

# Comparison of $2D$ similarity and $3D$ superposition. Application to searching a conformational drug database\*

Martin Thimm<sup>†</sup>, Andrean Goede<sup>‡</sup>, Stefan Hougardy<sup>†</sup>, Robert Preißner<sup>‡</sup>

June 30, 2004

## Abstract

In a database of about 2000 approved drugs, represented by  $10^5$  structural conformers, we have performed  $2D$  comparisons (Tanimoto coefficients) and  $3D$  superpositions. For one class of drugs the correlation between structural resemblance and similar action was analysed in detail. In general Tanimoto coefficients and  $3D$  scores give similar results, but we find that  $2D$  similarity neglects important structural/functional features. Examples for both over- and underestimation of similarity by  $2D$  metrics are discussed. The required additional effort for  $3D$  superpositions is assessed by implementation of a fast algorithm with a processing time below 0.01 seconds and a more sophisticated approach (0.5 seconds per superposition). According to the improvement of similarity detection compared to  $2D$  screening and the pleasant rapidity on a desktop PC, full-atom  $3D$  superposition will be an upcoming method of choice for library prioritization or similarity screening approaches.

Keywords: similarity screening, superposition, drug database

## Introduction

The accessibility of large compound databases has changed from exclusive in-house databases of large pharmaceutical companies to inexpensive publicly available sources [1]. At this time about two million different compounds can be purchased from different vendors [2]. In this context established methods like  $2D$  similarity searching are increasingly applied to identify active compounds for experimental assays. It was generally accepted that similar compounds having Tanimoto coefficients larger than 0.85 will exhibit similar biological activity

---

\*Stefan Hougardy and Martin Thimm are supported by the DFG Research Center "Mathematics for key technologies", Andrean Goede and Robert Preißner are supported by the BMBF funded Berlin Center of Genome Based Bioinformatics (BCB)

<sup>†</sup>Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

<sup>‡</sup>Institut für Biochemie, Charité, Monbijoustr. 2, 10117 Berlin, Germany.

[3]. This assumption could be reaffirmed at a lower level of 80% similar activity [4]. In different assays the fraction of active 0.85 similars declined to 60% – 40% [5]. In a recent analysis considering more than a hundred different assays the estimation of the chance that a compound that is  $> 0.85$  Tanimoto similar to an active is itself active was further reduced to 30% [6]. The resulting risk of missing attractive compounds gives rise to a number of analyses comparing different molecular descriptors and similarity metrics for different purposes [7, 8]. But nevertheless even the accuracy of the prediction of one of six drug classes remains at 66% [9]. Descriptors representing 3D information [10, 11] and pharmacophore based approaches [12, 13] are opportunities to overcome the weaknesses of 2D descriptors. However a lot of experience and intuition has to be invested to achieve reasonable results [14]. The superposition of 3D structures is a time-consuming task.

For an extensive review on methods for structural alignment see [15] and the references therein.

Structural flexibility has to be taken into account. The latter problem can be approached either during comparison [16] or prior to comparison [17]. Different approaches for the generation of conformers exhibit strengths and weaknesses. Knowledge based procedures explore configuration data from crystallographic databases [18], others emphasize the distinct features of bound ligand conformations especially for the muscarinic acetyl-choline receptor [19]. Simulated annealing for difference minimisation [20] or clustering procedures [21] for better coverage of the low-energy conformational space [22] are applied.

The selection of the right features for the prediction of bioactivity requires compound class specific techniques to obtain reasonable performance [23]. It was shown that the inclusion of 3D information via 3D field descriptors generates further biologically relevant hits [24].

Typical 3D QSAR studies in this field are restricted to a limited set of compounds matching the pharmacophore model [25, 26]. To complement this technique shape-based approaches are implemented [27] and successfully applied to similar problems as considered in the analysis of this paper [28].

The similarity is in the eye of the beholder, as Kubinyi illustrated [29, 30]. The scoring of the 3D similarity remains difficult because the balance between geometrical and physicochemical [31] terms may influence the results towards scaffold hoppers [32] or R-group similarities [33]. For this analysis we selected the drug class of the neuroleptics because these compounds are known to have a number of potential side effects like extrapyramidal adverse events. The therapeutic action of neuroleptics is mediated by their interaction with transmitter receptors in particular with the sub-types of the dopamine-receptor. Here we focus on side effects that can be explained by the affinity to further receptors: histamine-, serotonin-, adrenergic, and muscarinic receptor [34]. This can roughly be estimated by the similarity with compounds from indication classes directly addressing these receptors like antipsychotics, psychoanaleptics or antihistamines. The drug classification scheme according to the WHO recommendation [35] was utilised to this end.

In this analysis we examine

- the correlation between  $2D$  and  $3D$  similarity,
- the fitness of Tanimoto coefficients for the drug class recognition
- whether a simple geometric score will be useful as  $3D$  similarity measure
- whether  $3D$  superposition will be useful to detect similar actives in a drug database.

## Methods

### Drug classification

Recently, the recommendations of the WHO Expert Committee responsible for updating the WHO Model List of Essential Medicines were published [35]. For the first time, a list of all items on the Model List sorted according to their 5-level Anatomical Therapeutic Chemical (ATC) classification codes was given. As the therapeutic subgroup is determined by the second level and the chemical component describes the lower level(s) of classification it is useful for this type of analysis, correlating structural similarity with similar therapeutic action.

The pharmacological action of neuroleptics is mediated by their interaction with transmitter receptors in particular with the sub-types of the dopamine-receptor. Therefore (particular) neuroleptics are known to have a number of side effects and are dubbed "dirty drugs" [36]. Here we focus on effects that can be explained by the affinity to further receptors: histamine-receptor (H1), serotonin-receptors (5-HT<sub>2A/B</sub>, 5-HT<sub>3</sub>), adrenergic receptor (alpha 1) and muscarinic acetyl-choline receptor (M) [34]. This can roughly be estimated by the similarity with compounds from indication classes directly addressing these receptors: N05A (antipsychotics), N06 (psychoanaleptics), D04 / R06 (antihistamines, systemics / dermatologicals).

### Data

All comparisons are performed on a database of 2086  $3D$ -structures of drugs extracted from our inhouse data base. This complies with the number of approved drugs included in the ChemIDplus database [37], which contains a total of 177000 chemical structures. To improve the conditions for the  $3D$ -comparisons 85800 conformers were computed with Catalyst [38] according to the algorithm of Smellie [39]. The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of drugs. It is controlled by the WHO Collaborating Centre for Drug Statistics Methodology [40], and was first published in 1976. The data base covers 218 ATC-major classes (like N05A). 185 ATC-major classes are represented by at least 3 structures, this meets 98 per cent of all such classes containing at least 3 different actual low molecular weight compounds – without e.g. combinations, bandages or proteins. Table 1 shows the 13 members of the ATC-class N05A (antipsychotics) that were used for the data base search.

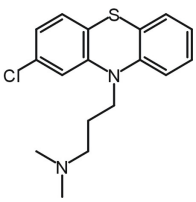
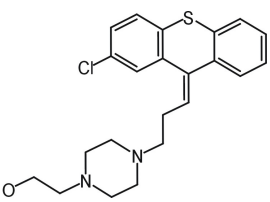
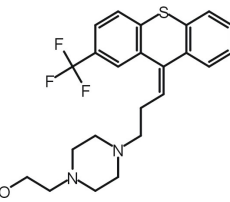
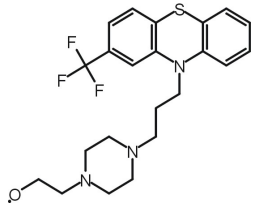
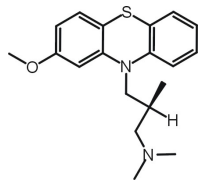
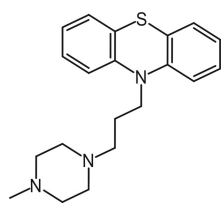
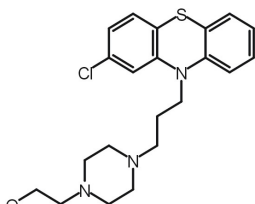
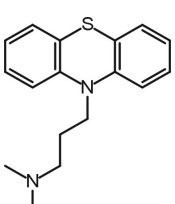
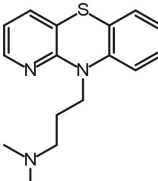
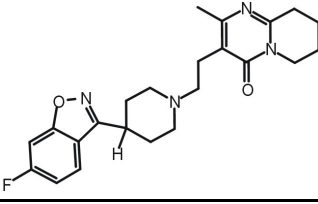
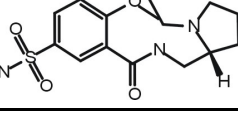
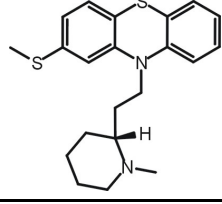
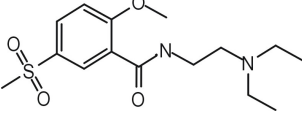
chlorpromazine N05AA01 	clopenthixol N05AF05 	flupenthixol N05AF01 
fluphenazine N05AB02 	methotrimeprazine N05AA02 	perazine N05AB10 
perphenazine N05AB03 	promazine N05AA03 	prothipendyl N05AX07 
risperidone N05AX08 	sulpiride N05AL01 	thioridazine N05AC02 
tiapride N05AL03 		

Table 1: Antipsychotics used for the data base search: name, ATC code and chemical structure

## 2D-Comparison

The 2D-comparison of the molecules was carried out using the Tanimoto coefficients [41] computed by the corresponding procedure of Accord from Accelrys

[38]. For this reason the fingerprints of the structures are calculated using the Daylight algorithm [42] and compared by the Tanimoto similarity measure for bit strings. Fingerprints are Boolean arrays of a given length. To evaluate the fingerprint each pattern of the molecule is generated. Such patterns are

- atoms of a special type
- bonds (single, double,...) between atoms of special types
- paths of different lengths (2 to 7) between atoms of the same type and same order of the bonds, e.g. C=CN, CC=N for path of length 2, or O=CC=N for path of length 3.

Because of the limited length of the fingerprint it is not possible to assign a special bit for only one pattern. Instead of this each pattern is assigned a small number of positions (say 4 or 5) along the fingerprint which are set to 1. Therefore the fingerprints of two molecules can be the same while the molecules are different. Additionally, the positions corresponding to a special pattern account for the occurrence of the pattern. Multiple appearances of the same pattern give the same fingerprint. Therefore featureless molecules (such as C<sub>20</sub>H<sub>22</sub> or C<sub>30</sub>H<sub>32</sub>) give the same fingerprint. Nevertheless the fingerprints indicate whether a compound can be a substructure of another molecule. The Tanimoto coefficient between two fingerprints is the proportion of the bits in common and the bits in at least one fingerprint:

$$TC = \frac{BC}{B1 + B2 - BC} \quad ,$$

where  $BC$  is the number of bits which are 1 in both fingerprints, whereas  $B1$  and  $B2$  are the number of bits which are 1 in the first or the second fingerprint, respectively.

### 3D score

The 3D–superposition–algorithms investigated here are designed to find spatial similarity between molecules. We follow the paradigm that a necessary condition for functional similarity is similar geometry. For this reason the scoring function is built to measure spatial similarity only. By this it is possible to find superpositions which we would not have found by simultaneously trying to incorporate physicochemical features. So it may happen that few of the geometrically found hits turn out to be biologically irrelevant, but we considerably lower the risk of losing most interesting and relevant hits that come from different chemical structures.

Similarity of molecules is measured by superposition. In general the superposition problem may be decomposed into two subproblems: First, we have to find an assignment (or matching) of the atoms of one molecule to the atoms of the other molecule, that tells us which atom on one side has to be superimposed with which atom on the other side (or not superimposed at all). Second, a rigid

motion is computed to optimally perform this action. There are two competing objectives:

1. The more atoms are actually superimposed the better the superposition should be scored.
2. The distances of the matched atoms should be as small as possible.

A way to balance these two goals is the following scoring function: Consider two molecules  $A$  and  $B$  with  $m$  and  $n$  atoms resp. ( $m \leq n$ ). Given an assignment  $M$  that maps the atoms  $a_i^M$  of  $A$ ,  $i = 1, \dots, k$ ,  $k \leq m$ , to the atoms  $b_j^M$ ,  $j = 1, \dots, k$ . The resulting superposition has then score

$$score(A, B, M) = \frac{k}{m} e^{(-rmsd(M))}.$$

The first term,  $\frac{k}{m}$ , measures the proportion of actually superimposed atoms of the smaller molecule, the second term, the root mean square distance of these atoms,

$$rmsd(M) = \sqrt{\frac{1}{k} \sum_{i=1}^k dist(a_i^M, b_i^M)^2},$$

controls the distance of matched atoms.

By this definition we try to find a superposition with *many atoms* superimposed *with small distance*. The first term increases with the number of superimposed atoms, but in the majority of cases this will make the second term decrease since more assigned atoms will result in higher rmsd.

Observe that the value of the scoring function is always between 0 and 1, where larger values mean higher similarity.

The scoring function is not restricted to molecules of the same size. It is also possible to compare molecules of quite different size, since the first term of our scoring function allows us to find smaller molecules inside larger ones.

### 3D Comparison

Since no polynomial time algorithm is known to solve the superposition problem as described above to optimality, we have to use heuristic methods to get good solutions in reasonable time. We have implemented two different approaches, a fast one and a more sophisticated one.

In general the superposition problem may be decomposed into two subproblems: First, find the assignment for the two atom sets, i.e., which atom in one molecule should be superimposed with which atom in the other molecule. Second, find the rigid motion that gives the optimal superposition of the atom sets, given the assignment. Our two approaches mainly differ in the effort made to solve the first subproblem.

One good news is that the second subproblem is known to be solvable in polynomial time [43]. A way to solve the superposition problem is to enumerate

all possible assignments and to compute the rigid motion for each of them. But here the bad news is that the number of possible assignments grows highly exponential in the number of atoms. With this naive approach only instances with very few atoms (less than 10) may be handled.

With the help of a branch-and-bound approach, a widely used technique in optimization, we are able to reduce the number of assignments to be tested dramatically. Doing this we are able to solve the superposition problem up to optimality quite fast (10 atoms: few seconds, 16 atoms: 1–2 minutes).

Since drug-like molecules are often larger and since the above mentioned running times are still far too slow we have to find ways to overcome these difficulties. We can no longer hope to solve the problem exactly, i.e., the solution of the algorithm described next will not be guaranteed to be best possible, but we will get it very quickly without losing much quality. In what follows we describe the two algorithms in more detail.

### **Fast 3D-superposition**

The algorithm presented here can roughly be sketched by the following steps:

1. superposition of the centers of mass
2. orientation according principal moments of inertia
3. atom pair assignment
4. improvement

The first orientation (in step 2) is of course independent of transformations of the coordinate system, and quite stable for small alterations of the atomic positions. The normalisation of the atomic sets is unique except for possible rotations (original arrangement and rotations of  $180^\circ$  around  $x$ -,  $y$ - or  $z$ -axis). This means that the degree of freedom is strongly reduced and the assignment of pairs of atoms is relatively straightforward for identical and slightly modified atomic sets because only four possible normalisations have to be checked to identify related atoms: imagine determining the correct orientation of a credit card (magnetic strip: top surface, right; top surface, left; bottom surface right; bottom surface, left). In a first step the centres of mass of the two atomic sets are determined. All the coordinates of the atoms included are transformed to superimpose the centres of mass. To determine the least and largest (orthogonal) expansion, the plane and the straight line of minimal quadratic distance to all atoms have to be computed. The normal line of the plane gives the least expansion and the straight line of minimal quadratic distance points at the largest expansion. Using these directions one atomic set is rotated such that the major directions coincide. There are four possible normalisations for an atomic set that coincide with the exception of  $180^\circ$  rotations around  $x$ -  $y$ - or  $z$ -axes. In a further step all four normalisations are used to determine the pairs of atoms between the two atomic sets. This normalisation procedure is stable even if additional atoms are included in one of the sets. Therefore, the

normalisation of the atomic set can be used to identify pairs of corresponding atoms. Two atoms form a pair if they are mutually the nearest atoms, and their distance is lower than a given cut-off value. Different cut-off values were tested showing that a cut-off of 2.5 Å performed best for sets of densely packed atoms. For all four normalisations the number of atom pairs is chosen and the root mean square distance (rmsd) calculated for the related atomic pairs. The normalisations are weighted on the basis of these values. The best normalisation (largest number of pairs) is used in a further step to improve the alignment. For the given set of pairs the optimal superposition is estimated, followed by a new search of related pairs until the assignment of the atoms does not change.

### **Sophisticated 3D–superposition**

The second algorithm proceeds in three phases:

1. Reduce the given instance to a smaller new instance that is in a certain sense similar to the original one.
2. Solve this new instance optimally with the above mentioned branch–and–bound technique.
3. Lift the solution of the smaller, artificial instance to a solution for the original problem.

Next we describe the three phases in more detail:

#### **Phase 1: Reduction to a smaller artificial instance**

The running time of the exact algorithm mainly depends on the number of atoms of the two molecules; so the aim of this phase is to construct, starting with the original molecules, new, artificial pseudo–molecules with fewer pseudo–atoms, that are still spatially similar to the original molecules. This is done iteratively in the following way, sometimes called *hierarchical clustering*:

- Start with the original molecule, call every atom a pseudo–atom.
- While the number of pseudo–atoms is larger than a predefined number  $r$ 
  - look for those two pseudo–atoms with the smallest distance, merge them to a new pseudo–atom. The coordinates of this new pseudo–atom are given by the weighted center of gravity of the two merged ones, which are then deleted.

So in every such step the number of pseudo–atoms is decreased by one. As a remark we should say that we take into account how many original atoms are represented by a pseudo–atoms by attaching weights to them. The idea is that the new instance constructed in this way still carries the spatial information of the original molecule to a certain extent.

#### **Phase 2: Exact solution of artificial instance**

Now we solve our artificial instance with the exact algorithm. The solution of this step is the starting point of phase 3.



### Phase 3: **Lift of intermediate solution to the original instance**

The rigid motion which led to the solution in phase 2 may also be applied to the original instance. The idea of our approach is now that, since the smaller artificial instance is spatially similar to the original one, the position of the original atoms after this rigid motion is not far from a very good solution. We only have to refine the assignment to these atoms. This is done as follows: The distance of atoms that will be assigned to each other should now be already quite small, so a natural approach is the following:

- **step 0** Declare all atoms to be not fixed.
- **step 1** Sort those atoms of the first molecule that have not yet been assigned (fixed) increasingly by their distance to the nearest neighbour in the other molecule, that is still available. Take the first  $s$  of them ( $s$  is a small predefined number).
- **step 2** Enumerate all possible assignments of these  $s$  atoms to their nearest neighbours (including the possibility that an atom is assigned to have no matching partner), carry out the appropriate rigid motion and pick the one with the best score value (on this partial instance of already fixed atoms and these  $s$  new atoms).
- **step 2** Fix the assignment on these  $s$  atoms, perform the implied rigid motion to all atoms. Goto step 1 while there are atoms that are not yet fixed.

We conclude this section with some implementation details. Since the solution of Phase 1, the artificial pseudo-molecule, may look quite different for different numbers  $r$ , we perform this step for several different values of  $r$ . The optimal solution of Phase 2 may not be exactly what we want, since we observed in numerous tests of the exact algorithm that there are instances that have quite a number of solutions with very similar score values but very different assignments. To overcome this we store not only the best solution but the best  $n$  of them (seen during the branch-and-bound process). Phase 3 also depends on the predefined number  $s$ , so again as in Phase 1 we perform this phase for all several different values of  $s$ . As one would expect, the quality of the solution increases with the size of the  $r$ - and  $s$ -intervals and with  $n$ , but so does the running time. (Running time grows nearly proportional with  $n$  and the number of different  $s$ -values and superproportional in the number of different  $r$ -values, since larger  $r$ -values get more and more expensive.)

Our standard parameters for drug-like molecules are  $r \in \{3, 4, 5\}$ ,  $s \in \{3, 4\}$ ,  $n = 40$ . We determined them as a result of numerous test trying to find an optimal tradeoff between quality and running time.

Although we cannot prove the optimality for our algorithms, we wanted to see how far away we are in the relevant cases. As a test we computed  $3D$ -scores for quite a number of instances which were known to have large  $3D$ -scores ( $> 0.70$ ) up to optimality with the branch-and-bound algorithm mentioned above. (As a remark we should say that this is possible in these cases - with still very

large running times - since branch-and-bound algorithms tend to find very good results - if they exist - quite "fast".) The results showed that in all cases the sophisticated approach was within five percent of the optimal solution.

## Results

To compare the two *3D*-superposition algorithms with the *2D*-approach we selected 13 antipsychotics (ATC code N05A) and computed both Tanimoto coefficients and *3D*-scores (fast and sophisticated) for each of them with all drugs in the database (with more than 14 non hydrogen atoms).

To compute the *3D*-score of two conformers of molecules of average size (25 atoms) we need about 0.01 sec. (fast) and 0.5 sec. (sophisticated) (on a 2GHz PC), resulting in a total running time of 25 sec. (fast) and 21 min. (sophisticated) to fully compare two molecules, both given by 50 structural conformers.

Algorithm	ATC code		$\Sigma$
	N05A,N06,R06,D04	others	
Tanimoto > 0.85	113	51	164
<i>3D</i> -soph. > 0.75	157	43	200
<i>3D</i> -fast > 0.75	131	25	156

Table 2: Number of hits

Unfortunately we can not compare the values of Tanimoto coefficients and *3D*-scores one-to-one. As mentioned in the introduction it is generally accepted that Tanimoto coefficients larger than 0.85 start to indicate similar activity. To find a corresponding value for the *3D*-score we counted the number of hits with Tanimoto coefficient larger than 0.85 and found that a *3D*-score of about 0.75 gives approximately the same number of hits.

We found 164 hits with Tanimoto coefficient larger than 0.85. The *3D*-superposition algorithms returned 200 (sophisticated) and 156 (fast) hits with *3D*-score larger than 0.75 (see Table 2).

Since we want to show that the *3D*-approach is appropriate to find molecules that have similar activity we first looked at the ATC codes of the hits with score value larger 0.75 and found that 78, 5% (157 out of 200) (sophisticated) and 84% (131 out of 156) (fast) of these can be found in drug subclasses that are known to have similar activity or similar adverse reaction. (ATC codes N05A, N06, R06, D04) The proportion of hits in these classes for Tanimoto coefficients larger than 0.85 is 69% (113 out of 164) (see Table 2).

To compare the results of the *2D*- and *3D*-approaches we have to look at three different sets of seemingly similar pairs of molecules:

- a. Large Tanimoto coefficient ( > 0.85), small *3D*-score ( < 0.75).
- b. Small Tanimoto coefficient ( < 0.85), large *3D*-score ( > 0.75).

c. Large Tanimoto coefficient ( $> 0.85$ ), large  $3D$ -score ( $> 0.75$ ).

Algorithm		ATC code		
$3D$ -soph.	Tanimoto	N05A,N06,R06,D04	others	$\Sigma$
$> 0.75$	$> 0.85$	74	11	85
$> 0.75$	$< 0.85$	83	32	115
$< 0.75$	$> 0.85$	39	40	79

Table 3: Number of hits: Tanimoto vs. soph.  $3D$

Algorithm		ATC code		
$3D$ -fast	Tanimoto	N05A,N06,R06,D04	others	$\Sigma$
$> 0.75$	$> 0.85$	66	10	76
$> 0.75$	$< 0.85$	65	15	80
$< 0.75$	$> 0.85$	47	41	88

Table 4: Number of hits: Tanimoto vs. fast  $3D$

Ad a. For this set of hits, comparing Tanimoto coefficients to both sophisticated and fast  $3D$ -superposition gives a similar picture (see Table 3 and Table 4 for the exact numbers). Approximately one half of this set of hits lies in the above mentioned relevant drug classes. A closer inspection for this subset (i.e. 3. column) shows that in most of the cases the  $3D$ -score for these hits is larger than 0.65. Looking at these hits, that are clearly relevant from a biological point of view, we can infer that already  $3D$ -score above 0.65 are worth while to look at. (See Fig. 1 a) for an example).

The second subset of hits (i.e. 4. column) with large Tanimoto coefficients and small  $3D$ -scores consists to a large proportion of those hits for which the Tanimoto coefficient highly overestimates the structural/functional similarity of the molecules. An examples can be seen in Fig. 1 b).

Ad b. For this set of hits the situation changes. The number of hits with large  $3D$ -score that are found by the sophisticated algorithm are significantly larger than those found by the fast algorithm, for both the relevant drug classes and the others. (see Table 3 and Table 4 for the exact numbers).

Since the  $3D$ -superposition algorithms are not designed to incorporate chemical features there are some hits that are clearly geometrically relevant, but perhaps their prediction about similar activity is quite limited. These hits can be found in the second subset (i.e. 4. column).

What we are really aiming for is the first subset (i.e. 3. column). Here we find hits that are both geometrically similar and relevant concerning prediction of similar activity and function. One reason why in these cases Tanimoto coefficients do not indicate similarity are slight changes in chemical structure. Furthermore there are hits with two molecules of somewhat different size. It is known (see [44]) that for these instances Tanimoto coefficients are more and

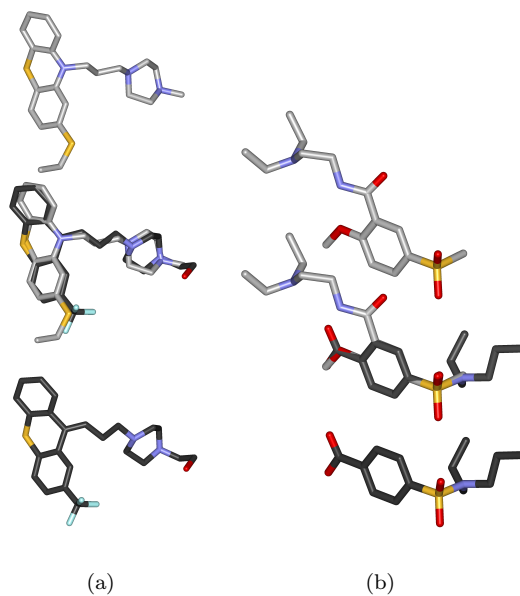


Figure 1: Comparisons with 2D similarity above threshold and 3D similarity below threshold.

(a) Superposition of flupentixol (bottom, ATC code: N05AF01) with thioethylperazine (top, ATC code: R06AD03) - with a 3D score of 0.63 and a Tanimoto coefficient of 0.89. The similarity is underestimated by the 3D score because the distortions in the tricyclic ring are not properly represented by the conformers.

(b) Superposition of tiapride (top, ATC code: N05AL03) and probenecid (bottom, ATC code: M04AB01) with a 3D score of 0.58 and a Tanimoto coefficient of 0.85. The resemblance to probenecid, an antigout drug, is overestimated by the Tanimoto coefficient because of the identical chemical subgroups (phenyl, sulfonyl).

more inefficient, while our  $3D$ -score is designed to also find these hits. (see some examples for both cases in Fig. 2).

For this type of hits the sophisticated approach is clearly superior to the fast algorithm. Comparing the numbers in Table 5 with those in Table 3 and Table 4 shows that most of the hits for which the two  $3D$ -approaches differ can be found in the class discussed here. The main reason for this is that the fast approach is not able to find hits for two molecules that have different overall geometry, in particular small molecules that are substructures of larger ones are not found (see Fig. 3).

Algorithm		ATC code		
$3D$ -soph.	$3D$ -fast	N05A,N06,R06,D04	others	$\Sigma$
$> 0.75$	$> 0.75$	130	22	152
$> 0.75$	$< 0.75$	27	21	48
$< 0.75$	$> 0.75$	1	3	4

Table 5: Number of hits: fast  $3D$  vs. soph.  $3D$

Ad c. In this set we find those hits, that are quite similar in both chemical structure and size. They are reported as relevant by both approaches. For this type of hits both strategies are most similar. Here again, as in case a., the fast and the sophisticated  $3D$ -algorithm perform comparably.

From our point of view we have therefore seen several strong arguments in favour of the  $3D$ -superposition algorithms. It can be clearly seen that the  $3D$ -approach is able to detect similar activity and similar adverse reaction, even with this seemingly simple, purely geometry-based scoring function.

For large data sets a fast  $3D$ -superposition algorithm combined with Tanimoto coefficients helps to increase the set of relevant hits.

If one aims to really find all, at least geometrically relevant hits – this may be important for smaller and more specific sets of molecules – it is worth while to follow the sophisticated  $3D$ -approach (with a somewhat smaller threshold for relevance). We were able to find really relevant hits that can not be found by simple  $2D$ -methods or by the fast  $3D$ -algorithm.

## Discussion

In agreement with our results it is shown in [27] and [28] that  $3D$  similarity searches retrieve compounds with more diverse topology while  $2D$  similarity works best when the query molecule contains relatively rare and distinct topological features that are responsible for the biological activity.  $2D$  similarity works poorly when common functional groups as in peptides are considered. A similar fragment- or topomer-based steric shape screening was shown to be more selective than  $2D$  similarity [13], especially advantageous "lead-hopping" was observed. A reasonable speed for the in silico screening of large compound libraries can be achieved by full-atom superposition procedures as presented in

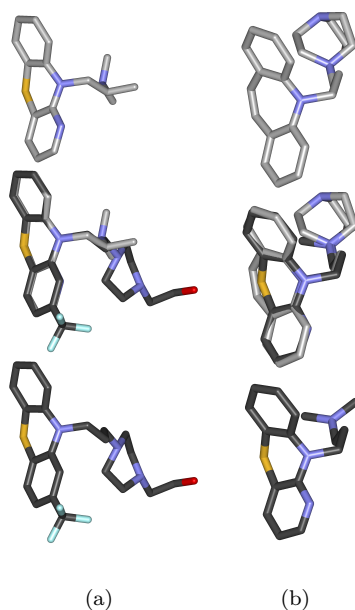


Figure 2: Comparisons with low Tanimoto coefficients and 3D scores above threshold.

(a) Superposition of fluphenazine (bottom, ATC code: N05AB02) with isothipendyl (top, ATC codes: D04AA22, R06AD09) with a 3D score of 0.81 and a Tanimoto coefficient of 0.69. The resemblance to isothipendyl, an antihistaminic agent, is neglected by the 2D similarity measure because of missing chemical groups (trifluoromethyl, piperazin) and quite different sizes of the molecules.

(b) Superposition of prothipendyl (bottom, ATC code: N05AX07) and pipramol (top, ATC code: N06AA05) with a 3D score of 0.76 and a Tanimoto coefficient of 0.72. The similarity to pipramol, an antidepressant, is missed by 2D comparison because the middle ring is seven membered in pipramol (dibenzazepine derivative) and six membered in prothipendyl (azaphenothiazine derivative).

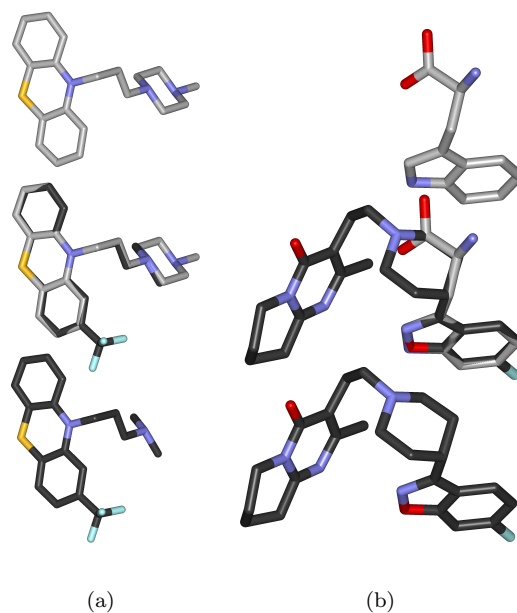


Figure 3: Differences between fast and sophisticated superpositions.

(a) Superposition of perazine (top, ATC code: N05AD10) with triflupromazine (bottom, ATC code: N05AA05) with a  $3D$  score of 0.77 (sophisticated), 0.50 (fast) and a Tanimoto coefficient of 0.84. The resemblance between the two neuroleptics is neglected by the fast superposition algorithm because the centers of gravity do not fit.

(b) Superposition of tryptophan (top, ATC code: N06AX02) and risperidone (bottom, ATC code: N05AX08) with a  $3D$  score of 0.77 (sophisticated), 0.50 (fast) and a Tanimoto coefficient of 0.44. The similarity to tryptophan, an antidepressant, is missed by the fast superposition algorithm because of the very different overall geometry of the molecules.

this analysis.

With receptor structures available ligand-docking programs have been shown to enrich hit lists of in silico screening approaches [45] but in the case of psycholeptics a number of structurally unknown receptors are engaged. Most of the processes involved in ADME are driven by rather unspecific interactions between drugs and macromolecules but drug transporters and cytochromes gained increased interest in early ADME profiling via similarity based structure activity relation (SIBAR) [46]. The increased predictive power of the 3D- vs. 2D-similarity for side effects demonstrated in this analysis gives rise to the hope that improvements in ADME and toxicity profiling will be possible.

Limitations of the fast 3D superposition approach are spherical compounds for which it might fail to find proper assignments. The known size bias and size limitation of 2D similarity measures [44] also may cause problems for the fast algorithm.

The conformer generation is a general problem because the 3D similarity between two structural ensembles depends critically on the original structures, the conformer generation [22] and clustering [47] algorithm, the parameters like energy threshold, and the number of conformers per compound. In particular the number of rotatable bonds will restrict the 3D similarity approach or will require new algorithms [48].

## References

- [1] Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J Chem Inf Comput Sci* 2001, 41, 702-712.
- [2] Bradley, M. P. An overview of the diversity represented in commercially-available databases. *Mol Divers* 2002, 5, 175-183.
- [3] 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.
- [4] Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ Res* 1998, 8, 23-39.
- [5] Delaney, J. S. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol Divers* 1996, 1, 217-222.
- [6] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J Med Chem* 2002, 45, 4350-4358.



- [7] 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.
- [8] Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of similarity measures for searching the dictionary of natural products database. *J Chem Inf Comput Sci* 2003, 43, 449-457.
- [9] Dixon, S. L.; Merz, K. M., Jr. One-dimensional molecular representations and similarity calculations: methodology and validation. *J Med Chem* 2001, 44, 3795-3809.
- [10] Wildman, S. A.; Crippen, G. M. Three-dimensional molecular descriptors and a novel QSAR method. *J Mol Graph Model* 2002, 21, 161-170.
- [11] Turner, D. B.; Willett, P. The EVA spectral descriptor. *Eur J Med Chem* 2000, 35, 367-375.
- [12] Hecker, E. A.; Duraiswami, C.; Andrea, T. A.; Diller, D. J. Use of catalyst pharmacophore models for screening of large combinatorial libraries. *J Chem Inf Comput Sci* 2002, 42, 1204-1211.
- [13] Cramer, R. D.; Jilek, R. J.; Andrews, K. M. Dbtop: topomer similarity searching of conventional structure databases. *J Mol Graph Model* 2002, 20, 447-462.
- [14] Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J Comput Aided Mol Des* 2002, 16, 653-681.
- [15] Lemmen, C; Lengauer, T. Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des* 2000,14,215-232.
- [16] Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J Med Chem* 1998, 41, 4502-4520.
- [17] Kramer, A.; Horn, H. W.; Rice, J. E. Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J Comput Aided Mol Des* 2003, 17, 13-38.
- [18] Klebe, G.; Mietzner, T.; Weber, F. Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *J Comput Aided Mol Des* 1999, 13, 35-49.
- [19] Furukawa, H.; Hamada, T.; Hayashi, M. K.; Haga, T.; Muto, Y.; Hirota, H.; Yokoyama, S.; Nagasawa, K.; Ishiguro, M. Conformation of ligands bound to the muscarinic acetylcholine receptor. *Mol Pharmacol* 2002, 62, 778-787.

- [20] Mills, J. E.; de Esch, I. J.; Perkins, T. D.; Dean, P. M. SLATE: a method for the superposition of flexible ligands. *J Comput Aided Mol Des* 2001, 15, 81-96.
- [21] Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J Chem Inf Comput Sci* 1996, 36, 563-571.
- [22] Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J Comput Chem* 2003, 24, 10-20.
- [23] Weston, J.; Perez-Cruz, F.; Bousquet, O.; Chapelle, O.; Elisseeff, A.; Scholkopf, B. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics* 2003, 19, 764-771.
- [24] Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J Chem Inf Comput Sci* 2000, 40, 295-307.
- [25] Bostrom, J.; Bohm, M.; Gundertofte, K.; Klebe, G. A 3D QSAR study on a set of dopamine D4 receptor antagonists. *J Chem Inf Comput Sci* 2003, 43, 1020-1027.
- [26] Bostrom, J.; Gundertofte, K.; Liljeforsa, T. A pharmacophore model for dopamine D4 receptor antagonists. *J Comput Aided Mol Des* 2000, 14, 769-786.
- [27] Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible. *J Chem Inf Comput Sci* 1997, 37, 80-86.
- [28] Guner, O.F.; Hahn, M.; Li, H.; Hassan, M. 2D versus 3D shape similarity: use of molecular shape-based 3D searching techniques for identifying novel compounds. Case study, Accelrys; [http://www.accelrys.com/cases/2dvs3di\\_full.html](http://www.accelrys.com/cases/2dvs3di_full.html)
- [29] Kubinyi, H. Molecular similarity. 1. Chemical structure and biological action. *Pharm Unserer Zeit* 1998, 27, 92-106.
- [30] Kubinyi, H. Molecular similarity. 2. The structural basis of drug design. *Pharm Unserer Zeit* 1998, 27, 158-172.
- [31] Iwase, K.; Hirono, S. Estimation of active conformations of drugs by a new molecular superposing procedure. *J Comput Aided Mol Des* 1999, 13, 499-512.
- [32] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew Chem Int Ed Engl* 1999, 38, 2894-2896.

- [33] Holliday, J. D.; Jelfs, S. P.; Willett, P.; Geddeck, P. Calculation of inter-substituent similarity using R-group descriptors. *J Chem Inf Comput Sci* 2003, 43, 406-411.
- [34] Maurer, I.; Volz, H. P. Cell-mediated side effects of psychopharmacological treatment. *Arzneimittelforschung* 2001, 51, 785-792.
- [35] The selection and use of essential medicines. Report of the WHO Expert Committee, 2002 (including the 12th Model list of essential medicines). World Health Organ Tech Rep Ser 2003, 914, i-vi, 1-126.
- [36] Uhl, G. R.; Vandenberg, D. J.; Miner, L. L. Knockout mice and dirty drugs. *Drug addiction. Curr Biol* 1996, 6, 935-936.
- [37] Specialized Information Services (SIS) Division of the National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, <http://chem.sis.nlm.nih.gov>
- [38] Accelrys Inc., San Diego, CA. <http://www.accelrys.com>
- [39] Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J Comput Chem.* 2003 Jan 15;24(1):10-20.
- [40] WHO Collaborating Centre for Drug Statistics Methodology, <http://www.whocc.no>
- [41] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998, 38, 983-996.
- [42] Daylight Chemical Information Systems, Santa Fe, NM. <http://www.daylight.com>.
- [43] Umeyama, S. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991, 13, 676-681
- [44] 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.
- [45] Jenkins, J. L.; Kao, R. Y.; Shapiro, R. Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. *Proteins* 2003, 50, 81-93.
- [46] Klein, C.; Kaiser, D.; Kopp, S.; Chiba, P.; Ecker, G. F. Similarity based SAR (SIBAR) as tool for early ADME profiling. *J Comput Aided Mol Des* 2002, 16, 785-793.
- [47] Raymond, J. W.; Blankley, C. J.; Willett, P. Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J Mol Graph Model* 2003, 21, 421-433.

- [48] Raymond, J. W.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Smoothed Bounded Distance Matrices. *J Chem Inf Comput Sci* 2003, 43, 908-916.