Journal of Medicinal Chemistry

Subscriber access provided by American Chemical Society

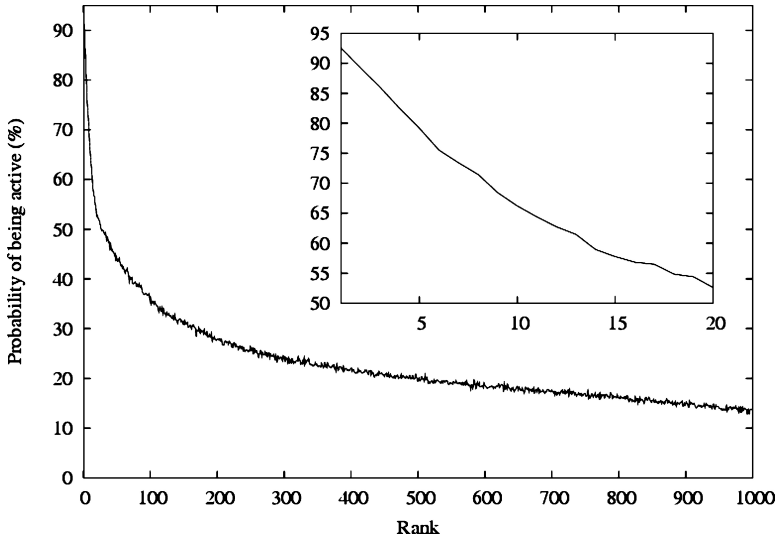
Article

# Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information

Jrme Hert, Peter Willett, David J. Wilton, Pierre Acklin,
Kamal Azzaoui, Edgar Jacoby, and Ansgar Schuffenhauer

*J. Med. Chem.*, **2005**, 48 (22), 7049-7054• DOI: 10.1021/jm050316n • Publication Date (Web): 01 October 2005

**Downloaded from http://pubs.acs.org on March 29, 2009**




## More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 6 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

ACS Publications
High quality. High impact.

Journal of Medicinal Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

# Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information

Jérôme Hert,[†] Peter Willett,*[,†] David J. Wilton,[†] Pierre Acklin,[‡] Kamal Azzaoui,[‡] Edgar Jacoby,[‡] and Ansgar Schuffenhauer[‡]

*Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K., and Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland*

*Received April 7, 2005*

We test the hypothesis that fusing the outputs of similarity searches based on a single bioactive reference structure and on its nearest neighbors (of unknown activity) is more effective (in terms of numbers of high-ranked active structures) than a similarity search involving just the reference structure. This turbo similarity searching approach provides a simple way to enhance the effectiveness of simulated virtual screening searches of the MDL Drug Data Report database.

## Introduction

Virtual screening involves scoring the molecules in a chemical database in order of decreasing probability of biological activity to ensure that potential hits are synthesized and tested at as early a stage as possible in a lead-discovery program.[1−3] There is much interest in structure-based approaches for virtual screening, where a 3D structure is available for the biological target.[4,5] When such information is not available, ligand-based approaches such as pharmacophore searching,[6] substructural analysis,[7] and similarity searching[8] can be used. Here, we focus on the last of these approaches. Similarity searching involves taking a molecule with the required activity, such as a weak hit from a high-throughput screening program, and then searching this target or reference structure against a database to find the molecules that are most similar to it.

There are many different types of similarity measure;[9−12] however, by far the most widely used are those based on 2D fingerprints and an association coefficient, most commonly the Tanimoto coefficient.[8] These were found to be effective in operation, and they are also extremely efficient, involving just simple logical operations on binary strings to compute the number of bits common to a pair of fingerprints. Fingerprint-based similarity searching is now some 20 years old;[13,14] however, the technique is of continuing interest,[15−22] and this paper provides a further contribution to the development of this popular approach to virtual screening. Specifically, we report a simple way of enhancing the effectiveness of similarity-based virtual screening, using information about the nearest neighbors (NNs) of the initial target structure in a similarity search. We refer to this approach as *turbo* similarity searching; a turbocharger increases the power of an engine by using the engine's exhaust gases, and here, we increase the power of a similarity searching procedure by using the reference structure's nearest neighbors. In the following, we refer to similarity searching and turbo similarity searching as *Sim* and *TurboSim*, respectively, with the latter being based on two observations: (1) the general applicability of the similar property principle and (2) our recent work on the use of multiple reference structures for similarity searching.[23−25]

The similar property principle was first presented explicitly by Johnson and Maggiora[26] and states that molecules that are structurally similar are likely to have similar properties (an idea that also underlies structural approaches to molecular diversity and chemogenomics[27,28]). If the principle applies to a particular biological activity and set of compounds, then the NNs of a bioactive reference structure are also likely to possess that activity. There are many exceptions to the principle,[29] with even very minor structural variations having a drastic effect on the levels of activity in a set of analogues. However, if the principle was not of general applicability, then it would be difficult to develop systematic approaches for the identification of novel bioactive molecules, and there is now substantial evidence to support its use in lead-discovery programs.[20,22,30−32]

Most studies of similarity searching have considered the use of only a single bioactive reference structure. The availability of published competitor compounds or hits from high-throughput screening (HTS) means that multiple reference structures may be available, and this has spurred interest in similarity searching methods that can make use of such information.[15,33,34] We recently reported a detailed comparison of several different search algorithms that can be used when multiple reference structures are available and showed that the best of these algorithms, a technique we call *group fusion*, results in a level of retrieval effectiveness that is noticeably superior to that obtainable from the use of a single reference structure. Subsequent studies have demonstrated the general applicability of this approach when used with a wide range of different types of 2D fingerprint and of different types of similarity coefficient.[23−25] The power of the approach was demonstrated by the finding that fusing the rankings from as few as 10 randomly selected reference structures was

* To whom correspondence should be addressed. Phone: +44-114-2222633. Fax: +44-114-2780300. E-mail: p.willett@sheffield.ac.uk.
[†] University of Sheffield.
[‡] Novartis Institutes for BioMedical Research.

10.1021/jm050316n CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/01/2005

Input the reference structure *R*

Compute the similarity of *R* with every molecule in the database *D*

Sort *D* in decreasing order of the calculated similarity values to give a sorted database *SD*(0)

Identify the *k* NNs of *R* from the top of the list *SD*(0)

For each such nearest-neighbour, *NN*(*i*)

    Compute the similarity of *NN*(*i*) with every molecule in *D*

    Sort *D* in decreasing order of the calculated similarity values to give a sorted database *SD*(*i*)

Fuse the sorted lists *SD*(0)-*SD*(*k*) to give the final output from the turbo similarity search

**Figure 1.** Use of nearest neighbors for turbo similarity searching.

more effective than the very best similarity search possible even when there were many hundreds of individual active molecules to choose from.[23]

Previous work has hence shown that using multiple active reference structures in a similarity search is more effective than using a single active reference structure and that the NNs of an active reference structure are also likely to be active. The combination of these two observations suggests that a turbo similarity search, i.e., one involving a reference structure and its NNs (referred to subsequently as TurboSim), is likely to be more effective than a similarity search involving just that reference structure on its own (referred to subsequently as Sim). If this can be shown to be the case, then we have an extremely simple way of enhancing the effectiveness of a conventional similarity searching system by adopting the strategy summarized in Figure 1. The appropriateness of the strategy is investigated in the remainder of this paper.

## Methods

If the TurboSim strategy in Figure 1 is to be effective, i.e., if it is to yield a greater number of high-ranked actives than the number obtainable from the original reference structure *R* on its own, then we must first identify an appropriate way of combining the sorted lists *SD*(0)−*SD*(*k*) and, second, demonstrate that the NNs of *R* are also able to retrieve active structures from the database *D*.

*Data fusion* is the name given to a body of techniques that are used to combine the results of different rankings of a database in response to a reference structure (the name *consensus scoring* is often used when a database is ranked using a docking algorithm). Previous studies of data fusion have used a single reference molecule that is characterized by several different representations or matched with the database using several different similarity coefficients.[16,35,36] The alternative group fusion approach has a single representation and a single similarity coefficient but involves combining the search outputs obtained with several different reference structures.[25] Specifically, assume that some database molecule *j* yields similarity scores of $s_1, s_2, ..., s_k$ with *k* different reference structures. Then an effective similarity search can be obtained by ranking the database molecules on the basis of the largest of these scores, i.e., $\max\{s_1, s_2, ..., s_i, ..., s_{n-1}, s_k\}$.[23] The similarity scores in all of the experiments reported in this paper were computed using the Tanimoto coefficient, with the reference and database structures characterized by Scitegic ECFP_4 fingerprints. These fingerprints encode circular substructures centered on each non-hydrogen atom in a molecule by a string of extended connectivity values that are calculated using a modification of the Morgan algorithm. ECFP stands for extended connectivity fingerprint. At this level of description, each atom is initially coded by an integer describing that atom's number of connections, element type, charge, and mass, and the 4 denotes the diameters of the circular substructures that are encoded in the fingerprint. The Scitegic software represents a molecule by a list of integers each describing a molecular feature and each in the range $-2^{31}$ to $2^{31}$. Here, the integers describing a molecule were hashed to a string of 1024 bits as described previously by Hert et al.[24]

**Table 1.** Mean Recall at 5% for Conventional Similarity Searching (Sim) with Just a Single Reference Structure and Turbo Similarity Searching (TurboSim) Using Different Numbers of NNs

| | | | number of NNs in TurboSim | | | | |
|---|---|---|---|---|---|---|---|
| activity class | actives | Sim | 5 | 10 | 20 | 50 | 100 |
| 5HT3 antagonists | 752 | 31.7 | 34.8 | 36.8 | 38.6 | 42.1 | 44.0 |
| 5HT1A agonists | 827 | 26.3 | 28.1 | 29.6 | 31.8 | 34.5 | 36.2 |
| 5HT reuptake inhibitors | 359 | 21.6 | 23.4 | 24.0 | 23.8 | 24.3 | 24.1 |
| D2 antagonists | 395 | 25.1 | 25.8 | 26.9 | 27.5 | 29.1 | 30.3 |
| renin inhibitors | 1130 | 90.4 | 91.2 | 92.1 | 93.1 | 94.3 | 94.7 |
| angiotensin II AT1 antagonists | 943 | 77.4 | 80.8 | 83.5 | 86.7 | 90.2 | 92.0 |
| thrombin inhibitors | 803 | 44.5 | 45.6 | 47.1 | 48.3 | 51.0 | 50.7 |
| substance P antagonists | 1246 | 28.6 | 30.5 | 31.7 | 32.2 | 33.3 | 34.1 |
| HIV protease inhibitors | 750 | 51.6 | 51.9 | 52.6 | 53.3 | 54.5 | 55.2 |
| cyclooxygenase inhibitors | 636 | 13.7 | 14.6 | 15.0 | 15.3 | 15.1 | 14.4 |
| protein kinase C inhibitors | 453 | 21.0 | 21.1 | 21.1 | 21.1 | 20.9 | 20.6 |
| average over all classes | 754 | 39.2 | 40.7 | 41.9 | 42.9 | 44.5 | 45.1 |

The extent to which the procedure shown in Figure 1 can enhance retrieval effectiveness was studied using the simulated virtual-screening system employed in our previous studies of similarity searching using multiple reference structures.[23,24] Specifically, we used a version of the *MDL Drug Data Report* (MDDR) database from which we removed duplicates and molecules that could not be processed using local software to give a total of 102 514 molecules that were available for searching. This database was searched using the 11 sets of bioactive compounds noted in Table 1, with Sim and TurboSim searches being carried out for all of the 8294 active molecules in these 11 classes.

## Results and Discussion

An initial experiment was carried out to confirm that the data sets studied here satisfy the similar property principle, as this is an inherent assumption underlying TurboSim. The probability that a particular database compound would be active was plotted as a function of its rank, i.e., the part of a ranked list in which it would occur during a conventional similarity search. This probability was obtained by averaging the number of times a compound at a given rank was active when averaged over all of the 8294 individual similarity searches. The results are shown in Figure 2, which confirms the potential of NNs for TurboSim searching. Inspection of the inset shows that the first 15 NNs have a probability of about 0.6 or more of being active. Analogous relationships between similarity scores (rather than rank positions) and probability of activity have been noted by previous workers.[31,32,37,38]

The searches here are evaluated by the recall at 5%, i.e., the percentage of the database molecules belonging to the same activity class as the reference structure that is retrieved in the top 5% of a ranking of the database.

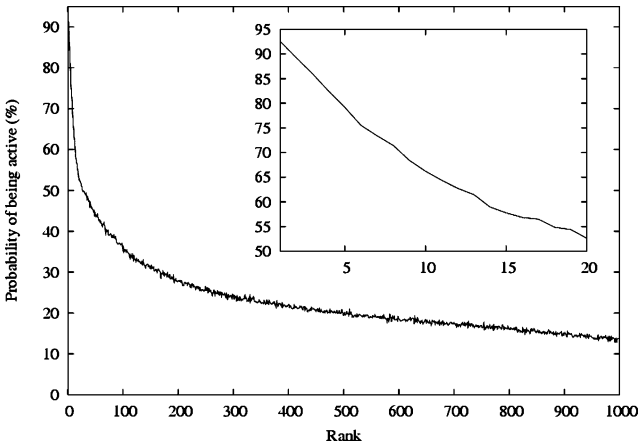



**Figure 2.** Average probability of a compound being active as a function of its rank. The inset shows the left-hand part of the plot (right at the top of the ranking) in greater detail.

The principal results of the study are listed in Table 1, which lists the mean recalls averaged over all of the individual active molecules in each activity class when a TurboSim search is carried out using the specified number of NNs (5, 10, 20, 50, or 100). The table also contains the comparable results for a conventional similarity search (Sim). Inspection of this table shows that TurboSim is nearly always superior to Sim in its ability to identify active molecules, with the only exceptions being the protein kinase C TurboSim-50 and TurboSim-100 searches. With some of the other activity classes, the increases in performance are really quite marked, most notably the 5HT3 and 5HT1A agonists where TurboSim-100 has mean recalls of 44.0 and 36.2 that are 38.8% and 37.6% higher than the corresponding Sim results. Even a small number of NNs is generally sufficient to bring about a noticeable increase in the number of actives retrieved; for example, the TurboSim-10 searches have a mean recall of 41.8 that is some 6.9% higher than the corresponding Sim result.

It is perhaps surprising that the best results are generally obtained with the largest number of NNs, since Figure 1 shows clearly that the probability that an NN is active drops off rapidly as one moves down the ranking of a data set. However, the fact that the average recall does increase, even with 100 NNs, means that these molecules are providing useful information. This situation is analogous to that described in a very recent paper by Klon et al.,[39] who considered high-scoring molecules in a docking study to be active (irrespective of whether this was the case) and then used this information in a subsequent substructural analysis study. At some point, one would assume that the inclusion of further NNs would start to affect the retrieval performance. For the activity classes studied here, we found that use of 100 NNs was significantly better than use of 200 NNs and we have hence not included results for the latter number of NNs in Table 1.

In a conventional similarity search, the search output reflects the direct relationships (as encoded in the molecules' 2D fingerprints) that exist between the reference structure and the database that is being searched. As a result of the fusion that has taken place, a TurboSim search output additionally accounts for the indirect relationships (as encoded in the NNs) that the reference structure has with the database. Because one is bringing this additional information to bear in a TurboSim search, it seems reasonable that the search performance should increase. However, as one includes more and more NNs in a TurboSim search, the structural relationships that the reference structure has with the NNs will become more and more tenuous. Some of them may have very low similarities with the reference structure and hence (following Figure 1) a low probability of activity. There is hence the possibility that the use of many NNs will result in the inclusion of less and less useful structural information; i.e., one is progressively adding background noise rather than any meaningful signal. A natural modification to the basic TurboSim procedure is hence to employ a user-defined similarity threshold in addition to the user-defined number of NNs; however, while this may result in less noise, it may also result in there being insufficient NNs with this degree of similarity with the reference structure. That there is such a tradeoff is illustrated by the results in Table 2, which lists the mean results (averaged over all of the activity classes) obtained when Tanimoto thresholds of 0.3, 04, 0.5, 0.6, and 0.7 were used in TurboSim-100 searches and also lists the mean number of NNs (averaged over all of the searches) when that threshold was used. It will be seen that the best results are obtained with the lowest threshold of 0.3 and that the average mean of the recalls with this threshold is better than when no threshold is defined. The fact that the best results are obtained with a Tanimoto value as low as 0.3 may appear surprising given previous studies suggesting that two compounds require a Tanimoto similarity of about 0.7−0.8 to have a similar bioactivity.[37,38] However, the numbers of NNs in Table 2 demonstrate clearly that there are normally far too few close neighbors at this similarity level using these particular fingerprints to enable the group fusion to be effective.

**Table 2.** Mean Recall at 5% for Turbo Similarity Searches Using 100 NNs and a Threshold Value for the Tanimoto Similarity

| | similarity threshold | | | | |
|---|---|---|---|---|---|
| | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
| average number of NNs | 81.6 | 44.1 | 20.3 | 7.8 | 2.9 |
| average over all classes | 45.2 | 44.0 | 42.0 | 40.2 | 39.5 |

The statistical significance (or otherwise) of the differences in recall performance of Sim and TurboSim was assessed using the sign test, as advocated by van Rijsbergen for the comparison of pairs of database search results.[40] Specifically, the null hypothesis $H_O$ was tested that there was no difference in recall between Sim and TurboSim, as detailed in Table 3. First, we considered the activity classes as a whole, giving a total of 11 cases where TurboSim could be superior (in terms of mean recall), equal, or inferior to Sim. In the sign test, the numbers of nonequal cases are used in a calculation based on the binomial distribution.[41] $H_O$ could be rejected with $p \leq 0.006$ for all sets of TurboSim searches listed in Table 3. The Sign Test was then repeated but considering each of the individual active molecules in turn, giving a total of 8294 cases where TurboSim could be superior, equal, or inferior to Sim. Here, the large-sample version of the sign test was used, involving a calculation based on the normal distribu-

**Table 3.** Data Used as Input to the Sign Test Analysis of the Significance of the Differences in Recall Performance between Sim and TurboSim

| | all 11 activity classes | | | all 8294 active molecules | | |
|---|---|---|---|---|---|---|
| NNs | TurboSim > Sim | TurboSim = Sim | TurboSim < Sim | TurboSim > Sim | TurboSim = Sim | TurboSim < Sim |
| 5 | 11 | 0 | 0 | 4804 | 195 | 3295 |
| 10 | 11 | 0 | 0 | 5361 | 157 | 2776 |
| 20 | 11 | 0 | 0 | 5815 | 128 | 2351 |
| 50 | 10 | 0 | 1 | 6378 | 100 | 1816 |
| 100 | 10 | 0 | 1 | 6451 | 78 | 1765 |
| 100 (0.3) | 11 | 0 | 0 | 6547 | 80 | 1667 |

**Table 4.** Mean Recall at 5% of Two Types of Upper-Bounds and Lower-Bounds for TurboSim

| | upper-bounds | | lower-bounds | |
|---|---|---|---|---|
| activity classes | ref + actives among the 100 NNs | ref + 100 active NNs | inactives among the 100 NNs | 100 inactive NNs |
| 5HT3 antagonists | 49.4 | 65.7 | 33.5 | 32.1 |
| 5HT1A agonists | 38.0 | 55.3 | 31.7 | 31.9 |
| 5HT reuptake inhibitors | 27.8 | 62.8 | 21.7 | 21.7 |
| D2 antagonists | 30.6 | 68.6 | 28.7 | 28.8 |
| renin inhibitors | 95.2 | 96.6 | 90.2 | 89.8 |
| angiotensin II AT1 antagonists | 91.6 | 95.2 | 90.9 | 92.2 |
| thrombin inhibitors | 58.6 | 71.6 | 37.4 | 33.9 |
| substance P antagonists | 42.2 | 53.8 | 20.4 | 15.8 |
| HIV protease inhibitors | 59.8 | 76.1 | 50.8 | 49.0 |
| cyclooxygenase inhibitors | 17.5 | 49.2 | 12.5 | 12.0 |
| protein kinase C inhibitors | 23.2 | 58.1 | 18.4 | 18.3 |
| average over all classes | 48.5 | 68.4 | 39.6 | 38.7 |

tion.[41] $H_O$ could be rejected with $p \leq 0.00003$ for all the TurboSim searches listed in Table 3. A sign test was also carried out to compare the 8294 TurboSim-100 searches with the corresponding TurboSim-100 searches when a threshold of 0.3 is applied. This test showed the latter to be significantly better ($p < 0.00003$).[41]

For comparison with the results in Tables 1 and 2, Table 4 lists the results from sets of searches using two lower-bounds and two upper-bounds. The lower-bounds were obtained by carrying out a TurboSim search using the inactives in the set of 100 NNs or using the top-ranked 100 inactive NNs. Although both values are noticeably less than the basic recall of 39.2% in Table 1, they are also much greater than the recall value that would be obtained by random selection, i.e., 5%. The effectiveness of these lower-bound searches may appear rather surprising; however, it simply means that even when inactive molecules are used in TurboSim, the molecules still contain sufficient relevant substructures in common with the reference structure to enable the identification of further active molecules. The two upper-bound searches demonstrate the performance available with full knowledge of the actives; when one uses a set of 100 active NNs, the search performance is very much greater than with TurboSim-100 (where one assumes that all of the top-100 NNs are active). When just the true actives in the top-100 NNs are used, the performance is much closer to TurboSim-100 (where one includes these actives and further molecules that are assumed to be active but that are actually inactive).

The data in Table 4 help to explain why TurboSim is effective in practice. The NNs of the reference structure, all of which are assumed to be active in a TurboSim search, are of two types: those that really are active and those that really are inactive. The upperbound searches demonstrate that use of the former has a considerable positive effect on search performance, while the lowerbound searches demonstrate that the use of the latter has only a marginal effect on search performance. Taken together, these two types of behavior yield the overall enhancements demonstrated in Table 1. A further, more qualitative reason may explain why TurboSim seems to work well in practice. The inclusion of NN-based similarity searches in TurboSim results in a broadening of the focus that characterizes a conventional Sim search. This means that the search can better explore the chemical space if the desired active molecules are not tightly clustered together (as in the normal expectation in Sim). Such islands of activity are known to exist.[42−44] If this is the case, then we would expect that TurboSim would become progressively better than Sim as the diversity of the actives increases. Recent work in Sheffield demonstrates that this behavior is observed in practice, validating the use of TurboSim for scaffold hopping applications.[45]

**Table 5.** Mean Recall at 5% (Averaged over all 11 Activity Classes) of Active Ring Systems for Conventional Similarity Searching (Sim) with Just a Single Reference Structure and Turbo Similarity Searching (TurboSim) Using Different Numbers of NNs

| ring definition | Sim | Turbo-Sim-5 | Turbo-Sim-10 | Turbo-Sim-20 | Turbo-Sim-50 | Turbo-Sim-100 |
|---|---|---|---|---|---|---|
| cyclic systems | 40.9 | 42.6 | 43.7 | 44.8 | 46.3 | 46.8 |
| skeletal cyclic systems | 42.2 | 43.8 | 44.9 | 45.9 | 47.3 | 47.6 |
| reduced-aryl cyclic systems | 43.4 | 44.8 | 45.8 | 46.6 | 47.7 | 47.8 |
| reduced-skeletal cyclic systems | 45.0 | 46.3 | 47.3 | 48.1 | 48.9 | 48.9 |

The results in Tables 2 and 3 involved the use of a fixed number of NNs and the use of a similarity threshold for TurboSim. However, it is possible to consider more complex ways of using the NN information, for example, by scaling the results so that ranked lists resulting from NNs that are strongly similar to the reference structure contribute more to the final fused ranking than do the lists from the less similar NNs. Variations of this idea for ranking databases of text documents have been discussed by Croft et al.;[46] however, we found that such variations were not particularly effective in the current context, and they are hence not discussed here. Similar comments apply to TurboSim-5 to TurboSim-100 searches based on sets of compounds obtained by applying a MaxMin diversity selection algorithm to the set of 200 NNs for each reference structure.

In the final set of experiments, as in our previous study[24] and as advocated recently by Good et al.,[47] we looked at the ability of TurboSim to identify not just active molecules but active ring scaffolds, i.e., scaffolds that occur in the sets of bioactive molecules for each of the 11 MDDR activity classes. Table 5 lists the recall at 5% for four different levels of ring specificity defined in the MEQI software from Pannanugget Consulting, as exemplified in Figure 3: cyclic systems, skeletal cyclic systems, reduced-aryl cyclic systems, and reduced-skeletal cyclic systems. The trends in this table mirror closely those in Table 2, with TurboSim again resulting in noticeable increases in recall. The largest increases are associated with the most detailed level of description, i.e., the cyclic systems.

Thus far, we focused on the effectiveness of the process in terms of its ability to retrieve active mol-

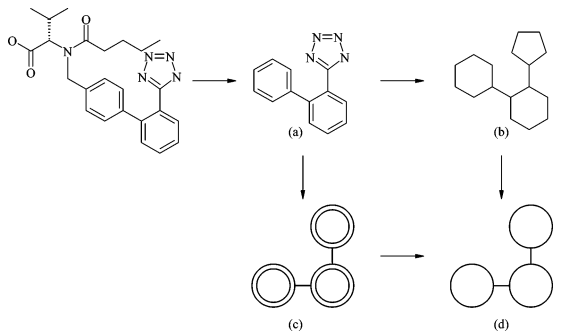



**Figure 3.** Hydrogen-free example of cyclic system (a), skeletal cyclic system (b), reduced-aryl cyclic system (c), and reduced-skeletal cyclic system (d) for Diovan.

ecules; however, no database-searching procedure will be of practical utility unless its efficiency enables it to be implemented on large files on a routine basis. If large numbers of NNs are to be used, then search times will necessarily be much extended, as inspection of Figure 1 would suggest that use of $k$ NNs in the TurboSim will require the additional calculation of $k$ times as many similarity coefficients as will Sim. In fact, search times can be minimized by the following procedure: execute the search of the database for the original reference structure; identify its NNs and note the value of the similarity coefficient for each of the database molecules; carry out a second search of the database, with each database molecule being matched against all of the $k$ NNs, updating its associated maximum score if appropriate. This will require just two scans of the database, rather than the $k + 1$ scans that would be required by a straightforward implementation of Figure 1. In our implementation with C programs running under Unix on a Linux PC, the TurboSim-100 searches were only about 5 times slower than the basic SS searches.

In conclusion, we briefly compare TurboSim with other ways of using NN information in a similarity search. A basic similarity search, referred to here as Sim, involves simply matching the reference structure against each of the molecules in the database to find the NNs. There are at least three obvious extensions, the first of which is iterative, or sequential, similarity searching. Here, the NNs resulting from the initial search are assayed, and those that prove to be active are then used as reference structures in their own right for database searches. Rather than using these active NNs individually, a second extension is to use the set of them in a group fusion procedure in which the database is ranked against all of the active NNs and the resulting rankings are fused using a fusion rule such as the sum or the maximum of the similarity scores or the similarity ranks. In the third extension, the active NNs can be combined with those NNs that proved to be inactive in the assay, and the resulting set of actives and inactives can be used as a training set for a machine-learning procedure (such as substructural analysis, a support vector machine, or binary kernel discrimination). While such procedures are known to be highly effective, they all require some biological testing that is carried out on the ranked output resulting from the initial similarity search to identify those NNs that are truly active. The TurboSim procedure, conversely, *assumes* that the NNs are active and uses this assumption to maximize the effectiveness of the ranking resulting from the initial search based on the single active reference structure. Once this ranking has been generated, the NNs are assayed and any of the procedures above can be invoked in the normal way. The only previous work of which we are aware that is related to the TurboSim approach is a recent study of reduced graphs by Harper et al.[48] Here, an initial similarity search is carried out using reduced-graph representations of the reference structure of the database molecules. The NNs resulting from this search are inspected manually and then some of them selected for use in a subsequent group-fusion similarity search based on Daylight fingerprint representations; i.e., manual selection is used for the second-stage search rather than the bioassays required for the three extensions noted previously.

## Conclusions

In this paper we showed that it is possible to increase the effectiveness of similarity-based virtual screening by carrying out multiple database searches using the nearest neighbors resulting from an initial similarity search. The approach requires no modifications to existing similarity software other than the ability to fuse the outputs of the multiple searches to give a single combined ranking of the database structures. These increases in search effectiveness are achieved at minimal computational cost, and we hence conclude that turbo similarity searching provides a very simple way of increasing the cost effectiveness of similarity-based virtual screening systems. Our experiments have used 2D fingerprints and the Tanimoto coefficient, but there is no reason in principle why this approach could not also be used with any other type of similarity measure that satisfies the similar property principle.

**Acknowledgment.** We thank Novartis Institutes for Biomedical Research for funding; MDL Information Systems Inc. for the provision of the MDDR database; and Pannanugget Consulting, the Royal Society, Scitegic Inc., Tripos Inc., and the Wolfson Foundation for software and laboratory support.

## References

(1) Bohm, H.-J., Schneider, G., Eds. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, Germany, 2000.
(2) Klebe, G., Ed. *Virtual Screening: An Alternative or Complement to High Throughput Screening*; Kluwer: Dordrecht, The Netherlands, 2000.
(3) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(4) Lyne, P. D. Structure-Based Virtual Screening: An Overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.
(5) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein−Small Molecule Docking Methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.
(6) Warr, W. A.; Willett, P. The Principles and Practice of 3D Database Searching. *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; American Chemical Society: Washington, DC, 1997; pp 73−95.
(7) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1974**, *17*, 533−538.
(8) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(9) Dean, P. M., Ed. *Molecular Similarity in Drug Design*; Chapman and Hall: Glasgow, U.K., 1994.
(10) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(11) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(12) Nikolova, N.; Jaworska, J. Approaches To Measure Chemical Similarity. A Review. *Quant. Struct.−Act. Relat. Comb. Sci*. **2003**, *22*, 1006−1026.
(13) 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.
(14) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(15) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.
(16) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(17) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.
(18) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.
(19) 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.
(20) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.
(21) Xue, L.; Stahura, F. L.; Bajorath, J. Similarity Search Profiling Reveals Effects of Fingerprint Scaling in Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2004**, 2032−2039.
(22) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-Directed Nearest-Neighbor Searching. *J. Med. Chem.* **2005**, *48*, 240−248.
(23) 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.
(24) 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.
(25) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Losel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest-Neighbour Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.
(26) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
(27) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(28) Schuffenhauer, A.; Jacoby, E. Annotating and Mining the Ligand−Target Chemogenomics Knowledge Space. *BIOSILICO* **2004**, *2*, 190−200.
(29) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 225−252.
(30) 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.
(31) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(32) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activities? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(33) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746−753.
(34) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(35) 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.
(36) Salim, N.; Holliday, J. D.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.
(37) Delaney, J. Assessing the Ability of Chemical Similarity Measures To Discriminate between Active and Inactive Compounds. *Mol. Diversity* **1995**, *1*, 217−222.
(38) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(39) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743−2749.
(40) van Rijsbergen, C. J. *Information Retrieval*; Butterworth: London, 1979.
(41) Siegal, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, 1988.
(42) Oprea, T. I. Chemical Space Navigation in Lead Discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384−389.
(43) Jacoby, E.; Davies, J.; Blommers, M. J. J. Design of Small Molecule Libraries for NMR Screening and Other Applications in Drug Discovery. *Curr. Top. Med. Chem*. **2003**, *3*, 11−23.
(44) Dobson, C. M. Chemical Space and Biology. *Nature* **2004**, *432*, 824−828.
(45) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Manuscript in preparation.
(46) Croft, W. B.; Lucia, T. J.; Cringean, J. K.; Willett, P. Retrieving Documents by Plausible Inference: An Experimental Study. *Inf. Process. Manage.* **1989**, *25*, 599−614.
(47) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD Technique Performance: A Virtual Screening Case Study in the Design of Validation Experiments. *J. Comput.-Aid. Mol. Des.* **2004**, *18*, 529−536.
(48) 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.

JM050316N