J Comput Aided Mol Des (2008) 22:479–487
DOI 10.1007/s10822-007-9164-5

LASSO—ligand activity by surface similarity order: a new tool for ligand based virtual screening

Darryl Reid · Bashir S. Sadjad · Zsolt Zsoldos · Aniko Simon

Received: 1 October 2007 / Accepted: 18 December 2007 / Published online: 18 January 2008

Abstract Virtual Ligand Screening (VLS) has become an integral part of the drug discovery process for many pharmaceutical companies. Ligand similarity searches provide a very powerful method of screening large databases of ligands to identify possible hits. If these hits belong to new chemotypes the method is deemed even more successful. eHiTS LASSO uses a new interacting surface point types (ISPT) molecular descriptor that is generated from the 3D structure of the ligand, but unlike most 3D descriptors it is conformation independent. Combined with a neural network machine learning technique, LASSO screens molecular databases at an ultra fast speed of 1 million structures in under 1 min on a standard PC. The results obtained from eHiTS LASSO trained on relatively small training sets of just 2, 4 or 8 actives are presented using the diverse directory of useful decoys (DUD) dataset. It is shown that over a wide range of receptor families, eHiTS LASSO is consistently able to enrich screened databases and provides scaffold hopping ability.

Keywords Conformation independent QSAR descriptor · Scaffold hopping · Virtual screening · Ligand based screening

Electronic supplementary material The online version of this article (doi:10.1007/s10822-007-9164-5) contains supplementary material, which is available to authorized users.

SimBioSys Inc.—http://www.simbiosys.ca

D. Reid · B. S. Sadjad · Z. Zsoldos · A. Simon (✉)
SimBioSys Inc., 135 Queen's Plate Drive, Suite 520,
Toronto, ON, Canada M9W 6V1
e-mail: aniko@simbiosys.ca

Introduction

Computational methods and tools are now an essential part of the drug discovery process for virtually every major pharmaceutical or biotechnology company. The challenge of finding the next 'best-in-class' drug has made companies look towards these computational methods as a way to identify isosteric molecular scaffolds that cover new chemical space. This idea of 'scaffold-hopping' (also referred to as leap-frogging, scaffold searching or lead hopping) involves searching for new lead compounds that exhibit the same (or similar) biological activity but have significantly different topologies.

Virtual screening of compound databases is one of the main applications of computational methods in lead discovery. Virtual screening can be considered as the in silico equivalent of High Throughput Screening (HTS). Virtual screening does have several advantages over experimental HTS, for instance, virtual screening requires a fraction of the time and cost of HTS, depending on the computational method used and computational resources available. In addition, virtual screening does not have the physical limitations implicit in experimental HTS, that is to say, in virtual screening, the size and content of the database screened is essentially limitless (limited only by the time the chemists wishes to spend on the process). Virtual screening is not a replacement for physical screens but has however become an essential partner to augment the process.

Virtual screening can be broadly divided into two classes, structure based virtual screening (SBVS) and ligand based virtual screening (LBVS). In SBVS a crystal structure or homology model of the target receptor or protein is used to model how potential ligands might interact or bind with the receptor. There are many tools available for SBVS, including flexible ligand docking tools such as eHiTS [1], DOCK [2], Glide [3, 4], and GOLD [5]. All these tools model how a ligand might bind within the receptor and attempt to estimate the binding affinity (the docking score).

LBVS does not explicitly consider the target receptor when trying to find new molecules that might exhibit biological activity towards a given target. Instead in LBVS, knowledge of known bioactive molecules is used to estimate the activity of new molecules. Simple similarity searches based on a single reference ligand are probably the most common use of LBVS. Here a single known bioactive molecule is used to search a database for molecules that are similar to it, based on any number of different criteria. When more data are available, for example several active ligands, methods such as pharmacophore mapping may be applied. Pharmacophore approaches attempt to create a 3D model of the key interactions necessary for binding to a receptor by looking for common substructural features between known actives. Pharmacophore methods typically rely on the overlay of actives to find common potential interaction sites and thus having the correct conformation of the ligands is critical. While there are methods for automating conformation generation and automatic pharmacophore identification, such as those used in Phase [6] and GASP [7], most of these programs rely on relatively low energy conformations. It has been shown that the bound conformations of ligands within the active site of receptors often do not correspond to energetic minima of an unbound ligand [1, 8]. Therefore, the task of conformation generation is even more complicated than just creating sets of low energy conformers.

Increasingly machine learning methods are being applied in situations where multiple bioactive ligands are known. Machine learning methods, as applied to LBVS, work by taking a set of known active ligands and a set of known or presumed inactive ligands as a training set. The machine learning algorithm then analyses the training set to develop a set of decision rules which can later be applied to classify new molecules. The more abstract nature of machine learning methods can allow them to overcome some of the challenges of more simplistic overlay pharmacophore techniques.

In recent years, several tools have been developed that apply machine learning methods to computer aided molecular design (CAMD). Tools using support vector machines (SVM) [9], decision trees [10], naïve Bayesian classifiers (NBC) [11], and binary kernel discrimination (BKD) [12] have all recently been published.

In this work we will focus on a new tool for ligand based virtual screening, eHiTS LASSO (electronic High Throughout Screening Ligand Activity by Surface Similarity Order). eHiTS LASSO uses a neural network machine learning approach combined with a new molecular descriptor based on interacting surface points to screen databases of potential ligands. The performance of the method is analyzed using the recently published directory of useful decoys (DUD) [13] dataset. This dataset has been designed as a benchmark for structure based molecular docking programs. However, we feel that the wide ranging scope and data size of this database make it appropriate for evaluating ligand based screening methods as well.

Computational methods

eHiTS LASSO

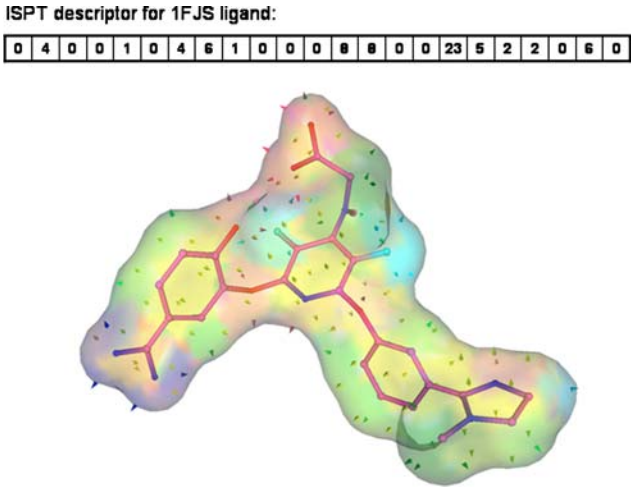
Surface point molecular descriptor

In this work we introduce a new ligand based virtual screening tool that combines a surface property based molecular descriptor with a non-linear learning method that extracts key binding patterns from the chemical interaction properties. The new molecular descriptor uses surface properties to capture the essence of bioactive molecules that are necessary for binding. The Interacting Surface Point Type (ISPT) descriptor is built upon the scoring function of the eHiTS docking program, eHiTS Score. eHiTS Score uses interacting surface point types to describe the properties of ligands and receptors that are responsible for binding. The ISPT descriptor uses the surface point types defined in eHiTS Score to characterize a molecule.

The use of surface properties has been applied to Quantitative Structure Activity Relationship (QSAR) methods before. Clark [14] noted that while much information can be extracted from a 2D drawing of a molecule and a 3D representation adds conformational information, a surface drawing contains far more information about the ability of a molecule to interact with other molecules. Ligand binding can be looked at as matching surface properties of the ligand with complementary properties of the active site [15]. In the eHiTS docking program, ligand binding is largely estimated by considering how well the surface points on the ligand interact with the surface points on the receptor. eHiTS Score, and thus the ISPT descriptor, uses 23 different types of surface points, listed in Table 1. Each surface point has associated chemical properties which affect how the point interacts with its environment. The ISPT descriptor is simply a feature vector which is made up of the counts of the different types of surface points on the ligand. Figure 1 shows the feature vector of the ZK-807834 inhibitor of Factor Xa [16]. A surface has been generated on the ligand with the surface points placed in the appropriate locations. Each surface point is colored based on the surface point type. The ISPT descriptor is a count of the occurrence of each surface point type.

Table 1 23 Surface point types with description

#	Surface point type name	Description
1	Metal	Positively charged metal ion interaction point
2	Charged_Hplus	Positively charged hydrogen bond donor, e.g., Arginine
3	Primary_Amine_Hlp	Primary amine hydrogen/lone-pair, e.g., $-NH_3^+$or $-NH_2$
4	Hdonor	Strong (primary) hydrogen bond donor H (polar-atom-H)
5	Weak_Hdonor	Weak (secondary) hydrogen bond donor H (polarized C–H)
6	Charged_Lonepair	Lone pair of negatively charged group, e.g., PO_3^-
7	Acid_Lonepair	Lone pair of an acidic functional group, e.g., carboxolate
8	Lonepair	Strong (primary) hydrogen bond acceptor lone pair
9	Hydrophob	H on sp^3 hydrophobic carbon
10	H_Arom_Edge	H on hydrophobic carbon in aromatic ring (non-polarized)
11	Ws_Lipo	H on weak secondary hydrophobic atom (e.g., carbon next to polar)
12	Neutral	H/Lp on neutral atom (no recognized activity)
13	Ambivalent_Hlp	Donor H or acceptor Lp depending on protonation state
14	Rotatable_H	Rotatable-hydroxy donor H
15	Rotatable_Lp	Rotatable-hydroxy acceptor Lp
16	Weak_Lonepair	Weak (secondary) hydrogen bond acceptor lone pair
17	Pi_Sp2_Polar	π electron on sp^2 polar atom (N/O)(non-resonating, non-aromatic)
18	Pi_Sp2_Carbon	π electron on sp^2 carbon atom (non-resonating, non-aromatic)
19	Halogen	Lone electron pair of a halogen atom (F, Cl, I, Br)
20	Sulfur	Lone electron pair of a sulfur atom
21	Pi_Aromatic	π electron of an aromatic ring
22	Pi_Reson_Polar	π electron on polar atom (N/O) in resonance chain, e.g., amide
23	Pi_Reson_Carbon	π electron on carbon atom in resonance chain, e.g., amide



ISPT descriptor for 1FJS ligand:

0	4	0	0	1	0	4	6	1	0	0	0	8	8	0	0	23	5	2	2	0	6	0

Fig. 1 The ZK-807834 inhibitor as complexed with Factor Xa in the 1FJS PDB file. The surface points shown are colored by surface point type and are used to create the ISPT descriptor vector shown above. Each column represents one of the 23 surface point types

Conformation independence and scaffold hopping

The ISPT descriptor captures the necessary surface properties of bioactive molecules without a direct reliance on the topology of the molecule. Therefore, molecules with similar ISPT feature vectors could have very different scaffolds. Figure 2 shows two molecules with obviously different 2D scaffolds, each molecule generates the same ISPT descriptor. While the 3D positions of the surface points generated depend on the conformation of the molecule, the simple count of each type does not. The reason is that the surface points in eHiTS are generated systematically based on the hybridization and local connectivity of each atom. An sp^3 hybridized atom has one surface point in the direction of each of the tetrahedral coordination of its electron arrangement except for those directions occupied by bonds to other heavy atoms, i.e., for each Hydrogen atom and each lone electron pair it has exactly one surface point of a specific type. An sp^2 atom similarly has surface points generated, but in addition to those (up to three directions), it also gets two additional surface points in the directions perpendicular to the plane of its bonds (or σ electrons) to represent the interaction possibilities of the π electrons above and below the plane. Thus, the number and type of interacting surface points is fully determined by the hybridization and number of heavy atom connections of each atom. These quantities are invariant during conformation changes, therefore the LASSO QSAR descriptor vector, which contains the total counts of each type for all the atoms of the molecule, is also independent of the conformation.

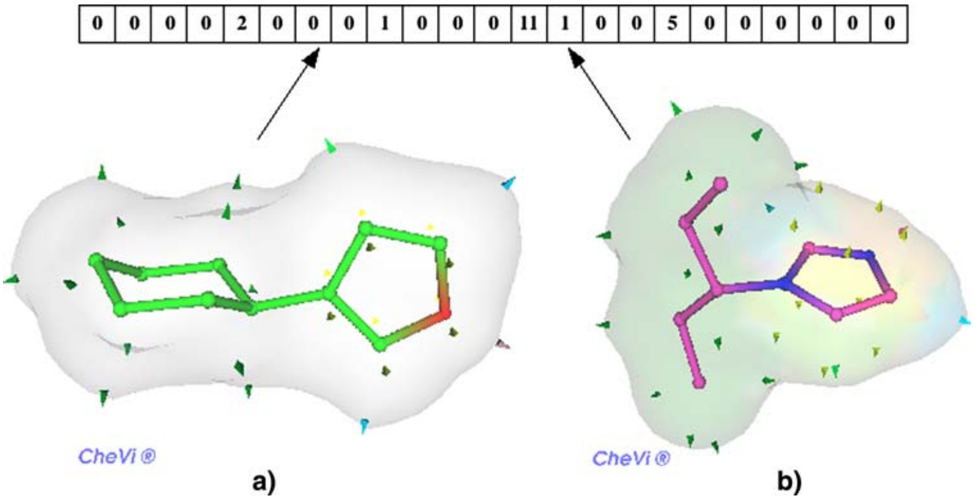


Fig. 2 The ISPT descriptor, shown above, is the same for both molecule (**a**) and molecule (**b**), each with obviously different 2D scaffolds

Machine learning using neural network

eHiTS LASSO uses a feed-forward neural network with a single hidden layer, with training performed using the Stuttgart Neural Network Simulator (SNNS) program [17]. The hidden layer has five nodes and the output layer just one. All nodes of one layer are connected to all nodes of the next and each edge has a weight. The number of input nodes is equal to the number of surface point types in the ISPT descriptor (i.e., 23). To train the neural network, two datasets are required, a set of known active molecules and a set of known or presumed inactive molecules. Each set is converted into ISPT descriptors by eHiTS LASSO and fed into the neural network. The training set is then internally divided into two sets, an internal training set and an internal validation set, as a mean to reduce overtraining. The output is a trained neural network file that can be used to screen test databases of molecules. The result is a normalized score ranging from 0 to 1, with 0 having no similarity to the active molecules and 1 having the highest level of similarity. The assumption is that molecules with high eHiTS LASSO scores are more likely to have similar bioactivity.

Experimental details

To examine the performance of eHiTS LASSO, we decided to use the newly assembled directory of useful decoys (DUD) [13] dataset. The DUD dataset is a carefully constructed benchmarking dataset designed for molecular docking. The dataset spans 40 different receptor families and contains sets of known active ligands for each. A key difference of DUD and some of the other recently published datasets [18–20] is the focus on the decoys in the dataset. A total of 2,950 active ligands were collected for 40 different protein families, each family having tens to hundreds of actives. For each active in each set, 36 ‘drug-like’ molecules were chosen from the ZINC database of commercially available compounds to act as decoys. The decoy molecules were chosen to have similar physical properties (molecular weight, logP, number of hydrogen bond acceptors/donors, etc.) but differed from the ligand topologically. The first necessary step was to clean up the DUD datasets to remove any duplicate molecules, based on ZINC IDs. This resulted in 93,161 decoy molecules which were combined with the actives for each target family for the testing below.

Machine learning approaches like that employed by eHiTS LASSO are often at their best when there is a lot of training data available. However, this is not always the case in industrial settings, especially at the beginning of a lead-discovery program when often there are only a few known active molecules. The experiments here are designed to evaluate the ability of eHiTS LASSO to retrieve bioactive molecules from a screened database when minimal training sets are available. Therefore, for each of the 40 DUD targets, five sets of 2, 4 and 8 active ligands were selected to be the basis of the training sets, referred to as SET2, SET4 and SET8 here. The training step was repeated five times to reduce the influence of abnormal training sets, for families with less than 5 × 2, 5 × 4 or 5 × 8 actives only N runs were completed, where N equals the number of actives divided by 2, 4, or 8. A trained neural network file (referred to as a trained net file) was generated by eHiTS LASSO for each training set.

Test sets were composed of the actives for each target and 93,161 DUD decoy molecules, with the training set actives and decoys removed for each run.

Results

The ability of eHiTS LASSO to retrieve the bioactive ligands for each family is shown in the graphs in Fig. 3. Here we see the average enrichments for the training sets of 2 (shown in yellow), 4 (shown in green) and 8 (shown in red) actives. In most cases this is the average over five runs, however in some datasets with less than 20 or 40 actives, fewer runs were performed for the 4 or 8 active training sets. It is also important to note that the active ligands and decoys used in the training sets were removed from the test set for each run so as to not bias the results. The published results for the DOCK docking program [13] are included in these plots as a point of reference and are shown in dashed black. Due to the removal of the training sets from the test sets for eHiTS LASSO the results are not on exactly the same test database. However, as the eHiTS LASSO results are averages over several runs and the total number of molecules in the dataset is much larger than in the training sets, we do not believe that this affects the results in any significant way.

All molecules were converted into ISPT descriptor format and the descriptors were used for the training and testing. eHiTS LASSO took under 5 sec to screen each 90k dataset on a standard CPU.

Looking at the average enrichments achieved over all target families (Fig. 4), it is clear to see that eHiTS LASSO outperforms DOCK on these datasets. This is consistent with several recent comparisons of ligand based screening method compared with docking programs, which have shown that ligand based methods outperform docking in enrichment applications [21, 22]. However it is also important to note that eHiTS LASSO performs well for virtually every family. On average, results did improve as the size of the training set increased, from 2 actives to 8 actives. On average 30, 47 and 52% of the actives were recovered in the top 10% of the screened database for SET2, SET4 and SET8 runs, respectively. To better judge the ability of eHiTS LASSO to rank active compounds early in the screened database we will use the BEDROC metric, recently published by Truchon and Bayly [23], for the purpose of early recognition. Similar to the ROC metric [24] BEDROC is bounded by the interval [0,1] and can be interpreted as the probability that a ranked active, randomly selected, will be positioned before a randomly selected compound distributed following an exponential parameter α. For this study, α has been assigned a value of 20.0, in accordance with the recommendations of Truchon and Bayly. The average BEDROC results for SET2, SET4 and SET8 are 0.21, 0.33 and 0.4, respectively. This compares to a value of 0.25 for DOCK.

It is clear that for a machine learning approach to create a predictive model of activity having more data available for training should increase the chance that generalizations can be made. When training on just two active molecules there is typically not enough information to adequately represent all the activity of the target. However, when 8 actives are present in the training set there is more diversity represented which helps in the training of the neural network.

It is also interesting to note that in general, as the size of the training set is increased, the variability of enrichment within runs of each set is reduced. Stated another way, when training on just two actives there can be a rather large range of enrichment results between different runs as compared to the range of enrichment when training using 8 actives. This again reflects the ability of the neural network to generalize from a smaller training set. As the training set increases it is more likely that the training set and thus the generated trained network file will more accurately represent the activity of the target set. Figure 5 shows a typical run in which the variability is reduced going from 2 to 4 to 8 actives in the training set.

We also wanted to investigate the reliance of eHiTS LASSO on 2D similarity and thus on its ability to perform scaffold hopping. To perform the structural similarity measurements, all molecules were first converted from MOL2 to MOL format using ChemAxon's molconvert version 4.0. Then the structures were converted into chemical fingerprints using GenerateMD [25]. The Compr utility [25] was then used to calculate the dissimilarity of each dataset, meaning that larger values represent more diverse structures than smaller values.

Mean pairwise dissimilarity values were calculated for the actives in each training set in SET2, SET4 and SET8 and the rest of the actives for that target. This gave a measure of the diversity between the training set actives and the test set actives. The plots (Fig. 6) shows the relationship between the 2D structural diversity and the enrichment achieved by eHiTS LASSO, as measured with the BEDROC metric. Each plot has been divided into four quadrants; Quadrant #1: high diversity, low enrichment, Quadrant #2: high diversity, high enrichment, Quadrant #3: low diversity, low enrichment and Quadrant #4: low diversity, high enrichment. It is reassuring that we see virtually no correlation between enrichment and structural diversity. Since the descriptor used in eHiTS LASSO does not consider the molecular skeleton, we hoped this would be the case. As we go from SET2 to SET4 to SET8 we see the points shift to the right, into Quadrants #2 and #4, illustrating the fact that we get better enrichment with larger training sets. While we are still able to get some very strong enrichment with just 2 actives in the training set, the results are not consistent. However, using 8 actives there are virtually no points in Quadrant #3. Thus, when training on 8 actives in this dataset, all highly similar datasets

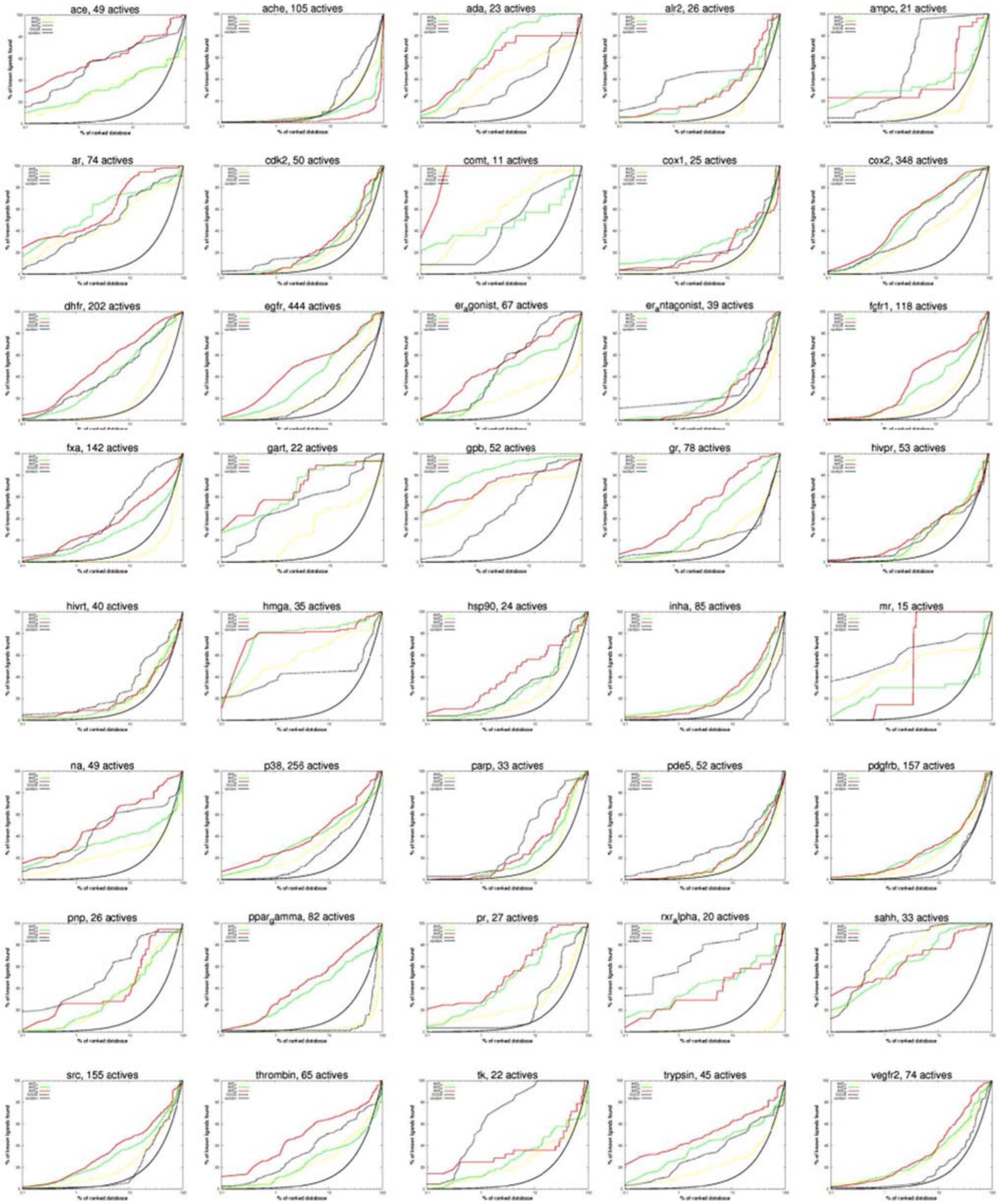


Fig. 3 Enrichment plots over 40 target families. Results for SET2 (yellow), SET4 (green) and SET8 (red) are averages over the repeated training runs. The dotted black lines are the DOCK results. Enrichment plots show the ability to rank known active molecules higher than presumed inactive molecules. Results are shown on a log scale to highlight early enrichment, that is to say enrichment in the top 1–5%. "Random" enrichment (or no enrichment) is shown in solid black

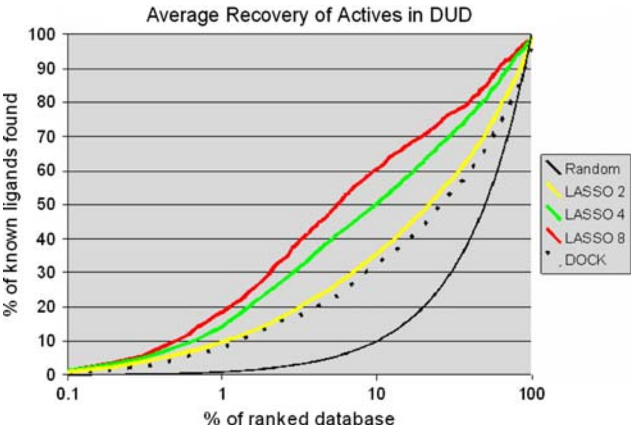


Fig. 4 Average enrichment for SET2 (yellow), SET4 (green), SET8 (red) and DOCK (dotted black) over all 40 targets

produced good enrichment results as do most of the structurally diverse datasets as evidenced by the population in Quadrant #2.

To further look at the effect, or lack thereof, of 2D structural diversity on eHiTS LASSO results, we looked at the diversity of the training set actives compared to the test set actives in terms of both 2D similarity (as measured above) and the average pairwise RMSD of the ISPT feature vector descriptor. The RMSD was calculated between each active in the training set and the test set and an average RMSD was determined. Figure 7 shows the relationship of 2D structural diversity and the LASSO RMSD. The resulting scatterplot shows no correlation between 2D similarity and the feature vector space as determined using the ISPT descriptor.

To illustrate a scaffold hopping event, Fig. 8 shows the actives used in the training set for the catechol *O*-methyltransferase (COMT) family. The training set shown consists of just two molecules, both containing tri-alcohol aromatic moiety and a flexible chain. The compound shown in Fig. 9 was recovered in the top 0.3% (rank 286) of the screened database and obviously has a different and unique scaffold from that in the training set. In this example it is clear that the ISPT descriptor can identify different scaffolds which have similar surface properties.

Discussion

The molecular descriptor used in eHiTS LASSO is independent of ligand conformation and has been shown to successfully enrich screened databases across a wide range of target families. In comparison to Dock, eHiTS LASSO was able to improve enrichment of the screened databases. It would be very interesting to see how other ligand based methods compare on the publically available DUD dataset and we encourage others to report such results.

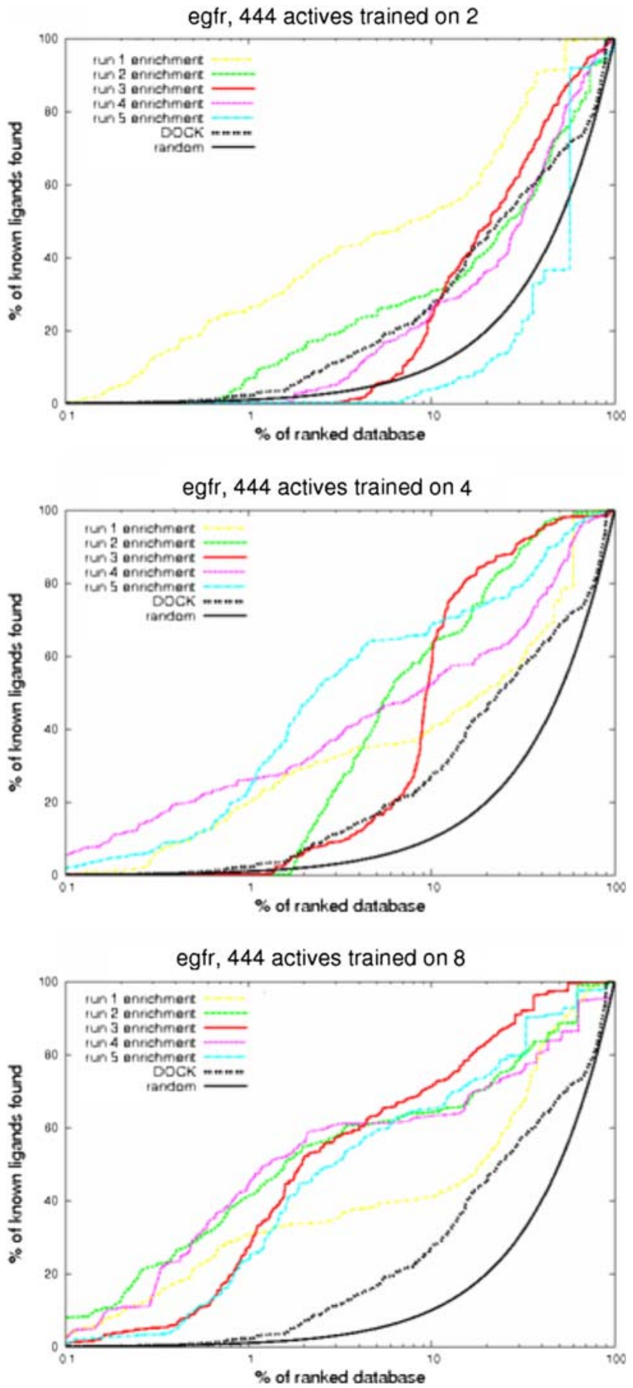


Fig. 5 Variability of enrichment results due to training set size. The average for the five training runs for SET2, SET4 and SET8 are shown in dark red. The variability reduces as the training set size increases from 2 to 8. Results are shown for EGH target

The ISPT descriptor used in eHiTS LASSO captures the surface properties required for activity and eHiTS LASSO is able to sufficiently generalize those properties to be able to discover potential lead molecules with similar surface properties. Lying somewhere between a 2D and a 3D descriptor the ISPT descriptor does not contain any shape

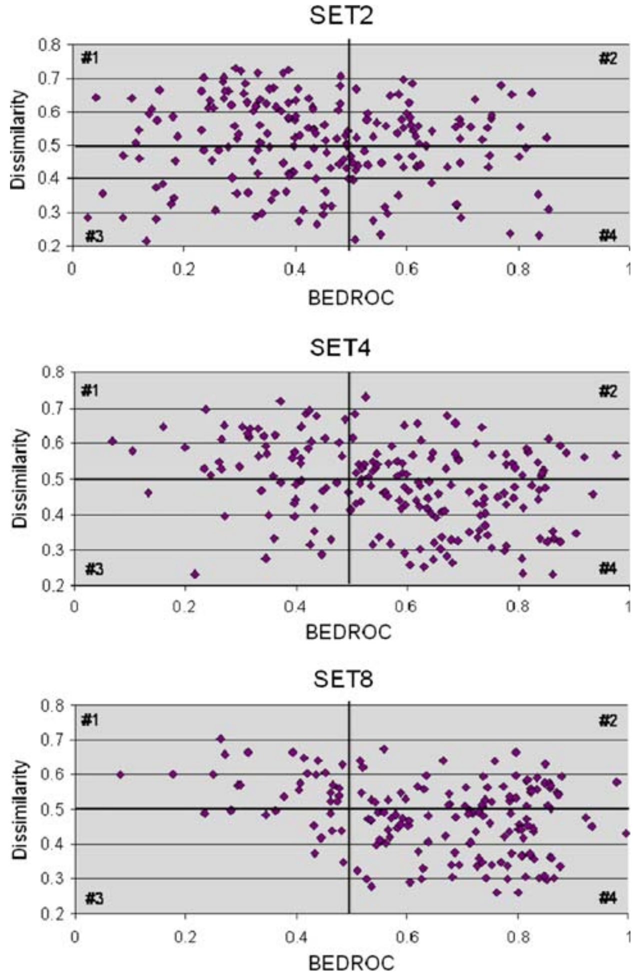


Fig. 6 The effect of diversity of the active ligands on the enrichment results, shown using the BEDROC metric for SET2, SET4 and SET8. Each plot is divided into four quadrants

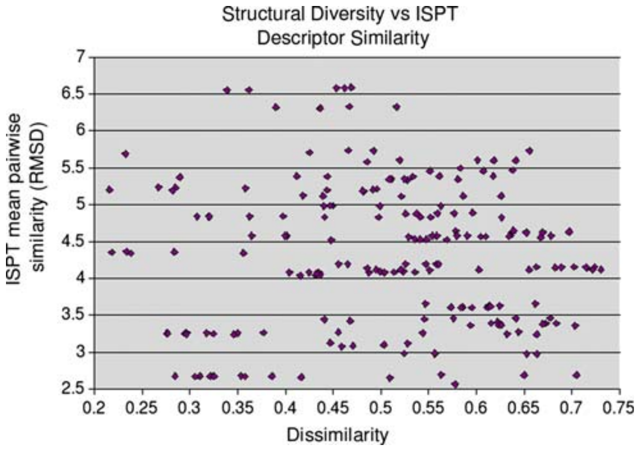


Fig. 7 Correlation of ISPT descriptor Average RMSD and the 2D Structural dissimilarity measure for the actives used in training and testing

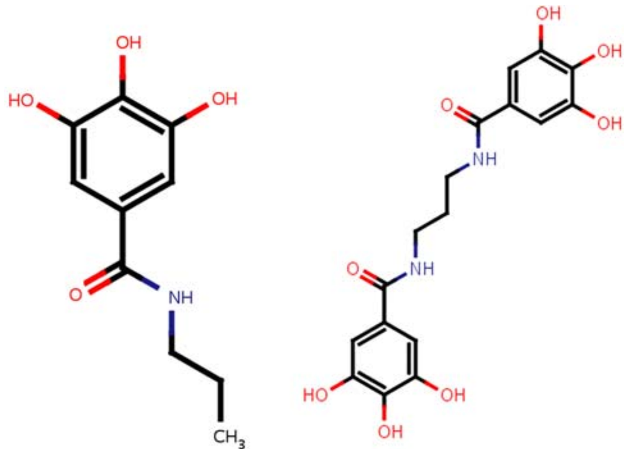


Fig. 8 Example of active molecules used as a training set for COMT target family. Note the similar scaffolds within the training set

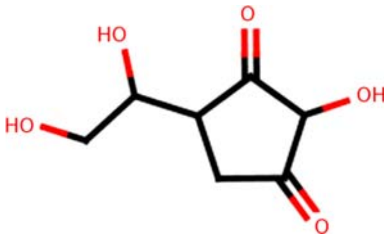


Fig. 9 Example of scaffold hopping, shown COMT active molecule was recovered in the top 0.3% of the screened database

or 2D connectivity information. There may however be some size information implicit in the descriptor due to capturing the counts of surface points and larger molecules will have more surface points than smaller molecules and eHiTS LASSO may be somewhat sensitive to this.

The conformation independence of eHiTS LASSO overcomes any problems of ligand conformation generation. In addition, there is no need for overlaying ligands, as is often the case in pharmacophore approaches. It has also been shown that eHiTS LASSO has very little correlation to the 2D similarity of the molecules used for training and testing meaning that eHiTS LASSO is able to recover structurally diverse actives. In a prospective study this would lead to scaffold hopping events in lead discovery applications.

The incredible speed of eHiTS LASSO, 1 million structures screened in less than 1 min on a single CPU, makes it an ideal tool to be used as a pre-docking screen. eHiTS LASSO will return a high percentage of false positives, due to the disregarding of 3D relationships of surface properties, however because of this it will also return a higher percentage of different scaffolds. Taking the results of eHiTS LASSO and feeding the top *N*% into a docking program would allow the docking program to weed out many of the false positives. eHiTS LASSO is fully integrated with the eHiTS docking tool and it is highly recommended that it be used as a pre-docking screening tool for large virtual screens.

References

1. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2006) eHiTS: a new fast, exhaustive flexible ligand docking system. J Mol Graph Model 7:421–435
2. Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, Rizzo RC (2006) Development and validation of a modular, extensible docking program: DOCK 5. J Comput Aided Mol Des 20(10–11):601–619
3. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749
4. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47(7):1750–1759
5. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52(4):609–623
6. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20(10–11):647–671
7. Jones G, Willett P, Glen RC (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J Comput Aided Mol Des 9(6):532–549
8. Perola E, Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem 47(10):2499–2510
9. Saeh JC, Lyne PD, Takasaki BK, Cosgrove DA (2005) Lead hopping using SVM and 3D pharmacophore fingerprints. J Chem Inf Model 45(4):1122–1133
10. Wagener M, van Geerestein VJ (2000) Potential drugs and non-drugs: prediction and identification of important structural features. J Chem Inf Comput Sci 40(2):280–292
11. Chen B, Harrison R, Papadatos G, Willett P, Wood D, Lewell X, Greenidge P, Stiefl N (2007) Evaluation of machine-learning methods for ligand-based virtual screening. J Comput Aided Mol Des 21(1):53–62
12. Harper G, Bradshaw J, Gittins JC, Green DV, Leach AR (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. J Chem Inf Comput Sci 41(5):1295–1300
13. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801
14. Clark T (2004) QSAR and QSPR based solely on surface properties? J Mol Graph Model 22(6):519–525
15. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41(3):856–864
16. Brandstetter H, Kuhne A, Bode W, Huber R, von der Saal W, Wirthensohn K, Engh RA (1996) X-ray structure of active site-inhibited clotting factor Xa. Implications for drug design and substrate recognition. J Biol Chem 271(47):29988–29992
17. Zell A Stuttgart neural network simulator. University of Stuttgart. http://www-ra.informatik.unituebingen.de/SNNS/ (04/05/2005)
18. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
19. Pham TA, Jain AN (2006) Parameter estimation for scoring protein-ligand interactions using negative training data. J Med Chem 49(20):5856–5868
20. Zhang Q, Muegge I (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 49(5):1536–1548
21. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47(4):1504–1519
22. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50(1):74–82
23. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47(2):488–508
24. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) 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 48(7):2534–2547
25. JChem v 3.2.4. http://www.chemaxon.com/jchem/ (06/20/2007)