Comparison of Topological, Shape, and Docking Methods in Virtual Screening

Georgia B. McGaughey,*,† Robert P. Sheridan,‡ Christopher I. Bayly,§ J. Chris Culberson,† Constantine Kreatsoulas,† Stacey Lindsley,† Vladimir Maiorov,‡ Jean-Francois Truchon,§ and Wendy D. Cornell‡

Department of Molecular Systems, WP53F-301, Merck Research Laboratories, West Point, Pennsylvania 19486, Department of Molecular Systems, RY50SW-100, Merck Research Laboratories, Rahway, New Jersey 07065, and Medicinal Chemistry, Merck Frosst, Canada H9H 3L1

Received February 7, 2007

Virtual screening benchmarking studies were carried out on 11 targets to evaluate the performance of three commonly used approaches: 2D ligand similarity (Daylight, TOPOSIM), 3D ligand similarity (SQW, ROCS), and protein structure-based docking (FLOG, FRED, Glide). Active and decoy compound sets were assembled from both the MDDR and the Merck compound databases. Averaged over multiple targets, ligand-based methods outperformed docking algorithms. This was true for 3D ligand-based methods only when chemical typing was included. Using mean enrichment factor as a performance metric, Glide appears to be the best docking method among the three with FRED a close second. Results for all virtual screening methods are database dependent and can vary greatly for particular targets.

INTRODUCTION

Virtual screening is used in drug discovery for lead finding, lead optimization, and scaffold hopping. In the past decade, the discovery of the marketed drug, Aggrastat,[1] can be traced to a virtual screen of Merck's corporate database. The Arg-Gly-Asp motif of the endogeneous peptide was mimicked using distances between the two charged centers and was used to search for nonpeptides in the sample collection. There are numerous examples of virtual screening evaluations[2−8] and recent review articles[9,10] that illustrate the advantage of virtual database screening prior to testing compounds in a biological assay. Nearly a century ago, the concept of similarity based screening was applied in the discovery of phenytoin, which is marketed as dilantin in the U.S.A.[11,12] The compound, initially synthesized in 1908, was identified by Merritt and Putnam as an antiseizure medicine using what we would now call a 2D similarity method. Even though computers were not available at the time, Merritt and Putnam took known narcotic drugs and tested their derivatives for activity against a known target.

Current computational methods are more sophisticated than those employed by Merritt and Putnam and are able to virtually screen typical corporate compound collections in less than a day. This can be attributed not only to the availability of software for this purpose but also to parallel computing and increased compute power per processor. Papers on virtual screening typically focus on whether a method can select novel actives from a database. This is useful for the lead-finding stage of a drug discovery project. However, other aspects such as the ability to reproduce the crystallographic pose are often measured in the case of docking. The results of virtual screening experiments are dependent on numerous variables such as the choice of actives, the choice of decoys (druglike compounds which are not active against the target but are included to test the ability of the method to differentiate actives from inactives), the targets, the software version, and the various options imposed therein.

Merck has been interested in virtual screening methods for the past three decades. Two 3D methods most extensively used in-house are FLOG[13] and SQ.[14] These methods are described in-depth in the cited references. Both methods use precalculated conformations which are computed through the use of either an in-house knowledge-based conformation generator[15] or the application of distance geometry.[16,17] In this paper we compare these methods to commercially available software. In particular, we compare our SQ algorithm to the ligand/shape-based screening method ROCS[18] distributed by OpenEye Scientific Software, and we evaluate our structure-based FLOG algorithm against both Glide[19,20] (Schrödinger) and FRED[21] (OpenEye Scientific Software). In addition, since 2D similarity measures are widely used for virtual screening, we include our in-house method TOPOSIM[22] and the de facto industry standard, Daylight.[23]

METHODS

A retrospective evaluation such as ours requires a set of targets and databases to search. The databases, in turn, consist of a carefully chosen set of actives and a large number of decoys. Here we use two sources for actives and decoys, the MDDR[24] (a licensable database compiled from patent literature) and the MCIDB (Merck's corporate database). We screened both the MDDR and MCIDB databases for two main reasons. We wanted to know if a difference existed in enrichment rates for the MCIDB relative to the MDDR due to differences in particular structural series contained in them. Furthermore, much of the published literature involves searching over proprietary databases, making it impossible

* Corresponding author e-mail: georgia_mcgaughey@merck.com.
† WP53F-301, Merck Research Laboratories.
‡ RY50SW-100, Merck Research Laboratories.
§ Merck Frosst.

10.1021/ci700052x CCC: $37.00
Published on Web 06/26/2007

Table 1. Targets

target	PDB code	MDDR activities	actives in MDDR set	actives in MCIDB set
CA_I	1azm	"Carbonic Anhydrase Inhibitor"	80	241
CDK2	1aq1	"Protein Kinase C Inhibitor"[a]	77	104
COX2	1cvu (1cx2)[b]	"Cyclooxygenase Inhibitor" "Cyclooxygenase-2 Inhibitor"	257	100
DHFR	3dfr	"Dihydrofolate Reductase Inhibitor"	26	0
ER (alpha)	3ert	"Antiestrogen" "Estrogens"	74	233
HIV-pr	1hsh	"HIV-1 Pr Inhibitor"	136	226
HIV-rt	1ep4	"Reverse Transcriptase Inhibitor"	149	218
NEURAMINIDASE	1a4q	"Antiviral" + similar to inhibitors in PDB	12	0
PTP1B	1c87	"Antidiabetic" + similar to inhibitors in PDB	8	103
THROMBIN	1mu6 (1dwc)	"Thrombin Inhibitors"	200	510
TS	2bbq	"Thymidylate Synthetase Inhibitor"	31	0

[a] There are no known "CDK2" inhibitors in the MDDR. The closest related kinase is "Protein Kinase C" and that key word was used to search for actives. [b] When the source of the ligand is different than the source of the protein target, the source of the ligand is given in parentheses.

for others to reproduce the results. Hence, reporting results for the MDDR and MCIDB separately would allow interested groups to reproduce our claims for at least one database. It should be noted that generally the MDDR is a good resource for retrospective studies in that it contains many diverse molecules in multiple therapeutic activities, but it does suffer from some limitations, the most severe being that of "false negatives", i.e. MDDR compounds are tested for a few activities, but they might actually be active on some other target if only they were tested there. This is a potential issue for any retrospective study. However, since our aim is to compare virtual screening methods, and each method "sees" the same list of actives for a given target, we expect that this will not be a major problem.

Targets. *Protein Targets.* We chose a set of 11 targets based on the following criteria: (1) the existence of at least one high-resolution, publicly available crystal structure complexed with a ligand, (2) the presence of a large number of structurally diverse actives in the MDDR and/or the MCIDB, (3) the inclusion of only a single representative from the same family of enzymes (e.g., only one serine protease, only one acid protease, etc.), and finally, (4) the selection of a diverse set of active sites (some hydrophobic, some hydrophilic, some large, some small).

In most cases, there exist many high-resolution crystal structures of each target (e.g., HIV-protease, HIV-pr). It is known that the crystal structure one uses affects the apparent effectiveness of docking methods.[25] However, the point here is to compare methods for a set of selected targets. The list of structural targets, the respective Protein Data Bank (PDB) codes, and the number of actives present in either the MDDR or MCIDB database is in Table 1. Historically, we have no known, in-house actives for dihydrofolate reductase (DHFR), neuraminidase, or thymidylate synthase (TS), and as a consequence, the field for the number of actives in the MCIDB for those targets is zero.

Typically, crystal structures were taken as is. However, in the case of HIV-1 reverse transcriptase (HIV-rt), missing loops, none of which were within 15 Å of the bound ligand, were added. For all targets, the ligand was separated from the protein, and all waters were eliminated. The entire protein was kept, and there was no attempt to carve out an active site. Alternative side-chain positions present in the PDB file were deleted, and residues with missing side chains were named ALA (alanine).

Ligand Targets. In most cases, the ligands from our protein targets seemed to be reasonably druglike and would make good targets for ligand-based screening (2D and 3D). However, in two cases (COX2 and THROMBIN), we used a ligand from a different crystal structure other than the one used for the docking methods as the cognate ligand was too small to be considered druglike. In those two cases, the alternative source is noted in Table 1. The bond orders of the cocrystallized ligands were corrected by hand. The structures of the ligand targets are in Figure 1.

Database and Actives. Our aim was to construct diverse databases where close analogs were eliminated. This would make the selection of each active in a virtual screen a quasi-lead-hopping event and would eliminate a potential bias where 2D methods would appear better than 3D methods by finding close analogs to the target. Note that we are assuming molecules are different "leads" if their global similarity is below a certain threshold as defined in the next paragraph. The molecules may still contain common ring systems, which some could consider a common "scaffold". Hence, we refer to "lead-" rather than "scaffold-hopping".

Source for MDDR Database. We clustered the entire MDDR version 2003.2 (~129 000 compounds with structures) using the Butina[26] algorithm with a cutoff of 0.7 using the Dice definition of similarity and the Atom Pair[27] (AP) descriptors. Our database consisted of ~25 000 cluster centers. This was about the size of database we wanted—large enough to give a realistic test but small enough to search quickly. Preliminary searches showed that large MDDR molecules caused some of the docking programs to fail, and so molecules with greater than 80 non-hydrogen atoms were deleted. Our final MDDR database consisted of 24 580 compounds.

Most of the actives were defined by keywords in the "Activity Type" field in Table 1. CDK2 is problematical because the MDDR contains only two protein kinase related activities: "Protein Kinase C Inhibitor" and "Tyrosine-Specific Protein Kinase Inhibitor". Neither is particularly specific to CDK2-type kinases, but we found better results with the first. Since there are no NEURAMINIDASE inhibitors or protein tyrosine phosphatase, nonreceptor type 1 (PTP1B) inhibitors explicitly in the MDDR, we generated

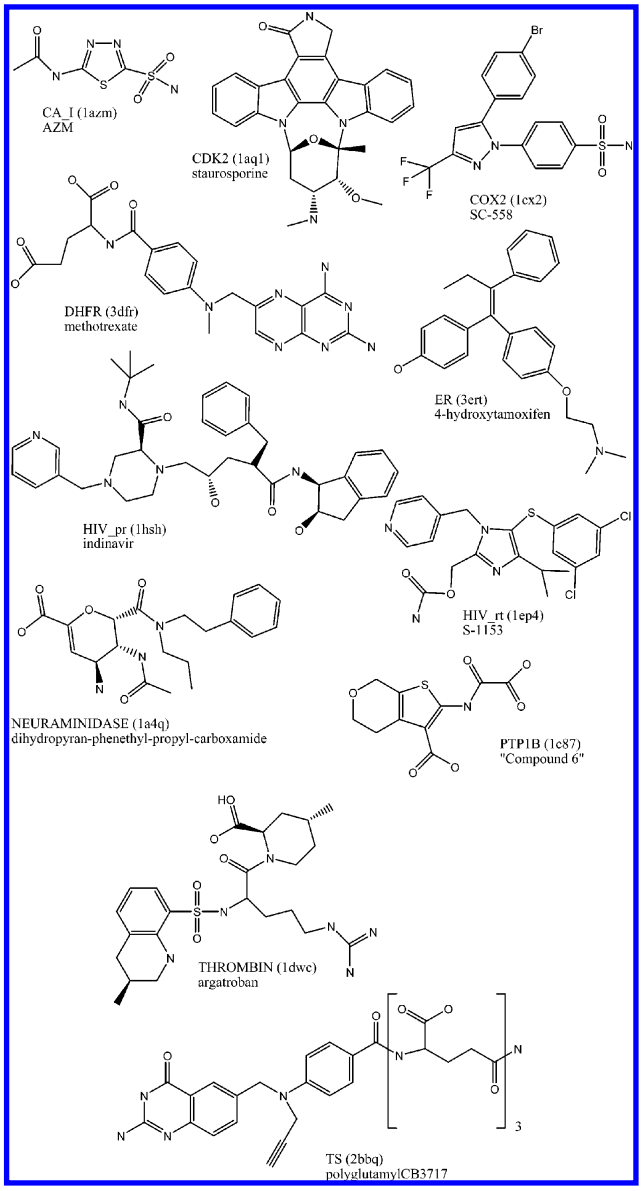


Figure 1. The ligand targets used for the ligand-based methods.

our own lists. For NEURAMINIDASE, we looked at "Antiviral" compounds which had an AP Dice similarity of 0.6 to any of the neuraminidase inhibitors in the PDB. For PTP1B we looked at "Antidiabetic" compounds similar to PTP1B inhibitors in the PDB using an AP Dice similarity of 0.6.

Source for MCIDB Database. The MCIDB database was constructed with randomly selected compounds from the MCIDB plus a combined set of actives chosen by in-house experts for each target. Usually the actives were those molecules where the $IC_{50} < 1\ \mu M$ for the primary in vitro assay. Again, molecules with greater than 80 non-hydrogen atoms were eliminated, and the molecules were clustered with the Butina algorithm as described above. The total number of MCIDB compounds was 9869. We have no known in-house actives for DHFR, NEURAMINIDASE, and TS, so those searches were not done.

Topological Ligand-Based Screening Methods. These methods use only the connection table of the target molecule and ignore the coordinates. Both methods make use of a database of precomputed descriptors.

TOPOSIM. Our favored method of topological similarity searching is TOPOSIM. The method calculates the Dice similarity between a ligand target and every molecule in the database. The frequency of the descriptors is taken into account. Here we use the AP descriptor.[22]

Daylight. Similarity using Daylight fingerprints is a de facto standard because of the great popularity of the tools provided by Daylight Chemical Information Systems (Sante Fe, NM, www.daylight.com). Daylight uses Tanimoto similarity based on the presence or absence of path-based substructures. Here we used the 3/10 paths (3−10 bonds or 4−11 contiguous atoms), which is considered good for calculating similarities of drug-sized molecules.[28] However, we also looked at the 0/7 paths (1−8 contiguous atoms), which is the default in the Daylight toolkit.

3D Ligand-Based Screening Methods. These methods use the 3D coordinates of the ligand and calculate a similarity to the 3D structures in the database. Here we use the probe ligand in its receptor bound conformation from the X-ray crystal structure. For all the 3D ligand-based methods described in this manuscript, pregenerated, explicit conformations of each molecule are examined. Different methods use a different default number of conformations, and we kept the default in each case. At run time, the similarity to the target was calculated for each conformation, and the similarity assigned to a database molecule was the similarity of the conformation most similar to the target.

SQW. SQW is an update of our SQ method.[14] Non-hydrogen atoms in molecules are assigned as one of seven atom types which represent the following chemical features: cation, anion, hydrogen bond donor, hydrogen bond acceptor, polar, hydrophobic, and "other". The SQ score reflects the superposition between two molecules and is the sum of a matching function over each pair of atoms from the two molecules. The matching function takes into account the similarity of the types of each atom and the distance between them. SQW works in two stages: (1) given a target and a conformation of a candidate in the database, a clique-matching algorithm finds many initial orientations of the candidate onto the target, (2) then, the Nelder-Mead simplex algorithm, starting with the best initial orientation, moves the candidate so as to maximize its score. SQW can use "essential" points to specify atoms in the target that must be matched. For this paper we did not use any essential points to be consistent with the other methods which do not have any such constraints.

For SQW we use a database with a maximum of 25 conformations per molecule, which we refer to as a flexibase. The procedure to generate a flexibase is discussed in detail elsewhere.[29] Currently, we generally use a knowledge-based method[15] instead of a distance geometry method to generate diverse, low-energy conformations.

In our original publication we used the raw SQ score to rank database entries against the target. In this work we calculated a Dice-like similarity score:

$$\text{similarity} = \frac{2 * \text{SQscore(target,candidate)}}{\text{SQscore(target,target)} + \text{SQscore(candidate, candidate)}}$$

This makes SQW more like the other similarity methods, which use some type of normalization.

SQW-Shape. This is the same as above, except that all atoms are set to the same type ("other"). Thus, the specific chemistry of molecules is masked, and the similarity includes only the steric overlap, i.e., the shape.

ROCS. The ROCS program is a shape-similarity method from OpenEye Scientific Software (www.eyesopen.com) based on the Tanimoto-like overlap of volumes.[18] We used version 2.2 with the default parameters. The default maximum number of conformations per molecule is ~400, generated using Omega 1.8. However, for our MDDR and MCIDB databases, we found that, on average, 190 conformers per molecule were generated using the default options.

ROCS-color. This version of ROCS[18] uses, in addition to a shape component, a "color force field" score based on defined atom types: cations, anions, hydrogen bond donors, hydrogen bond acceptors, hydrophobes, and rings. Thus, this method is very much like SQW in philosophy. ROCS-*color* uses the same OMEGA-generated conformation database as for ROCS. We used the default parameters.

Docking Methods. It has been noted by Cole et al.[30] and Kellenberger et al.[31] that comparing docking methods fairly is beset with difficulties. Aside from different programs having different native scoring functions, the major problem is that it is almost impossible to pose the same type of query for any two methods, there being different ways of representing databases, introducing constraints, normalizing the scores, etc. Also, all such programs contain important adjustable parameters an expert user can modify for a particular target to improve the results, and one may add target-specific constraints (required hydrogen bonds, specific water molecules, etc.). Philosophically the question becomes whether one should follow the developer's recommendations as closely as possible for each method or make all the methods act as similarly as possible. Is it more realistic to make target-specific adjustments and constraints (as an expert might on a well-known target) or not to make them (as if for a new target)? Our choice was to follow the out-of-the-box defaults and recommendations for each method as closely as possible but, on the other hand, not to make target-specific adjustments or add target-specific constraints even in the case where that is the usual practice. Our main reason for that is because the topological and shape-based methods usually do not employ such constraints.

FLOG. FLOG[13] is our in-house docking routine. Since our original publication it has undergone a number of implementation changes and minor improvements. For instance, we now use the same computational engine as SQW, and we use a softer Morse potential function for the van der Waals instead of the Lennard-Jones potential function. However, the philosophy has not changed in that the binding site is still represented as a set of grids encoding the score value for each of the seven previously described SQ atom types. Also, one must have a set of "match centers" representing the volume of the binding cavity. These are used by a clique-matching algorithm to generate initial orientations of candidate conformations from the database. One feature of the FLOG scoring grid, unique to the docking methods used here, is that the position of polar hydrogens (e.g., hydroxide groups in serine) and/or the tautomeric state of some side chains (e.g., histidine) can be left ambiguous if not enough information is available to specify them. In addition, FLOG is not very sensitive to formal charges.

FLOG grids were generated to extend 5 Å around the maximum extent of the ligand, and match centers were generated at the grid maxima. The general practice with FLOG is to assign one or more match centers as type "essential". For example, one could require that a pocket be occupied by a cation from a candidate molecule in the database. However, to be consistent with the philosophy described above not to use target-specific constraints, we made one match center (usually at the global maximum for the grid) a generic "essential" (e.g., it was required to be matched by any atom in the candidate). This condition biases which part of the active site is to be occupied but does not specify by what kind of ligand atom.

The same multiconformer database was used for FLOG as for SQW. Since most docking functions score larger molecules more favorably, most docking programs use a method to normalize the score. Here we divide the raw FLOG score by the cube root of the number of non-hydrogen atoms in the ligand as suggested by Pan et al.[32]

Glide. Glide is a docking program distributed by Schrödinger (www.schrodinger.com).[19,20] We used version 3.0. Because it uses explicit electrostatics, Glide is very sensitive to formal charges, and we made some appropriate modifications regarding the formal charges and hydrogen positions of the receptor (for instance, the two aspartates in HIV-pr were made to have a single negative charge instead of two negative charges). Schrödinger recommends that the protein be refined in the presence of a ligand so that any close contacts are removed and that polar hydrogen atoms are required to engage in the best hydrogen-bonding network. We used the Schrödinger Maestro interface for this. Then a series of scoring grids were generated 5 Å around the ligand (the use of larger grids did not affect the results). Hydrogen-bonding constraints can be added during the generation of the grid, but we added none.

Databases were prepared by the Schrödinger Ligprep routine. Since Glide flexes candidates "on the fly", there is only one conformation per molecule stored in the database. However, Ligprep may prepare more than one tautomer or charge state per molecule.

Glide can be run in a number of modes that differ in the balance of speed versus thoroughness. We chose to run in a "high-throughput" mode, which is appropriate for database searches, being 7−10 faster than SP (single precision) mode which puts it on par with FRED in speed. Generally, Glide scores do not correlate highly with molecular weight, so normalization is not part of the Glide procedure.

FRED. FRED is a docking program by OpenEye Scientific Software (www.eyesopen.com).[21] We used version 2.0.2. FRED uses grids to score the ligand conformations. We set the grids to 5 Å around each ligand in the target of choice.

FRED uses a precomputed database of conformations; we used the same OMEGA-generated database as for ROCS. We used the consensus function with the default settings. The recommended normalization protocol for FRED is MASC.[33] This involves docking each compound in the database to multiple targets and normalizing the score of a given compound on a given target against the range of scores of that molecule on the multiple targets. In our case the targets in Table 1 constitute the "multiple targets".

Processing Scores and Goodness Measures. Let us assume n actives in a database of N molecules. The most common approach to measuring the relative merits of screening methods is to create a series of accumulation curves. This involves sorting the scores for the whole database (in ascending or descending order depending on whether good scores are more negative or positive, the best being rank 1, the next best being rank 2, etc.) and then graphing the total number of actives found as a function of the total number of database molecules screened. To do a proper comparison between methods, all the methods have to have screened N molecules. This is an issue because not all methods can score all molecules. For instance, the default settings for Glide cause many database molecules to be skipped because they have too many rotatable bonds. Our convention is to give unscored molecules an arbitrarily unfavorable score before sorting.

If the scoring method is perfectly predictive of activity, the accumulation curve will have a slope of 1 until the number of molecules screened reaches n, and the slope will be zero thereafter. If the scoring method is of no use, the actives will be randomly scattered in the sorted list, and the curve will be a diagonal of slope n/N. One hopes the curve will be hyperbolic, indicating the front of the list is enriched in actives.

The most common way of quantifying the enrichment of an accumulation curve is to take a cutoff fraction f of the database and ask how many more actives are present in that database than would be expected by a random scattering of actives. This is called the "enrichment factor" or "enhancement factor" (EF)—the larger, the better the method. Calculating EF has a number of potential issues, one being that the maximum EF is limited by N/n and also by $1/f$, whichever is smaller. For example, at $f = 10\%$, the maximum possible EF is 10. An alternative to calculating an enrichment is to measure the area under a Receiver Operator Characteristic[34] (ROC) curve (not to be confused with the method ROCS). The major difficulty with the ROC curve is that it weights all parts of the accumulation curve equally, while in practice it is the very beginning of the curve which is important in a virtual screening experiment. Earlier[35] we suggested a "smoothed" version of the EF called the "robust initial enhancement" (RIE). Two of the authors (Bayly and Truchon) have a paper[36] that points out that the RIE has advantages relative to ROC and EF. They also propose a new metric, BEDROC, which has the advantage of both RIE and ROC. We found that the RIEs and EFs in this study were highly correlated, and therefore it makes sense here to stick with the simpler, more conventional EF.

Given that we will use some type of EF, we must choose a value of f, and this decision is for the most part arbitrary. Values of $f = 1\%$, 2%, 5%, and 10% have all been used in the literature, 10% being especially common. If f is small, then the maximum EF can be higher, giving more discrimination between methods. On the other hand, the number of compounds over which to find actives is small when f is small and EF becomes very sensitive to small changes in rank. We would argue for a small f for the following reason: In real-life virtual screening, where the databases to be searched are very large, say $N \sim 1\,000\,000$ and the number of compounds that can be submitted for screening is much smaller, say ~1000, it seems realistic to use a cutoff of 0.1%, much smaller than the conventional 10%. For the small databases we use here ($N \sim 25\,000$ and ~10 000), that would be too small (looking for actives in only a few tens of compounds), so we use 1% as a compromise.

Diversity of Actives. Given that we identify actives among the highest-scoring compounds, how should we measure their diversity? Since diversity depends on a notion of similarity, there is an unavoidable circularity when applying diversity to similarity-based virtual screening methods. Here we will define diversity for a search as the mean pairwise AP Dice similarity of actives in the top-scoring 1% of the database: the smaller the mean pairwise similarity, the more diverse the actives. One should note that in diverse databases we use here the maximum pairwise similarity is 0.7.

RESULTS

Enrichments. We ran retrospective virtual screening experiments using all methods for 11 targets over two diverse databases, the MDDR and MCIDB. Figure 2 depicts the accumulation curves for all methods over the MDDR for HIV-pr, a well-studied target. This example qualitatively demonstrates the importance of looking not at the performance over the database as a whole but putting emphasis at the very beginning of the list as we do with the EF. The shape of the Glide curve, in particular, needs some comment. Many of the actives were not scored by Glide because they contained too many rotatable bonds according to the default cutoff in Glide version 3.0. Thus, the actives were given arbitrary poor scores and ended up at the end of the list. This is reflected in the large increase in slope after 20 000 compounds. Had we used an area under the ROC curve as a metric for enrichment, we would have concluded that the ability of Glide to find HIV-pr actives is slightly worse than random. Fortunately, because the EF considers only actives found in the front of the list, HIV-pr has a reasonable EF.

Table 2 shows the EFs (at 1%) for the ligand-based methods. Table 3 contains the EFs for the docking methods. We also calculated the EFs at 10%, the RIEs at alpha = 100, BEDROC[36] alpha = 20 and the area under the ROC curve (in the Supporting Information). In general, using any of the "early recognition" metrics[36] does not substantially change the conclusions.

Most of our conclusions are drawn by looking at the mean EF over all the targets for a given database, but it should be noted that, as is commonly observed in nearly all virtual screening studies, there is no one method that is superior on every target for any given database.

Ligand-Based Methods. For ligand-based methods, TOPOSIM slightly outperformed any 3D method when averaged over our targets using the MDDR database. ROCS-*color* also does extremely well, particularly on Merck's corporate database, MCIDB. We note that the 2D- (TOPOSIM, Daylight) and 3D-ligand (ROCS-*color* and SQW) methods all do well and are statistically indistinguishable from each other. Our in-house 3D similarity algorithm, SQW, achieved the same level of enrichment as ROCS-*color* but only on the MDDR database. It is clear from the results that 3D ligand-based methods are best when using some kind of chemical typing rather than just shape (i.e., ROCS-*color* versus ROCS or SQW versus SQW-shape). The two shape-based methods (SQW-shape and ROCS) behave similarly,

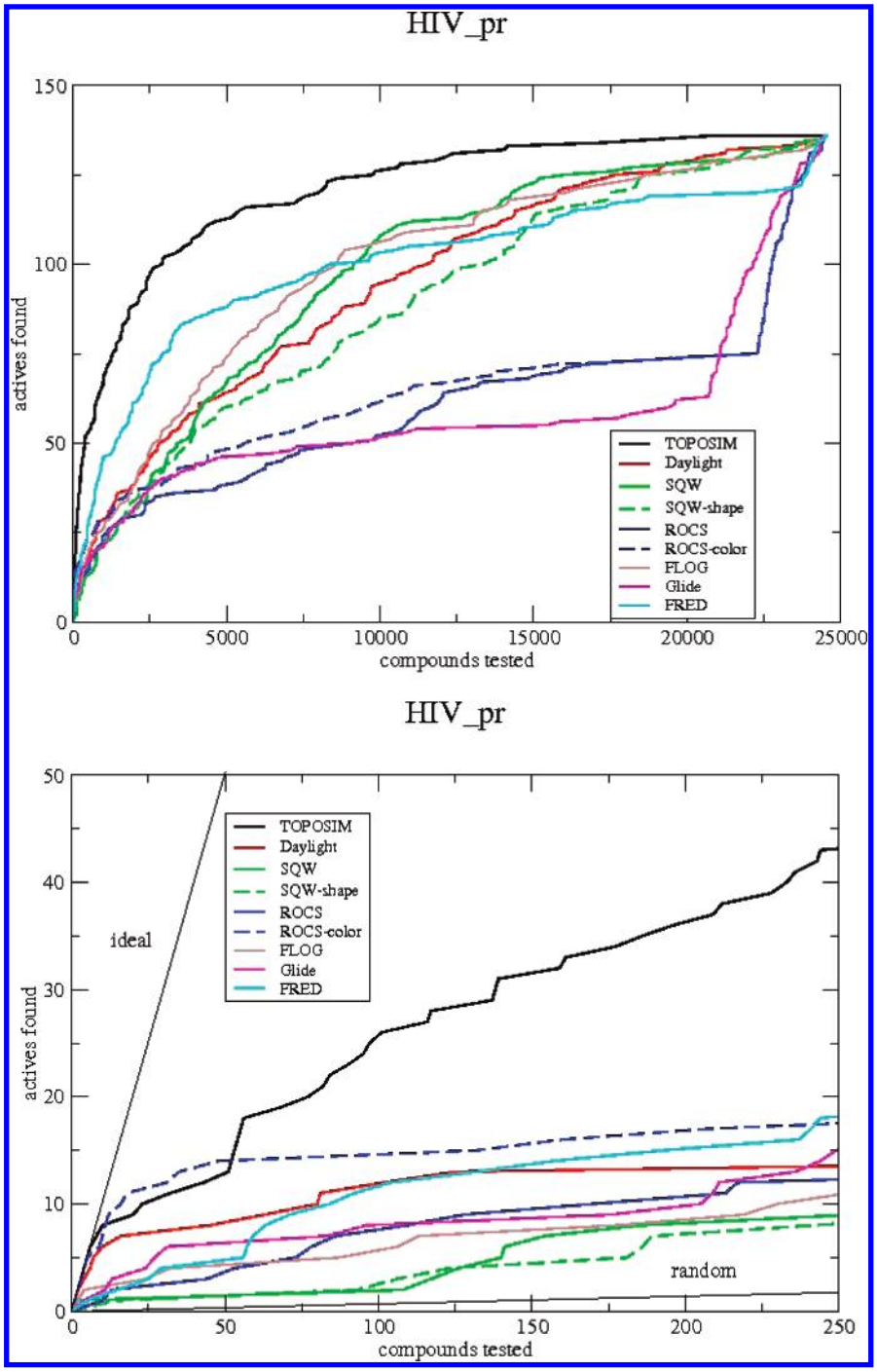


Figure 2. The accumulation curves for HIV-pr over the MDDR database. The upper figure shows the entire database, the lower the first 1% of the database. The "ideal" line shows what an accumulation curve would be for a perfect virtual screening method, with all the actives at the front of the list. The "random" line shows what the accumulation curve would be for actives distributed randomly in the list. Of particular interest in both figures are the Glide accumulation curves. In the top figure the accumulation curve has an unusual shape. The sharp increase after around 20 000 compounds reflects the fact that many HIV-pr actives in MDDR could not be scored because they contained more rotatable bonds than the default cutoff and were given arbitrary poor scores by our protocol.

not a surprise since they are different implementations of the same metric.

Given that ROCS-*color* is the best 3D ligand-based method, we examined the effect of using different numbers and sources of conformations. We applied ROCS-*color* to conformations taken from the in-house flexibases used for SQW and FLOG (using a maximum of 25 and 100 conformations) instead of its native OMEGA-generated database. Table 4 shows that the number and/or source of conformations do not seem to systematically affect the enrichment.

We note here that Daylight in its 3/10 configuration performs somewhat more poorly than TOPOSIM for most cases. There is one case (TS on MDDR) where Daylight does much better than TOPOSIM. This is explainable in retrospect. Most of the TS actives contain folate-like rings with only one or no glutamates. TOPOSIM counts descriptor frequency so molecules with one glutamate appear slightly different from the target, which has three glutamates. Daylight counts only presence or absence of descriptors, so the number of glutamates becomes less important. In its

Table 2. EFs for Ligand-Based Virtual Screening Methods

target	TOPOSIM	Day-light	SQW	SQW-shape	ROCS	ROCS-*color*
		MDDR				
CA_I	56.4	50.2	6.3	3.8	11.3	31.4
CDK2	22.2	10.4	9.1	11.7	11.7	18.2
COX2	21.1	5.1	11.3	15.2	17.2	25.4
DHFR	34.7	27.0	46.3	0.0	3.9	38.6
ER	14.9	12.2	23.0	16.3	10.8	21.7
HIV-pr	31.7	11.8	5.9	5.9	8.9	12.5
HIV-rt	2.7	3.4	5.4	3.4	4.0	2.0
NEURAMINIDASE	33.4	25.1	25.1	16.7	16.1	92.0
PTP1B	50.2	50.2	50.2	12.5	12.5	12.5
THROMBIN	28.6	13.0	27.1	4.5	6.0	21.1
TS	22.7	61.5	48.5	6.5	0.0	6.5
mean	29.0	24.5	23.5	8.8	9.3	25.6
median	28.6	13.0	23.0	6.5	10.8	21.1
		MCIDB				
CA_I	25.5	17.6	7.1	11.3	5.9	18.0
CDK2	1.9	8.7	0.0	4.8	1.9	1.9
COX2	17.1	0.0	4.0	16.1	16.1	28.2
ER	10.4	11.2	12.5	4.3	3.9	17.3
HIV-pr	19.2	14.7	11.6	3.6	9.4	18.7
HIV-rt	2.3	1.8	1.4	0.9	1.4	1.8
PTP1B	0.0	2.0	1.0	1.0	0.0	0.0
THROMBIN	8.7	3.4	6.5	1.4	0.4	5.1
mean	10.6	7.4	5.5	5.4	4.9	11.4
median	9.6	7.4	5.5	4.0	2.9	11.2

Table 3. EFs for Docking Methods

target	FLOG	Glide	FRED
	MDDR		
CA_I	1.3	1.3	2.5
CDK2	1.3	1.3	9.1
COX2	0.4	4.7	0.4
DHFR	15.4	7.7	15.4
ER	4.1	10.1	19.0
HIV-pr	7.4	10.3	13.3
HIV-rt	0.0	0.0	4.7
NEURAMINIDASE	0.0	25.1	8.4
PTP1B	12.5	62.7	12.5
THROMBIN	4.0	14.0	6.5
TS	9.7	9.7	12.9
mean	5.1	13.4	9.5
median	4.0	9.7	9.1
	MCIDB		
CA_I	0.0	0.0	0.8
CDK2	1.0	3.9	5.8
COX2	0.0	0.0	0.0
ER	3.0	7.3	7.8
HIV-pr	8.5	9.8	7.6
HIV-rt	0.8	1.4	2.3
PTP1B	17.6	38.1	2.9
THROMBIN	1.0	11.7	8.5
mean	4.0	9.0	4.5
median	1.0	5.6	4.5

default 0/7 configuration, Daylight does poorly compared to the 3/10 configuration (mean EF 19.8 for MDDR and 3.6 for MCIDB versus 24.5 and 7.4, respectively).

Docking Methods. Averaged over all the targets, the docking methods, Glide, FRED, and FLOG, are consistently worse than the ligand-based methods. Some docking targets (e.g., CA_I, COX2) seem particularly difficult. FLOG is definitely the poorest docking method among the three and Glide is superior when examining the mean EF values over all targets, but the mean EF value for Glide is skewed by its superb performance on PTP1B. If we eliminate PTP1B and NEURAMINIDASE from the MDDR analysis, then FRED appears better than Glide (mean EF 9.3 and 6.6, respectively). We note again, however, that in this evaluation Glide was run in HTVS, its fastest and least accurate mode. Glide performs better running in "single precision" (SP) mode (data not shown).

Individual docking methods can do better on specific targets. It is interesting that FRED does better on CDK2 and HIV-rt. Even FLOG is superior in at least one case, DHFR. That is due to the definition of atom types. FLOG is the only docking method here that allows an ambiguous atom type of either hydrogen bond donor or acceptor for the N1 atom in the pteridine and diaminotriazole rings of DHFR ligands. Thus the N1 can interact favorably with the (deprotonated) active site aspartate. On the other hand, the other methods require explicit protonation states; in those cases the N1 is deprotonated. For NEURAMINIDASE and PTP1B, Glide outperforms FRED and FLOG. Both of those targets contain very polar active sites, and Glide *may* perform better on such active sites. However, a larger sample size would need to be examined to draw a firm conclusion.

Differences between Databases. Finally, we note that the results for the two databases, MDDR and MCIDB, can be very different. MDDR often appears better for a number of individual targets and methods. The methods where the differences are the most extreme are ligand-based methods: TOPOSIM, Daylight, and SQW. The two targets for which there are the most extreme cases are CDK2 and PTP1B. For example the EF for PTP1B TOPOSIM is 50.2 for MDDR but 0.0 for MCIDB. These are explainable in retrospect. The target for CDK2 is staurosporine. Many of the MDDR actives are molecules that contain parts of staurosporine, while none of the MCIDB actives do. Similarly, the 3-carboxythiophene ring in the PTP1B target is seen in a number of the MDDR actives but not in any of the MCIDB actives.

Comparisons to Enrichments in Other Studies. The results from our docking study can be compared with those in the literature for common targets by calculating the EF at 10% for the MDDR (see the Supporting Information). Given that ours is the first published study reporting results for Glide in HTVS mode, we would expect our enrichments calculated for Glide to be systematically lower than those from earlier studies. In reality this is not necessarily the case.

Our own study controls the low influence of active and decoy selection on enhancements since our EFs at 10% for MDDR and MCIDB across a diverse set of targets are so similar, in contrast to what is seen for the ligand-based methods. Thus, we cannot attribute differences between our results and those in the literature to factors such as active and decoy selection in our study. For COX2 MDDR results, our study obtained enrichment factors of 1.9 and 2.9 for FRED and Glide, respectively, compared with 6.4 and 6.2 from Schulz-Gasch and Stahl.[37] For ER, our study found enrichment factors of 3.4 and 2.3 for FRED and Glide, compared with again systematically higher values of 7.9 and 7.5 from Schulz-Gasch and Stahl.[37] Chen et al. reported an enrichment of 3.9 for Glide, closer to our value.[2] For NEURAMINIDASE, our study found enrichments of 3.3 and 8.3 for FRED and Glide, respectively, compared with 8.5 and 9.7 from Schulz-Gasch and Stahl.[37] In this case, the difference in enrichments observed in our study between the two programs was not found by the other investigators. For thrombin, enrichment values of 2.8 and 5.2 were obtained for FRED and Glide, as compared to the systematically higher values of 7.8 and 8.7 from Schulz-Gasch and Stahl.[37]

Table 4. EFs for ROCS-*color* Using Different Sources for Ligand Conformations

target	MDDR OMEGA	MDDR 25 confs	MDDR 100 confs	MCIDB OMEGA	MCIDB 25 confs	MCIDB 100 confs
CA_I	31.4	36.4	37.6	18.0	24.2	24.7
CDK2	18.2	20.8	23.5	1.9	2.9	2.9
COX2	25.4	24.6	25.0	28.2	28.2	27.2
DHFR	38.6	27.0	30.9	–	–	–
ER	21.7	27.1	24.4	17.3	21.6	20.3
HIV-pr	12.5	11.8	11.1	18.2	15.2	16.9
HIV-rt	2.0	1.3	1.3	1.8	0.9	0.5
NEURAMINIDASE	92.0	83.6	83.6	–	–	–
PTP1B	12.5	37.6	37.6	0.0	2.0	2.0
THROMBIN	21.1	18.1	18.6	5.1	4.7	4.7
TS	6.5	32.4	19.4	–	–	–
mean	25.6	29.3	28.5	11.3	12.5	12.4
median	21.1	27.0	24.4	11.2	10.0	10.8

In light of the fact that, in our hands, active and decoy sets give overall similar results, it suggests that these differences can be attributed to the method. Further Cummings[6] et al. and Chen[2] et al. obtained enrichments of 8.0 and 9.8 using Glide. Finally, for PTP1B, our study yielded enrichment values of 1.2 and 6.2 for FRED and Glide, respectively. Other enrichments reported for Glide include 1.6 (Chen[2] et al.), 6.0 (Cummings[6] et al.), and 8.4 (Klon[3] et al.). The similar Glide results reported for Cummings and Klon relative to our work demonstrates that changing the default values is not a necessary means to achieve higher enrichments. Rather, we are confident that utilizing the default values is an acceptable procedure for evaluation purposes. Although Muegge and Enyedy[5] have demonstrated the strong effect of the underlying database on the calculated enrichments, we believe such differences are attributed to the inherent methodology, not to the database. One would hope that trends would be consistent across different studies, and this is clearly not always the case, even taking into account the use of Glide in HTVS mode in this study. Unfortunately the use of proprietary protein structures and ligands in most of these studies prevents reproducing the results, and thus it is hard to know the source of the disagreement.

Diversity of Actives. Finally we show the diversity of the actives selected by all the methods. As examples, we show in Figures 3−5 the first five and last actives in the top 1% of the database for COX2, DHFR, and HIV-pr for all methods. Even in the 2D similarity methods where one might expect that the actives at the front of the list should be "obvious analogs", we see different scaffolds.

One can quantitate the diversity by the mean pairwise similarities, which are in Table 5. We list them for only those target/method/database combinations where there are at least 10 pairs of actives in the top-scoring 1% of the database. We have included only those target/database combinations where there is at least one docking method that meets that criterion. For each target/database combination we list in the last column the mean pairwise similarity of all actives, which can be considered a lower limit to the pairwise similarity attainable by any specific method. The last column is almost always the lowest number in each row, indicating that no method is capturing all the possible diversity among the actives. Since we use the same similarity definition for TOPOSIM and for the diversity, it is not at all surprising that for each row the TOPOSIM number is often the highest; compounds more similar to the same ligand target will tend to be similar to each other. However, even the TOPOSIM mean pairwise similarities are far lower than the expected maximum of 0.7, far lower than the similarity expected for analogs would be (∼0.65), and seldom much greater than the other numbers in the same row. There seems to be no systematic trend for ligand-based 3D methods versus docking. Thus, at least for the databases we use here, where close analogs have been eliminated, there is no evidence that certain methods (e.g., docking) can systematically select much more diverse actives than others. This is in general agreement with the visual impression in Figures 3−5.

DISCUSSION

As with any retrospective virtual screening study, one must apply a number of general caveats to interpret the results as applied to real world applications. The first is that the results depend strongly on the exact choice of target and the database, such that it is very hard to compare one study with another. Also in practice, chemists do not blindly take the top-scoring hits from any one method but typically apply visual inspection to triage them, and this may greatly change the apparent enrichments. Having made these caveats we note the following trends:

Topological Methods and Lead-Hopping. As measured by EF, the 2D similarity methods (TOPOSIM, Daylight) perform well at lead-hopping when applied to a diverse database. Despite recent caveats raised around the idea that "similar molecules give rise to similar activities",[38,39] topological similarity is very useful in predicting activity, even in the situation here where "similar" does not necessarily mean "close analog". This is notable since such methods are by far the simplest to understand, the least computationally expensive, and require the least adjustment of parameters by the user. Similar observations have been made by us[40] and others.[41,42] One must raise the caveat that the 2D methods find actives that are not quite as diverse as found by 3D methods, but the difference in diversity is not necessarily large. One may ask how it is possible for 2D similarity methods to perform nearly as well as 3D methods at lead hopping.[40] Active molecules may be similar enough to the target by some 2D descriptor that they are more common than expected among the decoys but not be so similar that they are close analogs of the target. Thus it is possible for 2D similarity searches to hop to another series and have a high enrichment in a diverse database. Different 2D descriptors vary in this ability, and the AP descriptor seems good in this respect. This has consequences for the practice of

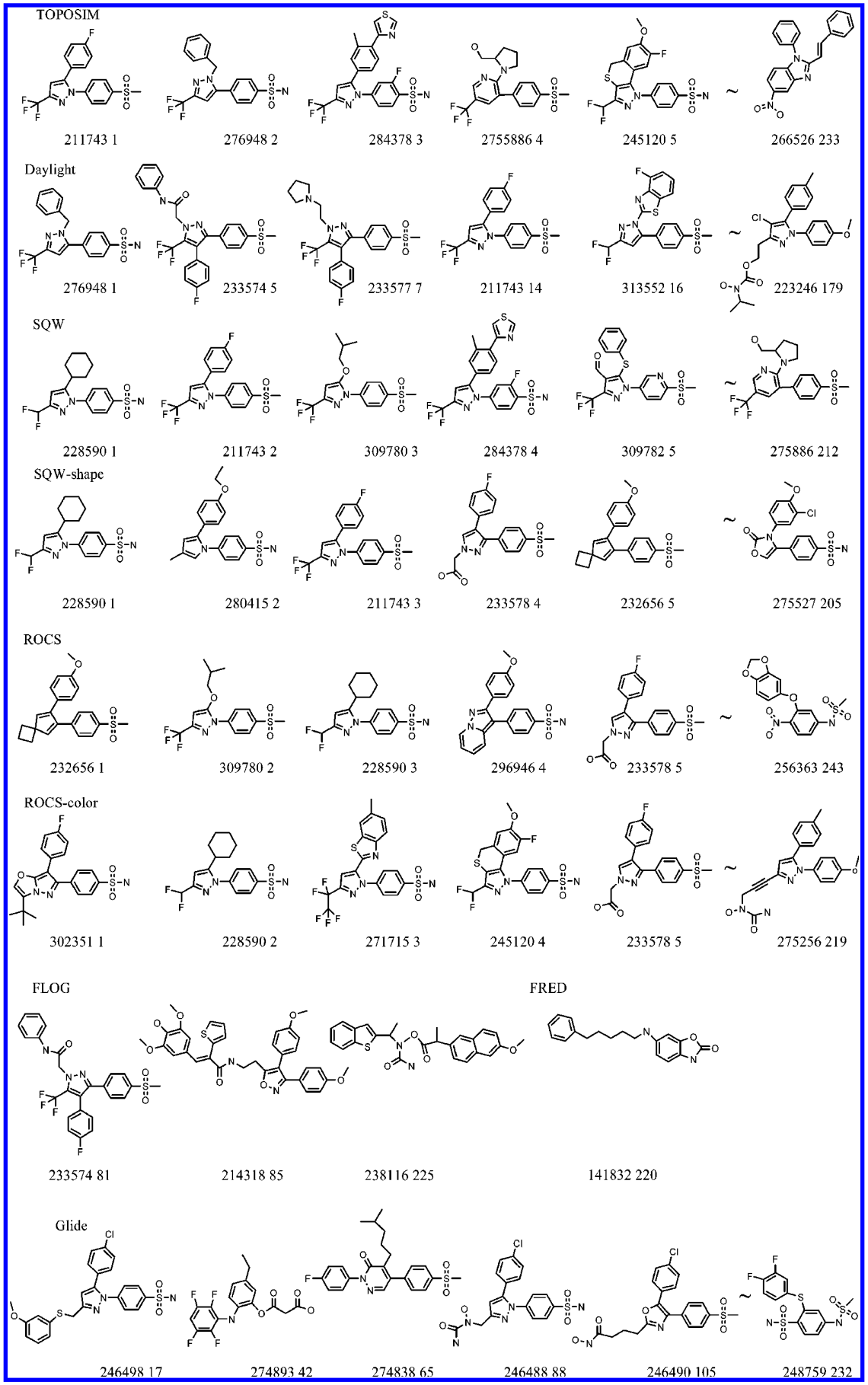


Figure 3. MDDR COX2 actives from the nine methods. We show the first five and the last actives in the top 1% of the database. The first number is the MDDR molecule identifier and the second number is the rank for the method. For some methods, there may be fewer than 6 actives in the top 1% of the database.

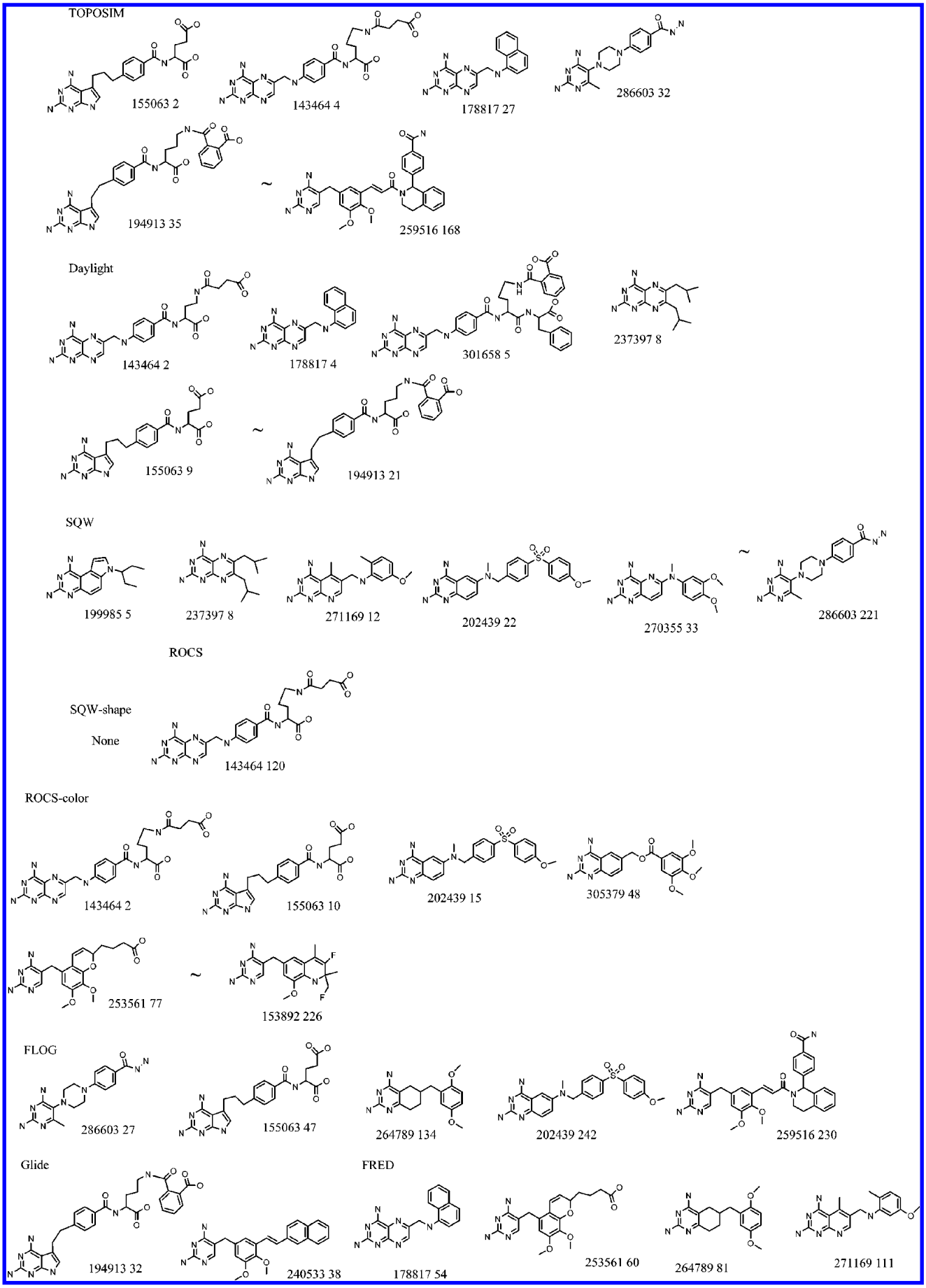
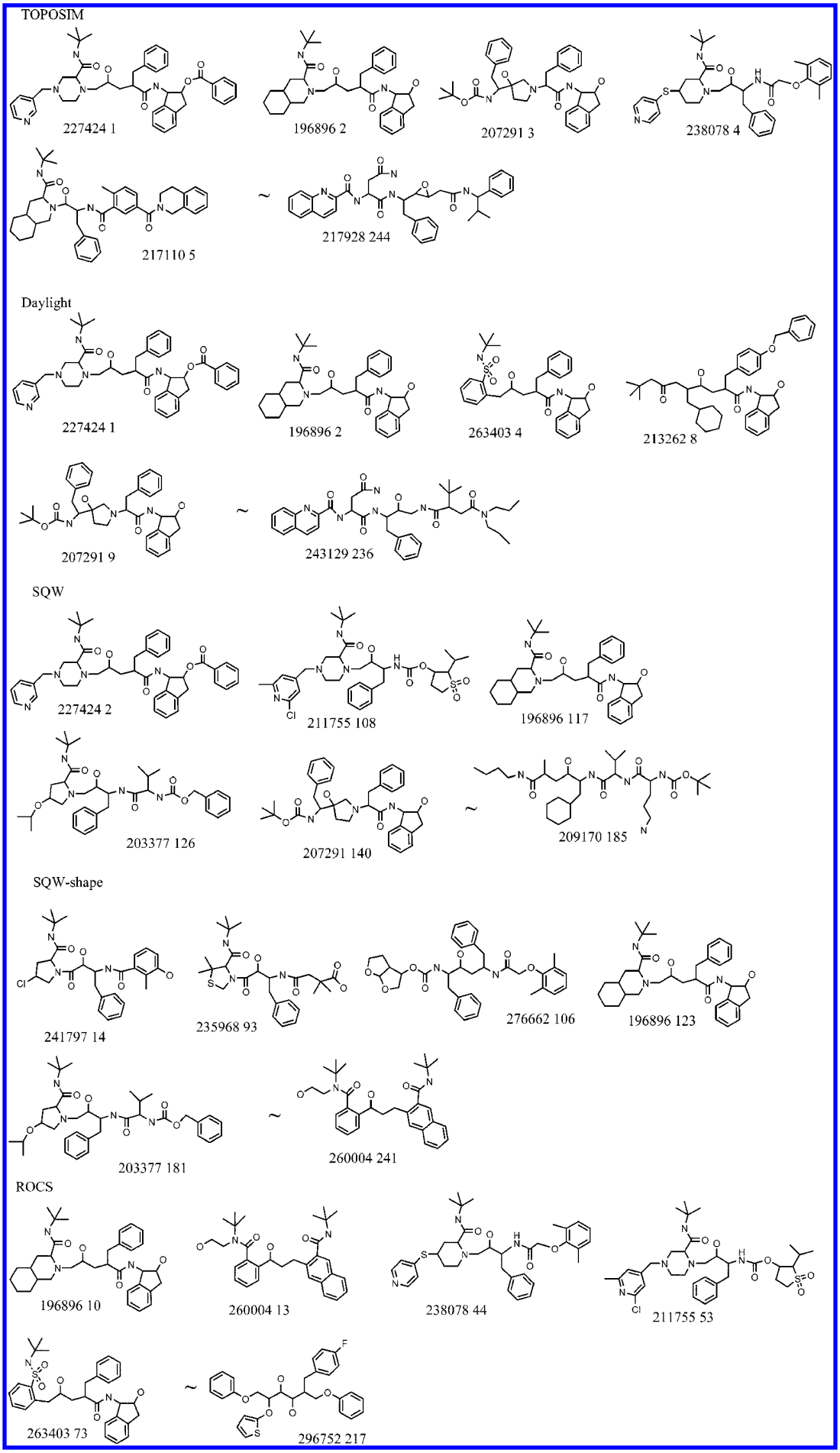


Figure 4. MDDR DHFR actives from the nine methods. It should be noted that all of the DHFR actives in the MDDR have the 2,4-diaminopyridine group.


TOPOSIM
227424 1
196896 2
207291 3
238078 4
217110 5
~
217928 244
Daylight
227424 1
196896 2
263403 4
213262 8
207291 9
~
243129 236
SQW
227424 2
211755 108
196896 117
203377 126
207291 140
~
209170 185
SQW-shape
241797 14
235968 93
276662 106
196896 123
203377 181
~
260004 241
ROCS
196896 10
260004 13
238078 44
211755 53
263403 73
~
296752 217

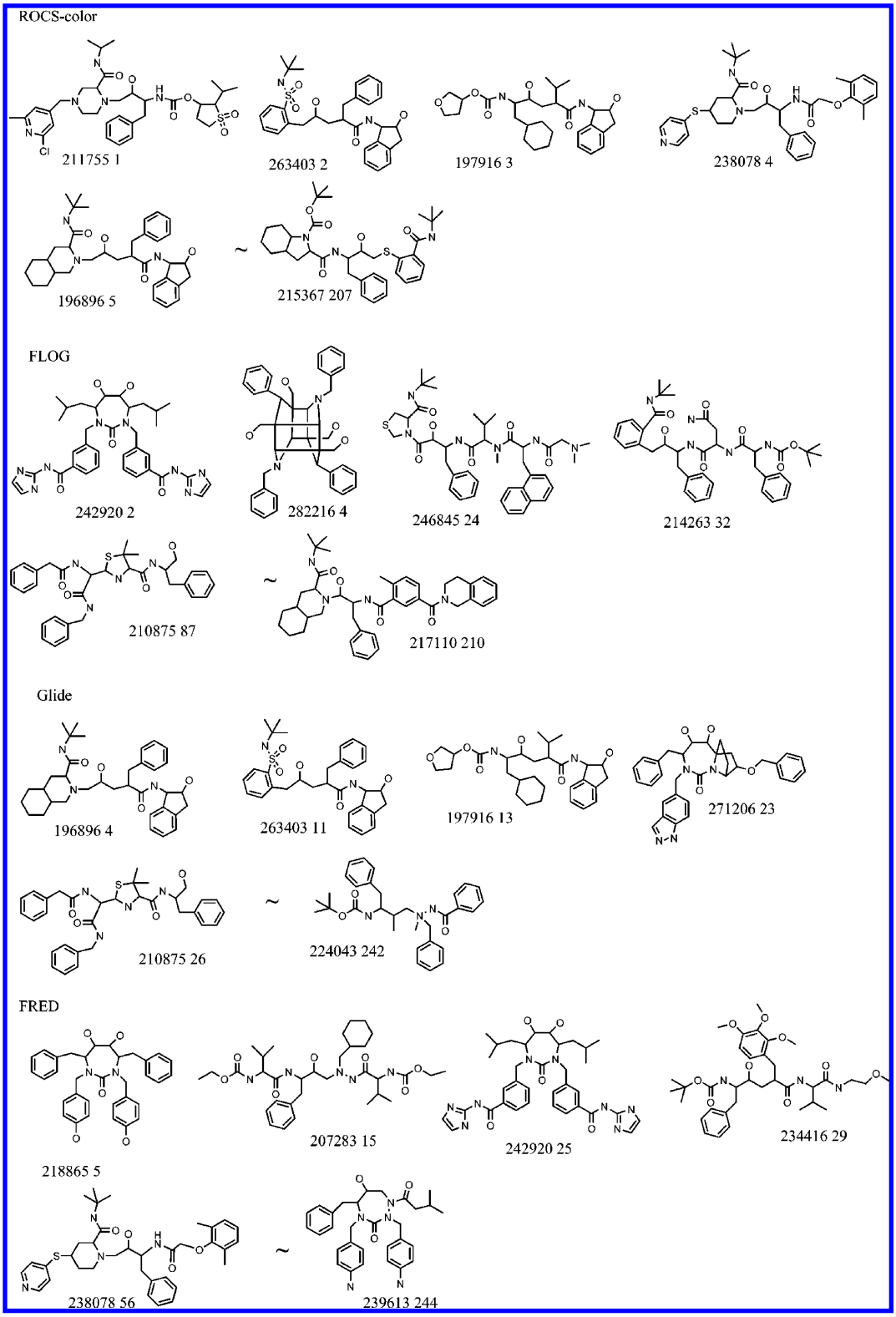


Figure 5. MDDR HIV-pr actives from the nine methods.

similarity searching. Whereas Merck chemists might typically not look at compounds that had a similarity to a target less than 0.6 by AP Dice (well below where the compounds are analogs), our results suggest reasonable enrichments as low as 0.5.

In this context we note that Daylight is often used in the literature as a standard to compare 2D versus 3D methods or to validate various variations of 2D methods (e.g., reduced graphs) aimed at enhancing lead-hopping. Often authors declare some new method superior because it can beat Daylight. However, some investigators have reported that Daylight does very poorly at lead-hopping.[40−42] During the course of this work we came to appreciate how sensitive the apparent goodness of Daylight is to the path length of the descriptors, and there are some configurations of Daylight (e.g., the default 0/7) that are too easy to outperform, making them poor standards for 2D similarity. Very often it is not clear what configuration is being used in any given study,

Table 5. Pairwise Similarities of Actives in the Top 1% of the Sorted Lists Where at Least 10 Pairs Are Present

target	TOPOSIM	Daylight	SQW	SQW-shape	ROCS	ROCS-*color*	FLOG	Glide	FRED	all actives
MDDR										
CDK2	0.46	0.48	0.51	0.41	0.41	0.38			0.45	0.24
COX2	0.44	0.46	0.39	0.42	0.41	0.41		0.35		0.28
ER	0.47	0.40	0.32	0.38	0.44	0.38		0.37	0.32	0.25
HIV-pr	0.48	0.45	0.50	0.49	0.44	0.43	0.41	0.41	0.38	0.34
PTP1B	0.54	0.54	0.54					0.41		0.35
THROMBIN	0.46	0.43	0.39	0.38	0.46	0.42	0.38	0.38	0.40	0.33
mean	0.48	0.46	0.44	0.42	0.43	0.40	0.40	0.38	0.39	0.30
MCDIB										
CDK2	0.65	0.50		0.50				0.35	0.49	0.35
ER	0.49	0.38	0.35	0.46	0.52	0.42	0.53	0.35	0.33	0.30
HIV-pr	0.50	0.43	0.46	0.47	0.42	0.42	0.43	0.44	0.45	0.40
HIV-rt	0.49	0.33						0.23	0.38	0.28
PTP1B		0.64					0.37	0.36		0.31
THROMBIN	0.49	0.38	0.38	0.36		0.42	0.33	0.40	0.41	0.36
mean	0.52	0.44	0.40	0.45	0.47	0.42	0.42	0.36	0.41	0.33

and we urge authors using Daylight to be explicit in this regard.

Despite our observations in the past that 2D methods yield higher enrichment rates than any 3D method,[40] in this paper we note that ligand-based 3D methods (SQW and ROCS-*color*, especially the latter) are nearly on a par with TOPOSIM. It has been suggested that TOPOSIM using AP descriptors may have a slight unfair advantage relative to the other ligand-based methods in retrieving actives given the fact that we used AP similarity in preselecting the actives and decoys. However, the selection was done within the database only, without regard to any of the targets. Also TOPOSIM with the TT descriptors[22] gets very similar results (not shown) and Daylight does well also. A more reasonable case could be made that selecting NEURAMINIDASE and PTP1B actives for the MDDR using topological similarity could make the topological methods look better, but the best EF for NEURAMINIDASE in the MDDR is for SQW and Glide does spectacularly well on PTP1B. It is probable that TOPOSIM (or any 2D method) has an advantage when many close analogs are present in a database, and using a diverse database has removed that advantage. Of course, individual cases vary greatly from the general rule. In the MDDR database results, the TOPOSIM EF for CA_I is much better than SQW, but in DHFR, ER, and NEURAMINIDASE MDDR, SQW is better than TOPOSIM. Among the 3D ligand-based methods, clearly shape alone does not encode enough information to discriminate actives from inactives; one must add chemical character to the atoms.

3D Ligand-Based Methods and Conformers. In regard to ligand-based 3D methods, we see that the source of conformers, or the number of conformers, has little effect on the EF for ROCS-*color*. This is consistent with some of our unpublished observations for SQW as well. Others have observed that enrichment is not sensitive to the number of conformations for docking methods, in particular FRED[43] and DOCK.[44]

Comparing Docking Methods. We concur with Cole et al.[30] and Kellenberger et al.[31] that comparing docking methods is a problematic endeavor. Results reported in the literature are very sensitive to the exact target, the presence or absence of constraints, the method of normalization, the composition of the database, etc., and different methods may differ in their sensitivity. In this paper we chose not to use problem-specific constraints, but we note in passing that such constraints can have a very different effect depending on the method. For instance, FLOG performs much better, on par with FRED, when "essential point" constraints are added. Glide results on CA_I did improve when we used a constraint to the catalytic Zn; however, most other targets did not show great improvements with hydrogen bond constraints. The default filters that affect which molecules can be scored are also important. For instance some of the defaults (number of rotatable bonds and atoms) in Glide 3.0 negatively impact the results for HIV-pr. Good FRED enrichments are dependent on MASC[33] normalization; uncorrected FRED scores result in much lower enrichments in most cases.[45] On the other hand, FLOG scores, once normalized for molecular weight, show no further improvement in enrichment with MASC-like correction.[45]

Ligand-Based versus Docking Methods. Our results seem to support the idea that docking is generally poorer at selecting actives than most ligand-based methods (2D or 3D) as measured by enrichment. This is especially significant given the computational cost of docking compared to ligand-based methods. At least some of the literature appears to agree on this point. Zhang and Muegge[41] comparing AP similarity, 3D fingerprints, and Glide on a different set of targets and databases reached the same conclusion. Also, Hawkins et al.[46] showed that ROCS-*color* was superior to the docking method SURFLEX, again on a different set of targets and database. An apparent disagreement is by Chen et al.[2] who found that ROCS performed on par with Glide. The results in this paper agree that ROCS and docking methods can have similar enrichments, but we also note that had they used ROCS-*color* the conclusions of these authors would have probably favored ligand-based methods. The conclusions in the Chen et al. paper should be read with an accompanying letter to the editor by Perola et al.[47]

We do wish to raise a qualification about the above conclusion. In our original study design, we first chose the protein as the docking target, took its cocrystallized inhibitor as a ligand-based target, and used these as the basis of comparison for ligand-based versus docking methods. This is consistent with the usual practice in the literature. However, an anonymous reviewer raised the valid point that linkage is somewhat arbitrary and, just as docking results for any given target are sensitive to exactly which crystal structure is used,[25] which ligand one uses might change the relative goodness of ligand-based methods versus docking

drastically. In a study to be published elsewhere we chose up to 5 total protein−ligand complexes of the same target and compared TOPOSIM and ROCS-*color* on the ligand to Glide on the protein. While it is true that for any given protein−ligand pair for a given target it is nearly impossible beforehand to guess whether a ligand-based or docking method will do better, ligand-based methods still do better averaged over a number of targets. (A similar situation is seen in the examples presented here.) Thus we believe our conclusions hold overall. We agree with the concluding statement of Zhang and Muegge[41] that "knowledge about active ligands for a drug target [query molecule] can be as valuable as a crystal structure for obtaining novel scaffolds from virtual screening". It should be noted that we are running the ligand-based methods in the simplest manner, using only one target ligand per search. It has been shown that better enrichments and better diversity can be obtained using more than one target simultaneously, at least for topological similarity.[48]

One would naively expect that having a crystal structure for a protein would give a chemist all the information needed for predicting what molecules would bind, much more than knowing the ligand. So why is this not true in practice? Tirado-Rives and Jorgensen recently suggested why scoring functions currently used in docking are fundamentally problematic at predictive binding to a real receptor.[49] Scoring functions do not account for what they call "conformer focusing" which is the free-energy change for the ligand adhering to the protein bound conformation.[49] Additional manuscripts concerning docking and scoring have recently appeared in the *Journal of Medicinal Chemistry*, and we encourage interested readers to review the multiple papers presented therein.[50] We do want to raise some additional points lest we give the false impression that we believe docking is not a worthwhile endeavor:

1. Here we are concerned only with the retrieval of actives via virtual screening and are ignoring other potential virtues of docking methods such as predicting the binding mode of a ligand.

2. Here and in many other papers, target-specific constraints are not used. Docking methods can achieve much better enrichments by tuning adjustable parameters, adding specific receptor-bound water molecules, by requiring that certain hydrogen bonds be formed, etc. Once enough additional information about a target is available, docking can make use of it.

3. Docking methods are in a constant state of development, for instance see ref 51, and it is possible that the limitations of current scoring functions will eventually be overcome.

Differences between Databases. One aspect we did not appreciate before is how different the results from two different databases could be. These differences are large for ligand-based methods, especially the ones that give a great deal of weight to local chemistry, e.g., TOPOSIM and Daylight. The difference in the docking methods is not as large. The differences are seen where MDDR contains compounds more closely related to the target than does the MCIDB. Given that the MDDR contains compounds from a large number of corporate sources, it would not be so surprising that the MDDR would sometimes have compounds closer to the target. In any case, the sensitivity of virtual screening methods to the exact composition of the database makes comparing studies from the literature very unreliable because the studies use different databases. Huang, Shoichet, and Irwin have compared their DUD database to several others (including the MDDR) and found that a poorly designed database can skew the results to make a docking method look better than it would with an appropriately designed database.[52] Our study uses two completely independent sources of both actives and decoys, specifically the MDDR and our corporate database, to examine this issue; while these databases have not been deliberately engineered to be difficult they do represent two independent and realistic examples of databases for use with docking methods. That a given docking method gives similar results for the same target using two different databases is strong support that we are truly testing the docking methodology without "skewing" from poor database design. We propose this as a good general concept: to evaluate a docking methodology based on appropriately designed active and decoys sets from two completely independent sources and show, by their similar results, that the results are not skewed by either database.

CONCLUSION

In closing, one may be led to believe that simpler is better. A simple topological method is clearly best if the only criterion is the number of actives found in the least amount of computational time. This paper justifies the utility (and necessity) of including chemical typing in ligand-based 3D searching. Additionally, we have structured our study to include both our in-house, proprietary compounds and publicly accessible compounds (MDDR external registry numbers (regnos) and rank order for each method against all targets are listed in the Supporting Information) in the hope that others can reproduce and expand on our results. In this study, we have demonstrated that sophisticated algorithms (e.g., 3D ligand and docking methods) are able to select more diverse actives than 2D methods. This result justifies the use of these computationally more expensive methodologies. Finally, 3D ligand based methods exhibit superior overall performance compared to docking methods due to their higher enrichment of actives with comparable diversity.

ACKNOWLEDGMENT

We thank Schrödinger for providing the special HTVS version of Glide for this evaluation, which was not available in the commercial release at the time this work was done. As of Glide version 4.0 this feature is available in the general release. We also thank OpenEye Scientific Software for allowing us to utilize their software for this study. The coauthors are grateful for the expert opinions rendered by the following computational chemists who were members of a company-wide (Merck) working group focused on docking and scoring: Brad Feuston, Kate Holloway, Peter Hunt, Toshiharu Iwama, Simon Kearsley, Uwe Koch, Kiyean Nam, Becky Perlow-Poehnelt, Stephane Spieser, and Chris Swain. We also thank Rich Bach, Gene Fluder, Joseph Forbes, Dennis Schuchman, Joseph Shpungin, Jerry Verlin, and Matt Walker for systems and software support. In addition, Paul Hawkins of OpenEye Scientific Software and Woody Sherman of Schrödinger are kindly acknowledged for their assistance in assuring the software was used correctly and a

special thanks to Tom Rush of Merck (MRL, Boston) for a careful read of the manuscript and suggestions.

Supporting Information Available: List of MDDR compounds forming the database; for MDDR, the ranking of each active for each target for each virtual screening method; and tables of the MDDR and the MCIDB results expressed as RIE, enrichment factor at 1% and 10%, BEDROC@α=20%, and the area under the ROC curve. For the ROC area, 0.5 indicates "random" performance and 1.0 indicates perfect performance. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hartman, G. D.; Egbertson, M. S.; Halczenko, W.; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D.; Lynch, R. J.; Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *J. Med. Chem.* **1992**, *35*, 4640−4642.

(2) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.

(3) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem*. **2004**, *47*, 2743−2749.

(4) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.

(5) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem*. **2004**, *11*, 693−707.

(6) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem*. **2005**, *48*, 962−976.

(7) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem*. **2005**, *49*, 5912−5931.

(9) *Virtual Screening in Drug Discovery*; Shoichet, N., Alvarez, J., Eds.; CRC Press: 2005.

(10) Cornell, W. D. Recent Evaluations of High Throughput Docking Methods for Pharmaceutical Lead Finding − Consensus and Caveats. In *Annual Reports in Computational Chemistry (ARCC) Volume 2;* Spellmeyer, D. C., Ed.; Elsevier: Amsterdam, 2006.

(11) Merritt, H. H.; Putnam. T. J. A new series of anticonvulsant drugs tested by experiments on animals. *Arch. Neur. Psych.* **1938**, *39*, 1003−1015.

(12) van Drie, J. Pharmacophore-based virtual screening. Abstracts of Papers, 231st ACS National Meeting, Atlanta, GA, United States, March 26−30, 2006.

(13) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.

(14) Miller, M. D.; Sheridan, R. P.; Kearsley, S. L. SQ: A program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505−1514.

(15) Feuston, B. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 754−763.

(16) Crippen, C. M.; Havel, T. F. *Distance Geometry and Molecular Conformation;* Bawden, D., Ed.; Research Studies Press, Wiley: New York, 1988.

(17) Kuszewski, J.; Nilges, M.; Brunger, A. T. Sampling and efficiency of metric matrix distance geometry: A novel partial metrization algorithm. *J. Biomol. NMR* **1992**, *2*, 33−56.

(18) Hawkins, P. C. D. A comparison of structure-based and shape-based tools for virtual screening. Abstracts of Papers, 231st ACS National Meeting, Atlanta, GA, United States, March 26−30, 2006.

(19) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem*. **2004**, *47*, 1750−1759.

(20) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem*. **2004**, *47*, 1739−1749.

(21) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.

(22) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−27.

(23) Daylight Chemical Information Systems, Inc. *Version 4.82*; 120 Vantis, Suite 550, Aliso Viejo, CA 92656, 2003.

(24) *MDL Drug Data Report, version 2005.1*; licensed by Molecular Design Ltd., San Leandro, CA, 2005.

(25) Kairys, V.; Fernandes, M. X.; Gilson, M. K. Screening drug-like compounds by docking to homology models: A Systematic Study. *J. Chem. Inf. Model*. **2006**, *46*, 365−379.

(26) Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(27) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(28) Bradshaw, J., personal communication.

(29) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565−582.

(30) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325−332.

(31) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225−242.

(32) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−72.

(33) Vigers, G. P. A.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47*, 80−89.

(34) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the 'receiver operating characteristic' curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.

(35) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.

(36) Truchon, J.-F.; Bayly, C. I., Evaluating Virtual Screening Methods: Good and Bad Metrics for the Early Recognition Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(37) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model. (Online)* **2003**, *9*, 47−57. Enrichment values were extracted from the histograms depicted in the manuscript.

(38) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem*. **2002**, *45*, 4350−4358.

(39) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity-a review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(40) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity methods. *Drug Discovery Today* **2002**, *7*, 903−911.

(41) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *J. Med. Chem*. **2006**, *49*, 1536−1548.

(42) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des*. **2004**, *18*, 529−536.

(43) Knox, A. J. S.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in compound database preparation − hidden impact on virtual screening results. *J. Chem. Inf. Comput. Sci*. **2005**, *45*, 1908−1919.

(44) Knegtel, R. M. A.; Wagener, M. Efficacy and selectivity in flexible database docking. *Proteins* **1999**, *37*, 334−345.

(45) McGaughey, G. B.; Sheridan, R. P.; Bayly, C.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D., unpublished results.

(46) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(47) Perola, E.; Walters, W. P.; Charifson, P. Comments on the Article "On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors". *J. Chem. Inf. Model.* **2007**, *47*, 251−253.

(48) Willett, P. Searching techniques for databases of two- and three-dimensional structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.

(49) Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880−5894.
(50) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.
(51) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(52) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

CI700052X