FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set

Simon Cross,*,† Massimo Baroni,‡ Emanuele Carosati,‡ Paolo Benedetti,‡ and Sergio Clementi‡

Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, United Kingdom, and Laboratory for Chemometrics and Cheminformatics, Chemistry Department, University of Perugia, Via Elce di sotto 10, I-06123 Perugia, Italy

Received June 8, 2010

The performance of FLAP (Fingerprints for Ligands and Proteins) in virtual screening is assessed using a subset of the DUD (Directory of Useful Decoys) benchmarking data set containing 13 targets each with more than 15 different chemotype classes. A variety of ligand and receptor-based virtual screening approaches are examined, using combinations of individual templates 2D structures of known actives, a cocrystallized ligand, a receptor structure, or a cocrystallized ligand-biased receptor structure. We examine several data fusion approaches to combine the results of the individual virtual screens. In doing so, we show that excellent chemotype enrichment is achieved in both single target ligand-based and receptor-based approaches, of approximately 17-fold over random on average at a false positive rate of 1%. We also show that using as much starting knowledge as possible improves chemotype enrichment, and that data fusion using Pareto ranking is an effective method to do this giving up to 50% improvement in enrichment over the single methods. Finally we show that if inactivity or decoy data is incorporated, automatically training the scoring function in FLAP improves recovery still further, with almost 2-fold improvement over the enrichments shown by the single methods. The results clearly demonstrate the utility of FLAP for virtual screening when either a limited or wide range of prior knowledge is available.

I. INTRODUCTION

Virtual (high-throughput) screening of chemical structures is routinely used for drug discovery as a way to reduce the number of compounds that are experimentally screened and to increase the likelihood of finding active hits.[1] Typically virtual screening approaches are divided into those that are ligand-based (LBVS), where known active compounds are used as sensors, and those that are receptor-based (RBVS), where the three-dimensional structure of the drug target is used as the sensor. In both approaches virtual structures are compared to the sensor and scored. The scores are then used to rank the virtual structures and either the top N structures are selected or a score threshold is used. Many computational approaches are available to do this, ranging from chemical similarity[2] to pharmacophore[3] and shape similarity[4] to docking.[5] Like all methods, there are limitations. Chemical similarity, most obviously, returns chemically similar structures that may be active but are less interesting from an intellectual property perspective. Pharmacophore similarity requires the time-consuming generation and validation of a pharmacophore model. Docking requires a crystal structure or comparative model of the target of interest, which must then be prepared accordingly for the docking algorithm, preferably in the same fashion as that used to validate the algorithm. Being a static snapshot, problems may arise from any induced fit that biases screening against all compounds except those closely related to the cocrystallized ligand.[6] More generally, it is not always obvious whether to choose one method over another, and it is more likely that using multiple approaches to select compounds is more appropriate.[7]

Data fusion takes information from multiple sources with the aim of combining it to yield inferences that are superior to any single source; Willett et al. have shown it improves performance of chemical fingerprint similarity metrics for virtual screening,[8] and various others have used it to combine data from different docking algorithms in RBVS (also called consensus scoring),[9−11] and LBVS.[12] Combining ligand and receptor information into a single sensor has also been shown to improve performance in docking.[13] If multiple sensor data is available, then the problem is also a multiobjective one, and multiobjective optimization functions can also be applied. Pareto ranking is one such method, and in the field of drug discovery, it has been applied in combination with Genetic Algorithms for library design and pharmacophore hypothesis generation.[14,15] To our knowledge, it has not been applied until now to the analysis of virtual screening.

GRID molecular interaction fields (MIFs)[16,17] have been applied to many areas of drug discovery, including pK_a and tautomer modeling,[18,19] structure-based drug design,[20] scaffold-hopping,[21,22] 3D-QSAR,[23,24] ADME and pharmacokinetic modeling,[25,26] and metabolism prediction[27,28] (for comprehensive reviews, see refs 29 and 30).

We[12,31−34] and others[35−37] have previously described molecular field-based approaches to virtual screening; however, here we describe a range of virtual screening approaches based on the GRID molecular interaction field approach FLAP[31,32] and demonstrate that excellent results

* To whom correspondence should be addressed. E-mail: simon@moldiscovery.com.
† Molecular Discovery Limited.
‡ University of Perugia.

10.1021/ci100221g
Published on Web 08/06/2010

can be achieved with varying levels of prior knowledge, using the challenging DUD data set described by Huang et al.[38]

II. METHOD

A. Data Set. The Directory of Useful Decoys (DUD) has recently been compiled as a benchmark data set, specifically for docking methods. DUD contains data on 40 relevant typical and interesting targets; for each target crystal structures, known active structures, and decoys are provided. The decoys for each target were chosen specifically to fulfill a number of criteria to make them relevant and as unbiased as possible; for each active compound structure on average 36 decoys were chosen from the druglike subset of the ZINC database of commercially available compounds, each of the 36 decoys chosen to resemble the active structure in terms of physical properties (molecular weight, clogP, number of hydrogen bond donors and acceptors) but different in terms of topological structure (Tanimoto similarity <0.9 using CACTVS type 2 fingerprints). The DUD decoy set thus comprises the target specific "own decoys" (typically a few thousand compounds) and the full decoy set against all targets: the "entire database" (∼100 000 structures). In the original study by Huang et al.,[38] a comparison of the effects of the decoy set was performed keeping the targets the same and varying the decoy sets between the DUD entire database, DUD own decoys, and decoys from Jain,[39] Rognan,[40,41] and MDDR,[42] and applying docking with the DOCK method[43,44] to score compound placement in the binding site. Typically, the Rognan decoys gave the best enrichments, followed closely by MDDR, then Jain, then DUD entire database, then DUD own decoys. Given that the own decoys set was chosen carefully to reflect the physical properties of the known actives, it is unsurprising that this set is the most difficult for scoring methods to discriminate between the known actives and decoys. The results described also highlight how critical the choice of decoys is; enrichments were in general at least half a log unit better against the uncorrected databases (MDDR, Jain, Rognan), hence judging screening methods on enrichment scores alone when different decoy sets are used is flawed. In this study, we elected to use the DUD own decoys as the most stringent and unbiased decoy set from the above list. A common criticism of data sets for retrospective analyses is that they contain many structural analogues, hence if a method finds one it is highly likely to find all of them, and DUD is no exception. The cox2 data set contains over 100 compounds that all possess the same reduced graph scaffold, demonstrated by Good using reduced graph clustering.[45] Release 2 of the DUD data set and the cluster parents from Good were downloaded from the DUD Web site.[46] We have used the actives subjected to Good's lead-likeness filtering (AlogP < 4.5, MW < 450) and clustering approach. Structures were used directly as downloaded, with the exception that they were ionised using the pK_a modeling software MoKa[18,19,47] at pH 7.4, and their 3D coordinates generated using an in-house implementation of MM3.[48−50] To compare with Cheeseright et al.,[37] we have also focused on the 13 data sets that contain at least 15 clusters, treating each cluster as a different chemotype (see Table 1 for details of this DUD subset) with the aim of demonstrating chemotype enrichment.

Table 1. Subset of the DUD Data Set Used in This Study (See Text for Details)[a]

target	PDB code	resolution (Å)	number of actives	number of decoys	number of chemotypes
ace	1o86	2.0	46	1728	18
ache	1eve	2.5	100	3732	18
cdk2	1ckp	2.1	47	1780	32
cox2	1cx2	3.0	212	12491	44
egfr	1m17	2.6	365	14914	40
fxa	1f0r	2.7	64	5102	19
hivrt	1rt1	2.6	34	1439	17
inha	1p44	2.7	57	3043	23
p38	1kv2	2.8	137	8399	20
pde5	1xp0	1.8	26	1810	22
pdgfrb	1t46	model	124	5625	22
src	2src	1.5	98	5801	21
vegfr2	1fgi	2.4	48	2647	31

[a] Abbreviations: ace, angiotensin-converting enzyme; ache, acetylcholinesterase; cdk2, cyclin-dependent kinase 2; cox2, cyclooxygenase 2; egfr, epidermal growth factor receptor; fxa, factor Xa; hivrt, HIV reverse transcriptase; inha, enoyl ACP reductase; p38, P38 mitogen activated protein kinase; pde5, phosphodiesterase 5; pdgfrb, platelet derived growth factor receptor kinase; src, tyrosine kinase SRC; vegfr2, vascular endothelial growth factor receptor kinase.

B. Scoring the Compounds. In this study, we have used the software FLAP (Fingerprints for Ligands and Proteins) as our virtual screening method, which has been described previously as a method for LBVS, RBVS, and also for measuring alignment independent receptor-receptor similarity.[31,32] Since the method has evolved since it was previously described, we will describe the methodology used in detail.

i. FLAP Ligand-Based Screening. For each small molecule (template sensor and test ligands), GRID Molecular Interaction Fields[16,17] (MIFs) are calculated, typically using small molecule probes representing common interactions, such as hydrophobic, hydrogen bond donor and acceptor, and shape. The MIFs are then condensed into discrete points describing the most favorable interaction locations, using a weighted energy-based and spaced coverage function. From these discrete points are produced all combinations of 4-point pharmacophores. Additionally, the GRID atom types for each heavy atom can also be used to produce the 4-point pharmacophores, so both MIF-based and atom-based quadruplets can be produced. We will refer to both cases as FLAP "hotspots". The quadruplets of hotspots for the small molecule test ligands are compared with the quadruplets of hotspots for the sensor (see Figure 1). For the small molecule test ligands, multiple conformations are stored in a database; each associated with their quadruplet description and filtered MIF fields. When a search is performed, all of the quadruplets from the test ligand are searched against all of the quadruplets in the sensor; quadruplets that match (within user-specified distance tolerances) are then used to overlay the test ligand onto the sensor. For each alignment, the tanimoto field similarity is calculated for each of the MIF types, between the test molecule MIF and the sensor molecule MIF. The best similarity value for each MIF type produced by a single conformer is retained for each molecule.

A range of scores is produced depending on the MIFs used in the calculation in addition to a global FLAP distance score (see Figure 2).

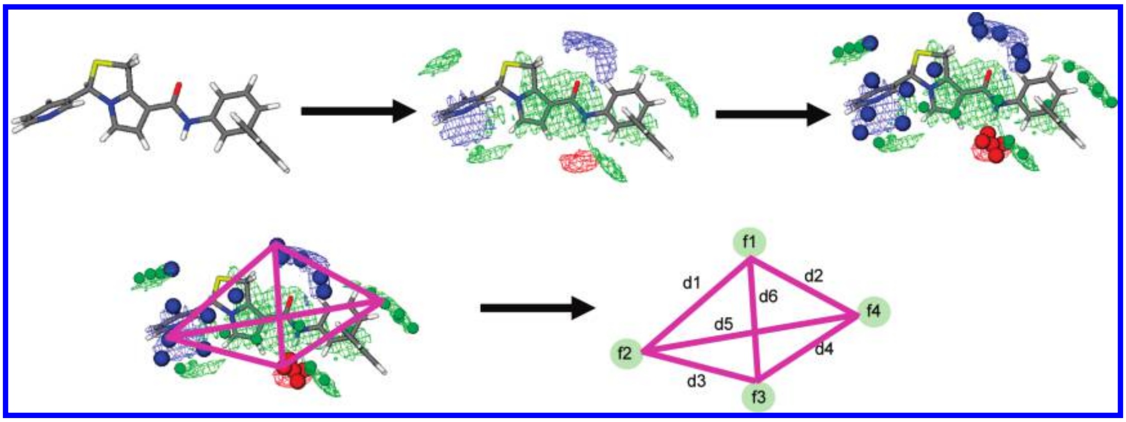
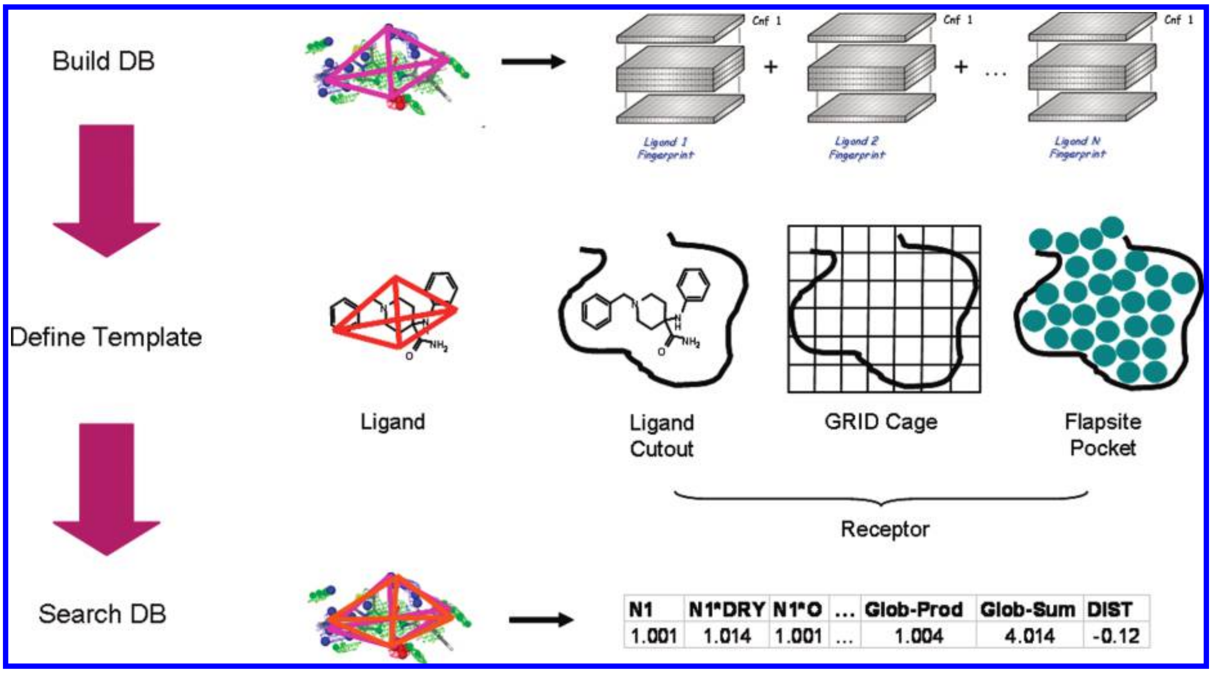


Figure 1. GRID molecular interaction fields are calculated around each molecule and condensed into pharmacophoric points. Quadruplets are defined with four pharmacophoric features (hotspots) at the vertices and the six distances between them; a final volume flag indicates chirality. Quadruplets for each combination of four hotspots are produced per conformer. The process is repeated for all conformers representing the ligand and the quadruplets and fields for each ligand stored in a database.



N1	N1*DRY	N1*O	...	Glob-Prod	Glob-Sum	DIST
1.001	1.014	1.001	...	1.004	4.014	-0.12

Figure 2. Schematic illustrating the FLAP workflow. Once the quadruplets and GRID molecular interaction fields have been calculated for each molecule they are stored in a database. The template can be a ligand or receptor; for each ligand conformer each quadruplet is compared the template quadruplets. If the quadruplets match then the test ligand is aligned to the template using the matching quadruplet and the various filtered field similarities calculated. The best similarity is then reported for each ligand in addition to an overall distance score.

To summarize: the hotspot quadruplets are used to search quickly for the best conformer matches; the hotspots in these matches are used to align the test ligand onto the template sensor along with their filtered GRID fields; the overlapping of the filtered GRID fields of the test ligand over the template sensor are used to score each conformer; each molecule is represented by the best scores and a global distance score.

In this study, descriptors for the data sets to be screened were prepared automatically by the FLAP executable using conformational analysis (random search saving up to 50 conformers, eliminating those with an root mean squared deviation of <0.2 Å), and 5 GRID probes (DRY, O, N1, C3, H) at a grid spacing of 0.7 Å. The compounds were then compared to a single conformer of the template, generated either by using an in-house 2D to 3D conversion program (DUD actives cluster parents), or by directly using the X-ray coordinates of the cocrystallized ligands or receptor site (see section F below). The results were ranked according to the FLAP distance score.

ii. FLAP Receptor-Based Screening. For receptor-based screening, the approach is analogous to the ligand-based approach described above, with the following differences. The template is no longer a small molecule, but a receptor, hence the fields are calculated in the receptor site and the quadruplets produced from these. As shown in Figure 2, the receptor site region can be manually defined by specifying a grid cage, by specifying an existing cocrystallized ligand and a distance tolerance around this ligand, or automatically determined by a pocket-finding algorithm that will be described in a future publication.

For the test molecules, the heavy atoms are assigned to their corresponding GRID atom types and the quadruplets constructed using these as the hotspots (since the receptor fields describe regions in space where corresponding interacting atoms should be placed). Once the initial searching and alignment is performed, the field comparison is performed to score the overlap. In this case, a pseudofield centered on the heavy atoms of the test ligands is compared with the receptor field description for the similarity calculation.

C. Data Fusion. When multiple ranked lists are available from different sensors or scoring methods (in this case virtual screening), there are many ways to combine the data to generate a single ranked list to prioritise the results (in this case selecting the best compounds to take forward to experimental screening). Typically these come from some form of voting or combining the ranks, consensus fingerprints, neural networks, or conditional probability. In this study, we compare the performance of Pareto ranking with a more simplistic consensus rank sum approach to combine the sorted lists produced using different known active ligands as sensors. The consensus rank sum approach simply sums the ranks each test molecule achieves in each list; the resulting rank sum list is then resorted with those molecules scoring lowest at the top of the list (a molecule coming first in 5 lists would therefore score 5, one coming second in 5 lists would score 10, and so on). The Pareto approach counts for each molecule the number of times other molecules achieve a better rank in all of the lists, thus the best molecules will receive a Pareto score of 0. This frequently gives rise to ties, which we then break by secondary sorting on the consensus sum ranks score. Finally, we also used a "recursive" Pareto approach, where the ties were broken by performing another round of Pareto ranking counting the number of times molecules achieved a better rank in all but one of the lists. A third iteration was performed on the ties of the second iteration, counting the number of times molecules achieved a better rank in all but two of the lists. Final ties were broken with the consensus sum ranks score.

D. Optimal Template Selection. Given some known active and inactive compounds, FLAP can automatically select optimal templates by training the scoring function. To do this FLAP uses an error function that simultaneously minimizes the proportion of false positives and false negatives by modifying both the field similarity coefficients and the FLAP global distance threshold. In this so-called "best-ranking" approach the molecules are reranked according to the minimum distance obtained for more than one template. In order to compare the ranked lists for different templates, all of the distances versus the templates are normalized and are inversely related to the similarity. However, this assumption is valid only when the "optimal template selection" process takes place.

E. Judging Performance. To facilitate comparison with other methods, we decided to use the Receiver-Operating Characteristic (ROC) curve as a performance metric. Since DUD contains a number of analogues, and it is usually of more interest to recover different chemotypes, we used the arithmetic weighting modification (awROC).[51] In this method, each true positive's contribution to the ROC curve is inversely proportional to its cluster size. Using the awROC method, we report the Area Under the Curve (AUC) which provides a measure of performance of the entire data set. Since virtual screening usually relies on early enrichment, we report the ROC Enrichment (ROCE) which is the ratio of the true positive rate to the false positive rate.[52] As recommended, we report this at false positive rates of 0.5%, 1%, 2%, and 5%, and we also report the percentage of chemotypes recovered. Before calculating the performance metrics we removed the ligand used as the template sensor (for the data fusion methods this includes all cluster parents), since these would come at the top of each list and artificially inflate the results. Additionally, any compounds missing scores (either from removal or software error) were placed at the bottom of each list to ensure fair comparison. To estimate errors, we used a bootstrapping procedure as performed by Cheeseright et al.;[37] for each result set, we randomly removed 20% of the data and recalculated the metrics, repeating this 10000 times and using the standard deviations as the error estimates.

F. Virtual Screening Scenarios. Since a variety of initial knowledge is provided by the DUD set, we elected to test a number of ligand-based and receptor-based scenarios (for convenience short-hand labels are given in bold type):

(1) The simplest ligand-based scenario occurs when only a single known active is available, so we elected to use each of the DUD known active chemotype cluster parents from Good[45] as single template sensors and calculate the average enrichment factors that were obtained (**LBt**). In addition, we also used the provided cocrystallized ligand as a single template (**LBX**).

(2) If multiple active compounds are known, the data fusion methods can be applied, and here, we combined the individual chemotype cluster parent searches using the rank sum consensus (**LBt SumRanks**), Pareto scoring (**LBt Pareto**), and recursive Pareto (**LBt Pareto R**) approach to give an indication of how performance would be affected as more information is known for the target.

(3) If multiple active and inactive compounds are known, the FLAP optimal template selection algorithm described above can be applied (**LBopt**). In this case, we used the chemotype cluster parents as the known actives, and randomly selected 10% of the 'own' decoys as inactives, although this approach is biased since the decoys are not proven inactives.

(4) If a receptor structure is known, it can be used as a template for FLAP. Here we used the receptor structure provided from DUD, defining the ligand cavity as 5 Å around the cocrystallized ligand for consistency (**RB**). If a cocrystallized ligand is known, the GRID fields in the receptor site can be biased according to the cocrystallized ligand to focus the search in a hybrid approach, which we also tested (**RBLB**).

(5) Additionally, we examined the effects of combining the cocrystallized ligand template search with the receptor-based search using the Pareto method (**LBX RB Pareto**) and also the effects of combining the all of the individual ligand-based cluster parent searches with the receptor-based results (**LBt RB Pareto**). We also looked at combining the results of the trained ligand-based approach (**LBopt**) with the ligand-biased receptor-based (**RBLB**) approach using the Pareto data fusion method (**LBopt RBLB Pareto**).

(6) For comparison, the scores using DOCK[38] and FieldScreen[37] were obtained, and for all results sets, the performance metrics described above in section E were calculated. We only included the FieldScreen results that did not use the receptor as an excluded volume, as the authors mention that there was no statistical difference between the two methods.[37]

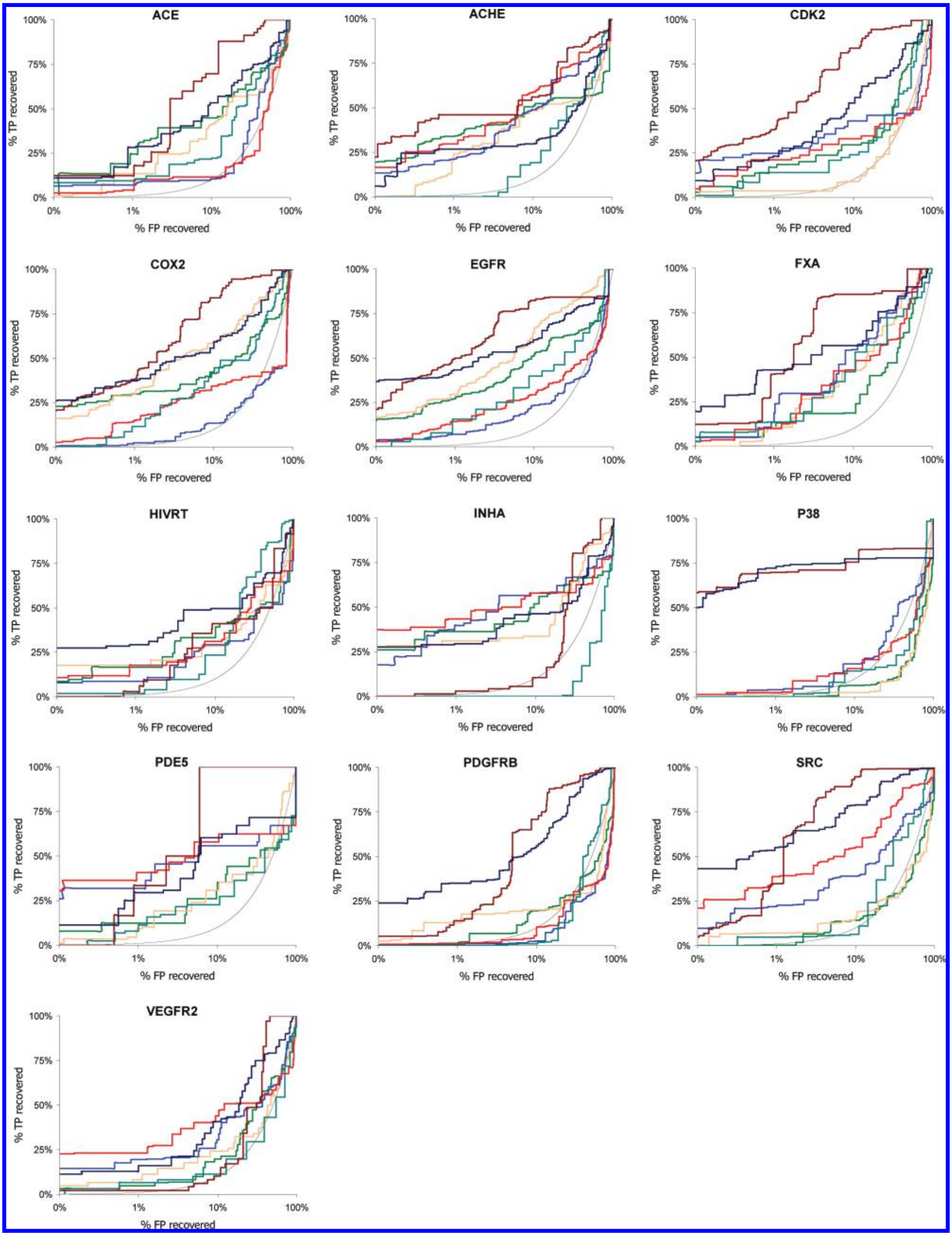


Figure 3. awROC enrichment plots for each of the thirteen DUD data sets using the FLAP LBX method (green), Fieldscreen (tan), FLAP RB (blue) and RBLB (red) methods, DOCK (teal), FLAP LBt Pareto R (dark red) and LBopt (dark blue), and random performance (gray line).

IV. VIRTUAL SCREENING RESULTS

The awROC enrichments for the 13 individual data sets, using the LBX, FieldScreen, RB, RBLB, DOCK, LBt Pareto R, and LBopt approaches is shown in Figure 3 (others are omitted for clarity). The awROC AUC values, awROC enrichments, and chemotype recovery, averaged across the thirteen targets are reported in Figures 4−6 respectively. Taking the AUC values for the different data sets in Figure

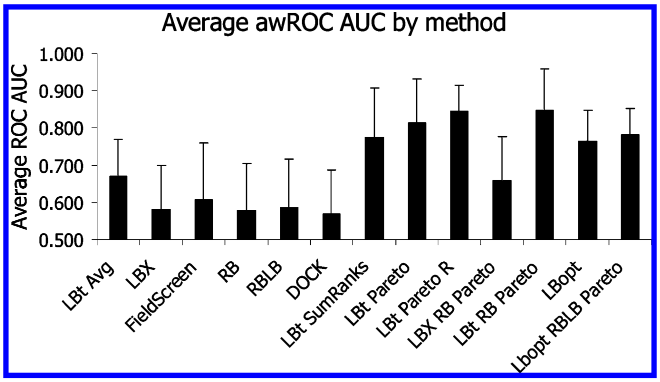


Figure 4. awROC AUC values averaged across the thirteen data sets for each method.

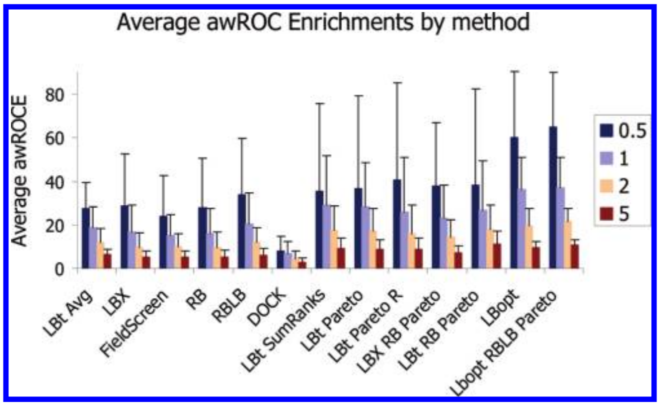


Figure 5. awROC enrichments at 0.5, 1, 2, and 5%, averaged across the thirteen data sets, for each method.

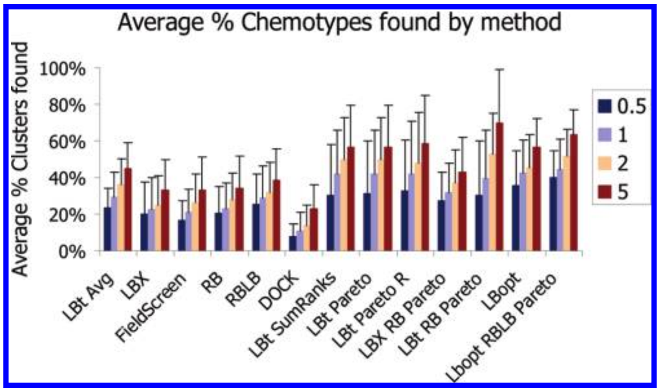


Figure 6. Percentage of chemotypes found, averaged across the thirteen data sets, for each method.

4 as a measure of overall performance, it is immediately obvious that the data fusion and training methods perform better than using only a single ligand or receptor, perhaps unsurprisingly. Pareto ranking (**LBt Pareto**) appears to give better results than the consensus sum ranks approach (**LBt SumRanks**), and the recursive Pareto method (**LBt Pareto R**) does better still. The only data fusion approach that does not perform quite as well is the **LBX RB Pareto** method, which is only using the cocrystallized ligand search and receptor-based search as input. The other data fusion and training approaches are all using the chemotype cluster parents as input, hence it is clear that the more information that is known in advance of performing a screen the better the performance.

In virtual screening however, it is much more desirable to focus on early enrichment, and to select the top fraction of possible compounds for assay. Figures 5 and 6 illustrate the performance at 0.5, 1, 2 and 5% false positive rates. In Figure 5, the **LBt Avg** method shows that, on average, choosing one of the chemotype cluster parents and using it as a template for FLAP will return approximately 20% of the actives with 1% of the decoys. It is interesting that this method performs very similarly to the **LBX** method using the cocrystallized ligand as a template; the cluster parents have been subjected to standard 2D to 3D conversion and minimization, whereas the cocrystallized ligand is present in its bioactive conformation. While perhaps unintuitive, others have reported that using consistently generated single conformations can be effective for virtual screening. Both FLAP methods for pure ligand-based screening yield similar ROCE and cluster enrichments to those reported for FieldScreen. Comparing with the receptor-based approaches, it appears that there is no real difference in enrichment between the FLAP receptor-based method, and the ligand-based methods. However, biasing the receptor fields with the cocrystallized ligand improves the enrichments, and unlike other docking methods without the need to manually define specific pharmacophore constraints. Both receptor-based approaches perform better than DOCK on average. For the data fusion approaches the methods seem fairly similar, with the recursive Pareto method performing slightly better at very early enrichment. It is worth noting that while the **LBX RB** Pareto approach did not appear to perform as well as the other data fusion methods according to the awROC AUC metric, looking at the early awROC enrichments, it shows very similar performance, and by using only the receptor and cocrystallized ligand as input. It is also performing better than doing a single ligand-biased receptor-based screen (**RBLB**), which is probably more restrictive in that it requires hits to be similar to the receptor fields that are in close proximity to the existing cocrystallized ligand, whereas the Pareto data fusion approach allows hits to be similar to either the existing ligand or the receptor fields. Finally, the training approaches show excellent early enrichments that are twice as effective as using the single methods alone, although this additional performance tails off once the false positive rate exceeds 2%. Figure 6 shows the percentage of total chemotypes found by each method averaged across the thirteen targets, at false positive rates of 0.5%, 1%, 2%, and 5%. In this case, rather than biasing the contributions of each active according to how many actives are present in its chemotype cluster, the first time a particular chemotype is found increments the total number of chemotypes found at a particular threshold. For the single template searches, a similar trend is seen as for the average early enrichments, with around 20% of the active chemotype classes being recovered with only 1% of the false positives. In this case though, the **LBt Avg** method shows even better performance of around 30% recovery, which again is surprising given that no knowledge of the bioactive conformation is known, and simply taking one of the chemotype cluster parents as a template for a FLAP ligand-based search on average performs better than all of the other single template methods. This highlights very strongly the ability of FLAP to perform lead-hopping or scaffold-hopping starting from a 2D structure of a single active compound. At a false positive rate of 5%, this method is finding 45% of the known active chemotypes, almost twice as many as the corresponding 23% recovered by DOCK. The **RBLB** ligand-biased receptor-based screen-

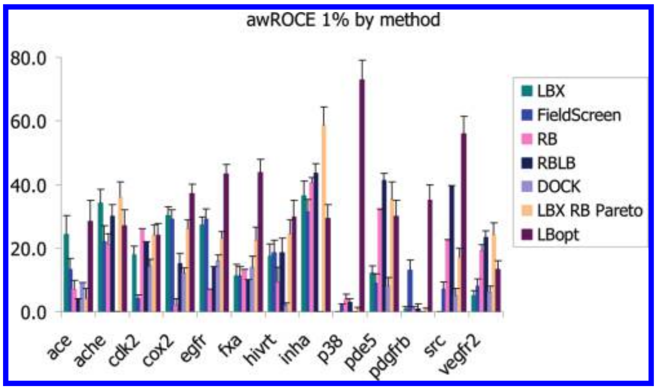


Figure 7. awROC enrichments at a false positive rate of 1% for selected methods. Error bars are taken from the standard deviations of the bootstrapped metrics.

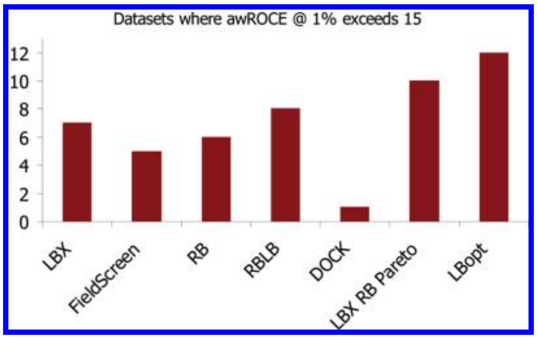


Figure 8. Number of targets for which each method, at 1% false positive rate, achieves an awROC enrichment of 15 or higher.

ing approach again performs better than the pure receptor-based screen (29% recovery versus 23% at false positive rate of 1%), again demonstrating that it is desirable to use known cocrystallized ligands to bias docking methods. The data fusion methods in general appear much better at finding different chemotypes, however all but the **LBX RB Pareto** approach have used the different chemotypes as input, so again this is perhaps unsurprising. The training methods are performing similarly as the data fusion methods, but not as well as shown in the awROC early enrichments in Figure 5; hence, the training is not improving the number of chemotypes found, but is improving the number of actives representing those chemotypes.

Figure 7 shows the awROC enrichments at a false positive rate of 1%, with selected methods shown for clarity, and Figure 8 shows the number of targets where each of these methods achieves an enrichment of 15 or higher at this level (15-fold enrichment over random).

The single template methods achieve this level of enrichment for around half of the targets, with the exception of DOCK that only manages it for a single target (egfr). Using both the receptor and cocrystallized ligand information, the **RBLB** approach manages slightly better (8 out of 13), however the Pareto data fusion using this information (**LBX RB Pareto**) achieves it for 10 out of 13 targets. The training approach improves upon this to achieve 15-fold enrichment or higher in all but one case. Interestingly, there are several cases where the **LBX** method significantly outperforms the **RB** method and vice versa. For the targets ace, ache, cox2, egfr, and hivrt, the ligand-based search performs better, whereas for the targets cdk2, pde5, src, and vegfr2 the receptor-based search gives better results. It is worth noting that the “cocrystallized” ligands for pdgfrb and vegfr2 actually came from different crystal structures, and that the receptor structure for pdgfrb is a homology model. So it is not surprising that we fail to get enrichments for any of the approaches with pdgfrb, with the exception of the training approach which started with the 2D structures of the known actives; in fact it is more surprising to us that the FieldScreen approach achieved any enrichment at all! For p38, as noted by Cheeseright et al.[19] the structure used contains a large ligand with the DFG loop in the DFG-out conformation when most published ligands bind to the DFG-in conformation. Unsurprisingly, this makes both the cocrystal ligand template and the receptor structure not well suited, and helps to explain why the **LBX** method finds no enrichment at 1%, and the **RB** approach finds only modest enrichment (4-fold greater than random). Again, avoiding the 3D crystal structure and starting with the 2D structures gives much better results in terms of the average single **LBt** approach (46-fold greater than random at 1%, data not shown) and the **LBopt** training approach (73-fold greater than random at 1%). Cox2 is a data set where the pure receptor-based approach is not giving good enrichments, so we looked at this in more detail. The ligand-binding here is primarily driven by lipophilic interactions, with the sulphonamide group providing some electrostatic interaction.[28] Examination of the hotspots used by FLAP shows that there are several donor and acceptor field point distributed in the cavity where the hydrophobic rings are placed, and it is possible this is adding too much noise for the FLAP method to be able to discriminate the actives from the decoys; improvements in the field filtering and sampling may eliminate this. However, the donor/acceptor interaction points where the sulphonamide is located also do not match well, and it is possible that the GLN192 is present in the incorrect orientation causing the fields to be incorrect in this region. It is worth noting that the ligand-biased receptor-based approach (**RBLB**) yields a much better enrichment, perhaps by focusing the field sampling in the relevant regions of the pocket and eliminating the noise. Comparing the Pareto data fusion approach using the information from the cocrystallized ligand and the receptor (**LBX RB Pareto**), to the ligand-biased receptor-based search (**RBLB**), it appears the Pareto approach is performing better in the majority of cases. The only data set where the **RBLB** approach is performing better is src, and for this case there is no signal coming from the **LBX** search, probably because the actives being retrieved are much smaller than the ligand template (largest 306.24 Da compared with the template at 481.59 Da). Hence the Pareto fusion in this case is diluting the **RB** screen with noise from the **LBX** screen. The training approach (**LBopt**) is performing well in all cases, with the worst example (vegfr2) giving an enrichment of 13.2-fold at the false positive rate of 1%, and the best example (p38) giving an enrichment of 72.8-fold (perhaps surprising given that all other approaches are performing poorly on this target).

V. CONCLUSIONS

In this work we have shown that the FLAP method of using GRID molecular interaction fields for virtual screening gives excellent performance, whether starting from some known active ligands, or a receptor crystal structure with or without a cocrystall ized ligand, known decoy data, or any combination of the above. The validation we have carried out uses the DUD data set after filtering and clustering, which has been shown to be a more challenging and robust test of screening approaches (compared to MDDR for example), and we have focused on early enrichment of diverse chemotypes, so we are not simply finding obvious structural analogues. Importantly, we have shown that FLAP is able to perform "lead-hopping" or "scaffold-hopping" into different chemical classes, making it extremely valuable alongside 2D similarity methods. Using individual "single template" approaches, both receptor-based and ligand-based, gave on average 17-fold improvements over random at a 1% false positive rate, which is extremely encouraging. Additionally, these screenings were finding different chemotypes, as exemplified by their different performance on the various targets. Biasing the receptor-based screening using the cocrystallized ligand to focus the results improved performance further to a 20-fold average enrichment. Surprisingly, using individual actives as templates in standardized 3D conformation gave, on average, better results than using the single bioactive conformation. Using data fusion in a range of approaches, performance was significantly improved to around 26-fold improvement at 1%. While it seems obvious that adding more information to the screen will improve results, and that there may be an element of overfitting, we have still demonstrated that using the Pareto data fusion approach on the receptor-based and cocrystallized ligand-based screens gives improved results over the ligand-biased receptor-based screen. Training FLAP using known actives and inactives (in this case decoys) is also possible and we have shown that this improves performance still further (36-fold at 1% false positive rate); this is analogous to producing target-specific scoring functions in docking. Of course, there is still room for improvement, and there are a number of areas that we are currently focusing on. The most obvious is to improve the receptor-based field sampling, which should reduce the noise and therefore improve these results further. It should also be possible to include water dynamically in the approach (GRID already comes with a water probe and an optimization routine for constructing hydrogen-bonding networks which could be used[17]), which is important for some targets. We are also investigating receptor flexibility, which is still not treated well by screening algorithms, and GRID already has this capability using the MOVE directive.[17] Another area we are investigating is the adaptation of the similarity function automatically depending on the receptor, for example up-weighting the hydrophobic similarity term in the overall score if the site is primarily hydrophobic. Aside from virtual screening, there are several areas where FLAP can be applied and that we are currently investigating including docking, alignment, pharmacophore hypothesis elucidation, and 3D QSAR.

ACKNOWLEDGMENT

We thank Ismael Zamora and Gabriele Cruciani for invaluable discussions.

Supporting Information Available: Ranked lists for each approach described for use in generating ROC data. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Clark, D. E. What has virtual screening ever done for drug discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 841–851.

(2) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(3) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. J. Fast and efficient in silico 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2006**, *47*, 2182–2196.

(4) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.

(5) 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.* **2006**, *49*, 5912–5931.

(6) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293–2302.

(7) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143–1152.

(8) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1–16.

(9) Feher, M. Consensus scoring for protein−ligand interactions. *Drug. Discovery Today.* **2006**, *11*, 421–428.

(10) Teramoto, R.; Fukunishi, H. Consensus scoring with feature selection for structure-based virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 288–295.

(11) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281–295.

(12) Carosati, E.; Mannhold, R.; Wahl, P.; Hansen, J. B.; Fremming, T.; Zamora, I.; Cianchetta, G.; Baroni, M. Virtual screening for novel openers of pancreatic K_{ATP} channels. *J. Med. Chem.* **2007**, *50*, 2117–2126.

(13) Cross, S. S. J. Improved FlexX docking using FlexS-determined base fragment placement. *J. Chem. Inf. Model.* **2005**, *45*, 993–1001.

(14) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2002**, *42*, 375–385.

(15) Richmond, J. R.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567–587.

(16) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(17) http://www.moldiscovery.com/soft_grid.php (accessed January 2009).

(18) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pK_a prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(19) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.* **2009**, *49*, 68–75.

(20) Von Itzstein, M.; Wu, W.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418–423.

(21) Ahlström, M. M.; Ridderström, M.; Luthman, K.; Zamora, I. Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J. Chem. Inf. Model.* **2005**, *45*, 1313–1323.

(22) Bergmann, R.; Linusson, A.; Zamora, I. SHOP: Scaffold hopping by GRID-based similarity searches. *J. Med. Chem.* **2007**, *50*, 2708–2717.

(23) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.

(24) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-Independent descriptors. *J. Med. Chem.,* **2005**, *48*, 2687–2694.

(25) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29–S39.

(26) Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting blood−brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204–2216.

(27) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the Chemist. *J. Med. Chem.* **2005**, *48*, 6970−6979.

(28) Ahlström, M. M.; Ridderström, M.; Zamora, I.; Luthman, K. CYP2C9 structure−metabolism relationships: Optimizing the metabolic stability of COX-2 inhibitors. *J. Med. Chem.* **2007**, *50*, 4444–4452.

(29) *Molecular Interaction Fields*, Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006.

(30) Cross, S.; Cruciani, G. Molecular fields in drug discovery: Getting old or reaching maturity. *Drug Discovery Today* **2010**, *15*, 23–32.

(31) Perrucio, F.; Mason, J. S.; Sciabola, S.; Baroni, M. FLAP: 4-Point Pharmacophore Fingerprints from GRID. In *Molecular Interaction Fields*; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006; pp 83−102.

(32) Baroni, M.; Cruciani, G.; Sciabola, S.; Perrucio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

(33) Carosati, E.; Cruciani, G.; Chiarini, A.; Budriesi, R.; Ioan, P.; Spisani, R.; Spinelli, D.; Cosimelli, B.; Fusi, F.; Frosini, M.; Matucci, R.; Gasparrini, F.; Ciogli, A.; Stephens, P. J.; Devlin, F. J. Calcium channel antagonists discovered by a multidisciplinary approach. *J. Med. Chem.* **2006**, *49*, 5206–5216.

(34) Carosati, E.; Budriesi, R.; Ioan, P.; Ugenti, M. P.; Frosini, M.; Fusi, F.; Corda, G.; Cosimelli, B.; Spinelli, D.; Chiarini, A.; Cruciani, G. Discovery of novel and cardioselective diltiazem-like calcium channel blockers via virtual screening. *J. Med. Chem.* **2008**, *51*, 5552–5565.

(35) Jilek, R. J.; Cramer, R. D. Topomers: A validated protocol for their self-consistent generation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1221–1227.

(36) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. Lead hopping. Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem.* **2004**, *47*, 6777–6791.

(37) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *J. Chem. Inf. Model.* **2008**, *48*, 2108–2117.

(38) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(39) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2005**, *49*, 5856–5868.

(40) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases: 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(41) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225–242.

(42) http://www.symyx.com/products/databases/bioactivity/mddr (accessed January 2009).

(43) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(44) http://dock.compbio.ucsf.edu (accessed January 2009).

(45) http://dud.docking.org/clusters/summary.pdf (accessed January 2009).

(46) http://dud.docking.org (accessed January 2009).

(47) http://www.moldiscovery.com/soft_moka.php (accessed January 2009).

(48) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8565.

(49) Lii, J.-H.; Allinger, N. L. Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575.

(50) Lii, J.-H.; Allinger, N. L. Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals potentials and crystal data for aliphatic and aromatic hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8576–8582.

(51) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.

(52) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 133–139.

CI100221G