groups of $n > 40$ will be the subject of future investigations.

Comparisons of the AVE and MAX rules and groups sizes of $n = 10$ and $n = 40$ highlight the variability in results between biological targets. Averaged over all targets the fusion rules increase recall rates, but each target is different; some targets show little improvement while others show dramatic improvement as the group count increases. Few targets show any change in ROC area with the AVE rule as n increases form 10 to 40, but there are many examples of recall gains using the MAX rule when n increases to 40.

The results over multiple targets also show that no one fingerprint of the four studied here is superior to all others in every case. The PCH and the MACCS fingerprints have comparable performance, and on average outperform the GpiDAPH fingerprints, with the TGT fingerprints producing the worst results. However, there is at least one biological target where one fingerprint outperforms all others. All fingerprints produce worse than random models (ROC area <50) for at least one target/fusion rule combination. The variability in recall performance between fingerprints

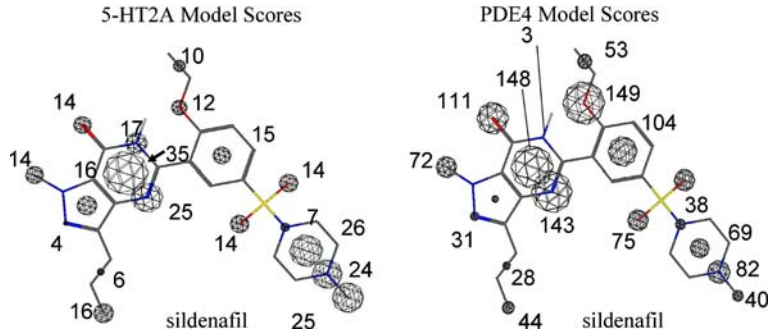


Figure 18. Sildenafil fragment scores using 5-HT2A and PDE4 activity models: The 5-HT2A and PDE4 model fragment scores for sildenafil. Sildenfil is moderately active against PDE4, with a pIC50 of 4.6 vs. a pIC50 of 8.6 against PDE5.

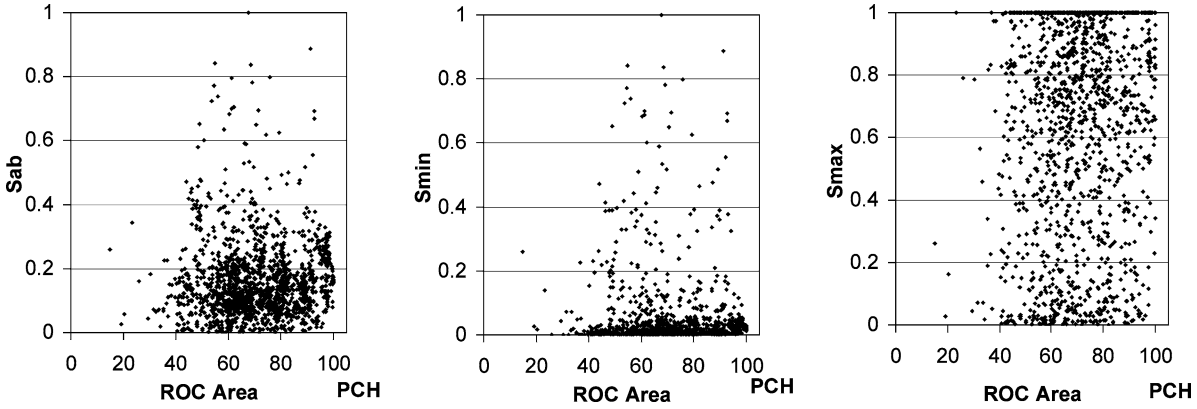


Figure 19. Plots of S_{ab}, S_{min} and S_{max} vs ROC Area: PCH fingerprint model S_{ab}, S_{max} and S_{min} values correlated with ROC area for all values of n.

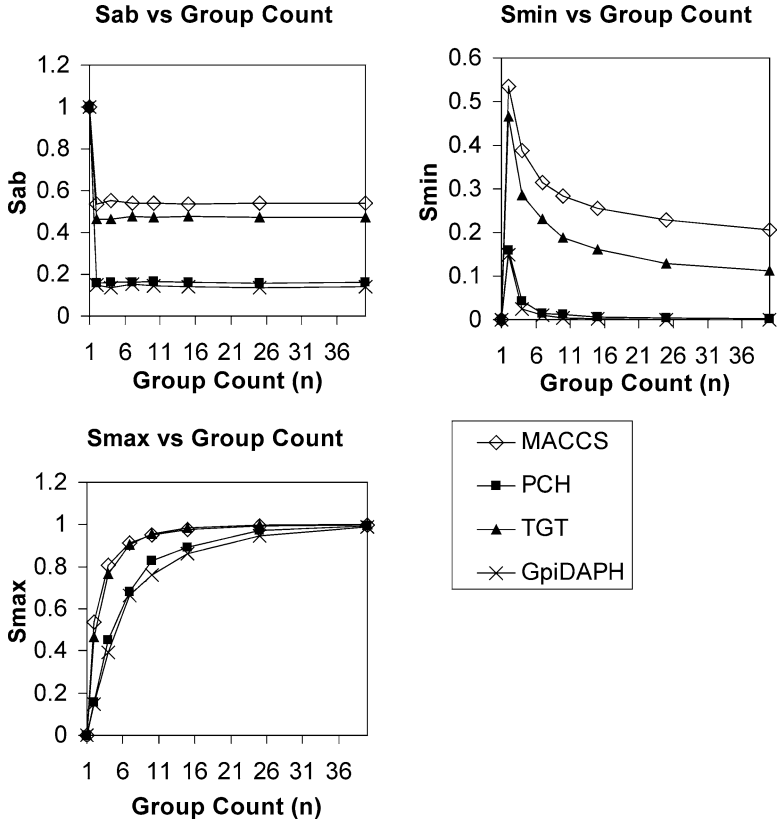


Figure 20. S_{ab}, S_{min} and S_{max} vs Group Count: The S_{ab}, S_{max}, and S_{min} values averaged over all models at each group count value n.

across biological targets seen here has been noted by others [36, 37], and suggests that no one fingerprint can be expected to outperform all others in all cases. Further studies into improving existing fingerprints and applying consensus scoring over multiple fingerprint schemes should be performed.

The training set groups were chosen randomly in the current study and we restricted ourselves to the Tanimoto coefficient, but the effects of using different training set selection techniques (such as diverse subset selection, clustering, etc) and different similarity coefficients should be investigated.

The study was also a first step in validating the newly implemented PCH fingerprint. The results showed the PCH fingerprint to be at least as good as the other three tested here, validating its use for the scaffold-hopping and fragment scoring portion of this study.

Group fusion: PDE5 scaffold hopping

The PDE5 inhibitor scaffold hopping experiment demonstrated, at least within the current data, scaffold hopping ability with the group fusion approach and the PCH fingerprints. The most reliable scaffold hops were shown to occur between chemotypes with minor scaffold modifications, as is to be expected with graph based similarity search. Major scaffold changes can only be achieved by hopping through intermediate modifications, a result congruent with the commonly held view of activity space that suggests islands of activity are separated by compounds with intermediate or no activity [57]. Hopping from one active chemotype to a significantly different chemotype requires traveling through many intermediate forms.

Reverse fingerprinting – fragment scoring and visualization

The reverse fingerprint fragment scores produced for the example compounds are encouraging. The scoring method for the most part isolated known important pharmacophore fragments in the query molecules. Fragment scores produced for DX-9065A and sildenafil using 'incorrect' activity models were for the most part small and spread more evenly over the molecule than scores produced by correct activity models. The range in sildenafil fragment scores produced by the three different target models – inactive (5-HT2A), poorly active (PDE4) and active (PDE5) – suggested the method may be sufficiently sensitive to distinguish between strongly and weakly active molecules, and possibly used in selectivity studies. Trends in the current results suggest that score magnitudes and relative score spreads may reflect the strength of the activity signal, and incorporating this information into scoring may improve fragment highlighting and lead to a method for activity scoring of the entire molecule. The current definitions of bit importance and fragment scoring can undoubtedly be improved upon for better results. These preliminary results suggest the fragment scoring approach has some merit, and should be investigated further.

Reference group similarity statistics

The lack of correlation between the S_{ab}, S_{max}, and S_{min} group statistics and the corresponding ROC areas was somewhat unfortunate, as it suggests these measures cannot be used *a priori* to predict the potential recall performance of a reference group. Although more work needs to be done on this, including assessing the performance of other similarity measures and other approaches to measuring reference group similarity statistics, it appears that at least in a general sense over all biological targets, the group similarity measures defined here are of little use in selecting reference group sets, or in predicting recall rates.

Acknowledgements

I would like to thank everyone with whom I have had discussions with on this subject. In particular I would like to thank Dr. Miklos Feher of Neurocrine, San Diego, Drs. Morten Langgård and Sune Askjaer Pedersen of Lundbeck, Copenhagen, and Drs. Paul Kowalczyk and Josh Du of Pfizer, Groton, for interesting discussions, helpful comments, literature references and general encouragement. I would also like to thank Dr. Suzanne Schreyer, Alain Deschenes and Paul Labute of Chemical Computing Group for useful suggestions, and Zovig Kevorkian for help with the manuscript.

References

1. Willett, P., *Chemical similarity searching*, J. Chem. Inf. Comput. Sci., 38 (1998) 983–996.
2. Sheridan, R.P. and Kearsley, S.K., *Why do we need so many chemical similarity search methods?*, Drug Discovery Today, 7 (2002) 903–911.
3. Miller, M.A., *Chemical database techniques in drug discovery*, Nat. Rev. Drug Discov., 1 (2002) 220–227.
4. Walters, P. et al., *Virtual screening – an overview*, Drug Discov. Today, 3 (1998) 160–178.
5. Johnson, M.A. and Maggiora, G.M., Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
6. Kubinyi, H., *similarity and dissimilarity – a medicinal chemists view*, Perspect. Drug Discovery Des., 11 (1998) 225–252.
7. Martin, Y.C., Kofron, J.L. and Traphagan, L.M., *Do structurally similar molecules have similar biological activity?*, J. Med. Chem., 45 (2002) 4350–4358.
8. Downs, G.M. and Willett, P., *Similarity searching in databases of chemical structures*, Rev. Comput. Chem., 7 (1995) 1–66.
9. Leach, A.R. and Gillet, V.J., An Introduction to Chemoinfomatics, Kluwer Academic, Boston, 2003.
10. Ginn, C.M.R., Willett, P. and Bradshaw, J., *Combination of Molecular similarity measures using data fusion*, Perspect Drug Discov Design, 20 (2000) 1–16.
11. Ginn, C.M.R., The Application of Data Fusion to Similarity Searching of Chemical Databases, Ph.D. thesis, University of Sheffield, 1998.
12. Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., *Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures to proteins*, J. Med. Chem., 42 (1999) 5100–5109.

13. Kontoyianni, M., McClellan, L. and Sokol, G.S., *Evaluation of docking performance: comparative data on docking algorithms*, J. Med. Chem., 47 (2004) 558–565.
14. Bissantz, C., Folkers, G. and Rognan, D., *Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations*, J. Med. Chem., 43 (2000) 4759–4767.
15. Stahl, M. and Rarey, M., *Detailed analysis of scoring functions for virtual screening*, J. Med. Chem., 44 (2001) 1035–1042.
16. Tong, W., Hong, H., Fang, H., Xie, Q. and Perkins, R., *Decision forest: combining the predictions of multiple independent decision tree models*, J, Chem. Inf. Comp. Sci., 43 (2003) 525–531.
17. Jurs, P.C., Kaufmann, G.W. and Mattioni, B.E., *Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble*, J. Chem. Inf. Comp. Sci., 43 (2003) 949–963.
18. Mozziconacci, J.C., Arnoult, E., Baurin, N., Chavatte, P., Marot, C. and Morin-Allory, L., 2-D *QSAR consensus prediction for high-throughput virtual screening; an application to cox-2 inhibition modeling and screening of the nci database*, J. Chem. Inf. Comp. Sci., 44 (2004) 276–285.
19. Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., Xie, Q. and Tong, W., *Three new consensus qsar models for the prediction of ames genotoxicity*, Mutagenesis, 19 (2004) 365–378.
20. Votano, J.R., Parham, M., Hall, L.H. and Kier, L.B., *New predictors for several adme/tox properties: aqueous solubility, human oral absorption, and ames genotoxicity using topological descriptors*, Mol. Divers., 8 (2004) 835–841.
21. Wang, R. and Wang, S., *How does consensus scoring work for virtual library screening? an idealized computer experiment*, J. Chem. Inf. Comp. Sci., 41 (2001) 1422–1426.
22. Feher, M., Baber, J.C., Shirley, W.A. and Gao, Y., *The use of consensus scoring in ligand-based virtual screening*, J. Chem. Inf. Comput. Sci., 46 (2006) 277–288.
23. Klon, A.E., Glick, M., Thoma, M., Acklin, P. and Davies, J.W., *Finding more needles in the haystack; a simple and efficient method for improving high-throughput docking results*, J. Med. Chem., 47 (2004) 2743–2749.
24. Hert, J., Willett, P., Wilton, D.J., Acklin, P.A., Azzaoui, K., Jacoby, E. and Schuffenhauer, A., *Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures*, Org. Biomol. Chem., 2 (2004) 3256–3266.
25. Willett, P., *Searching techniques for databases of two- and three-dimensional chemical structures*, J. Med. Chem., 48 (2005) 4183–4199.
26. Brown, R. and Martin, E., *Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection*, J. Chem. Inf. Comput. Sci., 36 (1996) 572–584.
27. Schuffenhauer, A., Floersheim, P., Acklin, P. and Jacoby, E., *Similarity metrics for ligands reflecting the similarity of the target proteins*, J. Chem. Inf. Comput. Sci., 43 (2003) 391–405.
28. Rarey, M. and Dixon, J.S., *Feature trees: a new molecular similarity measure based on tree matching*, J. Comput. Aided Mol. Des., 12 (1998) 471–490.
29. Xue, L., Godden, J.W., Stahura, F.L. and Bajorath, J., *Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a biniary classification scheme*, J. Chem. Inf. Comp. Sci. 43 (2003) 1151–1157.
30. James, C.A. and Weininger, D., *Daylight theory manual*, Daylight Chemical Information Systems, Inc., Irvine, CA, USA, www.daylight.com
31. Unity, Chemical Information Software, Tripos, Inc., St. Louis, MO, USA, www.tripos.com
32. Durant, J.L, Leland, B.A., Henry, D.R. and Nourse, J.G., *Reoptimization of mdl keys for use in drug discovery*, J. Chem. Inf. Comput. Sci., 42 (2002) 1273–1280.
33. ECFP*/FCFP*, Extended Connectivity Rings, Scitegic Inc., San Diego CA, USA 92123 www.scitegic.com
34. BCI – Barnard Chemical Information Ltd., Sheffield, UK, www.bci.gb.com
35. Xue, L., Godden, J.W. and Bajorath, J., *Database searching for compounds with similar biological activity using short binary bit string representations of molecules*, J. Chem. Inf. Comput. Sci., 39 (1999) 881–886.
36. Good, A.C.; Hermsmeier, M.A. and Hindle, S.A., *Measuring camd technique performance: a virtual screening case study in the design of validation experiments*, J. Comput.-Aided Mol. Des., 18 (2004) 529–536.
37. Good, A.C., Mason, J.S. and Cho, S.-J., *Descriptors you can count on? normalized and filtered descriptors for virtual screening*, J. Comput.-Aided Mol. Des., 18 (2004) 523–527.
38. MOE software (Version 2005.06) available from Chemical Computing Group Inc., 1010 Sherbrooke St. West, Montreal, Quebec, Canada www.chemcomp.com
39. Sheridan, R.P., Miller, M.D., Underwood, D.J. and Kearsley, S.K., *Chemical similarity using geometric atom pair descriptors*, J. Chem. Inf. Comput. Sci., 36 (1996) 128–135.
40. Clark, R.D., Fox, P.C. and Abrahamian, E.J., *Using pharmacophore multiplets fingerprint for virtual high throughput screening*. In: Alvarez, J., Shoichet, B. (Eds.), Virtual Screening in Drug Discovery, Taylor and Francis, New York, 2005, ISBN 0-8247-5479-4, pp. 207–224.
41. Schneider, G., Neidhart, W., Giller, T. and Schmid, G., "*Scaffold hopping*" *by topological pharmacophore search: a contribution to virtual screening*, Angew. Chem. Int. Ed., 38 (1999) 2894–2896.
42. Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., Bayesian Data Analysis, Chapman and Hall, New York, 1998.
43. Labute, P., *Binary-QSAR: a new method for quantitative structure-activity relationships*, in Biocomputing: Proccedings of the 1999 Pacific Symposium, pp. 444–455 . World Scientific Publishing, Singapore, 1999.
44. Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J. and Humblet, C., *Stigmata: an algorithm to determine structural commonalities in diverse datasets*, J. Chem. Inf. Comput. Sci., 36 (1996) 862–871.
45. MACCS keys: MDL Information Syetems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
46. Witten, I.H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, New York, 1999.
47. Holtje, H.-D., *Pharmacophore identification and receptor mapping*, In Wermuth, C.G. (Ed.), The Practice of Medicinal Chemistry, Academic Press, Boston, 2003, pp. 387–403.
48. Bramson, N.H. et al., *Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): design, synthesis, enzymatic activites and x-ray crystallographic analysis*, J. Med. Chem., 44 (2001) 4339–4358.
49. Norman, P., *PDE4 inhibitors: patent and literature activity 1999-mid 2000*, Exp. Opin. Ther. Patents, 10 (2000) 1417–1429.
50. Brandstetter, H., Kuhne, A., Bode, W., Huber, R., Von der Saal, W., Wirthensohn, K. and Engh, R.A., *X-ray structure of active site inhibited clotting factor Xa: implications for drug design and substrate recognition*, J. Biol. Chem., 271 (1996) 29988.
51. Rotella, D.P., *phosphodiestarase 5 inhibitors: current status and potential applications*, Nature Reviews: Drug Discovery, 1 (2002) 674–682.
52. Watanabe, Y., Usui, H., Shibano, T., Tanaka, T. and Kanoa, M., *Synthesis of monocyclic and bicyclic 2,4(1h,3h)-pyrimidinediones and their serotonin 2 antagonist activities*, Chem. Pharm. Bull., 38 (1990) 2726–2732.
53. Ketanserin patent, Janssen Pharmaceuticals N.V., European Patent Office. Kennis, L.E.J., Van der Aa, M.J.M., Van Heertum, A.M.A. and Jones, A.J. (1980) Nr. 001362, Appl. Nr. 803000-595.
54. Xu, R.X. et al., *Crystal structures of the catalytic domain of phosphodiesterase 4b complexed with amp, 8-br-amp and rolipram*, J. Mol. Biol., 337 (2004) 355–365.

55. Bode, W., Turk, D. and Karshikov, A., *The refined 1.9 Å X-ray crystal structure of D-Phe-Pro-Arg chloromethyl ketone-inhibited human alpha thrombin: Structural analysis, overall structure, detailed active site geometry and structure-function relationships*, Protein Sci., 1 (1992) 426–471.
56. Zhang, K.Y.J. et al., *A glutamine switch mechanism for nucleotide selectivity by phosphodiesterases*, Mol. Cell., 15 (2004) 279–286.
57. Schneider, G. and Fechner, U., *Computer-based de novo design of drug-like molecules*, Nature Reviews: Drug Discovery, 4 (2005) 649–663.