

ErG: 2D Pharmacophore Descriptions for Scaffold Hopping

Nikolaus Stiefl,^{*,†} Ian A. Watson,[§] Knut Baumann,[‡] and Andrea Zaliani[†]

Eli Lilly Research Laboratories, Essener Bogen 7, D-22419 Hamburg, Germany, Department of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Eli Lilly Corporate Center, 355 E Merrill Street, Indianapolis, Indiana 46225

Received October 18, 2005

An extended reduced graph approach (ErG) is presented that uses pharmacophore-type node descriptions to encode the relevant molecular properties. The basic idea of the method can be described as a hybrid approach of reduced graphs (Gillet et al. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 338–345) and binding property pairs (Kearsley et al. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 118–127). However, specific extension modifications to correctly describe the pharmacophoric properties, size, and shape of the molecules under study result in a very stable and good performance as compared to DAYLIGHT fingerprints (DFP). This is exemplified for 11 activity classes of the MDL Drug Data Report database, for which ErG performs as well or better than DFP in 10 cases. On the basis of the example data sets, the ability of ErG to switch from one chemotype to another (often referred to as “scaffold hopping”) is highlighted. Additionally, possible pitfalls of reduced graph approaches as well as suitable solutions are discussed with the help of example structures. Overall, it is shown that ErG is a widely applicable method capable of identifying structurally diverse actives for a given active search query. This diversity is achieved by a high degree of molecular abstraction, which in turn results in a low dimensional descriptor vector that allows very low computation times for similarity searches.

INTRODUCTION

In silico screening of compound databases has become a powerful tool for the identification of small molecule ligands. Especially, similarity searching based on one or multiple query structures to find other compounds with similar biological activity is a routinely applied method to search through corporate and third-party databases.^{1–4} Ligand-based approaches can be split into two major parts: the molecular description and the similarity metric, where the former is often represented in a substructural or atom-based way (e.g., MACCS keys⁵ and DAYLIGHT fingerprints⁶). One of the major problems of such a molecular description is the fact that, if the query structure has certain unwanted features (e.g., novelty or known metabolic or toxicity issues), the compounds retrieved from the databases are likely to exhibit similar deficiencies.

Hence, in recent years, for similarity searching, a shift toward more-general descriptions of the structures under study could be observed, many of which encode chemical features by more generic pharmacophoric properties^{4,7,8} as well as the molecular shape.^{9–11} With these novel techniques, the goal is to enable easy switching from one chemotype to another, which is often referred to as “scaffold” or “lead hopping”.^{12,13}

One such approach, a specific type of 2D representation, is that of reduced graphs.^{14–19} Reduced graphs are best

described as “...summary descriptions of the gross structural features of...substances...”. The most striking difference to published methods such as CATS¹² is that, in reduced graphs, not every single atom is used for the descriptor generation, but a more abstract description of the molecular features is employed. Recently, Gillet et al.¹⁷ and Barker et al.¹⁸ showed that their approach to reduced graphs is able to perform similarly well as standard techniques such as DAYLIGHT fingerprints (DFP). In this contribution, an extension and modification of the concept introduced in refs 17 and 18 is presented and its application to published data sets is performed.

Even though, in theory, the idea of omitting the 3D geometrical information of the molecules in a similarity search may seem counterproductive, so far, no general trend was observed which highlighted an advantage of 3D over 2D methods for similarity searching.^{20,21} As pointed out by Oprea, 2D and 3D methods rather complement each other.²² Hence, 2D structural descriptions still represent an interesting research target since they exhibit some advantages over 3D methods, such as the low computational cost.

This paper is organized as follows. First, the workflow and the theoretical background of the extended reduced graph (ErG) approach are described and differences from the work of Gillet et al.¹⁷ and Barker et al.¹⁸ are highlighted and discussed. Next, numerical and structural results obtained with ErG on standard data sets are given and compared to results obtained with DFP. On the basis of these results, advantages as well as disadvantages of the reduced graph approach are investigated and potential solutions are presented.

* Corresponding author. Current address: Novartis Pharma AG, Werk Klybeck, Postfach, CH-4002 Basel, Switzerland. Phone: +41 61 69 63068. Fax: +41 61 69 64870. E-mail: nikolaus.stiefl@novartis.com.

[†] Eli Lilly Research Laboratories.

[‡] University of Wuerzburg.

[§] Eli Lilly Corporate Center.

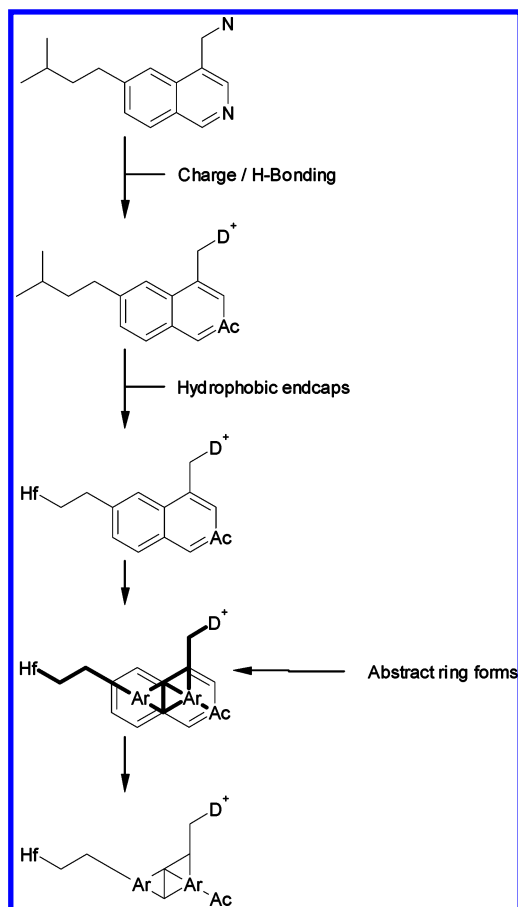


Figure 1. Conversion of the chemical graph of a sample structure into a reduced graph. D: H-bond donor; Ac: H-bond acceptor; Hf: hydrophobic group; Ar: aromatic ring system; +/-: positive/negative charge.

METHODS

ErG is best described as a hybrid approach of reduced graphs and binding property pairs.²³ It allows a chemically intuitive description and visualization of the chemical graph. As already mentioned, the reduced graphs in refs 17 and 18 are similar in spirit to ErG; however, important differences can be identified and will be highlighted in this section of the paper.

The most similar reduced graph approach to ErG is the R/F(4)/T method of ref 18. Briefly, in R/F(4)/T, aliphatic and aromatic ring systems (R) as well as H-bonding properties (F) are differentiated and terminal atoms (T) are explicitly encoded.

Reduced Graph Generation. The steps for the conversion of the chemical into the reduced graph are depicted in Figure 1.

First, atoms are formally charged to represent the molecule under physiological conditions. Next, H-bond donor and acceptor properties are assigned. These two steps are based on an in-house protocol; however, other simple assignment procedures could be applied as well (e.g., refs 17, 24, 25). Atoms that comprise donor and acceptor features are assigned a flip-flop flag, which is handled explicitly prior to descriptor generation (see below).

Second, so-called “endcap” groups are identified. These groups comprise lateral hydrophobic features with three atoms (e.g., isopropyl), which are often important for the shape and size of the molecules. Owing to the large size of

sulfur, thioethers are also encoded as endcap groups even though they are composed of only two atoms. It is important to note that, in contrast to ref 18 (R/F/T), not all terminal non-hydrogen atoms are retained. That way, a more general description is achieved and less weight is given to these atoms in the reduced graph.

In a third step, the ring systems of the structure are converted into their abstract forms. The overall abstraction procedure of ring systems can be briefly described as follows:

1. Add a centroid atom for each ring, and assign a corresponding feature (Ar/Hf).
2. Retain all ring atoms that are substituted, and create bonds from each of these atoms to the centroid.
3. Retain all bridgehead atoms, and create bonds from each of these atoms to the centroids of both rings.
4. Remove all nonsubstituted ring atoms, and retain all bonds between the atoms that are retained in steps 2 and 3.

In more detail, a centroid is generated for every ring system with a ring size of less than eight atoms; that is, macrocyclic structures are handled as found in the original structure. These centroids are then connected to all atoms of the ring that either are substituted or are assigned a charge, H-bond donor, or H-bond acceptor flag. All other ring atoms are deleted. The substituted atoms also include atoms that are members of more than one ring, for example, the two bridge atoms in naphthalene. As a result, it is possible to keep distances between features connected to ring systems as well as between ring centroids comparable to those of the original graph. Otherwise, the interfeature distances of compounds with and without ring systems may differ for para-substitution. This is especially important for potential “scaffold-hopping” applications and is exemplified for 6-amino-β-naphthol in Figure 2.

A special case of this molecular abstraction is that of highly fused ring systems. Here, the centroids of the fused system are collapsed into one point, unless the number of atoms of the larger fused ring is more than six and the number of atoms of the smaller ring is more than four. This is done to ensure that the number of reduced graph points is kept at a comparable level (see Figure 3).

Finally, the centroids are assigned either an aromatic (aromatic ring systems) or a hydrophobic (aliphatic ring systems) flag. Here, rings with more than 50% sp²-hybridized atoms, as well as aliphatic rings that are directly attached to aromatic rings, are also assigned the aromatic flag. That way, the reduced flexibility as well as the enhanced hydrophobicity of these ring systems is accounted for.

In summary, the result of the graph reduction procedure is a description in which H-bonding features, hydrophobic endcaps of a specific size, and aromatic as well as aliphatic ring systems are flagged as interesting. So-called “linker atoms”¹⁸ (sometimes referred to as “non-interacting” atoms¹⁹) are not taken into account during property assignment, since they do not contain any pharmacophoric information¹⁶ apart from the encoding of interfeature distances. For topological interfeature distances, all nonring atoms as well as ring atoms that are either flagged as a feature or substituted are kept in the graph. All other atoms are deleted.

The similarities of ErG to the reduced graph approaches published in refs 17 and 18 are as follows: in both approaches, only features thought to be relevant for ligand–receptor interactions are captured by the reduced graph (i.e.,

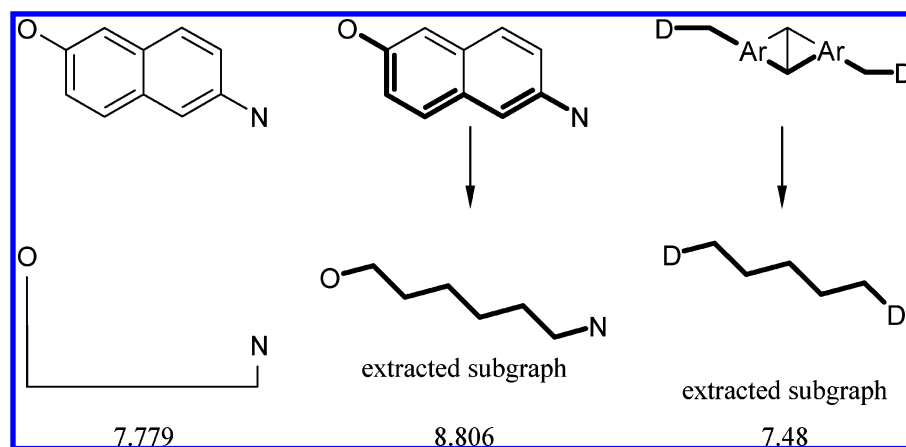


Figure 2. Geometrical interfeature distances in Å of the H-bond donor features of 6-amino- β -naphthol and the corresponding extracted subgraphs after minimization (bold). If the geometrical distance between these two features is important for biological activity, the distance corresponding to the molecule that represents the shortest path over the original graph (center) is too long. Put differently, compounds that are retrieved on the basis of this distance measure will be too large. However, the distance calculated for the molecule that represents the shortest path over the reduced graph (right) represents the geometrical distance much better.

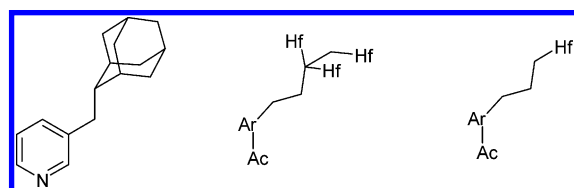


Figure 3. Molecular abstraction of highly fused ring systems. If all rings of the smallest set of smallest rings of the adamantane substructure on the left would be converted into pharmacophoric points (Hf: hydrophobic, middle), this feature would be over-represented with respect to the features of the pyridine substructure. To avoid this, highly fused rings can be collapsed into one hydrophobic feature (see text).

topological pharmacophores are generated). However, ErG attempts to produce a more general description of the original graph's features, and a heavy focus is put on the correct implicit handling of size and shape (different handling of encoding rings, substitution patterns, fused rings, and endcap groups).

Descriptor Calculation. For the conversion of the reduced graph into the descriptor, only so-called property points (PPs) with assigned features such as charge, H bonding, endcap, or abstract ring are used. These points are converted into a radial distribution function³⁴ of the form

Property Point 1 – topological distance – Property Point 2

where the topological interfeature distances are calculated over the reduced graph.

Hence, the descriptor vector \mathbf{v} is of size

$$\text{size}(\mathbf{v}) = \frac{n(n+1)}{2}(\text{maxDist} - \text{minDist} + 1)$$

where n is the number of properties used (default $n = 6$) and maxDist and minDist are the maximum (default = 15) and minimum (default = 1) distances between two PPs taken into account, respectively. With default settings, the size of the descriptor vector \mathbf{v} is 315, which is comparatively small as compared to the standard DFP, consisting of 2048 entries, or other fingerprint methods.²⁶

Fuzzy Incrementation. Each field of vector \mathbf{v} is then encoding a specific property–property-distance triplet (PPDT)

that is incremented if a corresponding triplet is found in the structure under study. Since it was shown before^{24,27} that a fuzzy incrementation may reduce the impact of the categorization error of the distance bins, this methodology was applied here as well. In this procedure, if a PPDT corresponding to a field of vector \mathbf{v} is identified, the respective field is incremented. Additionally, the first neighboring fields—in terms of distance—are incremented by *incr*. The amount of the fuzzy increment is a user-definable variable. If the user wants to see more closely related compounds, a smaller value would be chosen (usually 0.1). However, if the certainty about the interproperty distances is not too high, or if the user wants to retrieve compounds which are less similar to the query structure (referred throughout the manuscript as “needle”), the fuzzy increment would be set to a higher value (default = 0.3). The impact of different values of *incr* is given in the Results section.

With ErG, emphasis is placed on the concept that ring systems are handled separately from the H-bonding and charges features. This is a crucial step in the graph reduction because, that way, similarity can more easily be identified between compounds with different skeletons. If the H-bonding feature is part of the ring node, as in refs 17 and 18, a direct relation between H-bonding features within and outside a ring system or the similarity of the rings themselves may be missed (see Figure 4).

A direct consequence of this feature separation is that the overall number of encoded features is kept at a very low level, which in turn results in a rather small descriptor space.

Harper et al.¹⁶ also included the concepts of fuzzy incrementation and the description of the pharmacophoric similarity of similar groups (e.g., aromatic rings with and without acceptor properties) into their reduced graph approach. In terms of fuzzy incrementation, Harper et al. increment two bins: the bin corresponding to the actual distance and the bin that corresponds to the next shorter distance than the actual one. By doing so, the same weight is given to both distances, which is in contrast to the method applied here. With respect to feature separation, in ref 16, the similarity between properties is given by a mutation cost which is relatively low for the mutation of two similar properties (e.g., for an aromatic ring into an aromatic ring

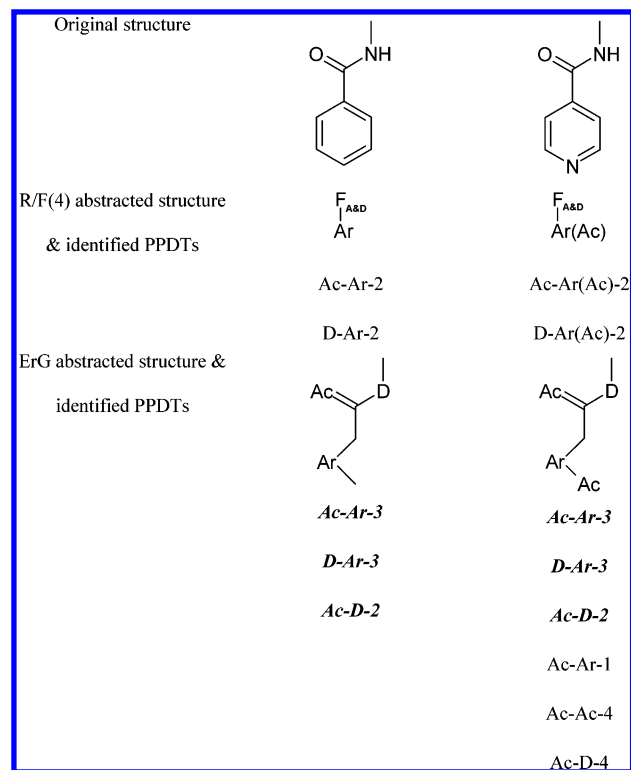


Figure 4. Example of feature separation. For the two structures given, the corresponding bits identified with (ErG) and without (R/F(4))¹⁸ feature separation are given. PPDs identified for both abstract structures are highlighted in bold/italics. It can be nicely seen that joint features lead to the encoding of similar pharmacophoric fragments into different bits of the descriptor vector, i.e., a different dimension. In contrast, with separated features, more common PPDs are identified for both structures.

with an acceptor). However, this concept requires that mutation costs are assigned to all possible feature combinations. Given the difficulty of estimating these costs, it seems advantageous to separate the features as in ErG. Then, estimating mutation costs is no longer necessary. In summary, given the differences between the two approaches, it is difficult to predict advantages and disadvantages theoretically.

Another important aspect of ErG is the way flip-flop features are handled. On the basis of the visual inspection of numerous protein side-chain interactions²⁸ as well as randomly chosen crystallized ligand–receptor complexes in the Brookhaven Protein data bank,²⁹ it was found that flip-flop atoms usually interact with the target as either H-bond donors *or* H-bond acceptors (as opposed to functional groups such as amides, which can act as donors *and* acceptors). To account for this behavior, the flip-flop feature is fully enumerated; that is, two vectors are generated for a structure that comprises a flip-flop atom, one vector encoding the flip-flop atom as a H-bond donor and the other one encoding it as a H-bond acceptor. This implementation of flip-flop-atom conversion differs from that in ref 18, where flip-flop atoms are encoded by incrementing both the H-bond donor and the H-bond acceptor at the same time. The advantage of the latter method is that it results in a compact descriptor with fewer entries than using a specific flip-flop property.¹⁷ However, it has deficiencies with respect to the search procedure. This is best explained with the help of an example. Assume we are given two structures that are identical except for one feature. The first compound carries a flip-flop feature (cpd1).

Table 1. Size of Activity Classes Used in This Study

activity class	abbreviation ^a	number of compounds
5-HT _{1a} agonists	5HT _{1a}	800
5-HT ₃ antagonists	5HT ₃	738
5-HT reuptake inhibitors	5HT _r	341
AT ₁ antagonists	AT ₁	932
cyclooxygenase inhibitors	COX	627
D ₂ antagonists	D ₂	389
HIV protease inhibitors	HIV	699
protein kinase C inhibitors	PKC	437
renin inhibitors	Ren	901
Substance P antagonists	SP	1203
thrombin inhibitors	Thr	760

^a Abbreviation used throughout the manuscript.

In the second compound, the flip-flop feature is exchanged for a H-bond donor feature (cpd2). When incrementing both fields H-bond donor and H-bond acceptor at the same time, the overall similarity measure of cpd1 and cpd2 cannot be 1 since cpd1 has an additional bits set which takes account of the H-bond acceptor feature of the flip-flop atom. However, if the flip-flop feature is fully enumerated, that is, two vectors are generated for cpd1, one of the generated vectors will have a similarity of 1 to cpd2.

Chemically speaking, in terms of ligand–receptor interactions, the method in ref 18 is handling flip-flop compounds as a moiety, which is simultaneously interacting as both a H-bond donor and acceptor; that is, a flip-flop is a superset of a H-bond donor and a H-bond acceptor feature, respectively. In contrast to that, the enumeration method implemented in ErG is describing this moiety as a H-bond donor *or* H-bond acceptor, which agrees with many crystal structures.²⁸

A problem of the explicit handling of flip-flop atoms is that of a possible numerical explosion of the database (for a compound with *n* flip-flop atoms, 2^{*n*} vectors are generated). However, usually, compounds with more than five to seven flip-flop atoms are not interesting, owing to their low drugability. Hence, for a “real-world” application, ErG can be applied with a default cutoff of five flip-flop atoms. For the database investigated here, approximately 0.38% of the structures would be rejected (~0.1% for a cutoff of 7), nearly all of which do not have a “common” druglike appearance.

For charged atoms, this enumeration scheme is not applied, since the true nature of interaction, charge or H-bonding, is often hard to identify and sometimes overlapping. Hence, for charged atoms, both properties are encoded simultaneously in the descriptor.

Data Sets. To evaluate the ability of ErG to identify biologically similar active compounds, a simulated virtual screening was performed on the MDL Drug Data Report database (MDDR).³⁰ Prior to conversion, salts were stripped from all compounds, and duplicates as well as compounds containing atoms typically not appearing in organic compounds were deleted from the database. Molecules with a molecular weight of more than 800 were not included in the analysis. This resulted in a set of 133 809 unique molecules. Next, the EXTREG numbers of the 11 activity classes used by Hert et al.³¹ were downloaded from the Internet.³² The number of active molecules in the respective activity class within the reduced MDDR is given in Table 1.

Of course, only those compounds in the respective activity classes were handled as active. Therefore, the results given

below may contain false negatives. However, this is the case for both competing descriptors (DFP and ErG). Hence, the error made is not overly important for this study.

Similarity Searching. Since the retrieval rates will vary for each query structure (“needle”), an average retrieval rate was calculated for each activity class. To achieve this, the similarity for each needle against the whole database (“haystack”) was calculated and the number of known active compounds q_{Act} in the first $x\%$ of the ranked database was identified. Throughout the manuscript, these compounds will be referred to as hits and the corresponding list of the $x\%$ will be referred to as the hit list.

This was done, in turn, for each needle. The average retrieval rate ($\text{RT}_{x\%}$) in the first $x\%$ of the database was then calculated according to

$$\overline{\text{RT}}_{x\%} = \frac{1}{p} \sum_{i=1}^p \frac{q_{i,\text{Act},x\%}}{p-1}$$

where p is the number of compounds in the respective activity class and $q_{i,\text{Act},x\%}$ is the number of actives found for needle i in the first $x\%$ of the database. Owing to the way H-bonding atoms with both, acceptor and donor properties, are handled within ErG (full enumeration), duplicate structures may appear within the ordered hit list. To take care of this, duplicates were removed, where the first occurrence of each compound (most similar structure) was kept in the list. If the needle structure comprised a “flip-flop” atom, for each enumeration, a search was performed, duplicates from these first hit lists were removed, and then the final lists were merged and the duplicates deleted and resorted.

Depending on the descriptor (DFP/ErG), different Tanimoto similarity coefficients were applied, where the binary form was used for DFP and the algebraic form was employed for ErG.^{2,33}

$$S_{A,B} = \frac{c}{a+b-c} \quad (\text{binary})$$

$$S_{A,B} = \frac{\sum_{i=1}^m n_{A,i} n_{B,i}}{\sum_{i=1}^m (n_{A,i})^2 + \sum_{i=1}^m (n_{B,i})^2 - \sum_{i=1}^m n_{A,i} n_{B,i}} \quad (\text{algebraic})$$

Here, a is the number of unique fragments in compound A, b is the number of unique fragments in compound B, c is the number of unique fragments shared by compounds A and B, m is the size of the ErG vector \mathbf{v} , $n_{A,i}$ is the i th entry of vector \mathbf{v} in compound A, and $n_{B,i}$ is the entry in vector field i in compound B.

Even though it is conceivable to use a binary or set-theoretic form² of ErG, this was not followed here because of the sparsity of the ErG descriptor matrix. Additionally, the quantitative (hologram) version encodes the size and shape of the molecules in an implicit way.³⁴

Scaffold-Hopping. Since one of the main aspects of this study was to identify to what extent ErG is able to switch among different chemotypes (i.e., “scaffold-hopping”), a scale for this capability was introduced. The basic concept

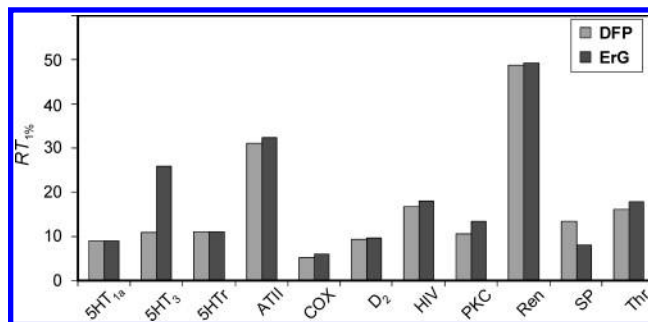


Figure 5. Average retrieval rates in the first 1% of the database. Over all data sets, the average absolute similarity value at 1% is 0.62 (standard deviation = 0.042) for ErG and 0.38 (standard deviation = 0.033) for DFP.

of this measure can be described as the identification of the number of structural families identified by ErG as compared to DFP.

Therefore, a k -means clustering was performed on the DFP/Tanimoto distances (distance calculated as $1 - \text{similarity}$) of all actives of one activity class. That way, each active was assigned to a structural “family” in DFP space. Next, for each active in turn, all hits from the DFP as well as the ErG similarity search and their respective families were identified. For both methods, it is then possible to calculate the number of families selected for each needle. These results were averaged over all needles. That way, a fast and easy way to compute the measure of the additional families hit by ErG as compared to DFP can be given. To gain more information about the similarities of ErG and DFP, an additional figure was computed that describes the number of families hit by both methods. This was repeated for each class. It is important to note that this clustering does not compare the two methods but only identifies whether ErG is actually able to identify structural classes different than those identified by DFP. To ensure that the results are not significantly dependent on k , the whole procedure was repeated for different values of k . It was found that, if k was chosen within a reasonable setting, the results did not change. Hence, here, only the results identified for $k = 15$ are presented.

In addition to these simple figures of merit, a visual inspection of identified hits was performed (see Results and Discussion). Here, a specific interest was given to those cases where ErG performed similarly to or was outperformed by DFP for specific needles. That way, it was attempted to better understand the differences between ErG and DFP and, if possible, to identify and reduce problems of the reduced graph approach.

RESULTS AND DISCUSSION

Retrieval of Actives. For the 11 activity classes, the average retrieval rates for the first 1% of the investigated database are given in Figure 5. Here, the cutoff value of 1% was chosen since it allows the retrieval of a number of compounds, which is quite similar to real-life applications. However, in addition, continuous retrieval rates over the first 10% of the database were calculated and are discussed below.

It can be seen that ErG is outperformed by DFP only in a single case, the Substance P antagonists (SP). For the other data sets, ErG is either similar in retrieval rates (5HT_T) or outperforms DFP by 0.7% (5HT_{1a}) to 140.7% (5HT₃). To

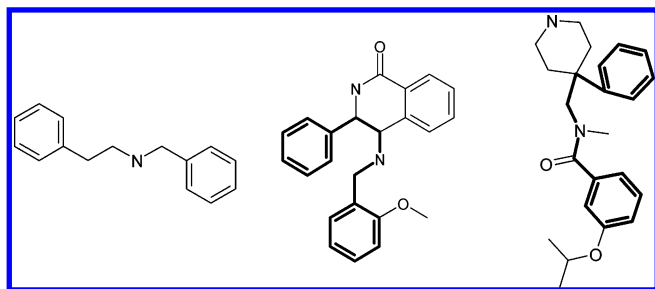


Figure 6. Common substructure of the Substance P antagonists (left), which is found in different actives (middle and right). However, the properties of the pharmacophoric features vary drastically.

gain a better understanding of the reasons for the rather different behavior of ErG and DFP on the SP as well as 5HT₃ data sets, both sets of actives were investigated on a structural level.

The 5HT₃ antagonist data set exhibits a rather uniform pharmacophore pattern in which an aromatic feature is found at a distance of approximately three to five bond lengths from a H-bond donor feature.^{35,36} For some of the structures, this arrangement is extended by an intermittent H-bond acceptor. The H-bond donor feature is found either within a highly fused ring system or an aromatic ring or without any direct ring connections. The aromatic feature is either a single ring of a different ring size or embedded within a larger ring system. Additionally, it can be noticed that the aromatic feature is fulfilled by a varying set of bioisosteres. On the basis of these findings, different explanations can be given as to why ErG performs so much better in this case. First, for this data set, a particular pharmacophoric pattern can be easily identified for a large portion of the data set. Since this pattern is fulfilled by a rather large number of compounds with different scaffolds, ErG's more generic approach makes it less sensitive to the scaffold changes, whereas DFP is much more affected. Second, the ratio between features fulfilling the pharmacophoric pattern and the overall number of features is comparatively well-balanced. Put differently, the pharmacophoric pattern is not hidden in a large number of features that are more or less varying. This might be an indication that the compounds within this set of active molecules are binding in a similar mode to that of the receptor.

For the Substance P antagonists, the opposite of the above can be observed. Here, in a large number of active compounds, the substructure given in Figure 6 is found.

However, this substructure is only found in structural terms, since the substitution patterns at the nitrogen atom vary in such a way that no common pharmacophoric pattern emerges apart from the two aromatic moieties.³⁷ In Figure 6, for example, the nitrogen is either a H-bond donor/positive charge feature (middle) or has no pharmacophoric property assigned (right). Accordingly, the more substructural-oriented DFP approach retrieves this pattern much better since both compounds are structurally more similar.

Additionally, the substructure in Figure 6 is a very general pattern that can be found in a large number of molecules. However, since this is the main part of the pharmacophoric properties found in most SP compounds, no additional information helps in differentiating SP from the rest of the database.

Table 2. Ratio of Number of Assigned Properties Per Compound against Its Number of Heavy Atoms

activity class	(# prop)/(# heavy atoms)	RT _{1%} (DFP)	RT _{1%} (ErG)
5HT _{1a}	0.323	8.82	8.88
5HT ₃	0.350	10.69	25.73
5HT _r	0.315	10.97	10.97
AT ₁	0.332	30.96	32.19
COX	0.306	5.07	5.85
D ₂	0.311	9.10	9.48
HIV	0.286	16.64	17.97
PKC	0.335	10.43	13.29
Ren	0.319	48.58	49.15
SP	0.278	13.32	7.84
Thr	0.353	15.92	17.61

Further inspection of the two data sets' properties allows the identification of two possible pitfalls of all reduced graph approaches, which can be avoided with descriptors that incorporate all atoms explicitly in their descriptions (e.g., DFP). In particular, for molecules with a little number of pharmacophoric features, minor changes in the pharmacophoric assignment owing to structural variations may result in drastic changes of the similarity measure between needle and database molecules. Hence, the resulting ranking may also change markedly. The impact of such changes is usually not as dramatic for compounds with higher numbers of pharmacophoric features. This behavior is similar in nature to the elsewhere-described behavior of sparsely populated fingerprints where the presence or absence of individual bits has a larger influence on the similarity coefficient than for more densely populated ones.^{38,39}

Additionally, since the pharmacophoric features can be fulfilled by a set of structural fragments (many-to-one correspondence), for molecules with very few pharmacophoric features, it happens more frequently that a high number of other molecules become comparatively similar. Put differently, if a molecule exhibits a small number of pharmacophoric features, the similarity distribution against the database is usually shifted to higher values and becomes narrower. As a consequence, actives that exhibit a high similarity value to the needle may be discarded more easily. When using DFP, there is no multiple-to-one correspondence of the structural features. Hence, such narrow distributions of the similarity values are usually not observed.

To describe the influence of the number of assigned properties on a numerical scale, the ratio of assigned properties was divided by the number of heavy atoms of the molecule. As can be seen in Table 2, a moderate correspondence of retrieval rates and this ratio with respect to DFP performance can be found. For all data sets where this ratio is higher than 0.33, ErG performs well as compared to DFP. In contrast to this, for the SP data set, this value drops to a comparatively low level, which is in line with the findings that especially this data set is problematic for ErG (even though it is, of course, just one possible influence factor for this behavior). However, the HIV data set also showed a comparatively low value for this ratio even though no obvious poor performance over the first 1% of the database was identified. A possible reason for this behavior is given below.

Interpretation of this ratio is rather difficult, since its magnitude can be influenced by different parameters (e.g., the ratio of ring systems to overall size, ring substitution

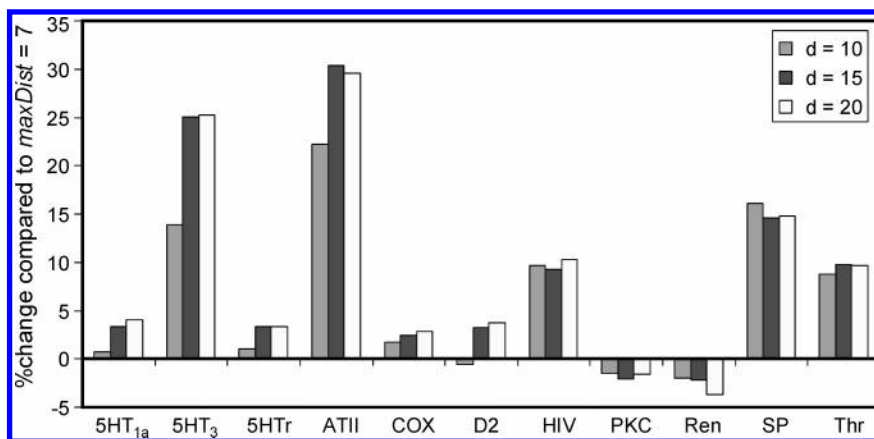


Figure 7. Percent change in average retrieval rates for ErG compared to $\text{maxDist} = 7$. Here, $\text{RT}_{1\%}$ for $\text{maxDist} = 7$ was set to 0, and the percentage increase/decrease for higher distances was recorded.

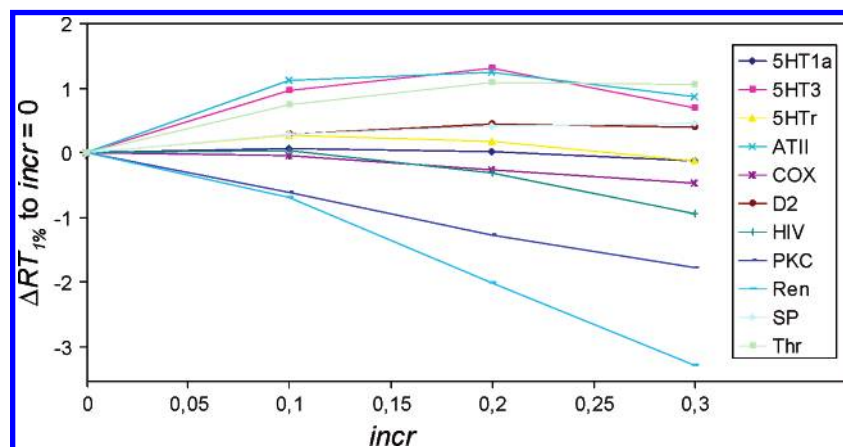


Figure 8. $\Delta\text{RT}_{1\%}$ of varying settings for incr compared to $\text{incr} = 0$. Beginning with $\text{incr} = 0.3$, retrieval rates for all data sets start to drop. The latter is probably a result of a too-unspecific description of the compounds under study.

patterns, etc.). Anyway, if for a given molecule of reasonable size comparatively many properties are assigned, it is described in more detail. This, in turn, will usually result in a higher absolute retrieval rate, either by the identification of more close analogues (if the data set is homogeneous) or by the reduction of nonmeaningful compounds further up in the hit list. On the other hand, if very few properties are assigned, minor changes to the molecules will result more easily in a higher fluctuation of similarity values. Nevertheless, this ratio can be used as a first indicator whether special caution should be given to the results of a reduced graph approach like ErG.

Investigation of Parameters. The two main parameters that can be changed by the user are the maximum interfeature distance maxDist and the amount of “fuzziness” described by incr . Hence, to understand the influence of these two parameters, retrieval rates for their variation were calculated.

First, the interfeature distance was investigated. Given the low default interatom distance used by other fingerprint techniques (e.g., DFP: seven bond lengths), it was most interesting to see whether a reduction in distance would impact retrieval rates for any of the activity classes. Therefore, retrieval rates for two smaller and one higher distance were calculated (see Figure 7).

For some activity classes, shorter distances than the default distance do not produce significant changes in retrieval rates. However, a general trend that the higher default distance results in better solutions as compared to $\text{maxDist} = 7$ can

be observed. For $\text{maxDist} = 20$, this behavior levels off. Given the descriptor vector sparsity of the reduced graph approach, this behavior is understandable, since the additivity of vector entries found in atom-based approaches (e.g., DFP) is not given. Hence, since a lower value of maxDist will not significantly increase the computational cost of the search procedure [the size of the descriptor varies between 315 ($\text{maxDist} = 15$) and 147 ($\text{maxDist} = 7$), respectively], the default maxDist of 15 is recommended.

Second, the “fuzziness” of the descriptor was varied. Theoretically, for a higher value of incr , the similarity values for compounds that show small distance differences between their pharmacophoric patterns should increase. However, from a certain value of incr on, the “real” information will be smeared out too much by the fuzzy incrementation. Usually, the aim of ErG is scaffold-hopping. Hence, it was interesting to identify to what level incr could be increased before a significant loss in performance was identified. Since a fuzzy increment incr of 0.3 is already very high, special interest was given to more strict descriptions. Accordingly, incr was set to values of 0 (crisp counts), 0.1, and 0.2 (see Figure 8).

In terms of average retrieval rates, no obvious trend can be identified for the different settings of incr . Overall, it seems as if, for these activity classes, the maximum setting of incr lies somewhere between 0.2 and 0.3. As stated before, an overall higher value of incr should, at least in theory, lead to a more fuzzy description of the interfeature distances.

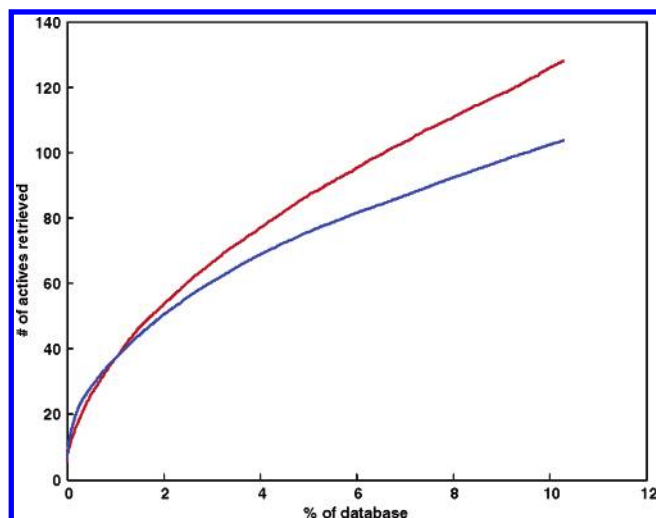


Figure 9. Cumulative number of actives retrieved over the first 10% of the database for the 5HT_r data set. Red line: ErG. Blue line: DFP.

Therefore, different settings of *incr* can be chosen by the user depending on the target (scaffold-hopping or not) of the respective study. In case compounds with more diverse interfeature distances are preferred, a value of *incr* = 0.2–0.3 should be used, whereas if *incr* is set to lower values, more closely related compounds will be obtained. Up to now, experience with the new descriptor suggests that the default settings (*incr* = 0.3) typically favor scaffold-hopping. Higher values of *incr* are not recommended since these could lead to too-similar descriptors for structurally dissimilar compounds (data not shown).

Diversity of Hits—“Scaffold-Hopping”. In a first step, the cumulative retrieval rates of ErG and DFP for the first 10% of the database were computed. Being a substructure-based descriptor, DFP should be able to better identify close analogues of the needle structure at the top of the hit list. With a larger portion of the hit list, the scaffold-hopping ability of ErG should become more and more important, and consequently, the search performance of ErG should improve. An example for such a change in performance is shown in Figure 9.

For the 11 data sets, the stated phenomenon can be observed in five cases (D₂, COX, 5HT_r, Ren, and HIV). For the other six test cases, a different behavior is found (see below). However, in two (HIV and Ren) out of the aforementioned five test cases, the secondary increase in performance of ErG is subsequently outperformed by DFP if a larger percentage of the database is retrieved (see Figure 10 for the most striking example).

A first indication that especially this latter data set is different in nature than the others was previously identified (ratio of the number of assigned properties per compound, see above). A closer investigation of the structural characteristics of the HIV and Ren data sets shows that they are the only ones of the 11 for which a large part of the compounds exhibits an extended peptidic backbone. This highlights another deficiency of reduced graph approaches if their encoding scheme is of quantitative nature (i.e., if the descriptor represents counts of a particular feature rather than the presence or absence of a feature which results in a bitstring). Since single-atom features such as H-bonding and

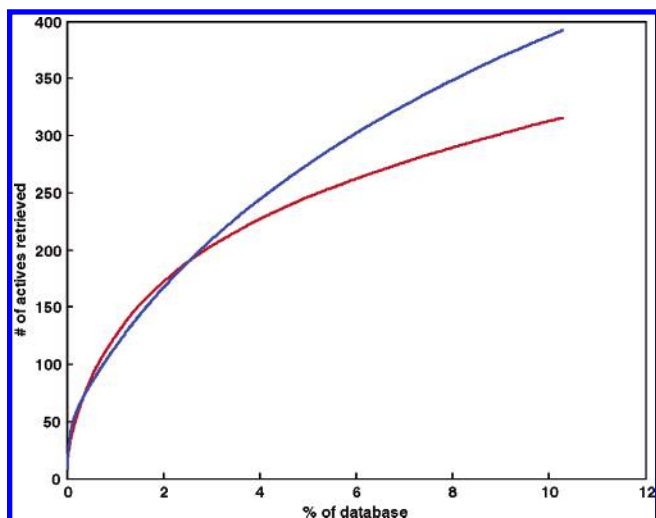


Figure 10. Cumulative number of actives retrieved over the first 10% of the database for the HIV data set. Red line: ErG. Blue line: DFP.

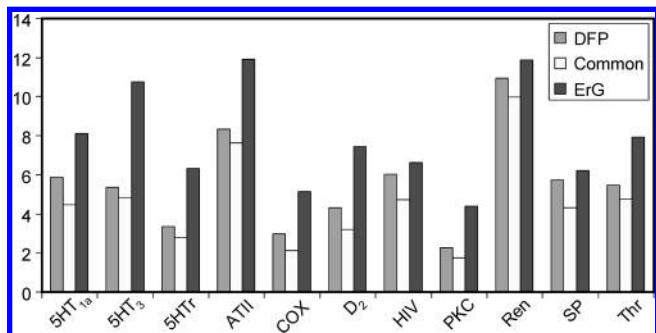


Figure 11. Results for the clustering of the hits retrieved at 1% of the database for the different activity classes. Displayed is the average number of clusters identified for DFP, ErG, and those common to both methods.

charge are usually much more frequent than multiatom features (hydrophobic, endcaps, and aromatic), compounds that comprise many of the former features (e.g., peptides) will generate an “information overflow”. Put differently, for peptides, the important pharmacophoric features are often hidden within the multitude of non-interacting “linker” atoms (amides) that are assigned a pharmacophoric property. Additionally, the side chains of most amino acids encode no or only one pharmacophoric property. Hence, the overall ratio of encoded interacting pharmacophoric points to assigned feature is comparatively small.

The curves for the other six data sets can be split into subgroups. In four cases (AT1, 5HT_{1a}, PKC, and Thr), ErG performs similarly to DFP over the close analogues, and then it starts to exhibit an increase in performance. For 5HT₃, ErG exhibits a dramatic performance superiority, which makes an interpretation of the curve difficult. The same is true for the SP data set, only that, in this case, DFP outperforms ErG.

The second step to investigate ErG’s ability to identify compounds different from DFP is based on the clustering of identified hits on the basis of DFP as described in the Methods section. Given the sometimes similar values of RT_{1%} for DFP and ErG, it was especially important to see whether ErG is actually finding other structural classes than those found by DFP. Figure 11 lists the overall number of clustered “structure families” for ErG and DFP.

It can nicely be seen that, in most cases, ErG is able to add a large number of structurally diverse actives to the clusters identified by DFP. In particular, the fact that even for data sets for which ErG is outperformed by DFP the number of clusters identified is higher (SP) shows that ErG is able to add new scaffolds to the DFP hits.

Interestingly, given that the SP data set is behaving completely different with respect to retrieval performance (see above) and, therefore, cannot be taken into account here, again, the HIV and Ren data sets exhibit different characteristics than the other data sets. For these more “peptidic” data sets, the number of additional structural families identified by ErG is comparatively small. Overall, this again can be ascribed to the peculiarities of the pharmacophoric reduced graph approach. However, this time it is not the amount of information that “hides” the real pharmacophoric properties but the fact that ErG will mainly identify other compounds with a high number of pharmacophoric points in a peptidic arrangement. That way, other peptides are identified more often as compared to small molecules.

Structural Examples. To further support the findings that ErG is, on the one side, behaving differently than DFP and, on the other, able to perform scaffold-hopping, specific cases (i.e., single needles) in which ErG performs similar to DFP or is outperformed by DFP were randomly chosen (to easier visualize the results, only examples with low numbers of hits are displayed here). The hits for both methods were investigated, and example structures are given in Figure 12.

The structures given in Figure 12 highlight the advantages and disadvantages of ErG compared to DFP. Independent of the needle used, nearly all DFP hits exhibit specific substructures, which are found in the respective needles (see Figure 13). That way, a large set of rather similar compounds is identified, which results in comparatively high retrieval rates. This property of DFP to identify similar compounds on the basis of a high degree of structural similarity is helpful if the user wants to find structurally near neighbors in a follow-up study. (It should be noted here that DFP would struggle if the needle was unique with respect to its distribution of substructures. In this case, the more generic ErG is more likely to retrieve similar compounds than DFP because of its “many substructures map to one pharmacophore” feature.)

However, if the target of the similarity search is to identify novel structures, this structural similarity usually has a lower priority. Here, the hits identified by ErG usually give a more diverse set of structures, as can be seen in Figure 12. Even though the ErG approach is still dependent on the patterns given in the needle, the abstraction of the ring systems as well as the inclusion of a fuzzy distance counting allows for a more generic identification of bioisosteric substructures. In particular, the way ring systems are encoded in ErG has a high impact on the diversity of the retrieved hits. As opposed to the reduced graphs introduced by Gillet et al.,¹⁷ ErG separates hydrogen-bonding properties of ring systems and the respective ring features (see Methods section). This more generic approach increases the similarity of aromatic rings independent of their substitution pattern. On the other hand, the uncoupling of hydrogen-bonding features from ring systems improves the similarity of ring- and nonring-containing compounds. Additionally, because ErG is not differentiating between distances of meta and para substitu-

tion (both show the same distance to the abstract ring centroid), the degree of diversity is further enhanced.

However, the latter two issues are not necessarily an advantage. For hetero-ring systems with a high degree of non-carbon atoms (e.g., EXTREG: 267812 in Figure 12c), the number of PPDTs incremented in the descriptor vector may lead to overemphasizing that particular ring system, which in turn may restrict the retrieved compounds to very similar ring systems. Another disadvantage may be the fact that substitution patterns necessary for biological activity are sometimes highly conserved. In such a case, not differentiating between meta and para substitution may result in fewer hits being identified.

Inspecting the results of both techniques led to another interesting observation: if compounds are selected on the basis of specific substructural moieties of the query structure, these structural moieties are not necessarily connected in the same way as in the query structure. Here, another advantage of the reduced graph approach could be exploited. Owing to the small number of graph nodes, a clique detection algorithm could be applied as proposed earlier by Takahashi et al.¹⁴

“ErG Normalization”. On the basis of the findings above, it was interesting to see whether it would be possible to reduce the problem of a high emphasis of very frequently appearing PPDTs by some sort of normalization procedure. Hence, the ErG descriptor was calculated for the MDDR in binary form (i.e., no fuzzy counts and no summation of entries), and the distribution vector **d** over the database was calculated according to

$$\mathbf{d} = \frac{\text{sum}(\mathbf{X})}{n}$$

where $\text{sum}(\mathbf{X})$ is the column-wise summation over the ErG descriptor matrix **X** and n is the number of compounds in the MDDR. This vector was used to weight the descriptor matrix **X** in order to give more weight to less frequently occurring vector entries. That way, the impact of less-informative variables should be reduced. The results of this procedure are given in Figure 14.

For most of the data sets, the weighting did not result in an obvious change of performance. However, an overall tendency toward a slight increase in retrieval rates can be observed, which is most prominent for 5HT₃, SP, and Thr. Even though the results obtained with this rather simplistic approach are not significant, they support the idea that, in principle, it may be possible to improve on retrieval rates if the information content of the database is taken into account. More extensive studies are, therefore, currently being undertaken in our groups.

CONCLUSION

ErG is a 2D description of chemical entities based on the concepts of reduced graphs introduced by Gillet et al.¹⁷ and Barker et al.¹⁸ In the latter publications, it was shown that this concept of reduced graphs to similarity searching allows for a simple and straightforward identification of biologic actives in large databases. This general applicability and performance of reduced graph approaches in ligand-based virtual screening was further validated in this contribution.

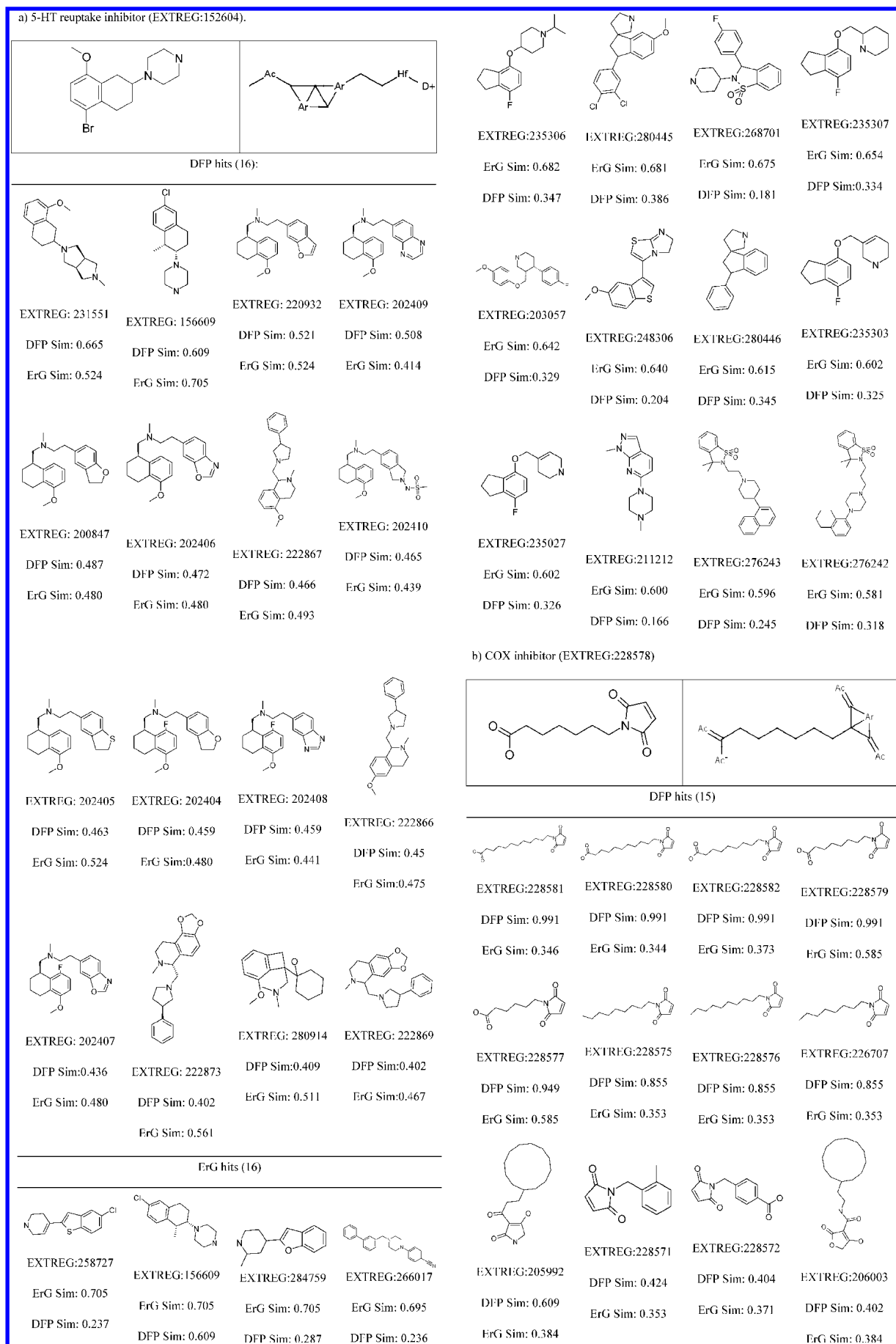


Figure 12.

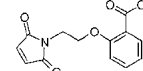

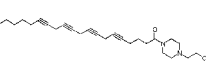
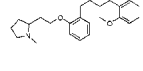
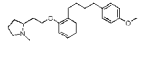
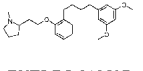
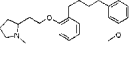
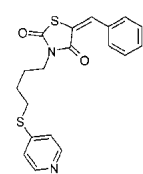
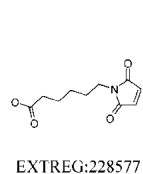
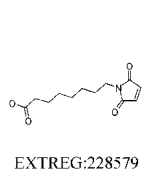
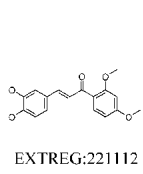
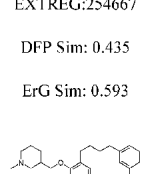
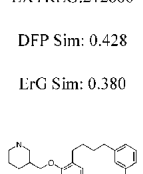
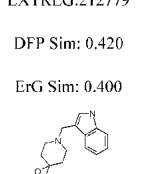
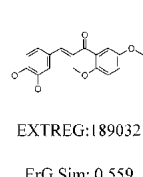
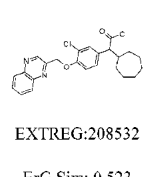
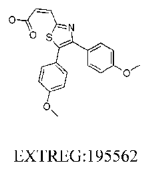
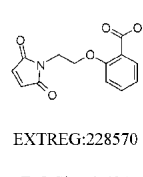
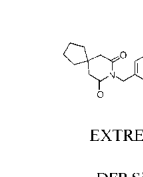
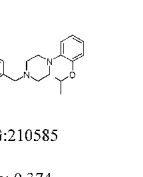
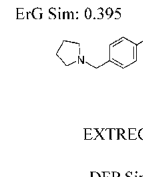
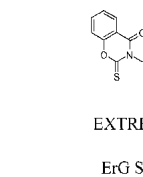
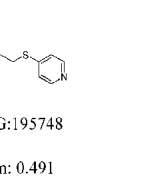
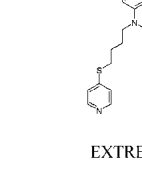
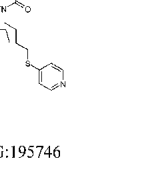
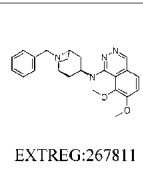
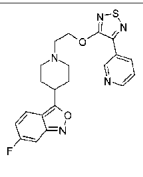
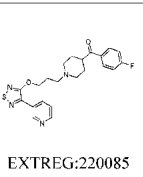
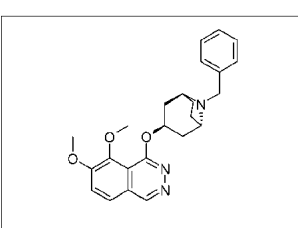
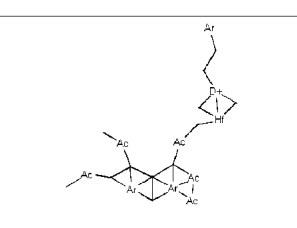
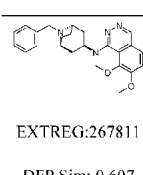
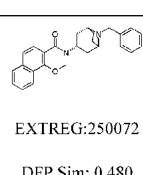
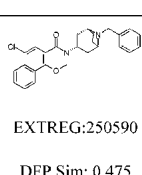
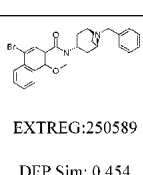
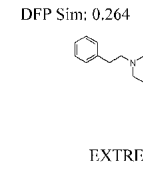
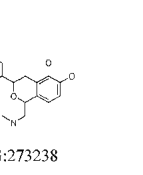
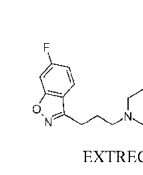
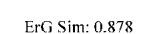
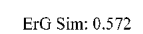
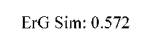
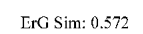



						
EXTREG:228570	EXTREG:144766	EXTREG:147062	EXTREG:212595	EXTREG:212596	EXTREG:212597	EXTREG:212594
DFP Sim: 0.398	DFP Sim: 0.378	DFP Sim: 0.376	DFP Sim: 0.444	DFP Sim: 0.442	DFP Sim: 0.440	DFP Sim: 0.436
ErG Sim: 0.494	ErG Sim: 0.261	ErG Sim: 0.25	ErG Sim: 0.376	ErG Sim: 0.380	ErG Sim: 0.439	ErG Sim: 0.380
ErG hits (10)						
						
EXTREG:195747	EXTREG:228577	EXTREG:228579	EXTREG:221112	EXTREG:254667	EXTREG:212600	EXTREG:212779
ErG Sim: 0.62	ErG Sim: 0.585	ErG Sim: 0.585	ErG Sim: 0.559	DFP Sim: 0.435	DFP Sim: 0.428	DFP Sim: 0.420
DFP Sim: 0.191	DFP Sim: 0.949	DFP Sim: 0.991	DFP Sim: 0.129	ErG Sim: 0.593	ErG Sim: 0.380	ErG Sim: 0.400
						
EXTREG:189032	EXTREG:208532	EXTREG:195562	EXTREG:228570	EXTREG:210585	EXTREG:210587	EXTREG:227653
ErG Sim: 0.559	ErG Sim: 0.523	ErG Sim: 0.507	ErG Sim: 0.494	DFP Sim: 0.374	DFP Sim: 0.374	DFP Sim: 0.377
DFP Sim: 0.133	DFP Sim: 0.153	DFP Sim: 0.117	DFP Sim: 0.398	ErG Sim: 0.385	ErG Sim: 0.385	DFP Sim: 0.388
						
EXTREG:195748	EXTREG:195746	EXTREG:210585	EXTREG:210587	EXTREG:210585	EXTREG:210587	EXTREG:210587
ErG Sim: 0.491	ErG Sim: 0.481	ErG Sim: 0.452	ErG Sim: 0.374	DFP Sim: 0.374	DFP Sim: 0.374	DFP Sim: 0.377
DFP Sim: 0.074	DFP Sim: 0.184	ErG Sim: 0.452	ErG Sim: 0.374	ErG Sim: 0.452	ErG Sim: 0.374	ErG Sim: 0.475
c) D ₂ antagonist (EXTREG: 267812)						
						
DFP hits (18)						
						
EXTREG:267811	EXTREG:250072	EXTREG:250590	EXTREG:250589	EXTREG:218095	EXTREG:220088	EXTREG:273240
DFP Sim: 0.607	DFP Sim: 0.480	DFP Sim: 0.475	DFP Sim: 0.454	ErG Sim: 0.716	ErG Sim: 0.716	ErG Sim: 0.714
ErG Sim: 0.878	ErG Sim: 0.572	ErG Sim: 0.572	ErG Sim: 0.572	DFP Sim: 0.264	DFP Sim: 0.264	DFP Sim: 0.260
						
EXTREG:267811	EXTREG:250072	EXTREG:250590	EXTREG:250589	EXTREG:273238	EXTREG:273241	EXTREG:273241
DFP Sim: 0.607	DFP Sim: 0.480	DFP Sim: 0.475	DFP Sim: 0.454	ErG Sim: 0.695	ErG Sim: 0.688	ErG Sim: 0.688
ErG Sim: 0.878	ErG Sim: 0.572	ErG Sim: 0.572	ErG Sim: 0.572	DFP Sim: 0.255	DFP Sim: 0.250	DFP Sim: 0.250

Figure 12. Needles randomly chosen from the activity classes and the respective hits found with DFP and ErG in the first 1% of the database. The number of compounds retrieved for each needle is given in parentheses. Needles in their original structure and the converted structure used for the ErG search are given in the respective boxes. Both similarity values (DFP and ErG) are shown for each of the hits. It can be nicely seen that the ErG hits are structurally much more diverse and that the ErG similarity values have higher overall values. Ac: H-bond acceptor; D: H-bond donor; Ar: aromatic; Hf: hydrophobic; +/-: positive/negative charge (if combined with D/A, both bins are incremented).

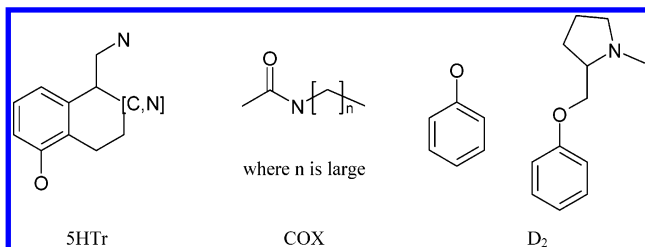


Figure 13. Common substructures found in the DFP hits of the respective needles. For D₂, two substructures are identified.

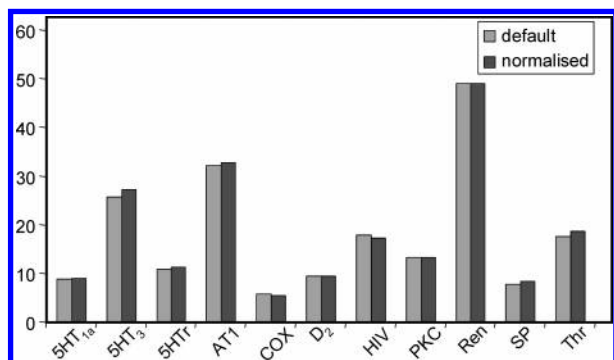


Figure 14. Average retrieval rates in the first 1% of the database for the default settings and the weighted ErG matrix.

However, in ErG, the introduced concept of reduced graphs was modified and extended to result in an approach that is not only able to produce comparatively good results but is also different in terms of its degree of abstraction, low in dimensionality, and therefore very fast to calculate.

With respect to average retrieval rates, ErG performs better than DFP for all but one data set under study. More importantly, however, ErG is able to “scaffold-hop” between different chemotypes of actives, which was shown with numerical and structural examples. This feature, in combination with the low number of descriptor entries of the ErG descriptor (315 bins) and its good interpretability of single vector entries, allows the expert as well as nonexpert user to understand why specific compounds were selected as similar.

With the latter properties, the weighting of specific property–property–distance combinations based on pharmacophores derived from target structures or other a priori knowledge of the target is a straightforward next step that was taken in our labs to improve on hit rates of virtual screening protocols.⁴⁰

ACKNOWLEDGMENT

The authors would like to thank Fred Bruns and Howard Broughton for fruitful discussions and a thorough review of the manuscript.

REFERENCES AND NOTES

- Bajorath, J. Virtual Screening in Drug Discovery: Methods, Expectations and Reality. *Curr. Drug Discovery* **2002**, *3*, 25–28.
- Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
- Sheridan, R. P.; Kearsley, S. K. Why do we need so many Chemical Similarity Search Methods? *DDT* **2002**, *7*, 885–931.
- MDL Information Systems Inc., San Leandro, CA.
- DAYLIGHT Inc., Mission Viejo, CA.

- Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144–6159.
- Mason, J. S.; Good, A. C.; Martin, E. J. 3-D Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.
- Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for their Self-Consistent Generation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1221–1227.
- Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. “Lead Hopping”. Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777–6791.
- Rush, T. S., III.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein–Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
- Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. ‘Scaffold-Hopping’ by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723–1740.
- Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639–643.
- Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Part 8. Reduced Chemical Graphs and Their Application in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.
- Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145–2156.
- Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
- Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346–356.
- Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
- Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- Martin, Y. C. 3D QSAR: Current State, Scope and Limitations. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1998; Vol. 3, pp 3–23.
- Oprea, T. I. On the Information Content of 2D and 3D Descriptors for QSAR. *J. Braz. Chem. Soc.* **2002**, *13*, 811–815.
- Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
- 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.
- Physiological ionization and pKa prediction. <http://www.daylight.com/meetings/emug00/Sayle/pkapedict.html>.
- 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.
- Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- Atlas of Protein Side-Chain Interactions. <http://www.biochem.ucl.ac.uk/bsm/sidechains/index.html>.
- The RCSB Protein Data Bank. <http://www.rcsb.org>.
- The MDL Drug Data Report database is available from MDL Information Systems Inc., San Leandro, CA.
- 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.
- <http://www.cheminformatics.org/>
- Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407–1414.

- (34) Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002**, 42 (1), 26–35.
- (35) Mahesh, R.; Perumal, R. V.; Pandi, P. V. Pharmacophore Based Synthesis of 3-Chloroquinoxaline-2-carboxamides as Serotonin₃ (5-HT₃) Receptor Antagonist. *Biol. Pharm. Bull.* **2004**, 27, 1403–1405.
- (36) Daveu, C.; Bureau, R.; Baglin, I.; Prunier, H.; Lancelot, J.-C.; Rault, S. Definition of a Pharmacophore for Partial Agonists of Serotonin 5-HT₃ Receptors. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 362–369.
- (37) Takeuchi, Y.; Shands, E. F. B.; Beusen, D. D.; Marshall, G. R. Derivation of a Three-Dimensional Pharmacophore Model of Substance P Antagonists Bound to the Neurokinin-1 Receptor. *J. Med. Chem.* **1998**, 41, 3609–3623.
- (38) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, DOI: 10.1021/ci050011h.
- (39) Flower, D. R. On the Properties of Bit String Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 379–386.
- (40) Stiefl, N.; Zaliani, A. A Knowledge-based Weighting Approach to Ligand-Based Screening. Manuscript in preparation.

CI050457Y