# Turbo Similarity Searching: Effect of Fingerprint and Dataset on Virtual-Screening Performance

Eleanor J. Gardiner, Valerie J. Gillet, Maciej Haranczyk, Jérôme Hert, John D. Holliday, Nurul Malim, Yogendra Patel and Peter Willett[1]

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Portobello Street, Sheffield S1 4DP, United Kingdom

**Abstract**. Turbo similarity searching uses information about the nearest neighbours in a conventional chemical similarity search to increase the effectiveness of virtual screening, with a data fusion approach being used to combine the nearest-neighbour information. A previous paper suggested that the approach was highly effective in operation; this paper further tests the approach using a range of different databases and of structural representations. Searches were carried out on three different databases of chemical structures, using seven different types of fingerprint, as well as molecular holograms, physicochemical properties, topological indices and reduced graphs. The results show that turbo similarity searching can indeed enhance retrieval but that this is normally achieved only if the similarity search that acts as its starting point has already achieved at least some reasonable level of search effectiveness. In other cases, a modified version of TSS that uses the nearest-neighbour information for approximate machine learning can be used effectively. Whilst useful for qualitative (active/inactive) predictions of biological activity, turbo similarity searching does not appear to exhibit any predictive power when quantitative property data is available.

## INTRODUCTION

Computer methods for the storage, retrieval and processing of chemical-structure information are well established [1] but it is only within the last few years that chemoinformatics (as it is now known) has become established as a key component in the search for novel agrochemicals and pharmaceuticals [2, 3]. An important aspect of chemoinformatics is *virtual screening*, the computer prioritisation of molecules in order of probability of biological activity (where the activity might be, e.g., killing some troublesome aphid in agrochemical research or lowering a person's cholesterol level in pharmaceutical research) so that attention can then be focused on those molecules that are most likely to exhibit the required activity [4-7]. There are many ways in which this can be achieved: in this paper we

[1] To whom all correspondence should be addressed: email p.willett@sheffield.ac.uk; tel. +44-114-2222633

focus on one of the most widely used, viz *similarity searching* [8, 9]. Given a molecule that exhibits some biological activity of interest (the *reference structure*) and a database of molecules that have not previously been tested for that activity, a similarity search procedure involves computing the similarity between the bioactive reference structure and each database structure, using some quantitative measure of structural resemblance. The database is then ranked in decreasing order of the computed similarities, and some fraction (e.g., the top-5%) of the ranked database passed on for further consideration, since these top-ranked molecules (the nearest neighbours) are expected to have the greatest *a priori* probabilities of bioactivity. Many different types of similarity measure have been discussed in the literature [10] but most current systems for lead discovery use measures based on 2D fingerprints and the Tanimoto coefficient, where a *fingerprint* is a binary vector encoding the presence or absence in a molecule of (typically a few hundred) small substructural fragments. Fingerprint-based similarity is clearly simple in concept but has proved to be very effective in operation [11-16]. Much of the popularity of similarity searching derives from the fact that it requires very little information for screening, specifically just a single known active, such as a competitor compound or a hit from an initial high-throughput screening experiments. It is thus normally used in the very initial stages of lead discovery, with more sophisticated screening procedures such as pharmacophore matching, CoMFA, SIMCA etc. becoming the methods of choice as training data becomes available in the lead-discovery and lead-optimisation stages.

Hert *et al*. have described an extension of similarity searching, referred to as *turbo similarity searching* (for reasons described below), that is based on two observations [17]. First, the *similar property principle* states that molecules that are structurally similar are likely to exhibit similar activities and properties [18, 19]. If the principle applies to a particular biological activity and set of compounds then the nearest neighbours of a bioactive reference structure are also expected to possess that activity. Second, recent studies have demonstrated the increased effectiveness of searching that can be obtained if not one but multiple bioactive reference structures are available, using a consensus approach called *group fusion* [16, 20-23]. Here, each reference structure in turn is used for a similarity search, and then the resulting rankings are combined to give a single consensus ranking. Turbo similarity searching (which we shall normally abbreviate to TSS) makes the assumption that the nearest neighbours of a reference structure are not just *likely* to be active (as suggested by the similar property principle) but actually *are* active, in which case these assumed active molecules can be used as reference structures in addition to the original reference structure (thus providing the multiple reference structures that are required for group fusion). These additional, nearest-neighbour searches are carried out automatically and thus the user of a TSS system

need do nothing more than is required for conventional similarity searching, i.e., the input of a bioactive reference structure.

The assumption that high-ranked structures are likely to be active has also been used in docking studies by Klon *et al*. [24] and by Fukunishi *et al*. [25]. It is, of course, only an assumption but the general applicability of the similar property principle means that it is likely to be a true assumption more often than not, in which case increases in the effectiveness of searching (when averaged over multiple searches) can be expected from the use of approach. We emphasise averaging over multiple searches since TSS cannot be expected to be superior to conventional similarity searching for each and every one of all the searches that might be carried out (as noted by Hert et al. in their original paper [17]).

The approach suggested by Hert *et al*. involves combining the ranked outputs from the set of similarity searches (i.e., searches for the original, active reference structure and for each of the additional, assumed actives) using group fusion. The final, combined search output is then expected to yield a better level of enrichment than a conventional similarity search (which we shall normally abbreviate to SS) based on just a single reference structure. Experiments with the *MDL Drug Data Report* (MDDR) database (available from Symyx Technologies at http://www.mdli.com/products/knowledge/drug_data_report/index.jsp) yielded favourable results and it was accordingly suggested that the approach provides a simple way of enhancing the effectiveness of current systems for virtual screening [17]. Hert *et al*. subsequently described an alternative form of TSS, in which the group-fusion stage was replaced by an approximate machine-learning procedure [26]. The choice of name – turbo similarity searching – is justified by analogy with an automobile: a turbocharger increases the power of an automobile engine by using the engine's exhaust gases, and a turbo similarity search increases the power of a chemical search engine by using the reference structure's nearest neighbours.

The original TSS experiments used the MDDR database with the molecules represented by one particular type of fingerprint (specifically the ECFP_4 Pipeline Pilot fingerprints available from SciTegic Inc. at http://www.scitegic.com). Here, we consider the effectiveness of TSS when used with other databases and other types of molecular representation to determine the generality of the approach for virtual screening.

## METHODS

***Group fusion*** A schematic outline of the basic TSS procedure is shown in Figure 1. As in a conventional similarity search, the user submits a reference structure and receives as output the database ranked in decreasing similarity order; however, in TSS the system carries out some number of additional similarity searches based on the computed nearest neighbours, and then fuses the resulting similarity lists prior to the generation of the final output. An algorithmic formulation of the procedure is shown in Algorithm 1.

Based on our previous studies, the fusion was carried out using the MAX rule. Assume that some database molecule has a similarity to the reference structure (either the original reference structure, *R*, or one of its nearest neighbours) of *Sim*(*i*) in the *i*-th similarity search ($0 <= i <= k$ where $k$ is the number of nearest neighbours used). Then the fused similarity for that database molecule is

$$\text{MAX}\{Sim(i)\},$$

with the final database ranking being based on the sorting of these fused similarity values in order of decreasing numeric value.

***Machine learning*** We have described previously [26] an alternative approach to TSS, where the nearest neighbours from the basic SS search are processed using a machine-learning technique, rather than group fusion as discussed thus far. Machine learning involves the analysis of a set of molecules of known activity or inactivity (the training-set) to yield a decision rule that can then be applied to molecules of unknown activity (the test-set) [27]. Hert *et al*. suggested that the nearest neighbours of the known reference structure could comprise the actives in a training-set, with the inactives being obtained by noting that the characteristics of inactives are approximated with a high degree of accuracy by the characteristics of the entire database that is to be searched [26]. This training-set is then used to generate a ranking of the molecules in the database that is the final output of the search.

There are several different methods for machine learning that could be used to generate a ranking [27]: based on our previous experiments, we have used substructural analysis (SSA). SSA assigns a weight to each bit (or substructure) in a fingerprint that describes that bit's differential occurrence in the active and inactive molecules constituting the training-set. The resulting weights are then used to rank the test database, with the score for a molecule being the sum of the weights for its constituent fragments. The molecules at the top of the resulting ranking are then judged as having the highest probabilities of activity. The weighting scheme used was the R2 weight, which has the form

$$R2 = \log\left(\frac{A_j / N_A}{I_j / N_I}\right).$$

Here, $A_j$ and $I_j$ are the numbers of active and inactive training-set molecules with bit $j$ set, and $N_A$ and $N_I$ are the numbers of active and inactive training-set molecules [28]. The resulting procedure – referred to subsequently as TSS-SSA – is shown in Algorithm 2.

***Evaluation of searches*** There is increasing interest in criteria for the evaluation of virtual-screening experiments [29-31]. Here, we have used a very simple criterion, *viz* the recall, i.e., the percentage of the active molecules that have been retrieved at some cut-off point in the ranking. In most cases, a cut-off of 5% was used so that, for example, a recall of 20% of the actives would correspond to a four-fold enrichment of the output as compared with random screening of the database. Some of the experiments additionally used a cut-off of 1%. Some number of the molecules in an activity class were used in turn as the reference structure, and the mean search performance averaged over all of the reference molecules for the class; the final measure of search effectiveness was then obtained by averaging over the activity classes, so that each class contributed equally to the overall performance.

***Databases*** Several databases have been used in our experiments. The largest number of experiments used the MDDR database mentioned previously. This database contains the structures and pharmacological class information for molecules that have been reported in patents, journals and conference proceedings as exhibiting biological activity. The bioactivity data in MDDR is qualitative: a molecule is noted as exhibiting a specific activity, and it is assumed to be inactive if that is not the case. The version used here contained 102,514 molecules, and searches were carried out for the eleven classes of active compounds that were first described by Hert *et al*. and that have been used in several subsequent studies both by ourselves and by others [21]. These activity classes are summarized in Table 1a and are collectively referred to subsequently as MDDR-A. Some of the experiments also used a set of ten activity classes that are known to be structurally diverse and that hence provide a tougher test of a search method's ability to discriminate between active and inactive molecules [26]. These activity classes are summarized in Table 1b and are collectively referred to subsequently as MDDR-B. Each row of Table 1a or 1b contains an activity class, the number of molecules belonging to the class, and an indication of the class's diversity. The diversity figures were obtained by matching each molecule with every other in its activity class, calculating similarities using the standard Unity 2D fingerprint (available from Tripos Inc. at http://www.tripos.com) and the Tanimoto coefficient and then computing the mean intra-set similarity. It will be seen, e.g., that the renin inhibitors form the most

homogeneous class in Table 1a and the cyclooxygenase inhibitors the most heterogeneous class; this is also the case for the WOMBAT classes in Table 1c.

Experiments were also carried out using the *World Of Molecular Bioactivity* database (WOMBAT, available from Sunset Molecular Discovery LLC at http://sunsetmolecular.com/products/?id=4). This contains the structures and experimental biological activity data for molecules described in several key drug-discovery journals such as *Journal of Medicinal Chemistry*, *Bioorganic Chemistry Letters*, *European Journal of Medicinal Chemistry* etc. The bioactivity data here is quantitative: a molecule has an associated IC50 etc, and it is assumed to be inactive if that value is too low or if it is absent. The WOMBAT activity classes mirror closely the MDDR ones. Specifically we have chosen for each activity that species for which there is the largest number of molecules with a measured pIC50 >= 5.0. These molecules are marked as active for that class; molecules with pIC50 <5.0 for that species are removed from the dataset, as are all molecules with the chosen activity but tested in species other than the chosen one. The resulting dataset contained a total of 138127 molecules, with the chosen activity classes being summarized in Table 1c.

Both the MDDR and WOMBAT databases have been extensively used for virtual screening but they do have one limitation, which is that they contain molecules that have been shown to be active in some particular test, call it A. Molecules that have not been tested in A are assumed to be inactive, i.e., the coding of a molecule as inactive represents a lack of knowledge as to its activity rather than a knowledge of its inactivity. The experimental results reported here (and in the many other papers that use these two datasets) hence involve some level of false negatives.

Further experiments used the NCI AIDS database (available from http://dtp.nci.nih.gov/), which contains molecules tested in the US government's anti-AIDS programme. The version used here contained 41,192 molecules, of which the 393 confirmed active and 1037 moderately active molecules were taken as the active set that was to be retrieved, i.e., in this database, the inactives are known explicitly and there is no counting of false negatives in the results.

Finally, experiments were carried out to assess the effectiveness of TSS for quantitative property prediction. The datasets studied here were: a set of 829 molecules with IC50 values provided by Janssen Pharmaceutica from a GPCR research programme; a set of 880 CDK2 inhibitors with IC50 values [32]; a set of 762 HIV protease inhibitors with Ki values [32]; a

set of 4449 molecules with melting points [33]; and a set of 927 molecules with aqueous solubilities [34].

***Structure representations*** There are many ways in which the structure of a chemical molecule can be represented in machine-readable form.

The most common type of representation for virtual screening is a fingerprint, i.e., a binary vector in which bits are set to denote the presence of substructural fragments in a molecule, and fingerprints were used in the majority of our experiments. Two main approaches have been developed for selecting the fragments that are encoded in a fingerprint [2, 3, 35]. In a *dictionary-based* approach, there is a pre-defined list of fragments, with normally one fragment allocated to each position in the bit-string. Here, a molecule is checked for the presence of each of the fragments in the dictionary, and a bit set (or not set) when a fragment is present (or absent). In a *molecule-based* approach, hashing algorithms are used to allocate multiple fragments to each bit-position. Here, a generic fragment type is specified, e.g., a chain of four connected non-hydrogen atoms, and a note made of all fragments of that type that occur in a given molecule. Each fragment is converted to a canonical form and then hashed using several (typically two or three) hashing algorithms to set bits in the fingerprint.

The original TSS studies [17, 26] used the ECFP_4 (for Extended Connectivity Fingerprint encoding circular substructures of diameter four bonds) fingerprints from the SciTegic Pipeline Pilot software (hashed to a fixed length of 1024 bits) and these fingerprints have also been employed here. In addition, we have used several other types of 2D fingerprints: SciTegic FCFP_4 (for Functional-Class Fingerprint encoding circular substructures of diameter four bonds) fingerprints (1024 bits), Tripos Unity fingerprints (988 bits), Digital Chemistry (formerly Barnard Chemical Information Limited, BCI) fingerprints (1052 bits, available from http://www.digitalchemistry.co.uk), Daylight fingerprints (2048 bits, available from http://www.daylight.com) and MDL key fingerprints (166 bits, available from http:/www.mdli.com). Of these, the BCI and MDL fingerprints are dictionary-based, the Daylight and SciTegic fingerprints are molecule-based (using linear chains and circular substructures, respectively), and the Unity fingerprints are based on both approaches, so as to cover the full range of commonly-available types of fingerprint. Details of these types of 2D fingerprint are provided by Leach and Gillet [3] while Hert *et al*. [36] discuss the use of these, and other types of fingerprints, for ligand-based virtual screening.

The MDDR and NCI searches were also carried out using a very different type of fingerprint generated from a 3D structure and encoding geometric pharmacophores. These PDT (for

pharmacophore distance triplet) fingerprints were obtained by computing CONCORD structures, generating 100 conformations for each such structure, and then calculating the distances between hydrogen-bond acceptors, hydrogen-bond donors and hydrophobes. The distances were binned from 3Å to 15Å at 1.5Å intervals, and the resulting 91,125-element bit-string folded to a length of 10,000 bits for searching. Similarities between pairs of fingerprints (of whatever kind) were computed using the Tanimoto coefficient.

Fingerprints provide a binary (presence/absence) representation of a chemical molecule. However, we have also studied three non-binary representations: 12 physicochemical properties (e.g., AlogP, logD, molecular weight, volume, and solubility) generated using the Pipeline Pilot software (available from SciTegic Inc.); 523 topological indices (e.g., molecular connectivity, kappa shape, and electrotopological state indices) generated using the Molconn-Z software (available from eduSoft at http://www.eslc.vabiotech.com); and 997-element molecular holograms generated using the Unity software (available from Tripos Inc.) with the default parameter settings, where the holograms denote not just the presence but also the frequency of occurrence of a substructural fragment in a molecule. The Molconn-Z and physicochemical data were processed using a Principal Components Analysis routine. Similarities between pairs of non-binary representations were computed using the Euclidean distance.

Finally, the last few years have seen interest in the use of *reduced graphs* for similarity searching. A reduced graph provides a summary representation of a molecule, with groups of connected atoms of the same type being conflated to single graph-nodes, the types of which are chosen to reflect characteristics that are likely to be of importance for biological activity, e.g., ring systems, charged groups and hydrogen donors and acceptors. There are many ways in which graph reduction can be carried out: here, we have used the procedures described by Barker *et al*. [37] in which all acidic and basic groups are first identified and in which the molecule is then partitioned into cyclic and acyclic fragments. Ring nodes may be aromatic or alicyclic, and are further sub-divided by their donor/acceptor characteristics (or lack thereof), as are acyclic feature nodes (with acyclic nodes having no donor/acceptor nature being referred to as linkers). The similarity between a pair of molecules is then computed using a Tanimoto-like similarity coefficient based on the maximum common subgraph (MCS) between the reduced graphs corresponding to those molecules, with the MCS being computed using the Bron-Kerbosch clique-detection algorithm [37]. The study of Barker *et al.* considered several ways in which reduced graphs could be used for virtual screening and we have used here that approach which performed best in their experiments (which they refer to as MCIS-FC).

## RESULTS AND DISCUSSION

***Initial fingerprint searches*** The first set of runs used the BCI, Daylight, ECFP_4, FCFP_4, MDL, PDT and Unity fingerprints in searches of the MDDR and NCI databases. In each set of searches, every active molecule in turn was used as the reference structure: 8294 in the case of MDDR-A, 8440 in the case of the diverse MDDR-B and 393 (the confirmed actives) in the case of the NCI database. As noted previously, the results were averaged over all the molecules within each activity class and then over all the classes. The results of the runs are shown in Tables 2-4, where SS denotes a conventional similarity search and where TSS-*x* denotes a turbo similarity search based on the original reference structure and the *x* nearest neighbours of that reference structure. The figures listed in these, and subsequent, tables are the mean percentages of the actives retrieved in the search.

Inspection of the MDDR searches in Tables 2 and 3 shows that the results obtained are analogous to those reported previously for the ECFP_4 fingerprints. Specifically, when the MDDR-A classes are used (Table 2) there is often a noticeable increase in the recall of the search, especially for the ECFP_4 fingerprints, as more nearest neighbours are included in a TSS, with the maximum recall typically being obtained with 50-100 nearest neighbours. The MDL and PDT fingerprints are different in behaviour, with SS consistently superior to TSS. Of the various types of fingerprint, ECFP_4 gives the best results, both in the initial SS and in the degree of enhancement when TSS is used: for this fingerprint, the maximum TSS recall corresponds to an increase of 15.1% of the recall of the conventional SS. Table 3 shows the results for MDDR-B. Here, the degree of enhancement is much less notable, even for ECFP_4, and for most of the fingerprints there would appear to be no advantage in using TSS. Similar comments apply to the NCI AIDS searches summarized in Table 4; indeed, here there is a noticeable decrease for the 3D PDT fingerprints.

The original TSS paper [17] used the ECFP_4 fingerprints on the MDDR database and the MDDR-A activity classes, and demonstrated that significant increases in recall could be achieved. The subsequent TSS paper [26] showed that such increases were not observed when the activity classes were chosen to be as diverse as possible, in which case the basic SS search is poor: the results in Table 3 show that this is also the case for the other types of fingerprint studied here. Indeed, for several of these it is not possible to obtain substantial improvements even with the MDDR-A classes in Table 2 (where the ECFP_4 fingerprints gave consistently the best performance). A reasonable conclusion from these observations would hence be that the best results will be obtained if the starting point for the TSS, i.e., the

basic SS search, is effective: this conclusion is supported by the results in Table 4 for the NCI database, where none of the SS or TSS searches are particularly effective.

Turning now to the WOMBAT searches, ten molecules were chosen at random from each activity class to be the reference structures for searching. The results of the WOMBAT searches using the top-5% are detailed in Table 5. Inspection of these figures shows that the effectiveness of TSS mirrors that observed in Table 2: there is a substantial increase in the effectiveness of the ECFP_4 searches (an increase of over 10% of the SS recall) and (to a lesser extent) the FCFP_4 searches, but little or no benefit resulting from the use of TSS with the other fingerprint-types. A similar pattern of behaviour is evident if the evaluation focuses on just the top-1% of the ranked outputs, as shown in Table 6.

Taken together, the results in Tables 2-6 suggest that TSS can bring about substantial enhancements in virtual-screening performance in some cases. However, the overall picture is rather less favourable to the approach than the initial results that were obtained using the ECFP_4 fingerprints on the MDDR-A dataset.

***Non-fingerprint searches*** The reduced graph searches used a subset of the MDDR database obtained from the application of filters based on molecular properties such as molecular weight, logP, and number of rotatable bonds, and on SMARTS rules obtained from a survey of medicinal chemists at AstraZeneca [38]. In all, this dataset contained 61902 molecules, of which there were 4713 actives across the MDDR-A activity classes. The results for the TSS searches are listed in Table 7, which again shows the effectiveness of TSS, with the best results being obtained from including 20-50 molecules in the second stage of the search. Thus the TSS-30 figures represent increases of 14.4% and 9.2% over the SS figures for the cut-offs of 1% and 5%, respectively, providing further evidence of the effectiveness of TSS when used on the MDDR-A database.

Given the generally positive results obtained thus far using MDDR-A, the results in Table 8 for the hologram, physicochemical property and Molconn-Z searches on this dataset came as a marked surprise. Specifically, it will be seen that little or no advantage accrued from the use of TSS, and that in many cases the recall decreased when compared with the basic SS. Similar negative comments apply to the diverse, MDDR-B results in Table 9.

We have probed this poor performance using a lower-bound and upper-bound analysis of the hologram and ECFP_4 searches that is reported in Table 10. The results here were obtained

by carrying out a TSS search based on fusing the rankings generated with either the top-ranked 100 inactive nearest neighbours as the reference structures (lower-bound) or the top-ranked 100 active nearest neighbours as the reference structures (upper-bound); full details of these bounding procedures are provided by Hert *et al*. [17]. Inspection of the fingerprint results, which are based on the use of the ECFP_4 fingerprints, in Table 10 shows that the lower-bound searches (with an average recall of 38.7%) are only slightly worse than the basic SS search (with an average recall of 39.2%). This may appear rather surprising but, as Hert *et al*. note, this simply means that even when inactive molecules are used in TSS, these nearest-neighbour molecules still contain sufficient relevant substructures in common with the reference structure to enable the identification of further active molecules. However, this will only be so if the similarity property principle applies. Whilst this would indeed appear to be the case for the ECFP_4 fingerprints, it is certainly not the case for the holograms, where the lower-bound recall (at 16.0%) is much worse than the basic SS search (with an average recall of 26.0%). If we now consider the upper-bound results, the ECFP_4 searches demonstrate clearly the performance gains that can be achieved when the principle applies. Here, the average recall increases from 39.2 (for SS) to 68.4 (for TSS), a rise of 74.5% of the SS recall. However, the increase is far less noticeable for the histograms: from 26.0 (for SS) to 33.9 (for TSS), a rise of 30.4% of the SS recall.

The original TSS paper used ECFP_4 fingerprints and the Tanimoto coefficient and concluded by noting that "...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". It is clear that the principle does not hold, or at least does not hold sufficiently well, for the holograms that were used, with consequent poor TSS performance. We presume that the Principle also does not hold for the two other non-binary representations (and for some of the binary fingerprints) used here.

***Use of machine learning*** All of the experiments thus far have used the original form of TSS, where the nearest neighbours of the original reference structure are used as reference structures in their own right as reference structures, and then the multiple rankings combined using group fusion. As noted previously, we have also described an alternative form of TSS [26], in which the group-fusion stage is replaced by a machine-learning procedure, specifically substructural analysis (SSA) in the experiments reported here (see Algorithm 2). The results of using this approach, referred to as TSS-SSA, on the MDDR-A, MDDR-B and NCI datasets are shown in Tables 11-13, which are thus comparable to the results in Tables 2-4 for "normal" TSS.

Tables 2 and 3 have demonstrated that TSS can result in substantial increases in recall when applied to the eleven activity classes in MDDR-A, but have little or no effect when applied to the diverse activity classes in MDDR-B. Tables 11 and 12 provide a dramatic contrast, with TSS-SSA behaving in exactly the opposite way, as there are substantial increases in recall for MDDR-B but no effect (or even reductions in recall) for MDDR-A. This behaviour had been observed previously using just the ECFP_4 fingerprints but is clearly more general in scope, with some of the effects being extremely large: the MDL TSS-SSA-10 recall with MDDR-A (21.9%) is 27.5% less than the SS recall (30.2%); whereas the Unity TSS-SSA-10 recall with MDDR-B (25.1%) is 51.2% more than the SS recall (16.6%). A comparison of Tables 4 and 13 show the differences for the NCI dataset: the former reveals little or no advantage from the use of TSS whereas the latter reveals that three of the fingerprints - ECFP-4, FCFP_4 and PDT - benefit greatly from the use of TSS-SSA. In summary, then, Tables 11-13 support our previous conclusions [26] in recommending the use of machine learning, rather than group fusion, for TSS when structurally diverse sets of active molecules need to be investigated. One final point to note is that there are some inconsistencies in the numbers of nearest neighbours required for maximum TSS performance: in particular, compare the TSS-SSA-10 and TSS-SSA-20 recalls for ECFP_4 and PDT in Table 12 as against those in Table 13.

***Overall effectiveness of fingerprint-based TSS*** The principal objective of the work reported in this paper was to assess the general effectiveness of TSS as a mechanism for virtual screening. However, the use of several different types of fingerprint has also enabled us to draw conclusions as to their relative performance.

The results in Tables 2-6 and 11-13 demonstrate very clearly the consistently higher level of recall achieved using the ECFP_4 fingerprints. Previous comparative studies have demonstrated their merits for group fusion and conventional SS searches [36], which was why we chose this type of fingerprint as the basis for our initial studies of TSS [17, 26]. The results obtained here show the (in retrospect) wisdom of this choice, since no other fingerprint (with the possible exception of the closely related FCFP_4 fingerprint) responds anywhere near as favourably as does ECFP_4 to the use of TSS. In addition, since our work was completed, a very recent comparative study has demonstrated the merits of the ECFP_4 fingerprint for establishing the similarity of drug targets [39].

The benefits that can be achieved from using TSS in ECFP_4-based similarity searching have been quantified by calculating the percentage increase over the SS recall for the TSS-*x* search with the highest recall (normally, but not consistently, TSS-100 or TSS-200). These increases are: 15.1, 7.7, 6.7, 10.4 and 14.2% for Tables 2-6 respectively, and 7.7, 36.4 and

35.2% for Tables 11-13 respectively. Thus, for this particularly effective fingerprint, it seems that a significant increase in the performance of similarity searching can be achieved by the use of the TSS approach; moreover, as noted in our previous papers and as is made clear from Figure 1, this improved performance is achieved without any effort on the part of the chemist carrying out the database search.

In concluding this section, although our experiments have not considered every type of fingerprint available, we believe that our findings are of general applicability. This is because the 2D fingerprints studied here include fingerprints based on the use of a fragment dictionary, based on hash coding the fragments in a molecule, and based on both types of approach; indeed, they include many of the fingerprints in current operational chemoinformatics systems.

***Quantitative property prediction*** All of the experiments thus far have used qualitative, i.e., active/inactive, bioactivity data for virtual screening. The final series of experiments sought to evaluate the use of TSS for the prediction of quantitative property values in the five datasets mentioned in the Methods section. The molecules here were represented by ECFP_4 and FCFP_4 fingerprints (with the exception of the Janssen corporate dataset, which was based on ECFP_6 fingerprints).

The basic idea in structure-based property prediction is a leave-one-out procedure in which the property value is assumed to be unknown for each of the dataset-molecules in turn. The predicted property value for each such molecule *X*, *P(X)*, is then taken to be the arithmetic mean of the observed property values of some number, *p*, of its nearest neighbours, i.e., those molecules that are structurally most similar to it. This procedure results in the calculation of a *P(X)* value for each of the *N* structures in a dataset, and an overall figure of merit is then obtained by calculating the product moment correlation coefficient between the sets of *N* observed and *N* predicted values. Extension of this SS-based method for property prediction simply involves fusing *k* nearest neighbour lists (as detailed in Algorithm 1) and then using the *p* nearest neighbours from the final ranking for computation of the *P(X)* value. Thus, *k* is the number of nearest neighbours that are used to produce the combined ranking that is the output from the TSS and *p* is the number of nearest neighbours from that combined ranking that are used to calculate the property value. Experiments were carried out with *p* set to 1, 2, 3, 5 and 10, and with *k* set to 0, 1, 2, 3, 5, and 10.

A typical set of results, for the aqueous solubility dataset, is shown in Table 14. The table lists the squared correlation coefficient values ($r^2$) for the correlation between the sets of

observed and predicted solubilities. The row with $k$=0 corresponds to the use of normal SS for property prediction, and it will be seen that the best correlation between observed and predicted solubilities is obtained using the three nearest neighbours (as measured in this case using the ECFP_4 fingerprints). A comparison of the values in this row of the table with the corresponding values in the rows for which $k$>0 shows that TSS brings about a consistent decrease in the correlations. Similar results were obtained for all of the other quantitative datasets studied here, and this was also the case when the FCFP_4 fingerprints were used or when a different fusion rule (the SUM rule [21]) was used. We hence conclude that TSS is not an appropriate tool for the prediction of quantitative property values.

## CONCLUSIONS

The computation of inter-molecular structural similarity is a vital component of modern approaches to virtual screening. Chemical similarity searching has normally involved the use of just a single bioactive reference structure to screen a database for molecules that have a high *a priori* probability of being active. In this paper, we have reported a detailed evaluation of an alternative approach (turbo similarity searching, or TSS) that additionally makes use of the reference structure's nearest neighbours, i.e., those that are structurally most similar to it, in the screening stage.

We have shown here that TSS based on group fusion can provide substantial enhancements in screening performance if the normal similarity search provides a good starting point, i.e., if the similar property principle holds and if the actives are well clustered using the chosen structure representation and similarity measure. This was particularly the case in the searches here that were based on the ECFP_4 fingerprints. If this is not the case (e.g., the MDDR-B and NCI AIDS searches), then an alternative approach to TSS based on an approximate form of substructural analysis (TSS-SSA) can provide enhancements in screening performance. That said, it must be emphasised that we have not been able to look at all possible combinations of dataset and structural representation, and even the results that we have obtained do exhibit some minor inconsistencies. There is no such inconsistency in the prediction experiments where TSS appears to be unsuited to the prediction of quantitative property values.

In more detail, our conclusions are as follows. First, the ECFP_4 fingerprints would appear to be the structure representation of choice for similarity-based virtual screening, whether using SS or TSS. Second, If the actives are indeed tightly grouped then TSS is likely to

provide a level of screening notably greater than does SS; if they are not tightly grouped then there is unlikely to be any difference in screening performance screening. Third, the use of TSS-SSA provides an alternative path to the identification of molecules for testing: here, the approach is effective for heterogeneous sets of actives, without notably decreasing performance when they are homogeneous. In practice, of course, one does not know the nature of the actives, and this hence suggests that both TSS and TSS-SSA searches should be carried out when identifying molecules for testing. Indeed, we hope in the future to study how SS, TSS and TSS-SSA outputs can best be combined to give a single output to the user who has provided the original reference structure.

Having established the general effectiveness of the TSS approach, our current work seeks to establish the best way of using the nearest-neighbour information in the second-stage of the search. Three developments are currently being studied. First, as noted in the previous paragraph we intend to investigate the combination of SS, TSS and TSS-SSA search outputs. Second, using the group fusion approach there are many different fusion rules that can be used to combine multiple rankings of a database and some of these may be superior to the MAX rule used thus far [40]. Third, rather than using, e.g., the 100 nearest neighbours for the reference structure; can better results – in particular for scaffold-hopping applications - be obtained by taking, e.g., just the ten nearest neighbours and then identifying the ten nearest neighbours for each of these? Other ways of using the reference structure's nearest neighbours include cluster-based approaches and the techniques for virtual screening described recently by Wale et al. [41]. This work will be reported shortly.

We conclude by noting the significance of the work reported here for the implementation of operational systems for similarity-based virtual screening. Such systems require the user to submit a reference structure, normally a known active, in response to which the system returns a list of nearest neighbours that are expected also to be active. TSS again requires only the submission of the reference structure but, as our results show, returns a list of nearest neighbours that is often richer in actives than is obtained using conventional similarity searching, particularly when the common circular substructure fingerprints are used. TSS hence results in an overall increase in effectiveness (as determined by the number of retrieved actives) without any decrease in efficiency (as determined by the degree of user effort); we hence believe that it provides an attractive tool for the implementation of ligand-based virtual screening.

**Acknowledgements**. We thank the following: Kristian Birchall for assistance with the preparation of the WOMBAT data; David Cosgrove for assistance with the reduced graph

searches; AstraZeneca, the British Council and the Polish Ministry of Science and Education Young Scientists Programme (WAR/342/86), the Government of Malaysia, Johnson and Johnson Inc., and the Novartis Institutes for Biomedical Research for funding; and Daylight Chemical Information Systems Inc., Digital Chemistry Limited, MDL Information Systems Inc., the Royal Society, SciTegic Inc., Sunset Molecular Discovery LLC, Tripos Inc. and the Wolfson Foundation for data, software and laboratory support.

REFERENCES

[1] P. Willett, "From chemical documentation to chemoinformatics: fifty years of chemical information science," *J. Inf. Sci.*, vol. 34, pp. 477-499, 2008.

[2] J. Gasteiger and T. Engel (eds.), *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH, 2000.

[3] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Second edition. Dordrecht: Kluwer, 2007.

[4] H.-J. Böhm and G. Schneider (eds.), *Virtual Screening for Bioactive Molecules*. Weinheim: Wiley-VCH, 2000.

[5] G. Klebe (ed.), *Virtual Screening: an Alternative or Complement to High Throughput Screening*. Dordrecht: Kluwer, 2000.

[6] F. L. Stahura and J. Bajorath, "Virtual screening methods that complement high-throughput screening," *Combin. Chem. High-Through. Screen.*, vol. 7, pp. 259-269, 2004.

[7] J. Alvarez and B. Shoichet (eds.), *Virtual Screening in Drug Discovery*. Boca Raton: CRC Press, 2005.

[8] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discov. Today*, vol. 11, pp. 1046-1053, 2006.

[9] H. Eckert and J. Bajorath, "Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches," *Drug Discov. Today*, vol. 12, pp. 225-233, 2007.

[10] P. Willett, "Similarity methods in chemoinformatics," *Ann. Rev. Inf. Sci. Tech.*, vol. 43, 3-71, 2009.

[11] P. Willett, J. M. Barnard and G. M. Downs, "Chemical similarity searching," *J. Chem. Inf. Comput. Sci.*, vol. 38, pp. 983-996, 1998.

[12] R. P. Sheridan and S. K. Kearsley, "Why do we need so many chemical similarity search methods?," *Drug Discov. Today*, vol. 7, pp. 903-911, 2002.

[13] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity - a review," *QSAR Combin. Sci.*, vol. 22, pp. 1006-1026, 2003.

[14] A. G. Maldonado, J. P. Doucet, M. Petitjean and B.-T. Fan, "Molecular similarity and diversity in chemoinformatics: from theory to applications," *Mol. Diversity*, vol. 10, pp. 39-79, 2006.

[15] R. C. Glen and S. E. Adams, "Similarity metrics and descriptor spaces - which combinations to choose?," *QSAR Combin. Sci.*, vol. 25, pp. 1133-1142, 2006.

[16] R. P. Sheridan, "Chemical similarity searches: when is complexity justified?," *Expert Opin. Drug Discov.*, vol. 2, pp. 423-430, 2007.

[17] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, "Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information," *J. Med. Chem.*, vol. 48, pp. 7049-7054, 2005.

[18] M. A. Johnson and G. M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity*. New York: John Wiley, 1990.

[19] Y. C. Martin, J. L. Kofron and L. M. Traphagen, "Do structurally similar molecules have similar biological activities?," *J. Med. Chem.*, vol. 45, pp. 4350-4358, 2002.

[20] M. Whittle, V. J. Gillet, P. Willett, A. Alex and J. Loesel, "Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: A comparison of similarity coefficients," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 1840-1848, 2004.

[21] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, "Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 1177-1185, 2004.

[22] C. Williams, "Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance," *Mol. Diversity*, vol. 10, pp. 311-332, 2006.

[23] Q. Zhang and I. Muegge, "Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring " *J. Med. Chem.*, vol. 49, pp. 1536-1548, 2006.

[24] A. E. Klon, M. Glick, M. Thoma, P. Acklin and J. W. Davies, "Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results," *J. Med. Chem.*, vol. 47, pp. 2743-2749, 2004.

[25] H. Fukunishi, R. Teramoto and J. Shimada, "Hidden active information in a random compound library: extraction using a pseudo-structure-activity relationship model," *J. Chem. Inf. Model.*, vol. 48, pp. 575-582, 2008.

[26] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, " New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching," *J. Chem. Inf. Comput. Sci.*, vol. 46, pp. 462-470, 2006.

[27] B. B. Goldman and W. P. Walters, "Machine learning in computational chemistry," *Ann. Reports Comput. Chem.*, vol. 2, pp. 127-140, 2006.

[28] A. Ormerod, P. Willett and D. Bawden, "Comparison of fragment weighting schemes for substructural analysis," *Quant. Struct.-Activ. Relat.*, vol. 8, pp. 115-129, 1989.

[29] S. J. Edgar, J. D. Holliday and P. Willett, "Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures," *J. Mol. Graph. Mode.*, vol. 18, pp. 343-357, 2000.

[30] J.-F. Truchon and C. I. Bayly, "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem," *J. Chem. Inf. Model.*, vol. 47, pp. 488-508, 2007.

[31] A. N. Jain and A. Nicholls, "Recommendations for evaluation of computational methods," *J. Comput.-Aided Mol. Design*, vol. 22, pp. 133-139, 2008.

[32] T. Liu, Y. Lin, X. Wen, R. N. Jorrisen and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand affinities," *Nucleic Acids Res.*, vol. 35, pp. D198-D201, 2006.

[33] M. Karthikeyan, R. C. Glen and A. Bender, "General melting point prediction based on a diverse compound data set and artificial neural networks," *J. Chem. Inf. Model.*, vol. 45, pp. 581-590, 2005.

[34] J. Delaney, "ESOL: estimating aqueous solubility directly from molecular structure," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 1000-1005, 2004.

[35] J. Gasteiger (ed.), *Handbook of Chemoinformatics*. Weinheim: Wiley-VCH, 2003.

[36] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, "Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures," *Org. Biomol. Chem.*, vol. 2, pp. 3256-3266, 2004.

[37] E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, V. J. Gillet, P. Kitts and P. Willett, "Scaffold-hopping using clique detection applied to reduced graphs," *J. Chem. Inf. Model.*, vol. 46, pp. 503-511, 2006.

[38] A. M. Davis, D. J. Keeling, J. Steele, N. P. Tomkinson and A. C. Tinker, "Components of successful lead generation," *Curr. Topics Med. Chem.*, vol. 5, pp. 421-439, 2005.

[39] J. Hert, M. J. Kaiser, J. J. Irwin, T. I. Oprea and B. K. Shoichet, "Quantifying the relationships among drug classes," *J. Chem. Inf. Model*., vol 48, pp. 755-765, 2008.

[40] P. Willett, "Data fusion in ligand-based virtual screening," *QSAR Combin Sci*., vol. 25, pp. 1143-1152, 2006.

[41] N. Wale, I. A. Watson and G. Karypis, "Indirect similarity methods for effective scaffold-hopping in chemical compounds," J, Chem. Inf. Model., vol. 48, pp. 730-741, 2008.

| Activity class | Active molecules | Mean pairwise similarity |
|---|---|---|
| 5HT3 antagonists | 752 | 0.351 |
| 5HT1A agonists | 827 | 0.343 |
| 5HT reuptake inhibitors | 359 | 0.345 |
| D2 antagonists | 395 | 0.345 |
| Renin inhibitors | 1130 | 0.573 |
| Angiotensin II AT1 antagonists | 943 | 0.403 |
| Thrombin inhibitors | 803 | 0.419 |
| Substance P antagonists | 1246 | 0.399 |
| HIV protease inhibitors | 750 | 0.446 |
| Cyclooxygenase inhibitors | 636 | 0.268 |
| Protein kinase C inhibitors | 453 | 0.323 |

(a)

| Activity class | Active molecules | Mean pairwise similarity |
|---|---|---|
| Muscarinic (M1) agonists | 848 | 0.206 |
| NMDA receptor antagonists | 1311 | 0.199 |
| Nitric oxide synthase inhibitors | 377 | 0.189 |
| Dopamine beta-hydroxylase inhibitors | 95 | 0.229 |
| Aldose reductase inhibitors | 882 | 0.232 |
| Reverse transcriptase inhibitors | 519 | 0.218 |
| Aromatase inhibitors | 513 | 0.229 |
| Cyclooxygenase inhibitors | 636 | 0.220 |
| Phospholipase A2 inhibitors | 704 | 0.224 |
| Lipoxygenase inhibitors | 2555 | 0.224 |

(b)

| Activity class (species) | Active molecules | Mean pairwise similarity |
|---|---|---|
| 5HT3 antagonists (rat) | 220 | 0.377 |
| 5HT1A antagonists (rat) | 592 | 0.399 |
| D2 antagonists (rat) | 910 | 0.367 |
| Renin inhibitors (human) | 474 | 0.592 |
| Angiotensin II AT1 antagonists (rat) | 724 | 0.443 |
| Thrombin inhibitors (human) | 421 | 0.418 |
| Substance P antagonists (human) | 558 | 0.427 |
| HIV protease inhibitors (human) | 1128 | 0.442 |
| Cyclooxygenase inhibitors (human) | 965 | 0.324 |
| Protein kinase C inhibitors (rat) | 142 | 0.565 |
| Acetylcholine esterase inhibitors (human) | 503 | 0.373 |
| Factor Xa inhibitors (human) | 842 | 0.394 |
| Matrix metalloprotease inhibitors (human) | 694 | 0.444 |
| Phosphodiesterase inhibitors (human) | 596 | 0.359 |

(c)

Table 1. Activity classes used in the virtual screening experiments (a) MDDR-A activity classes, (b) MDDR-B activity classes, (c) WOMBAT classes.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 32.8 | 33.8 | 34.2 | 34.7 | 34.9 | 34.8 |
| Daylight | 31.5 | 32.4 | 32.6 | 33.1 | 32.8 | 32.6 |
| ECFP_4 | 39.2 | 41.9 | 42.9 | 44.5 | 45.1 | 45.1 |
| FCFP_4 | 36.1 | 37.9 | 38.9 | 40.1 | 40.8 | 41.1 |
| MDL | 30.2 | 27.9 | 28.0 | 28.1 | 28.2 | 28.1 |
| PDT | 18.8 | 17.9 | 17.4 | 17.0 | 16.7 | 16.6 |
| Unity | 30.2 | 30.8 | 30.9 | 31.0 | 31.1 | 30.8 |

Table 2. SS and TSS searches of MDDR-A at 5% cut-off using fingerprints.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 20.7 | 20.9 | 20.6 | 20.2 | 19.6 | 19.6 |
| Daylight | 18.3 | 18.0 | 17.4 | 16.7 | 16.4 | 16.0 |
| ECFP_4 | 20.9 | 22.3 | 22.5 | 22.5 | 22.0 | 21.5 |
| FCFP_4 | 20.2 | 21.1 | 21.1 | 20.7 | 20.1 | 19.6 |
| MDL | 20.0 | 20.0 | 19.5 | 18.9 | 18.3 | 17.8 |
| PDT | 16.6 | 16.7 | 15.8 | 15.4 | 15.2 | 15.4 |
| Unity | 16.6 | 15.8 | 15.2 | 14.1 | 13.8 | 13.7 |

Table 3. SS and TSS searches of MDDR-B at 5% cut-off using fingerprints.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 12.1 | 12.3 | 12.3 | 12.5 | 12.8 | 12.9 |
| Daylight | 10.4 | 10.5 | 10.4 | 10.2 | 10.0 | 10.2 |
| ECFP_4 | 10.5 | 10.3 | 10.3 | 10.4 | 10.7 | 11.2 |
| FCFP_4 | 10.8 | 10.9 | 11.1 | 11.1 | 11.1 | 11.1 |
| MDL Keys | 11.9 | 11.9 | 12.0 | 12.1 | 12.3 | 12.4 |
| PDT | 13.8 | 10.3 | 10.1 | 10.1 | 10.1 | 10.3 |
| Unity | 11.5 | 11.5 | 11.6 | 11.8 | 11.7 | 12.0 |

Table 4. SS and TSS searches of NCI AIDS at 5% cut-off using fingerprints. The listed figures are the mean percentage of actives retrieved for the confirmed active reference structures.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 39.0 | 39.6 | 39.8 | 40.0 | 40.0 | 39.6 |
| Daylight | 35.1 | 35.9 | 36.0 | 35.6 | 36.2 | 35.7 |
| ECFP_4 | 47.2 | 48.6 | 49.5 | 50.6 | 51.9 | 52.1 |
| FCFP_4 | 42.2 | 43.0 | 43.9 | 44.7 | 45.1 | 45.6 |
| MDL Keys | 36.6 | 37.1 | 37.1 | 37.2 | 36.9 | 37.0 |
| Unity | 36.8 | 37.3 | 37.8 | 37.5 | 37.4 | 37.5 |

Table 5. SS and TSS searches of WOMBAT at 5% cut-off using fingerprints.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 23.6 | 23.8 | 24.1 | 24.4 | 24.8 | 24.2 |
| Daylight | 22.9 | 22.8 | 22.7 | 23.0 | 23.9 | 23.7 |
| ECFP_4 | 31.6 | 32.5 | 33.8 | 34.9 | 36.1 | 35.1 |
| FCFP_4 | 26.9 | 27.9 | 28.6 | 29.6 | 30.1 | 29.4 |
| MDL Keys | 21.3 | 22.0 | 22.2 | 22.6 | 22.4 | 22.5 |
| Unity | 22.7 | 22.8 | 23.3 | 23.5 | 23.2 | 23.1 |

Table 6. SS and TSS searches of WOMBAT at 1% cut-off using fingerprints.

| Cut-off | SS | TSS-10 | TSS-20 | TSS-30 | TSS-50 | TSS-100 |
|---|---|---|---|---|---|---|
| 1% | 22.1 | 24.9 | 25.1 | 25.3 | 25.1 | 24.6 |
| 5% | 38.1 | 40.9 | 41.3 | 41.6 | 41.1 | 39.7 |

Table 7. SS and TSS searches of a filtered version of MDDR-A at 1% and 5% cut-offs using reduced graphs.

| Representation | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 |
|---|---|---|---|---|---|
| Holograms | 26.0 | 23.7 | 22.7 | 21.9 | 21.2 |
| Molconn-Z | 18.5 | 18.7 | 18.6 | 17.9 | 17.5 |
| Properties | 24.1 | 24.4 | 24.3 | 23.9 | 23.5 |

Table 8. SS and TSS searches of MDDR-A at 5% cut-off using non-binary representations.

| Representation | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 |
|---|---|---|---|---|---|
| Holograms | 24.6 | 24.7 | 24.4 | 21.9 | 21.2 |
| Molconn-Z | 15.6 | 15.3 | 15.1 | 14.2 | 13.5 |
| Properties | 24.6 | 25.2 | 24.9 | 24.5 | 24.2 |

Table 9. SS and TSS searches of MDDR-B at 5% cut-off using non-binary representations.

| Activity class | Top-100 inactive nearest neighbours | | Top-100 active nearest neighbours | |
|---|---|---|---|---|
| | Holograms | Fingerprints | Holograms | Fingerprints |
| 5HT3 antagonists | 13.5 | 32.1 | 31.2 | 65.7 |
| 5HT1A agonists | 10.5 | 31.9 | 23.2 | 55.3 |
| 5HT reuptake inhibitors | 11.1 | 21.7 | 32.1 | 62.8 |
| D2 antagonists | 10.8 | 28.8 | 27.9 | 68.6 |
| Renin inhibitors | 40.1 | 89.8 | 77.4 | 96.6 |
| Angiotensin II AT1 antagonists | 40.7 | 92.2 | 52.1 | 95.2 |
| Thrombin inhibitors | 13.5 | 33.9 | 32.8 | 71.6 |
| Substance P antagonists | 7.5 | 15.8 | 22.3 | 53.8 |
| HIV protease inhibitors | 10.0 | 49.0 | 25.0 | 76.1 |
| Cyclooxygenase inhibitors | 8.8 | 12.0 | 22.4 | 49.2 |
| Protein kinase C inhibitors | 9.6 | 18.3 | 26.1 | 58.1 |
| Average over all classes | 16.0 | 38.7 | 33.9 | 68.4 |
| Average over all classes for SS | 26.0 | 39.2 | 26.0 | 39.2 |

Table 10**.** TSS lower-bounds and upper-bounds for recall of MDDR-A at 5% cut-off using ECFP_4 fingerprints and using molecular holograms. The bottom row contains the basic SS values for comparison purposes.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 32.8 | 29.2 | 28.5 | 28.2 | 28.5 | 28.8 |
| Daylight | 31.5 | 27.1 | 24.2 | 23.3 | 23.9 | 24.9 |
| ECFP_4 | 39.1 | 37.1 | 40.2 | 40.0 | 40.9 | 42.1 |
| FCFP_4 | 36.1 | 32.5 | 35.9 | 36.6 | 37.5 | 38.3 |
| MDL | 30.2 | 21.9 | 21.8 | 21.9 | 22.1 | 22.4 |
| PDT | 18.8 | 14.3 | 15.6 | 14.5 | 14.3 | 14.3 |
| Unity | 30.2 | 24.7 | 23.3 | 23.3 | 24.0 | 24.6 |

Table 11. SS and TSS-SSA searches of MDDR-A at 5% cut-off using fingerprints.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 20.7 | 27.1 | 27.1 | 26.0 | 24.8 | 23.8 |
| Daylight | 18.3 | 25.0 | 23.3 | 21.7 | 21.0 | 20.4 |
| ECFP_4 | 20.9 | 21.5 | 28.5 | 28.8 | 27.9 | 26.7 |
| FCFP_4 | 20.2 | 18.3 | 24.0 | 25.9 | 25.5 | 24.5 |
| MDL | 20.2 | 26.5 | 25.5 | 24.2 | 23.4 | 22.8 |
| PDT | 16.6 | 15.7 | 18.7 | 18.1 | 17.9 | 17.6 |
| Unity | 16.6 | 25.1 | 23.4 | 21.1 | 19.8 | 18.9 |

Table 12. SS and TSS-SSA searches of MDDR-B at 5% cut-off using fingerprints.

| Fingerprint | SS | TSS-10 | TSS-20 | TSS-50 | TSS-100 | TSS-200 |
|---|---|---|---|---|---|---|
| BCI | 12.1 | 12.9 | 11.9 | 11.2 | 11.5 | 11.9 |
| Daylight | 10.4 | 10.7 | 9.8 | 9.2 | 9.5 | 9.6 |
| ECFP_4 | 10.5 | 14.5 | 11.8 | 10.4 | 10.4 | 10.6 |
| FCFP_4 | 10.8 | 13.3 | 11.9 | 10.9 | 11.0 | 11.3 |
| MDL Keys | 11.9 | 10.9 | 10.7 | 10.7 | 11.0 | 11.5 |
| PDT | 13.8 | 18.4 | 14.4 | 10.0 | 9.4 | 9.5 |
| Unity | 11.5 | 10.4 | 9.9 | 9.8 | 10.1 | 10.6 |

Table 13. SS and TSS-SSA searches of NCI AIDS at 5% cut-off using fingerprints.

| *K* | *p* | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 5 | 10 |
| 0 | 0.547 | 0.612 | 0.635 | 0.606 | 0.570 |
| 1 | 0.424 | 0.480 | 0.538 | 0.532 | 0.525 |
| 2 | 0.381 | 0.484 | 0.518 | 0.524 | 0.517 |
| 3 | 0.346 | 0.470 | 0.512 | 0.525 | 0.496 |
| 5 | 0.288 | 0.434 | 0.474 | 0.505 | 0.494 |
| 10 | 0.167 | 0.292 | 0.374 | 0.441 | 0.472 |

Table 14. Computed $r^2$ values for the correlation between observed and predicted aqueous solubilities. *k* is the number of nearest neighbours used in TSS (so *k*=0 is conventional SS) and *p* is the number of nearest neighbours used for the prediction stage

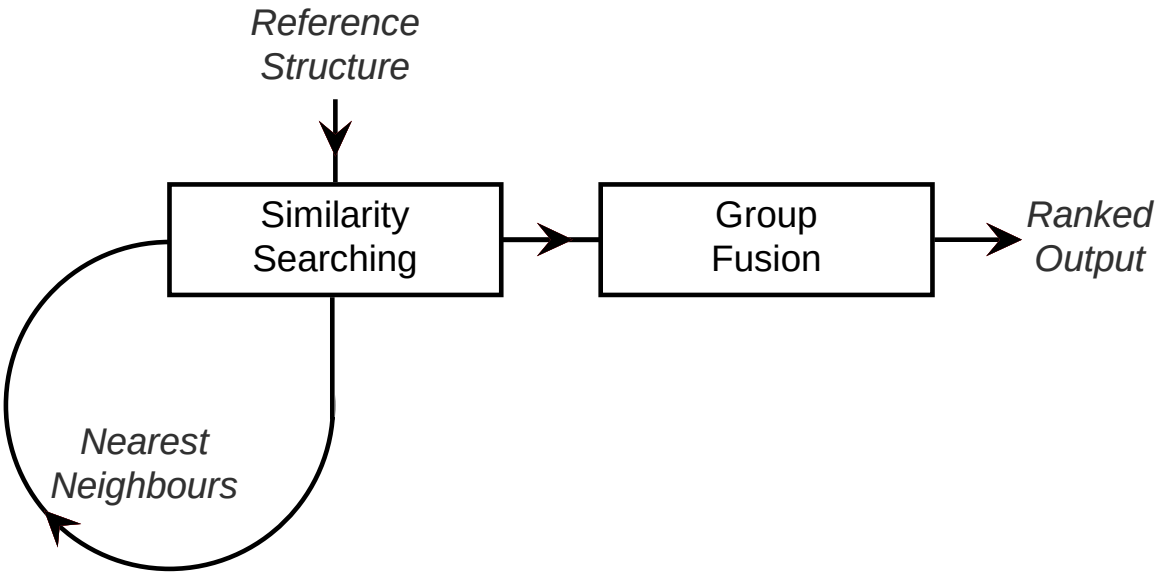


Figure 1. Schematic outline of a turbo similarity search. The user supplies a *Reference Structure* and receives the *Ranked Output* as would be the case in a conventional similarity search. However, the internal processing (not italicised in the figure) is more extensive in TSS, involving not one but multiple iterations of Similarity Searching and the additional Group Fusion step. The *Reference Structure* is matched against each of the database structures, the similarity computer in each case and the database ranked in decreasing similarity order so that the *Nearest Neighbours* can be identified (Similarity Searching). However, instead of outputting the *Nearest Neighbours* as the result of the search, each one is used in turn as the *Reference Structure* so that for, for *k* nearest neighbours, *k* rankings are produced (via Similarity Searching) in addition to that resulting from the initial similarity search. The *k*+1 rankings are combined into a single ranking (Group Fusion) using the MAX fusion rule (see text) and it is this fused ranking that is presented to the user (*Ranked Output*).

Input the reference structure *R*

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

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

Identify the *k* nearest neighbours 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*

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

Combine the sorted lists *SD*(0)-*SD*(*k*) with a fusion rule to give the final ranking

Algorithm 1. Turbo similarity searching using group fusion

Input the reference structure *R*

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

Rank *D* in decreasing order of the calculated similarity values

Assume that the *k* nearest neighbours at the top of the ranking are active, and that all the other molecules in *D* are inactive

Use *R*, the *k* nearest neighbours and the rest of *D* as the training-set for the calculation of *R2* weights for each of the fragments in *D*

Use these weights to score each molecule in *D* in turn

Rank *D* in decreasing order of the calculated scores to give the final ranking

Algorithm 2. Turbo similarity searching using substructural analysis