NIH Public Access
Author Manuscript
J Chem Inf Model. Author manuscript; available in PMC 2013 January 23.

Published in final edited form as:
J Chem Inf Model. 2012 January 23; 52(1): 16–28. doi:10.1021/ci2002507.

Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-based Pose Scoring and Physical Force Field-based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening

Jui-Hua Hsieh[*,≠], **Shuangye Yin**[†,≠], **Xiang S. Wang**[*], **Shubin Liu**[€], **Nikolay V. Dokholyan**[†,‡], and **Alexander Tropsha**[*,‡]

[*]Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products and Carolina Exploratory Center for Cheminformatics Research, Eshelman School of Pharmacy

[†]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

[€]Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Abstract

Poor performance of scoring functions is a well-known bottleneck in structure-based virtual screening, which is most frequently manifested in the scoring functions' inability to discriminate between true ligands versus known non-binders (therefore designated as *binding* decoys). This deficiency leads to a large number of false positive hits resulting from virtual screening. We have hypothesized that filtering out or penalizing docking poses recognized as non-native (i.e., *pose* decoys) should improve the performance of virtual screening in terms of improved identification of true binders. Using several concepts from the field of cheminformatics, we have developed a novel approach to identifying pose decoys from an ensemble of poses generated by computational docking procedures. We demonstrate that the use of target-specific pose (-scoring) filter in combination with a physical force field-based scoring function (MedusaScore) leads to significant improvement of hit rates in virtual screening studies for 12 of the 13 benchmark sets from the clustered version of the Database of Useful Decoys (DUD). This new hybrid scoring function outperforms several conventional structure-based scoring functions, including XSCORE::HMSCORE, ChemScore, PLP, and Chemgauss3, in six out of 13 data sets at early stage of VS (up 1% decoys of the screening database). We compare our hybrid method with several novel VS methods that were recently reported to have good performances on the same DUD data sets. We find that the retrieved ligands using our method are chemically more diverse in comparison with two ligand-based methods (FieldScreen and FLAP::LBX). We also compare our method with FLAP::RBLB, a high-performance VS method that also utilizes both the receptor and the cognate ligand structures. Interestingly, we find that the top ligands retrieved using our method are highly complementary to those retrieved using FLAP::RBLB, hinting effective directions for best VS applications. We suggest that this integrative virtual screening approach combining

[‡]To whom correspondence should be addressed: Alexander Tropsha, CB #7360, Beard Hall, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360, Tel: 919-966-2955, Fax: 919-966-0204, alex_tropsha@unc.edu, Nikolay V. Dokholyan, 120 Mason Farm Rd., Suite 3097, Genetics Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260, Tel: 919-843-2513, Fax: 919-966-2852, dokh@med.unc.edu.
[≠]These authors contributed equally to the paper.

Supporting Information Available: the distribution of poses generated from re-docking process for all 13 targets; the awROC curves for comparison of (a) MedusaScore, (b) MedusaScore+DistScore, (c) MedusaScore+Pose filter; and the awROCE values in Figure 4 and Figure 5. This information is available free of charge via the Internet at http://pubs.acs.org/.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cheminformatics and molecular mechanics methodologies may be applied to a broad variety of protein targets to improve the outcome of structure-based drug discovery studies.

Introduction

In recent years, structure-based virtual screening (VS) has become an increasingly more popular strategy for computer-aided drug design.[1, 2] Structure-based VS approaches explore available or synthetically feasible chemical databases to identify a relatively small number of high-scoring hits that can be validated experimentally. A successful structure-based VS method can be applied to large data sets of compounds, resulting in significant enrichment of true binders among the top ranking hits.

Employing rigorous scoring functions is essential to the success of structure-based VS campaigns since scoring functions play a critical role in both initial pose generation of compounds (docking) and further ranking of compounds (scoring). Scoring functions can be divided into several types.[3] Force field-based scoring functions predict binding affinity by explicitly accounting for intermolecular interactions such as electrostatic, van der Waals, hydrogen bonding, and hydrophobic interactions. However, due to the static nature of the underlying molecular models many important effects influencing the binding free energy such as entropy, micro-environment dependent polarization, π-stacking, and solvent effects, are often not taken into account. On the other hand, knowledge-based scoring functions employ various statistical parameters derived from experimentally determined protein-ligand structures that reflect their total physical interactions taking the molecular environment into account.[4, 5] Ideally, knowledge-based scoring functions may implicitly capture binding interactions that are difficult to model in force field-based scoring functions.

Despite the increasing popularity of structure-based VS, recent studies have shown that inaccuracy of scoring functions is the major bottleneck of structure-based VS.[6] It has been demonstrated that scoring functions often fail to recognize pose decoys, i.e., ligand poses that are geometrically different from the native binding orientation of the ligand in an experimentally determined crystallographic structure of the protein-ligand complex, yet score better than the native pose. In addition, known non-binders may also score better than true binders; such non-binders are designated as binding decoys.[6, 7] Obviously, the presence of both binding and geometrical pose decoys in an ensemble of compound poses resulting from computational docking studies will decrease the accuracy of structure based VS. Perhaps, in part for this reason, structure-based scoring functions are well-known for having target-dependent VS performances.[6]

Many studies have focused on the development of target-specific customized scoring functions[8] by adding expert-knowledge constraints (e.g., the hinge constraint in kinase targets[9]), or using native pose(s) as references to perform cheminformatics-based similarity ranking (e.g., SIFt[10, 11] and SIFt-variant methods[12, 13]) and to construct pose filters (e.g. pharmacophore filter[14]) to filter out undesired poses. These post-docking pose treatments can effectively improve the discrimination between true ligands and binding decoys in structure-based VS for the aimed target. Furthermore, several studies included pose decoys or poses of inactives in combination with native structures in order to help tuning the scoring functions against binding decoys, which consequently enhanced the accuracy of virtual screening.[15-20]

Herein, by combing the merits of concepts mentioned above, i.e., statistical and force-field based scoring functions we devise a target-specific knowledge based *pose* (-scoring) filter that is trained to distinguish native-like poses from pose decoys. The approach of our knowledge-based pose (-scoring) filter employs protocols that are routinely used in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cheminformatics research, e.g., binary quantitative structure activity relationship (QSAR) modeling, with the caveat that we use unconventional descriptors of the protein/ligand interface for pose scoring as opposed to using standard chemical descriptors of compounds. Unique to our approach is that this classifier is developed using both native-like and pose decoys generated from only one unique protein-ligand x-ray complex. The poses used in training are generated by multiple rounds of docking of this single cognate ligand to its binding target leading to an ensemble of different and diverse poses, where each pose can be defined either as native-like or a decoy based on the value of the pose's Root Mean Square Deviation (RMSD) from the known native pose. In addition, novel protein/ligand interfacial descriptors based on Delaunay tessellation approach and atomic properties derived from conceptual density functional theory (DFT) are applied to represent the poses in training the binary classifier. Unlike the majority of interaction fingerprints used in previous virtual screening studies, our protein/ligand interfacial descriptors represent not only isolated interactions between pairs of atoms, but also atom contact networks at the protein-ligand interface based on tessellation patterns.

Using this filter, we have developed a target-specific *hybrid* scoring function for structure-based virtual screening in an effort to combine the advantages of both knowledge-based pose scoring and force field-based hit scoring functions. For a VS campaign, multiple poses are generated initially for each compound using one of the standard docking approaches. Then, our filter is used to eliminate poses predicted as decoys; the remaining poses are predicted as native-like with a certain level of confidence assigned by the model. These remaining putative native-like poses are ranked with a hybrid score based on the combination of the pose confidence and the binding score assesses by the physical force field-based MedusaScore developed previously in our group.[21]

We test the performance of this novel, hybrid scoring function on several benchmark sets available from the Directory of Useful Decoys (DUD).[22] DUD is a specially designed data set including multiple targets, their known ligands, and decoys, which are compounds that are physically similar to yet topologically distinct from the known ligands. The recently refined DUD data sets include only lead-like compounds and have the true ligands clustered, making it an ideal benchmark set for testing scaffold hopping capability of VS methods.

We use Fred (OpenEye Scientific Software)[23] to dock all compounds to target structures and generate multiple poses for each compounds. We find that for most targets the combination of pose filter and MedusaScore leads to significant improvement in the enrichment of virtual screening hits, as compared with using the MedusaScore scoring function alone. Then we compare the VS performance of several established structure-based scoring functions (XSCORE::HMSCORE[24], Fred::ChemScore[25], Fred::PLP[26], and Fred::Chemgauss3[27]) and several novel VS methods without docking (FieldScreen[28], FLAP::LBX[29], and FLAP::RBLB[29]) that were recently reported to achieve good performances on the same DUD data sets. We find that our structure-based hybrid scoring function outperforms other structure-based scoring functions for majority of the targets. Furthermore, the retrieved ligands are less similar to the cognate ligand in comparison with ligand-based approaches (FieldScreen and FLAP::LBX), and are complementary to the ligands retrieved by another hybrid method (FLAP::RBLB).

Methods

Selection of Targets and Data Sets

The data sets of true ligands and presumed binding decoys for each target in this study are collected from the publicly available Directory of Useful Decoys (DUD)[22]. The DUD data sets were designed to minimize the physical biases inherent in the benchmarking of virtual

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

screening schemes against different biological targets. Each ligand was matched with 36 decoy molecules that resemble the true ligands in physical properties, such as molecular weight, LogP, number of hydrogen bonding groups, and number of rotatable bonds but are distinct from the ligand topologically. In total, the DUD database consists of 40 data sets and for each true ligand there are typically 36 decoy molecules. Further refinement of the DUD data sets is done recently by applying a lead-like filter (MW < 450, AlogP < 4.5) to both ligands and presumed binding decoys[35] as well as the reduced graph cluster filter to ligands[30]. These two filters are intended to mimic the real-life virtual screening campaign and to reduce the analogue bias inflating enrichment in virtual screening. We employ all 13 data sets from the refined FUD set, each of which includes at least 15 ligand clusters, for our method validation. The detailed information about the data sets is shown in Table 1. Six of the 13 targets belong to the kinase family (CDK2, EGFR, p38, PDGFrb, Src, and VEGFr2), where the majority of known ligands occupy ATP binding region. The remaining targets include the class of metalloenzymes (ACE, PDE5), serine protease (FXa), and several other enzymes (AChE, COX-2, HIVRT, and InhA). In order to compare directly with other VS methods, we use the protein-ligand complexes provided in the original DUD for pose filter training. For VEGFr2 and PDGFrb targets, the complex structures provided in the DUD data sets are generated by docking ligands to *apo* protein structures.

Docking Methods for Pose Generation

For each target, we prepare the x-ray structure using utilities available within the Molprobity[31] server to add and optimize hydrogen atoms while correcting potential misinterpretations of amino acid (asparagine, glutamine, or histidine) terminal flips. The crystallographic water molecules located inside the binding pocket are removed in order to avoid biases when generating poses of molecules but cofactors (e.g., NAD in 1p44 protein model) or metal atoms (e.g., Zinc in the 1o86 protein model) are preserved if they are important for enzyme to function or are involved in interactions with the cognate ligand. We add hydrogen atoms of each small molecules using MOE software (version 2007.09)[52] under standard protocols.

We employ the Fred docking software (version 2.2.5) from OpenEye Scientific[23] to generate an ensemble of poses for each compound. The ensemble is generated by enumerating rigid rotations and translations of each conformer within the binding site. The conformers of each compound are generated by Omega (version 2.2.1)[23] based on default parameters and the binding site is defined by a 5 Å grid box centered on the cognate ligand. For kinase targets, it is well-known that hydrogen bond interactions with the protein hinge residues is necessary for both Type I and Type II kinase inhibitors.[32] Thus, this constraint is applied during pose generation to improve docking accuracy.

We apply default parameters provided by Fred during docking except for the number of output poses. For pose filter construction, we retain up to 1000 top-scoring poses generated by docking a single cognate ligand in order to ascertain the conformational diversity of poses. For virtual screening, the top 30 poses (ranked by the Fred's default scoring function, Chemgauss3) of each molecule are preserved for re-scoring by other scoring functions (e.g., MedusaScore).

Ligands *vs*. Binding Decoys and Native-like Poses *vs*. Pose Decoys

"Binding decoys" are defined as ligands that do not bind to a specific target experimentally (non-binders) but score as high as (or better than) true ligands. Similarly, we use the terms "pose decoys" to describe the poses generated by docking the cognate ligand against the protein target but score better than native-like poses. In our study, native-like poses are defined as those generated by docking with binding mode(-s) similar to the native pose. The

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

similarity between docking poses and the native pose is often measured using Root Mean Square Deviation (RMSD). For the purpose of pose-filter training, we define a RMSD threshold of 4Å to classify poses into native-like poses and pose decoys. The 4Å threshold is consistent with the observation that there is a gap on the distribution plot (MedusaScore *vs*. RMSD) of poses generated by re-docking the cognate ligand for most targets (Figure 1, Figure S1).

Novel Descriptors of the Protein-Ligand Interface Based on Conceptual DFT

Earlier, we developed the so called ENTess chemical geometrical descriptors[33] of the protein-ligand interface. These descriptors are obtained by using Pauling electronegativity (EN) as an atomic property and Delaunay Tessellation (Tess) to characterize the protein ligand interface as follows. When applied to protein-ligand complexes represented at the atomic resolution level, Delaunay tessellation partitions the protein ligand interface into an aggregate of space-filling, irregular tetrahedra, with both protein and ligand atoms as vertices. Each Delaunay quadruplet is characterized by its unique four-atom composition, which defines the descriptor type (certainly, the same four-body compositions may occur in different, or even the same, protein/ligand interfaces). Furthermore, for each quadruplet we calculate the sum of En values of the composing atom-vertices, which produces the descriptor value. In the previous study,[33] we used the ENTess descriptors to build successful quantitative structure-binding affinity relationship (QSBR) models for 264 x-ray characterized protein-ligand complexes with known binding affinity; the modeling approach followed our standard model development and validation workflow.[34]

In this study, we have developed and employed novel descriptors that are methodologically similar to ENTess descriptors but are theoretically more rigorous.[35] These new descriptors employ pairwise atomic potentials for the protein-ligand complexes (PL) based on maximal charge transfer (MCT)[36] in place of Pauling electronegativities, called here PL/MCT-tess. The values of PL/MCT-tess descriptors are calculated from the following equation (see also Figure 2):

$$\mathrm{PL/MCTtess_m}=\sum_{k=1}^{n}\sum_{p}^{1\sim3}\sum_{l}^{1\sim3}(\mathrm{MCT_p}*\mathrm{MCT_l}/\mathrm{d_{pl}})_k \quad (1)$$

where PL/MCT-tess$_m$ is the potential of the *m*-th tetrahedron type (i.e., individual descriptor type); *n* is the number of occurrences of this tetrahedron type in a given pose; *p* is the vertex index of a protein atom, *l* is the vertex index of a ligand atom, and d_{pl} is the distance between a pair of protein and ligand atoms found in the same Delaunay tetrahedron. Note that Delaunay tetrahedra at the protein-ligand interface can be classified based on the relative content of protein and ligand atoms, i.e., three protein and one ligand atoms, two from each, or one protein and three ligand atoms; this explains the tetrahedral type counts in the second and third sum in Equation 1.

The MCT characterizes the maximal electron flow between the donor and acceptor atoms at the protein-ligand interface. It is derived from the conceptual DFT[37, 38], which provides a theoretical basis for calculating the PL/MCT-tess descriptors. The MCT is calculated as follows, assuming that the total energy of the system is perturbed by the charge transfer up to the second order:

$$\Delta E=\mu\Delta N+1/2\eta\Delta N^2 \quad (2)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where ΔE and ΔN represent energy change and charge transfer, respectively. When the total energy is minimized with respect to the charge transfer, $d\Delta E/d\Delta N = 0$, we have

$$\Delta N_{max} = \mu/\eta \equiv MCT \quad (3)$$

where μ and η are the chemical potential (negative of electronegativity) and the chemical hardness respectively, defined by $\mu = (\partial E/\partial N)_v$ and $\eta = (\partial^2 E/\partial^2 N)_v$ with ν representing the external potential formed by the framework of atomic nuclei.

Knowledge-based Pose Scoring Filter

As described above, we classify poses generated by docking the cognate ligand against the protein target into native-like and decoy poses based on a certain RMSD threshold. The problem of separating native-like poses *vs*. pose decoys for a molecule can be treated as a binary classification problem where each pose is characterized by the descriptors of the protein-ligand interface (i.e., PL/MCT-tess descriptors in this study). This representation treats the problem of discriminating native-like from decoy poses as a conventional binary classification problem faced in many conventional cheminformatics investigations that employ binary QSAR modeling.

To train this knowledge-based pose scoring function for each target, we retain up to 1000 poses generated by docking a single cognate ligand against the target (Figure 3). For the VEGFr2 and PDGFrb targets, where the native pose is unavailable (only *apo* structure and a model structure are available, respectively), the pose with the lowest MedusaScore is considered as a native pose. This is a reasonable assumption since MedusaScore performed well in an earlier benchmarking exercise for the native pose prediction.[21] We classify the poses based on the 4 Å threshold as either native-like (RMSD ≤ 4Å) or decoys (RMSD > 4Å) except for the PDE5 pose set where the gap is observed at 3 Å and therefore the 3 Å threshold is used. For the poses from docking the cognate ligand of *1ckp* (CDK2), we do not observe a characteristic distribution (as, for example, in Figure 1). Therefore, we regenerate the poses using MedusaDock[39] instead of Fred.

For each pose, we generate PL/MCT-tess descriptors to characterize its interactions with the target protein. The degree of similarity of each pose to the native pose is quantified by the Euclidean distance in the PL/MCT-tess descriptor space. Therefore, the pose distribution of each target's modeling set can be characterized by three parameters: the Euclidean distance to the native pose in the PL/MCT-tess descriptor space (x-axis), the RMSD value (y-axis), and the MedusaScore (color bar). It is desirable that poses with lower RMSD value correspond to smaller distances to the native pose in the PL/MCT-tess descriptor space (e.g., Figure 1).

If a binary data set including native-like (class 1) and decoy (class 2) poses is balanced (the ratio between the two classes is less than two), we randomly exclude 20% poses as the external test set and construct models based on the remaining 80% poses (training set). In the case of imbalanced distribution, we downsize the major class by retaining only those poses that are similar to poses in the minor class, where the degree of similarity is assessed by Euclidean distance in the PL/MCT-tess descriptor space and then use the same model validation procedure. For example, the ACE target has 48 native-like poses and 952 decoy poses; after down-sampling, only 49 decoy poses most similar (in terms of Euclidean distance) to the native-like poses are retained for model building and validation (Table 2 and Figure S1). We note that this approach to down-sampling where the most similar instances of the opposite class are retained naturally makes the classification problem more difficult and therefore it increases the statistical significance of the resulting classification models.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We employ the Support Vector Machines (SVM) software implemented in the open-source LibSVM[40] package to build binary classification models (i.e., pose filters). We use all models with eligible CV accuracy for predicting poses in the external test set and evaluate model external predictive accuracy by its ability to classify native-like versus decoy poses in the external set. The validated models (i.e., pose filter) are used further to score any pose generated in virtual screening as either native like or decoy, expecting that binding decoys will be classified as non-native poses. For each pose, we calculate a FilterScore, which is the fraction of models that predict it as native-like. This process of deriving pose filter is followed for each target in the refined DUD dataset.

It should be emphasized again that only one cognate ligand is used for each target to develop a pose scoring function. However, due to the generic nature of PL/MCT-tess chemical descriptors the respective pose filters can be applied to score all poses of all diverse ligands used in target specific docking and VS studies.

MedusaScore Scoring Function

MedusaScore[21] is a physical force field-based scoring function that describes the major physical interactions between proteins and ligands, including van der Waals interaction, salt bridge, hydrogen bonding and solvation. MedusaScore is an extension of the Medusa force field,[41] which was developed originally to describe physical interactions within proteins. The original parameters of the Medusa force field were trained on 34 high-resolution protein crystal structures with diverse sequences. Thus, by default MedusaScore is expected to be transferable and applicable to virtual screening of a variety of chemical compounds. Notably there were no protein-ligand data used in the development of MedusaScore, but it still exhibits remarkable accuracy in both docking pose discrimination and binding affinity prediction.[21] During the pose rescoring by MedusaScore, we turn off van der Waals repulsion because this term has been shown to be sensitive to small deviation in ligand poses.[21] It is safe to remove the term in this case because all steric clashes have already been considered during the generation of docking poses in eth refined DUD dataset.

Fusion of MedusaScore and FilterScore

In order to combine the FilterScore and the MedusaScore, which are of different nature and scales, we utilize normalized Z-scores based on statistical distributions of respective scores for the ensemble of poses for each ligand. We start by applying the pose filter to all poses generated by docking, and discard poses that are predicted as decoys by all eligible models (i.e., FilterScore = 0). Based on the mean (μ) and standard deviation (σ) of each respective scoring function, the Z-score for each pose is calculated from the respective raw scores (X) using Equation 4.

$$Z=(X-\mu)/\sigma \tag{4}$$

If the filter is constructed based on the entire sampling space of poses from docking the cognate ligand (for balanced datasets), we apply the same weight for both FilterScore and MedusaScore, and the Z-score for each pose is derived as:

$$Z_{combined}=Z_{MedusaScore}-Z_{FilterScore} \tag{5}$$

We add a minus sign for $Z_{FilterScore}$ so that lower Z-score will correspond to better ranked pose, consistent with the MedusaScore convention.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

If the filter is constructed based on the poses after the down-sampling procedure, we employ a modified scoring strategy based on the concept of the applicability domain.[42] We predict poses within the applicability domain using Equation 5 and predict poses out of the applicability domain by adjusting the weight of FilterScore by DistScore (Equation 6).

$$Z_{combined}=Z_{MedusaScore} - 0.5 * (Z_{FilterScore} - Z_{DistScore}) \quad (6)$$

The $Z_{DistScore}$ is the Z-score of each pose based on the its Euclidean distance to the native pose in the PL/MCT-tess descriptor space; the mean and the standard deviation are derived from the distribution of PL/MCT-tess Euclidean distances to the native pose of all docking poses. Assuming a normal distribution of inter-pose distances similar to that for poses from docking of the cognate ligand, we regard docking poses to be within the applicability domain of the modeling set when $Z_{DistScore} < -1$. This threshold is defined by inspecting the distribution of distances between poses in the PL/MCT-tess descriptor space of the modeling sets for five targets (ACE, CDK2, COX-2, HIVRT, and VEGFr2 in Figure S1). The final score for each compound is based on the pose with the lowest sum of Z-scores among all poses retained for that compound.

Evaluation of Virtual Screening Performance

To examine the overall performance of our method for a target data set in virtual screening, we plot the Receiver Operator Characteristic (ROC) curve. We calculate the Area Under the Curve (AUC) value for each ROC curve to estimate the average performance of our method throughout the ranked list. On the other hand, to quantify the performance of each method at the early stage of virtual screening for a target data set, we employ the ROC Enrichment (ROCE) value. Unlike the conventional enrichment factor (EF) metric, ROCE values are independent of the ratio of binding decoys to ligands in a target data set, making them ideal metrics for comparing different methods.[43] The ROCE value is defined as the ratio of true positive rates to the false positive rates, for a given percentage of binding decoys (i.e., the slope at each point on the ROC plot). We report ROCE values at 0.5%, 1%, 2%, 5% as suggested[43-45] and employed in previous publications[28, 29, 46]. The meaning of ROCE value at 1% represents the fold enrichment over random performance. In order to emphasize the retrieval of diverse scaffolds, the above metrics (ROCE and AUC) are modified by applying an arithmetic weight to each ligand (awROCE and awAUC)[47], which is inversely proportional to the size of the cluster it belongs to.

We estimate the uncertainty of awROCE/awAUC values using the statistical bootstrapping procedure.[28] For each ranked list, we randomly exclude 20% of data points and recalculate the awROCE values. This is repeated 10,000 times and the standard deviation of awROCE values is used to estimate the error of awROCE. Due to the nature of pose filter, many true negatives (presumed binding decoys) and some false negatives (true ligands) are eliminated in several data sets (e.g., ACE, p38, and etc). For these data sets, we calculate the awROCE values based on the reduced list, resulting in a larger estimated error at the low percentages.

Comparison against structure-based scoring functions, FieldScreen, and FLAP

Several popular structure-based scoring functions, which are reported to have good docking pose discrimination and binding affinity prediction[21, 48, 49], are selected to compare against our hybrid scoring function. It is intriguing to test the performance of these scoring functions since it has been suggested that scoring functions should be tailored for virtual screening.[19, 50] In total, we have tested five scoring functions including MedusaScore, HMSCORE, Chemgauss3, ChemScore[25], and PLP[26]. HMSCORE is part of the XSCORE[51] scoring utility. Chemgauss3, ChemScore, and PLP are scoring functions implemented in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Fred. Moreover, we also compare our approach with some methods that have been applied to the same data set including FieldScreen[28] and FLAP (both LBX and RBLB protocols)[29]. FieldScreen[28] and FLAP::LBX[29] are two novel ligand-based virtual screening approaches using grid points derived from the cognate ligand as query; FLAP::RBLB approach[29] utilize grid points generated from protein target bound to the cognate ligand. It should be noted that the binding decoys in DUD are designed to be physically similar to, yet topologically distinct from the true ligands. Any ligand-based approaches applied to this data set might generate overly optimistic results.

2D Chemical Similarity to the Cognate Ligand

We generate the MACCS structural keys for each compound using MOE software (version 2007.09)[52] under standard protocols. We calculate the Tanimoto coefficient (Tc) as the similarity metric between the cognate ligand and compounds in the screening library.

Results

Native-like vs. Pose Decoys Classifier

The number of poses used in the construction of the pose filter and the number of models used in predicting external test set or in further virtual screening are shown in Table 2 along with the model statistics for each target. We also present the distribution of poses of modeling set for each target in Figure S1. Depending on the target, the distributions of the native-like poses and pose decoys are either balanced (AChE, EGFR, FXa, InhA, p38, PDE5, PDGFrb, and Src), or shifted towards pose decoys (ACE, CDK2, COX-2, HIVRT, and VEGFr2). The details of modeling techniques to address the imbalanced classes have been described in the Methods. The results show that the overall accuracy for both the training set and the external test set exceeds 90% for all data sets except ACE, HIVRT, and p38. We predict the docking poses generated from each data set using the models which have CV accuracy greater than 90% except for the HIVRT data set which has no models with CV accuracy above 90%. In the latter case, a threshold of 80% is applied.

Initially, we used our standard modeling workflow[53], employing five-fold external cross-validation, to build the pose-filter. We found that models of each fold can achieve high and similar CV accuracy (0.92-0.98), and the VS performance using the models from each individual fold is the same (data not shown). This indicates that such data (poses from one protein-ligand complex) are easy to classify, and the 5-fold external CV modeling procedure does not bring extra reliability to resulting models. Therefore, in this study we adopt a simplified workflow to construct pose-filter models using one random external split (see Methods), and the validated models from this split are used in virtual screening.

It should be emphasized again that for each target-specific filter, we use only one cognate ligand to generate multiple docking poses for further model building. Nevertheless, the filter is applicable to diverse compounds during VS due to the generality of the chemical descriptors we use to characterize the protein-ligand interface. As demonstrated below, these single-ligand based pose filters can significantly improve the accuracy of virtual screening and true hit selection in combination with the MedusaScore force field.

MedusaScore after pose filtering generates better results than MedusaScore alone

We compare the VS performance of MedusaScore and MedusaScore plus pose filter. We apply the protocols to all the 13 targets in the DUD clustered data set. We measure the VS performance of the two scoring protocols using the awROC curves (Figure S2). More specifically, we use awAUC values to measure the overall ligand retrieval of the protocols,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and use awROCE values at 1% to measure the ligand retrieval at the early stage of VS (Figure 4a).

We find that VS performance is remarkably improved over the benchmark set by applying the MedusaScore plus pose filter (i.e., the hybrid scoring function). For all the 13 targets from the refined DUD set, the awAUC values using the hybrid scoring function are consistently higher than using MedusaScore alone. The improvements are least significant for EGFR and VEGFr2 targets, where the awAUC value is improved only by about 0.02 in both cases. This is probably due to the fact that using MedusaScore alone already results in high awAUC values for these two targets (0.83 and 0.65, respectively). For the other targets, the average degree of awAUC improvement is 0.15, and we find the highest improvements are for AChE and FXa.

When comparing the awROCE values at 1%, we find the hybrid scoring function is better than using MedusaScore alone for all targets except Src (Figure 4b). The improvement of awROCE at 1% is most significant for target PDE5 and PDGFrb. For PDE5, we are not able to retrieve any active ligand using MedusaScore alone (awROCE@1% = 0), but the value is improved to approximately 26.5 fold over the random at 1% after combining MedusaScore with the pose filter. The pose filter also improves the ligand retrieval for PDGFrb (awROCE@1% = 43.18), even though the original awROCE value is already high (23.49) using MedusaScore alone. In addition, for the two targets (EGFR and VEGFr2) where the least improvement of awAUC is observed, the awROCE values at 1% are also improved significantly.

Therefore, by combining MedusaScore with pose filter, we not only improve the overall VS performance (as measured by awAUC), but also improve the early enrichment (as measured by awROCE values at 1%). The improvement seems to be more pronounced at the early stage, which is a desirable feature because in practice, only a small fraction of VS hits will be experimentally tested.

MedusaScore plus pose filter approach *vs*. other structure-based scoring functions

We also compare the VS performance of our hybrid scoring function with four popular pose scoring functions, including XSCORE::HMSCORE, Fred::ChemScore, Fred::PLP, and Fred::Chemgauss3. We apply those scoring function to the same docking poses and compare their VS performance at the early screening stage (Figure 5).

We find that our hybrid scoring function outperforms others for most of the targets. At a false positive rate of 0.5%, the hybrid scoring function has the highest enrichment for seven out of the 13 targets. In addition, the awROCE values for those targets vary from 21.66 to 86.46. In contrast, other scoring functions have the best performance at no more than 3 targets, with awROCE values varying from 12.07 to 43. We find a similar trend at the 1% level. In this case, our hybrid scoring function has the highest enrichment for six targets, with awROCE values varying from 22.88 to 43.18, while other scoring functions perform best for at most three targets, with awROCE values in the range of 9.67 to 26.56. This comparison demonstrates that our hybrid scoring function has better and more consistent VS performance than conventional scoring functions.

The hybrid scoring function has the worst performance for Src. We will analyze the possible reasons of Src failure in the Discussion part. For this target, using MedusaScore alone gives reasonably good enrichment factor of 24.77, close to that from using ChemScore (25.97). With the aexception of Src, the hybrid scoring function tends to have the best performance on targets where using MedusaScore alone also gives fairly good enrichment.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

MedusaScore plus pose filter approach *vs*. other novel VS methods

We select a few recently developed VS methods, for which the benchmark results have been reported on the same DUD Cluster data set. One of the methods available for comparison is FieldScreen[28], which is a ligand-based scoring VS method that utilizes molecular fields derived from the cognate ligand as query. Excellent VS performance has also been reported using FLAP[29] molecular field-derived pharmacophores. For FLAP, we compare with two different VS protocols: FLAP::LBX, similar to FieldScreen, which uses ligand-based molecular field, and FLAP::RBLB, which uses both receptor and co-crystallized ligand structure to derive the pharmacophore query. These methods represent the state-of-art VS methods that have been fully tested using the entire DUD clustered set.

The awROC curves of scoring methods for each target are shown in Figure 6 and the awROCE values at each stage are tabulated in Table S6-S10. We find that the VS performance of each scoring method is target-dependent. Our method has the best retrieval for HIVRT, p38, and PDGFrb. RBLB has clearly the best performance for PDE5, Src, and VEGFr2. On the other hand, ligand-based VS methods unquestionably outperform other structure-based methods for COX-2. A close examination of the COX-2 data set reveals that around 47% of true ligands belong to the same cluster as the cognate ligand used as query. To further investigate the chemical similarity of retrieved ligands to the cognate ligand from different scoring approaches, we compare the average Tanimoto coefficient (Tc) values of true ligands from top 20 ranking lists (Table 3). Not surprisingly, we find that ligands retrieved by ligand-based VS methods are chemically more similar to the cognate ligands, as measured by the average Tc values. And the average Tc value of the retrieved COX-2 ligands is 0.88 (e.g., FieldScreen), much higher than the average of those for other 12 targets (0.66). The high degree similarity of COX-2 ligands to the query should result in better performance of any ligand-based VS methods such as FieldScreen and FLAP::LBX methods in this case.

We further compare the early enrichment for our hybrid scoring function and FLAP::RBLB approach because these two methods seem to have the best VS performance at the early stage (in the 0.5% to 5% range). In addition, both methods take advantage of the 3D structures of the receptor and co-crystallized ligands, albeit using different approaches for VS. We want to identify if the different approaches might retrieve different ligands. In fact, we find the two methods seem to be complementary to each other. Among the top 20 hits retrieved by the two methods, we find little overlap of the ligand types (Figure 7). For example, FLAP::RBLB approach is able to retrieve only one cluster for target p38 and PDGFrb, and two clusters for target ACE. In contrast, the MedusaScore with filter approach can retrieve 4, 5, and 7 clusters, for these three targets ACE, p38 and PDGFrb, respectively. Interestingly, the additionally retrieved ligand clusters do not overlap with those obtained using FLAP::RBLB approach. This is also the case for target VEGFr2, where MedusaScore with filter approach retrieved additional five clusters with no overlap with ligands retrieved by FLAP::RBLB method. For other targets such as AChE, CDK2, EGFR, HIVRT, and InhA, only a small fraction of the newly retrieved clusters overlaps with those from FLAP::RBLB approach. Hence, although both methods used receptor and cognate ligand structures for VS, the resulting performance of FLAP::RBLB approach and our approach seem quite complementary for different targets. Combining the two methods shall result in most diverse ligands among the top hits for VS application.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Discussion

Ligand dependency

The atom types in PL/MCT-tess descriptors are defined based on their unique chemical names. This implementation makes PL/MCT-tess descriptors fairly sensitive to special cases (e.g., when tri-fluoro functional group is present in the ligand). However, using poses with such unique types of interactions to construct pose filter makes it too specific. For example, in the Src data set, the cognate ligand used to construct pose filter is ANP, which has a long phospho-aminophosphoric chain uncommon to any lead-like ligands. Unsurprisingly, the pose filter predicts almost everything as pose decoys and the hybrid scoring function deteriorates the VS performance of MedusaScore against the Src data set. When we employ another cognate ligand obtained from *1yol* protein-ligand complex to construct pose filter, the hybrid scoring function has slightly improved VS performance of MedusaScore against the Src data set (awROCE@1% = 27.6 *vs*. 25.5; awAUC = 0.66 *vs*.0.62).

Similarly, the hybrid scoring approach only marginally improves the VS performance of MedusaScore against the COX-2 data set, where the pose filter is constructed based on the cognate ligand with a tri-fluoro functional group. For this case, combining MedusaScore with pure DistScore can easier filter out ligands having the distinctive features of the query compound than using MedusaScore plus pose filter approach (awROCE@1% = 8.7 *vs*. 4.1; awAUC = 0.67 *vs*. 0.39).

Parent scoring function dependency

Theoretically, the target-specific pose scoring filter can be used in combination with any other structure-based scoring function since the definition of pose decoys is based on the RMSD threshold, independent from the scoring function's output. This hybrid scoring approach can improve the VS performance by eliminating binding decoys recognized by pose filter or by increasing weight for the ligands favored by the pose filter. If the ligands favored by the pose filter have relatively poorer scores predicted by the parent scoring function and the high-scoring binding decoys are not completely eliminated, then combining the pose filter and the parent scoring function gives limited improvement. For example, in the CDK2 data set, the Cluster #1, #2, #3, #7, and #8 are favored by both the pose filter and Chemgauss3 but are relatively disfavored by MedusaScore, resulting in better performance when combining pose filter with Chemgauss3 (awROCE@1% = 26.0 *vs*. 14.4; awAUC = 0.84 *vs*. 0.71). Another example is the FXa data set, where combining pose filter with Chemgauss3 has better VS performance (awROCE@1% = 15.4 *vs*. 4.8; awAUC = 0.80 *vs*. 0.72). However, docking programs/scoring functions are well-known for having inconsistent VS performances across diverse targets.[6] Therefore, from the practical point of view, it is more important to improve scoring function performance consistently rather than to achieve ideal results for a few targets. The proposed pose filter is designed with this view in mind.

The evaluation of the threshold to classify native-like versus decoy poses

We find that the 4 Å-threshold seems optimal considering the distribution of RMSD values of poses and the pose filter performance in virtual screening. Lowering the threshold results in fewer native-like poses included, which occupy a smaller portion of the descriptor space; this ultimately leads to a smaller applicability domain of the pose filter. As a result, using this pose filter in VS leads to poorer performance compared with using the pose filter built based on the 4 Å-threshold. The PDE5 data set is an exception, where a clear gap around 3 Å RMSD can be observed on the pose distribution plot. In virtual screening against the PDE5 data set, the performance of the hybrid scoring approach with 3 Å-threshold filter is better than with the filter based on the 4 Å-threshold (awROCE@1% = 26.5 *vs* 17.2; awAUC = 0.75 *vs*. 0.72). Moreover, it should be interesting to include the output of a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

scoring function into the definition of native-like poses and pose decoys (e.g., to train filter only on those native-like poses and pose decoys that are ranked high by the given scoring function), thus, building filters specifically adjusted for each scoring function.

Virtual screening using MedusaScore in combination with DistScore

As shown in Figure 1, poses with lower RMSD value typically have smaller distances to the native pose in the PL/MCT-tess descriptor space. It can be assumed that, for a given molecule, its likelihood to be a true ligand is directly related to how close its pose is to the native pose, which can be reflected by the DistScore. We have applied DistScore in combination with FilterScore to virtual screening for the data sets, where down-sampling during filter construction is necessary. Therefore, for the proper comparison, we also perform virtual screening against all data sets using MedusaScore in combination with DistScore alone (Figure S2). We find that using pose filter to eliminate/penalize pose decoys in virtual screening can consistently improve MedusaScore VS performance and the MedusaScore plus pose filter approach has the best performance for all data sets except for the outliers mentioned above.

Conclusions

We have developed a novel knowledge-based pose (-scoring) filter using concepts frequently employed in cheminformatics research such as chemical descriptors of the protein-ligand interface and machine learning techniques for deriving binary pose classification (native-like *vs*. pose decoys) models. We have combined this novel pose filtering procedure with a recently developed physical force field-based scoring function (MedusaScore) to score the docking poses in virtual screening applications. We have validated this hybrid scoring function using the refined subsets (13 targets) from the DUD database. The refined DUD sets consist of only lead-like compounds and ligands are clustered based on the reduced graph algorithm, making them suitable for testing scaffold hopping capability of VS methods. The validation results demonstrated that our method can consistently improve the VS performance of MedusaScore provided that the protein-ligand complex is suitable for filter training. Comparing with other established structure-based scoring functions, including XSCORE::HMSCORE, Fred::ChemScore, Fred::PLP, and Fred::Chemgauss3, the hybrid scoring function outperforms other methods in six out of 13 data sets at early stage of VS (1% decoys been screened). Moreover, we find that ligands retrieved by the hybrid scoring function are chemically more diverse than those by other two ligand-based VS methods (FieldScreen and FLAP::LBX) using the same DUD data sets. Interestingly, we have observed that our method is complementary to FLAP::RBLB, which is a high-performance VS method that also utilizes both the receptor and the cognate ligand structures.

In summary, we have demonstrated that the hybrid cheminformatics/molecular mechanics based scoring function affords good enrichments in VS experiments and allows for effective scaffold hopping, suggesting that it could be applied to virtual screening against novel pharmaceutically relevant protein targets to identify promising leads. An interesting and unique feature of our approach is that the pose classifier is formally trained to recognize geometrical decoys of a single ligand; yet, it correctly recognizes (and eliminates) most of the binding decoys of multiple test ligands because they are predicted as geometrical decoys. In particular, this method is suitable for protein targets when only limited ligand binding data is available. A single x-ray protein-ligand complex or, as we have demonstrated for PDGFrb target, a homology based protein model with a known binder is sufficient for constructing a successful target-specific pose filter. Additional improvements can be sought for both the pose (-scoring) filter (e.g., using more than one ligand for training, employing alternative atomic properties or potentials for ENTess-like scoring functions, or

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

incorporating other features, e.g., shape parameters, characterizing Delaunay tetrahedron generated at the protein-ligand interface), as well as approaches for the integration of knowledge-based and physical force field-based scoring functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Alexander Golbraikh for helpful discussions and constructive criticism in the course of this research project. We also thank Dr. Alexander Sedykh for proofreading the manuscript and thoughtful discussions. We are grateful to Chemical Computing Group and OpenEye Scientific for software grants. Finally, we acknowledge access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The studies reported in this paper were supported in part by the NIH research grants GM066940 and its ARRA supplement GM066940-06S1 (AT), the University Cancer Research Fund Innovation Award (AT), the NIH research grant R01GM080742 (ND), and the NC TraCS Institute Pilot Grant 2KR10813 (XW).

Reference List

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A Geometric Approach to Macromolecule-Ligand Interactions. J Mol Biol. 1982; 161:269–288. [PubMed: 7154081]
2. Schneider G, Bohm HJ. Virtual screening and fast automated docking methods. Drug Discov Today. 2002; 7:64–70. [PubMed: 11790605]
3. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004; 3:935–949. [PubMed: 15520816]
4. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem. 1999; 42:791–804. [PubMed: 10072678]
5. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000; 295:337–356. [PubMed: 10623530]
6. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem. 2006; 49:5912–5931. [PubMed: 17004707]
7. Graves AP, Brenk R, Shoichet BK. Decoys for docking. J Med Chem. 2005; 48:3714–3728. [PubMed: 15916423]
8. Jansen JM, Martin EJ. Target-biased scoring approaches and expert systems in structure-based virtual screening. Curr Opin Chem Biol. 2004; 8:359–364. [PubMed: 15288244]
9. Perola E. Minimizing false positives in kinase virtual screens. Proteins: Struc Funct Bioinf. 2006; 64:422–435.
10. Chuaqui C, Deng Z, Singh J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. J Med Chem. 2005; 48:121–133. [PubMed: 15634006]
11. Nandigam RK, Kim S, Singh J, Chuaqui C. Position Specific Interaction Dependent Scoring Technique for Virtual Screening Based on Weighted Protein-Ligand Interaction Fingerprint Profiles. J Chem Inf Model. 2009; 49:1185–1192. [PubMed: 19415918]
12. Perez-Nueno VI, Rabal O, Borrell JI, Teixido J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. J Chem Inf Model. 2009; 49:1245–1260. [PubMed: 19364101]
13. Mpamhanga CP, Chen BN, Mclay IM, Willett P. Knowledge-based interaction fingerprint scoring: A simple method for improving the effectiveness of fast scoring functions. J Chem Inf Model. 2006; 46:686–698. [PubMed: 16562999]
14. Muthas D, Sabnis YA, Lundborg M, Karlen A. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. J Mol Graphics Modell. 2008; 26:1237–1251.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

15. Huang SY, Zou XQ. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem. 2006; 27:1876–1882. [PubMed: 16983671]

16. Teramoto R, Fukunishi H. Structure-based virtual screening with supervised consensus scoring: Evaluation of pose prediction and enrichment factors. J Chem Inf Model. 2008; 48:747–754. [PubMed: 18318474]

17. Teramoto R, Fukunishi H. Supervised consensus scoring for docking and virtual screening. J Chem Inf Model. 2007; 47:526–534. [PubMed: 17295466]

18. Teramoto R, Fukunishi H. Consensus scoring with feature selection for structure-based virtual screening. J Chem Inf Model. 2008; 48:288–295. [PubMed: 18229906]

19. Englebienne P, Moitessier N. Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins. J Chem Inf Model. 2009; 49:2564–2571. [PubMed: 19928836]

20. Sato T, Honma T, Yokoyama S. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening. J Chem Inf Model. 2010; 50:170–185. [PubMed: 20038188]

21. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV. MedusaScore: An accurate force field-based scoring function for virtual drug screening. J Chem Inf Model. 2008; 48:1656–1662. [PubMed: 18672869]

22. Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem. 2006; 49:6789–6801. [PubMed: 17154509]

23. Fred, version 2.2.5 & Omega, version 2.2.1. OpenEye Scientific Software; Santa Fe, NM: 2009.

24. Wang RX, Lai LH, Wang SM. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput-Aided Mol Des. 2002; 16:11–26. [PubMed: 12197663]

25. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions. 1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput-Aided Mol Des. 1997; 11:425–445. [PubMed: 9385547]

26. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand-protein complexes. J Comput-Aided Mol Des. 2000; 14:731–751. [PubMed: 11131967]

27. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. Biopolymers. 2003; 68:76–90. [PubMed: 12579581]

28. Cheeseright TJ, Mackey MD, Melville JL, Vinter JG. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. J Chem Inf Model. 2008; 48:2108–2117. [PubMed: 18991371]

29. Cross S, Baroni M, Carosati E, Benedetti P, Clementi S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. J Chem Inf Model. 2010; 50:1442–1450. [PubMed: 20690627]

30. Good AC, Oprea TI. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput-Aided Mol Des. 2008; 22:169–178. [PubMed: 18188508]

31. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D: Biol Crystallogr. 2010; 66:12–21. [PubMed: 20057044]

32. Liu Y, Gray NS. Rational design of inhibitors that bind to inactive kinase conformations. Nat Chem Biol. 2006; 2:358–364. [PubMed: 16783341]

33. Zhang SX, Golbraikh A, Tropsha A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem. 2006; 49:2713–2724. [PubMed: 16640331]

34. Tropsha A, Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. Curr Pharm Des. 2007; 13:3494–3504. [PubMed: 18220786]

35. Geerlings P, De Proft F, Langenaeker W. Conceptual density functional theory. Chem Rev. 2003; 103:1793–1873. [PubMed: 12744694]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

36. Parr RG, Von Szentpaly L, Liu SB. Electrophilicity index. J Am Chem Soc. 1999; 121:1922–1924.
37. Liu SB. Conceptual Density Functional Theory and Some Recent Developments. Acta Phys Chim Sin. 2009; 25:590–600.
38. Parr RG, Yang WT. Density-Functional Theory of the Electronic-Structure of Molecules. Annu Rev Phys Chem. 1995; 46:701–728.
39. Ding F, Yin S, Dokholyan NV. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. J Chem Inf Model. 2010; 50:1623–1632. [PubMed: 20712341]
40. Chang, C-C.; Lin, C-J. *LIBSVM*: a library for support vector machines; ACM Transactions on Intelligent Systems and Technology. 2011. p. 27:1-27:27.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
41. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comput Biol. 2006; 2:725–733.
42. Golbraikh A, Tropsha A. Beware of q(2)! J Mol Graphics Modell. 2002; 20:269–276.
43. Nicholls A. What do we know and when do we know it? J Comput-Aided Mol Des. 2008; 22:239–255. [PubMed: 18253702]
44. Jain AN, Nicholls A. Recommendations for evaluation of computational methods. J Comput-Aided Mol Des. 2008; 22:133–139. [PubMed: 18338228]
45. Hawkins PCD, Warren GL, Skillman AG, Nicholls A. How to do an evaluation: pitfalls and traps. J Comput-Aided Mol Des. 2008; 22:179–190. [PubMed: 18217218]
46. Jahn A, Hinselmann G, Fechner N, Zell A. Optimal assignment methods for ligand-based virtual screening. J Cheminform. 2009; 1:14. [PubMed: 20150995]
47. Clark RD, Webster-Clark DJ. Managing bias in ROC curves. J Comput-Aided Mol Des. 2008; 22:141–146. [PubMed: 18256892]
48. Cheng TJ, Li X, Li Y, Liu ZH, Wang RX. Comparative Assessment of Scoring Functions on a Diverse Test Set. J Chem Inf Model. 2009; 49:1079–1093. [PubMed: 19358517]
49. Englebienne P, Moitessier N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins? J Chem Inf Model. 2009; 49:1568–1580. [PubMed: 19445499]
50. Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP. eHiTS: an innovative approach to the docking and scoring function problems. Curr Protein Pept Sci. 2006; 7:421–435. [PubMed: 17073694]
51. Wang RX, Lu YP, Wang SM. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem. 2003; 46:2287–2303. [PubMed: 12773034]
52. Molecular Operating Environment (MOE), [2007.09]. Chemical Computing Group (CCG); Montreal, Canada: 2007.
53. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inf. 2010; 29:476–488.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

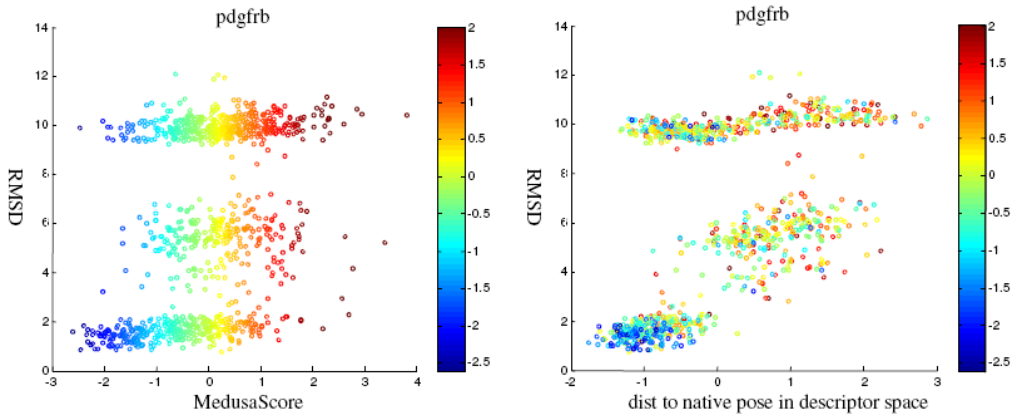

Figure 1.

The distribution of poses generated by re-docking the ligand structure obtained from the DUD website against the PDGFrb homology protein model. The pose with the lowest MedusaScore is served as the reference to calculate the RMSD value of poses (the lower MedusaScore values correspond to higher ranks). The left plot shows the pose distribution based on Z-score values of MedusaScore (x-axis) *vs*. RMSD values (y-axis). The right plot shows the pose distribution based on Z-score values of distance to the native pose in PL/MCT-tess descriptor space (x-axis) *vs*. RMSD values (y-axis). The data points are colored corresponding to their Z-score values of MedusaScore.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

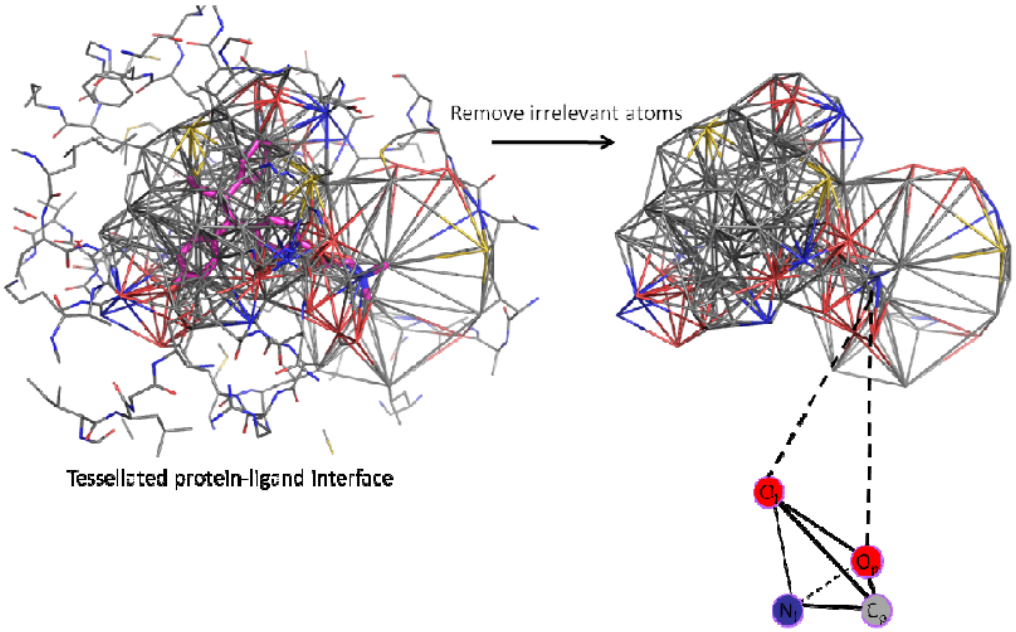

Figure 2.

Illustration of the method to derive PL/MCT-tess descriptors using the tesselated protein-ligand interface (e.g., 3ERT). The atom types for protein and ligand are treated differently. For instance, for the tetrahedron at the left corner, C_p and O_p are carbon and oxygen atoms from the protein while O_l and N_l are oxygen and nitrogen atoms from the ligand.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

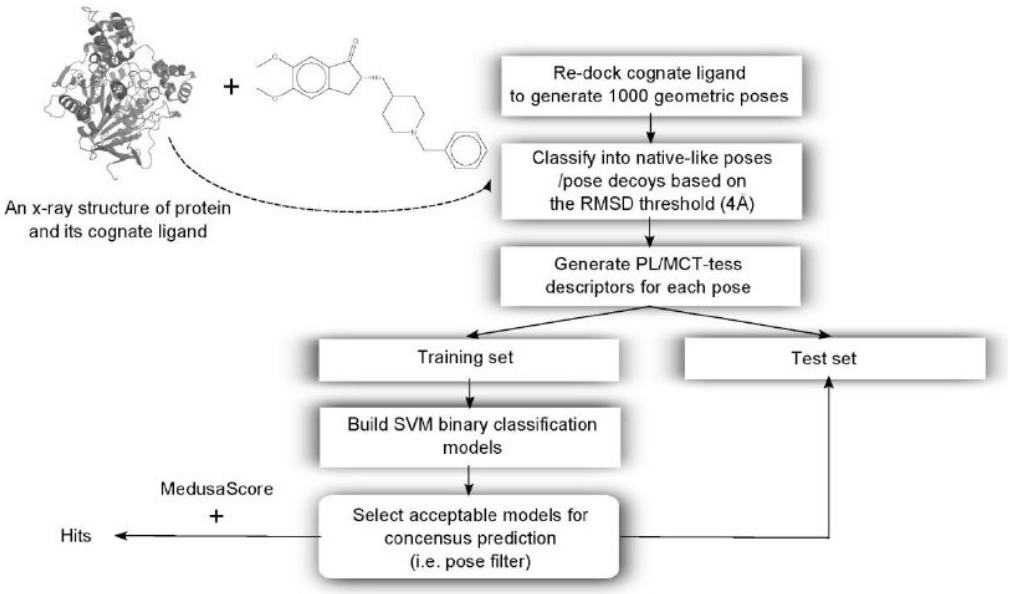

Figure 3.
Flowchart of the approach described in this paper for developing target-specific pose filters, and their use in combination with MedusaScore for VS.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

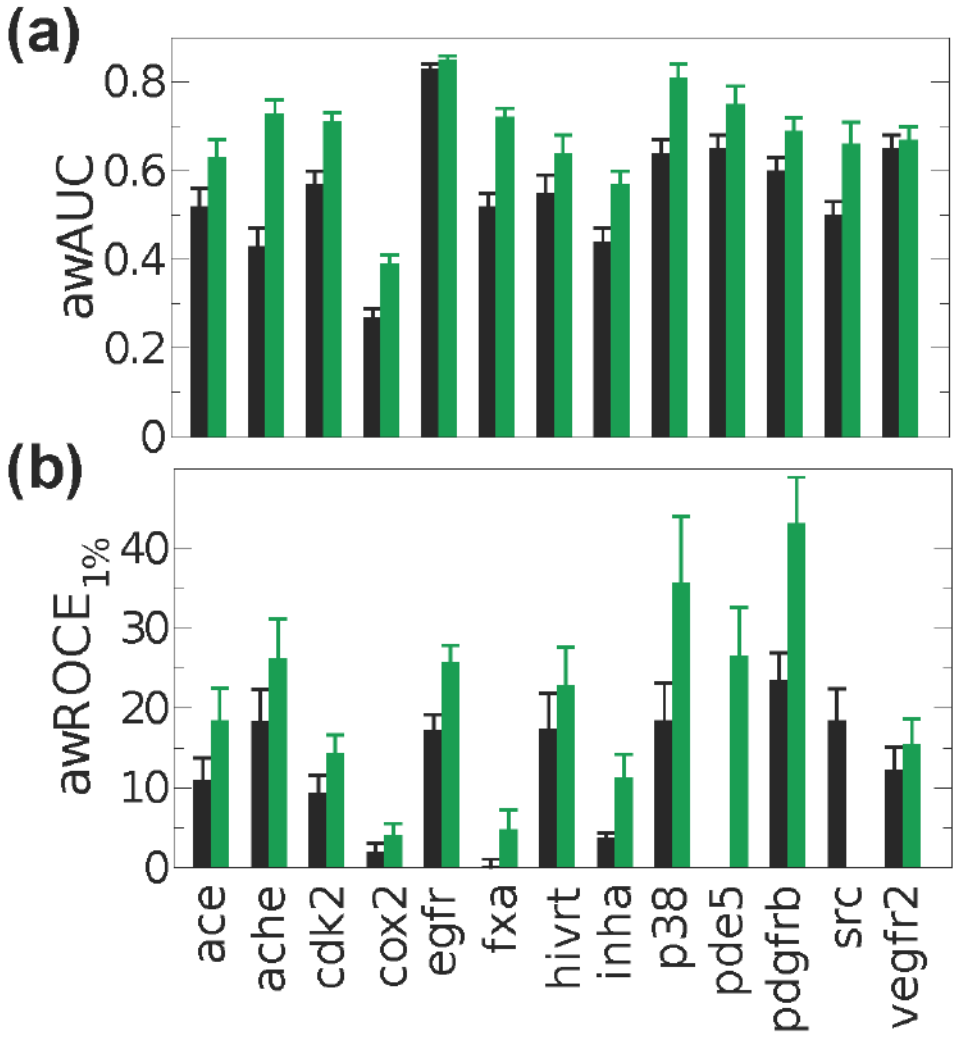

Figure 4.
The awROCE values at 1% (a) and awAUC values (b) of MedusaScore (black) and MedusaScore + filter approach (dark green) for each target.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

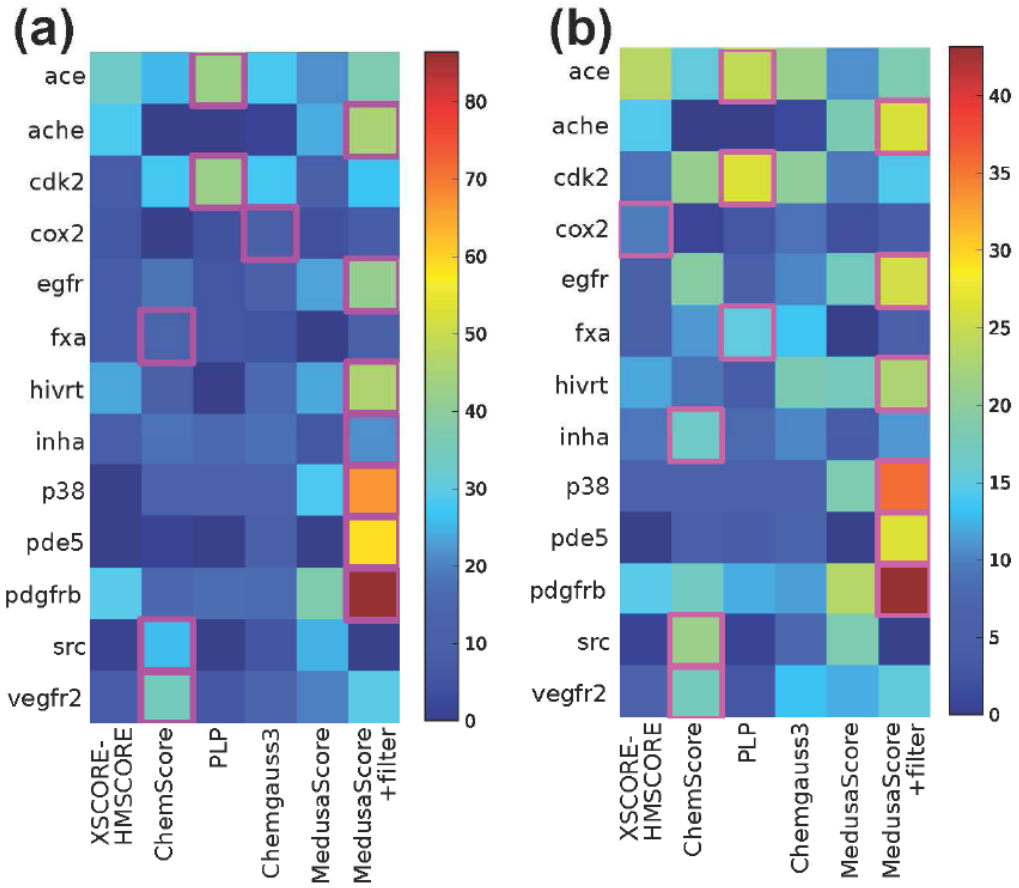

Figure 5.

The heat map of awROCE values at 0.5% (a) and 1% (b) of several popular structure-based scoring functions (XSCORE::HMSCORE, ChemScore, PLP, Chemgauss3, and MedusaScore) as well as MedusaScore plus Filter approach for each target. We highlight the highest awROCE values of a scoring method against a particular target (purple box)

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

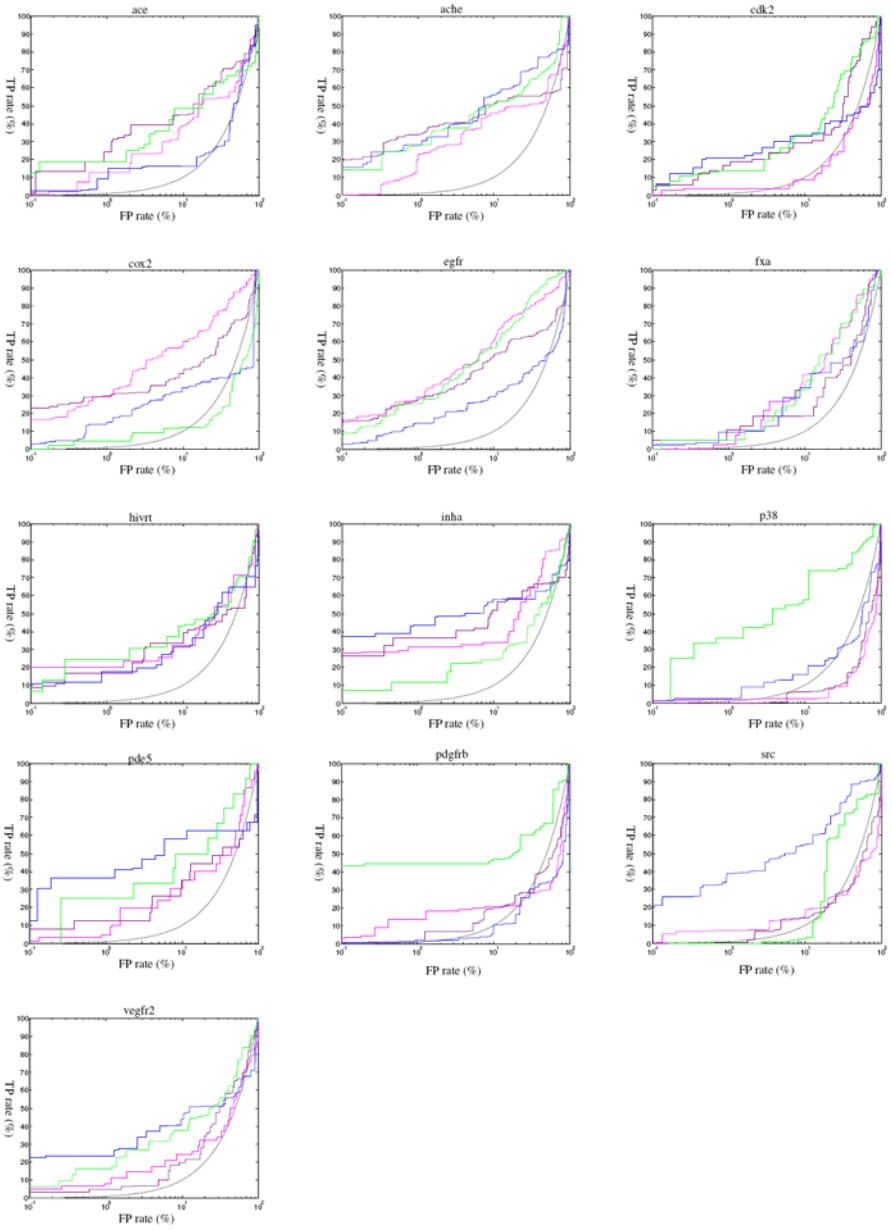

Figure 6.

The awROC curves of VS experiments for 13 DUD data sets. For each target, the true positive (FP) rate is plotted against the logarithmic false positive (FP) rate. Gray dot dash lines correspond to the random VS performance, magenta lines are from FieldScreen, purple lines are from FLAP (LBX), blue lines are from FLAP (RBLB), and green lines are from the MedusaScore + pose filter approach

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

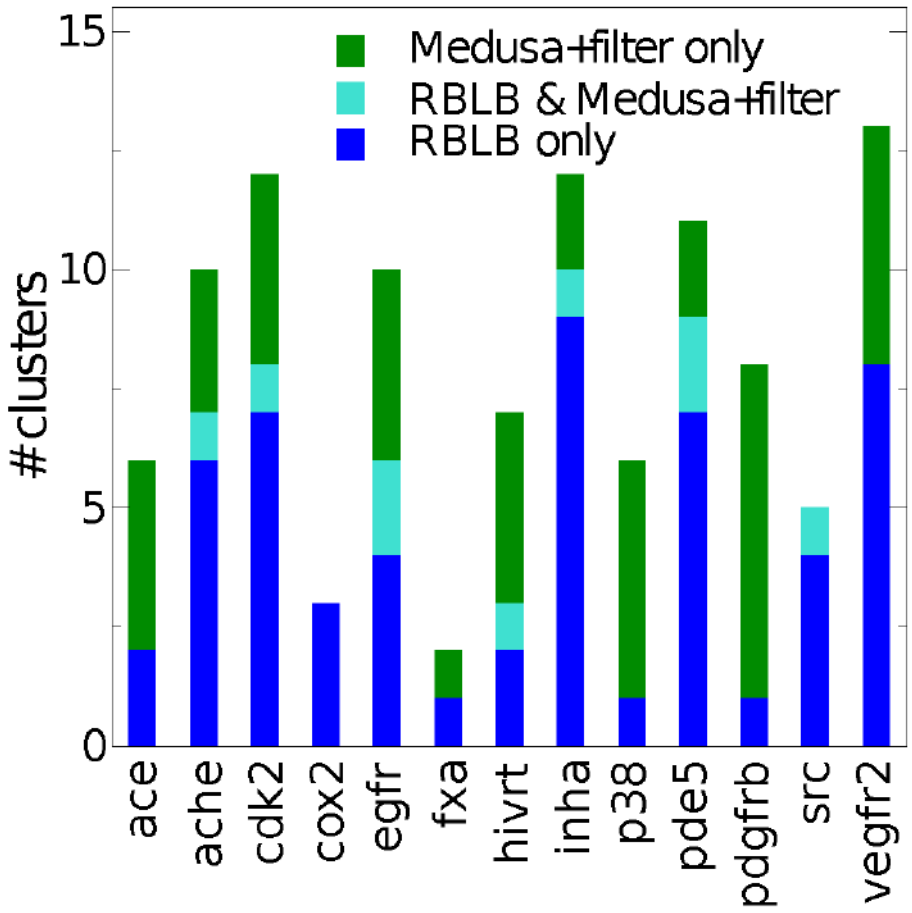

Figure 7.

The analysis of ligand cluster type retrieval of MedusaScore + filter approach and FLAP::RBLB approach from top 20 ranking list of each data set. We rearrange the retrieve clusters of each target based on a) the clusters only retrieved by MedusaScore + filter approach (green); b) the clusters only retrieved by FLAP::RBLB approach; c) the overlapping clusters of two approaches (cyan).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Summary of benchmark data sets used in studies described in this paper. The data sets are obtained from DUD website.

Target	Function	PDB	# of ligands	# of decoys	# of clusters
ace	metallopeptidase	1o86	46	1726	19
ache	acetylcholine esterase	1eve	99	3631	19
cdk2	serine/threonine kinase	1ckp	47	1776	32
cox2	cyclooxygenase	1cx2	212	11841	44
egfr	tyrosine kinase	1m17	365	14516	40
fxa	serine protease	1f0r	64	1888	19
hivrt	HIV reverse transcriptase	1rti	34	1415	17
inha	enoyl ACP reductase	1p44	57	2501	23
p38	serine/threonine kinase	1kv2	137	6230	20
pde5	phosphodiesterase	1xp0	26	1562	22
pdgfrb	tyrosine kinase	model[a]	124	5265	22
src	tyrosine kinase	2src	98	5216	21
vegfr2	tyrosine kinase	1vr2[b]	48	2479	31

[a]: protein structure is homology model, the ligand structure is taken from the DUD website

[b]: apo structure, the ligand structure is taken from DUD website

HIV: Human Immunodeficiency Virus; ACP: Acyl Carrier Protein

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Statistics of target-specific pose filters.

Targets	Training set			Num. models with CV accuracy ≥ 0.9[b]	Test set		
	Num. native-like poses	Num. pose decoys	CV accuracy[a]		Num. native-like poses	Num. pose decoys	Prediction accuracy
ace	48	49	0.93	99	13	12	0.89
ache	437	363	0.96	184	104	94	0.98
cdk2	168	245	0.97	188	44	60	0.96
cox2	125	96	0.96	148	36	20	0.99
egfr	296	504	0.94	157	74	126	0.97
fxa	384	416	0.94	150	100	100	0.94
hivrt	168	121	0.84	103	44	29	0.84
inha	296	504	0.96	116	78	122	0.96
p38	20	24	0.91	25	6	5	0.82
pde5	295	505	0.96	171	74	126	0.91
pdgfrb	276	524	0.96	149	73	127	0.95
src	444	356	0.94	157	112	88	0.94
vegfr2	132	103	0.93	129	35	27	0.95

[a] Average CV accuracy is derived from all eligible models with CV accuracy greater than 0.9 except for the HIVRT data set which has no models with CV accuracy above 0.9. Therefore, a 0.8 threshold is applied

[b] A 0.8 threshold is applied for the HIVRT data set

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Average 2D Tc of the active ligands retrieved from the top 20 ranking list of scoring approaches (FieldScreen, FLAP::LBX, FLAP::RBLB, and MedusaScore + filter).

Target	2D similarity			
	FieldScreen	**FLAP::LBX**	**FLAP::RBLB**	**MedusaScore + Filter**
ace	0.75	0.76	0.59	0.74
ache	0.75	0.81	0.52	0.48
cdk2	0.72	0.50	0.49	0.70
cox2	0.88	0.88	0.68	NA
egfr	0.58	0.50	0.45	0.64
fxa	0.49	0.95	0.45	0.45
hivrt	0.78	0.79	0.60	0.59
inha	0.82	0.82	0.69	0.81
p38	0.66	NA	0.45	0.57
pde5	0.74	0.67	0.57	0.69
pdgfrb	0.64	0.64	0.47	0.66
src	0.44	NA	0.47	0.45
vegfr2	0.59	0.67	0.48	0.47
Aver. Similarity	0.68	0.73	0.53	0.60