J Comput Aided Mol Des (2011) 25:1033–1051
DOI 10.1007/s10822-011-9482-5

Improving molecular docking through eHiTS' tunable scoring function

Orr Ravitz · Zsolt Zsoldos · Aniko Simon

Received: 27 April 2011 / Accepted: 31 October 2011 / Published online: 11 November 2011

Abstract We present three complementary approaches for score-tuning that improve docking performance in pose prediction, virtual screening and binding affinity assessment. The methodology utilizes experimental data to customize the scoring function for the system of interest considering the specific docking scenario. The tuning approach, which has been implemented as an automated utility in eHiTS, is introduced as a solution to one of the conundrums of the molecular docking paradigm, namely, the lack of a universally well performing scoring function. The accuracy of scoring functions has been shown to be generally system-dependent, and particularly lacking for binding energy and bio-activity predictions. In the proposed approach, pose and energy predictions are enhanced by adjusting the relative weights of the eHiTS energy terms to improve score-RMSD or score-affinity correlations. In a virtual screening context ligand-based similarity is used to rescale the docking score such that better enrichment factors are achieved. We discuss the algorithmic details of the methods, and demonstrate the effects of score tuning on a variety of targets, including CDK2, BACE1 and neuraminidase, as well as on the popular benchmarks—the Directory of Useful Decoys and the PDBBind database.

Keywords eHiTS · Docking · Screening · Binding affinity · Scoring function · Score tuning

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9482-5) contains supplementary material, which is available to authorized users.

O. Ravitz (✉) · Z. Zsoldos · A. Simon
SimBioSys Inc., 135 Queen's Plate Dr. Unit 520, Toronto, ON M9W 6V1, Canada
e-mail: ravitz@simbiosys.com
URL: http://www.simbiosys.com/

Abbreviations

HTS	High throughput screening
DUD	Directory of useful decoys
ISP	Interaction surface point
NA	Neuraminidase
RMSD	Root mean square deviation
ROC	Receiver operating characteristic
AUC	Area under the curve

Introduction

Molecular docking has consistently gained grounds in recent years as a major tool for computer aided drug discovery and drug design. It is utilized in three main scenarios: binding mode prediction, virtual screening, and binding affinity evaluation. Pose prediction is primarily used to rationalize observed experimental phenomena and trends, construct activity hypotheses and generate predictions for molecular modifications that may enhance or subdue certain properties such as binding affinity and specificity. In virtual screening, docking is commonly used to predict the likelihood of compounds to be hits in high-throughput screening (HTS) assays. Libraries of molecules are docked to a model of the target protein, and the docking score is used as a rough predictive measure for activity. In a more refined setting, docking is used to estimate the energy change upon binding of the ligand in the receptor active site. This binding affinity prediction is typically used to prioritize known hits and closely related compounds for subsequent steps in the drug discovery process.

Throughout the years progress in the molecular docking paradigm has been continuously tracked and measured by

researchers, and has been documented in numerous methodological and comparative studies that evaluated the various facets of the problem. While major advances have been observed, several problems have been haunting the field since its inception. Most notably, many studies point out to the following [1–9]: (1) pose prediction capability is target- and ligand-dependent and no tool is found to perform uniformly well across all protein families, (2) docking is lagging behind computationally cheaper ligand-based methods as a screening tool, and (3) correlation between docking scores and binding affinities is limited in general. Those problems should be attributed to shortcomings in scoring functions, and not to any difficulty in generating binding modes that are faithful to experimentally observed ones. Indeed, by adequate sampling of the conformational space one could generate poses that are very close to the native binding mode, but without a proper scoring function one would not be able to prune the conformational space in a meaningful way, and to identify the best solutions.

While method developers continue to pursue improvements in scoring capabilities, researchers employing docking programs often choose one or two tools that perform reasonably well for their systems of interest, or, less frequently, try to combine several tools using, for example, consensus scoring techniques [10]. In this paper we focus on an alternative approach, implemented in eHiTS, that takes advantage of experimental data to customize the scoring function to the system of interest, and to the specific task. Complementing ongoing efforts to improve the underlying scoring methods, this approach offers a framework for improving the scoring performance in a systematic and target-oriented way. The various types of tuning are utilized during the training of eHiTS' scoring function, and are also available to users as a complementary package. In a nutshell, score-tuning identifies and enhances those elements in the scoring function that contribute most to the proper modeling of the biological system, or rescales the score based on the ability of the ligand to promote key interactions in the binding pocket. The notion of fitting scoring parameters to reproduce experimental data is very well rooted in the scoring and docking field and the tuning approach may be best presented in this context.

Scoring functions in docking programs make assumptions and simplifications in an effort to find a balance between computational cost and accuracy of the results. Essentially there are three classes of scoring functions used in docking programs [11, 12]: force-field based, empirical and knowledge based. The types differ by the terms used to construct the overall score, and by the experimental data used for their parameterization.

Knowledge-based, or statistical scoring functions use statistics collected from structures of protein–ligand complexes to extract rules on preferred and non-preferred atomic interactions. The scoring function is trained to reproduce crystallographic or NMR binding modes. Common examples of knowledge based scoring functions include PMF [13], DrugScore [14] and PAS-Dock [15].

Empirical scoring functions are sums of uncorrelated terms, typically with a physical interpretation, such as Van der Waals or electrostatic interactions and hydrogen bonds. The terms and their relative weights are parameterized to reproduce experimental data, such as binding energies or conformations. ChemScore [16], X-Score [17], GlideScore [18], FlexX F-Score [19], PLP [20] and LigScore [21] are examples of empirical scoring functions.

Force-field based scoring functions are similar to empirical scoring functions in that they attempt to predict binding energies of ligands by adding individual contributions from different types of interactions. However, force-field based scoring functions use interaction terms derived from physical chemical phenomena as opposed to experimental affinities. Some examples of force-field based scoring functions include GoldScore [22], AutoDock [23], Glide-Emodel [24] and DOCK Energy [25].

On this crowded scale of scoring approaches, eHiTS Score combines elements from both the knowledge-based and the empirical approaches [26, 27]. It consists of several phenomenological terms that are derived from statistics collection on the PDB, and additional empirical terms that account for explicit physical interactions and effects like electrostatics, Van der Waals interactions and entropy loss upon binding. The relative weights given to each of the terms are used as adjustable parameters that can be tuned to reproduce different experimental properties. A traditional approach would adjust a single weight set to optimally reproduce crystal structures of a set of complexes, or to generate good correlation with binding affinity values. In this paper we describe how we use the abundance of experimental data to create multiple system- and goal-oriented weight sets rather than try to optimize a global (universal) weight set. Several similar efforts were reported in the literature. For example, Fradera et al. used a similarity metric between generated poses and a reference structure or a pharmacophore model to promote certain conformations and penalize others [28]. Mooij and Verdonk developed receptor-targeted scoring functions by modifying the atoms pair-wise interaction potential based on statistical analysis of complexes [29]. More recent studies used support vector machine and other approaches to develop target specific scoring functions [30, 31]. However, none of the published methods has been applied on a large scale to achieve extensive protein-family coverage, or has been integrated into a standard docking tool.

The three, complementary tuning approaches presented hereafter principally avoid the universality challenge and achieve improved binding mode and binding affinity

predictions as well as enhanced enrichment capabilities by utilizing experimental data for score customizations. Although the suggested methods are general, their implementation in eHiTS is particularly straightforward due to the structure and technical format of the scoring function, and due to the inherent knowledge-based orientation of the program. The approaches are used to equip the docking program with pre-tuned scoring schemes tailored to specific protein families and to the docking purpose. In addition, a special utility provides users with the tools to customize the scoring function with proprietary, recent, or otherwise selected data as they find fit.

The paper is organized as follows. We first provide a concise account of the eHiTS scoring function and the geometric family clustering concept which constitute the basis for the tuning process. Rank, enrichment and affinity tuning are then described in the theoretical section. Following the methods section we provide results and analysis for several test cases. BACE1 and CDK2 are used to examine several aspects of the rank-tuning, the type of score customization that is designed to improve the scoring of binding modes which are more faithful to crystal structures. Neuraminidase and the Directory of Useful Decoys (DUD) set are used to evaluate the enrichment tuning technique in virtual screening, and the PDBBind database is finally used as a benchmark for affinity tuning. Further insights and future directions are discussed in the conclusion.

Tuning approaches in eHiTS

The eHiTS score

The knowledge-based elements in eHiTS Score associate energy values with geometric patterns found in the PDB using the Boltzman principle. The primary device to identify and model key interactions in protein–ligand complexes in eHiTS is the concept of interaction surface points (ISPs). ISPs are placed on the surface of the ligand and the binding pocket during the stage of chemical perception of the system. Twenty-three different ISP types are used to represent a broad spectrum of chemical interactions, those include hydrogen bond donors and acceptors classified into various strengths, various types of π interactions contributors, hydrophobic elements, metal ions etc. (full list of ISP types is provided in the supporting materials, online resource 1). The pairwise interactions between ISPs are both distance- and direction-dependent, and the statistics collection from the PDB is carried out accordingly. Statistics is collected using a subset of the PDB that includes only structures of protein–ligand complexes solved at resolution of 2.5 Å or better. In addition, various filters are employed to guarantee a reasonable level of reliability by removing structures with clear errors or ambiguities using in-house tools as well as publicly available resources, primarily WHAT_CHECK [32] and the Electron Density Server [33]. Roughly 13,000 structures currently provide the basis for statistics collection, and those offer many millions of ISP coordinates that are used to generate this part of the eHiTS energy function. The ISPs are used to model interactions between the ligand and the receptor, as well as ligand intramolecular interactions.

Additional phenomenological terms in the eHiTS Score include a burial factor, hydrophobicity-weighted binding pocket surface coverage by the ligand, and other terms. In addition, there are several empirical terms modeling Coulomb and Van der Waals interactions, de-solvation effects, ligand strain energy, and loss of entropy upon binding.

In total, the various contributions to the score are given as a weighted sum of 20 terms:

$$E = \sum_{i=1}^{20} \omega_i e_i \quad (1)$$

where the e_i is the energy term associated with an empirical or knowledge-based contribution, and the ω_i is a weight parameter.

Protein family recognition and clustering

A certain metric is required to determine whether experimental data have enough in common to justify customization of the scoring function based on them. The concept of the protein family is used in eHiTS to designate structures that can be used collectively to customize a scoring scheme, and also to match targets to such clusters of structures and use the pre-tuned scoring during docking runs. The protein family classification is automated in eHiTS and is based on the geometric pattern of residues in the active site. The types of residues forming the active site is collected for every protein, and the 3D coordinates of the alpha carbon of each residue is used to compute a distance matrix between the binding pocket residues. Clustering of proteins into families is performed based on the similarity of the distance matrices. The clustering is an on-line process in which each receptor is compared to existing characteristic matrices. If similarity to any of the existing matrices falls short of the minimal similarity criteria (see below), the receptor will form a new family with its own geometry as the characteristic matrix. Otherwise, the receptor will be matched to the most similar family, and a new characteristic matrix will be generated for the family, based on the average across family member of the coordinates of the alpha carbons in the binding pocket.

The minimum criteria for two proteins to fall into the same family is to have at least 5 residues with all their pairwise distances compatible, the default tolerance for the distance difference is 3.0 Å. The tolerance and the minimal number of compatible residues are adjustable control parameters. The default values of those parameters have shown the greatest agreement with the stated biological categorization of receptors in the comment and header sections of the PDB entries. In other words, with these criteria, one finds that in many cases geometric families as defined by eHiTS correspond to biological families. There are, however, cases where two biological families are clustered together geometrically, as in the case of thrombin and trypsin, and there are cases where a biological family is split to two geometric families. The latter is commonly found in kinases where the in/out orientation of the DFG motif is a geometrically strong enough dividing factor.

As an example, the PDB contains many types of neuraminidase (NA) from a variety of organisms. Most notably, viral NA is prime target for the development of influenza treatment [34], and many NA structures are therefore associated with influenza viral strains. 90 crystal structures that were collected from the PDB for this study (see below) span 14 geometric families of various sizes, and two additional receptors are singletons. The biggest cluster in this set has 40 complexes, while most of the other clusters consist of 3–4 structures. It is interesting to note that the geometric family clustering generates good correlation with organisms and strains. For example, the largest family consists of influenza A Australia/G70C/75 and Tokyo/3/67 structures, while the Beijing/1/87 and Lee/40 strains of influenza B virus cluster each separately.

The choice of binding pocket geometry as the exclusive family defining criterion stems from a few reasons. First, given the size of the dataset (the PDB) automation is essential. Since there is no consistency in the biological classification in the PDB files, relying on the information provided therein will require a significant curation work. Replacing the criterion with a sequence similarity parameter will most likely improve the biological classification, but will have the negative effect of clustering together significantly different conformations of the same protein, such as DFG in/out conformations in kinases, which for the tuning purpose is undesirable, since different conformations may manifest different key interactions and bind different ligands. Combining geometry with sequence similarity could enhance the accuracy of family classification, however, it may fragment the family space to a degree that will limit the ability to tune reliable scoring schemes for the families. With the continuous growth of the PDB, this option may be re-evaluated in the future, but at the present, the geometric classification is viewed as a reasonable option.

In the tuning process, data pertains to receptors of the same family will be used to customize a single scoring scheme. In addition, a geometric blueprint will be generated for the family. In subsequent docking runs receptors will be compared to available blueprints, and will be matched to the closest family if the similarity meets the 5 residues, 3 Å minimal criteria mentioned above. If matched to a pre-tuned family, the docking process will use the tuned scoring by default, otherwise a global scoring scheme will be used. The tuning techniques are outlined below.

Rank tuning

It is commonly observed that docking programs are capable of producing binding modes that are in good agreement with crystallographic structures, but the scoring functions fail to rank the most accurate modes at the top of the predicted conformations. Rank tuning attempts to achieve improved ranking for binding modes with higher experimental integrity and to improve score-RMSD correlation in general. In this tuning mode, a set of crystallographic complex structures is supplied by the user, and those complexes are clustered into geometric families. The following tuning steps are carried out separately for each family. For each complex eHiTS generates hundreds of binding modes. Starting from a random set of ω_i's the score weight set is optimized using a Powell algorithm [35] with a goal function that includes the top rank pose's RMSD, the rank of the closest (lowest RMSD) pose, and the energy difference between the top-rank and closest poses. During each iteration of the Powell algorithm, the energy difference is computed using the eHiTS score with the weight set generated in the previous iteration. The weights optimization is repeated with multiple initial random sets, and finally the best performing set is selected.

It should be emphasized that rank tuning does not affect the generated poses, but rather affects their rank ordering. If one were to inspect all the generated conformations before and after the tuning, one would find the same predictions. Consequently, the closest pose is conserved in this process.

Enrichment tuning

Enrichment tuning is a type of score customization that is designed for virtual screening scenarios. As opposed to rank tuning, in this case the scoring weight set is not affected by the process, which allows applying the two tuning types together. Enrichment tuning acknowledges the good performance of ligand-based methods in screening, and takes advantage of it by convoluting the docking score with a ligand similarity factor. To perform enrichment

tuning the user supplies ligand–protein complexes, and optionally additional known active molecules. By default, both rank tuning and enrichment tuning will be carried out simultaneously. Conceptually, however, in a pure enrichment tuning, the receptors are used only to generate a geometric family blueprint, and therefore even a single complex is sufficient for this purpose, provided that additional active molecules are supplied.

The active compounds and the complexed ligands are used to train a LASSO filter. LASSO is a ligand-based virtual HTS application that is based on eHiTS' ISP concept. The LASSO QSAR descriptors are feature vectors that represent the number of each ISP type present on the surface of the ligand. The feature vector contains information about the number of ISPs of each of the 23 types on the surface of the ligand, but does not include any information about the location of these points, nor any other structural information about the ligand. Since the placement of ISPs on the surface of the ligand is determined based on the hybridization and local connectivity of each atom, the count of each ISP type is invariant under conformational change, and hence the feature vector is conformation independent. Although simplistic, this descriptor has been shown to be an effective basis for screening, and has demonstrated scaffold hopping capabilities [36].

A LASSO filter is a trained neural network file that determines ligand similarity. The filter training process takes as input structures of known actives and does not require affinity information. The algorithm selects a set of likely inactive molecules from an internal library of compounds, and uses it as decoys in the neural network training. The training process identifies common motifs among the feature vectors of the active molecules in the training set that distinguish them from the feature vectors of the decoy molecules. The similarity metric for screened molecules is calculated by comparing their QSAR descriptors with the LASSO filter. Ligand similarity is measured in LASSO on a scale between 0 and 1, where 0 indicates significant dissimilarity, and 1 indicates high similarity, although not necessarily identity.

When LASSO is integrated into eHiTS docking, in a virtual screening run in which the tuned scoring is invoked, each of the ligands will be docked as usual to the target receptor, and in addition, a ligand similarity factor will be computed. For each screened compound, the docking score and the similarity factor are multiplied and the resulting score is used to rank-order the library. This convolution of docking and ligand-based predictions is synergistic, as will be shown in the results section.

By default, whenever the supplied complexes are clustered as one family, rank tuning is performed in addition to the enrichment tuning, and both a customized scoring weight set and a ligand similarity filter are generated. The ligands in the supplied complexes as well as additional actives that may be supplied constitute the set of known actives for the LASSO filter training process. In subsequent docking runs, a receptor that will be classified as a member of the tuned family will use both elements (the tuned weight-set and the LASSO filter) in the scoring of the docked poses.

It is important to note that enrichment tuning and rank tuning are completely orthogonal, and hence complementary. The conformation independence of the LASSO descriptors is a useful property when running a ligand-based screening that does not involve modeling bound conformations, since it removes possible bias stemming from predetermined 3D structures. It is also often the case that for a specific target, the number of known binding modes falls short of the number of known actives, and cannot cover the full spectrum of known active scaffolds, thus limiting the applicability of 3D or 2D/3D approaches. In the context of tuning the conformation independence is beneficial since it does not affect nor is it affected by the rank-ordering of poses of a specific ligand, allowing the scoring to achieve improved performance both in terms of enrichment and in terms of pose prediction. Convoluting the docking score and the ligand similarity factor provides the desired synergy between knowledge of ligand properties and of binding modes. As opposed to enrichment tuning, affinity tuning, the additional tuning approach discussed below, is not orthogonal to rank-tuning and has to be carried out and used separately.

Affinity tuning

Binding affinity correlation with docking scores can be improved in a similar manner to the way score-RMSD correlation is handled in rank tuning. Experimental values of binding affinity in units of Ki are associated in the input with receptor-ligand pairs and replace the role of RMSD values in the tuning process. The affinity tuning process begins by docking the ligands into their respective receptors. The generated poses are scored using the standard eHiTS score, using rank-tuned weight sets when applicable, i.e. when a receptor is matched to a pretuned family. For each ligand-receptor complex, n, different poses, j, are generated, and the score of each of them is composed of the scoring terms, i. Following Eq. 1:

$$E_j^{(n)} = \sum_i \omega_i^{(n)} e_{ij}^{(n)} \tag{2}$$

For each complex a single value for each of the eHiTS score terms is computed using all the generated poses for the ligand:

$$\in_i^{(n)} = f\left(\omega_i^{(n)} e_{ij}^{(n)}\right) \tag{3}$$

The function f is calculated as follows. A histogram of $\omega_i^{(n)} e_{ij}^{(n)}$ values over the different poses, j, is calculated, and the highest probability value, $\omega_i^{(n)} \tilde{e}_i^{(n)}$, is found. The score term is then:

$$\in_i^{(n)} = 0.3 \times \omega_i^{(n)} \tilde{e}_i^{(n)} + 0.7 \times \omega_i^{(n)} e_{iTop-Rank}^{(n)} \quad (4)$$

The affinity tuning process is aiming to find a new set of weights, Ω_i, such that an optimal correlation will be achieved between the Ki values for the complexes and their energies:

$$E^{(n)} = \sum_i \Omega_i \in_i^{(n)} \quad (5)$$

The tuned weight set is not family-specific. When the affinity score is invoked in a docking run, poses are first generated and ranked using the regular eHiTS score. An affinity score, $E^{(n)}$ is then calculated and assigned to the top rank pose. The affinity scores of the other poses are scaled relative to the top-rank affinity score according to the scaling of the docking scores.

Pre-tuned weight-sets

As part of the development of the eHiTS score, we generate pre-tuned weight sets for the entire array of PDB complexes that were used in statistics collection. The pre-tuning includes rank tuning and enrichment tuning based on the PDB ligands. In addition to the family-based sets, a global weight set was developed to handle receptors that are not classified into any of the families. The global set is tuned using a diverse set of close to 100 complexes. The family-based tuning process is summarized schematically in Fig. 1. Unfortunately, datasets for binding affinity tuning and benchmarking are scarce, and there is not enough assembled and curated information to carry out family-based affinity tuning. Consequently, a global binding affinity weight set was tuned, and an affinity value, also termed eHiTS-Energy, is computed during docking in addition to the eHiTS score.

The commercial version of eHiTS, therefore, already includes the tuning elements mentioned above. The tuned scoring is used by default, but can be overridden by the user, and a separate tuning utility offers the users the capability to apply the three tuning approaches on their own data. In the following section we demonstrate the effect of the methods and discuss their benefits and limitations.

Methods

The effect of the individual tuning methods, and the combination of rank- and enrichment-tuning were evaluated as described below. The presented work was intentionally carried out using only options that are included in the commercial distributions of eHiTS and its tuning utility, and that are readily accessible to the common user.

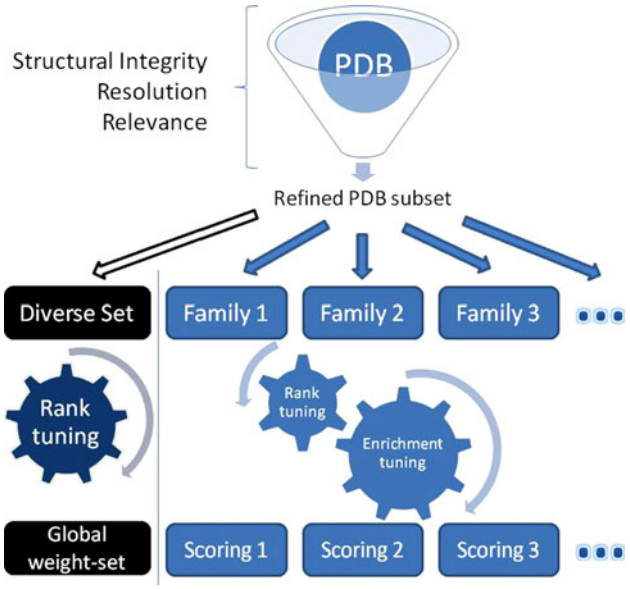

Fig. 1 Pre-tuning of the eHiTS scoring function. Rank- and enrichment-tuning is carried out for protein families. A default global set is rank-tuned with a diverse set of complexes

Validation runs

For the purpose of evaluating rank tuning we carried out cognate docking runs. Although the relevance of self-docking experiments to real usage of docking tools is limited [37], it provides a straightforward framework to evaluate the effect of tuning, avoiding the uncertainty associated with preparing and analyzing a cross-docking test. Two sets were developed for the purpose of evaluating rank tuning. The sets correspond to two targets of pharmaceutical significance that offer a large number of solved complexes in the PDB: β-secretase (BACE1), and cyclin dependent kinase 2 (CDK2). The datasets include crystal structures of complexes with resolution of 2.5 Å or higher, and where the ligand has a molecular weight below 800. There were 73 and 142 structures in the BACE1 and CDK2 sets, respectively. Each complex was split manually to receptor and ligand, and ions, ligands and cofactors were excised from the protein. The dataset preparation did not include assignment of charges or protonation states. When charges are not supplied as input, eHiTS uses its internal charges knowledgebase which has been developed using QM methods (RESP calculations at the B3LYP 6-31G* level of theory) on a large training set covering a large range of functional groups typical to biological complexes. Protonation states were handled using eHiTS' on-the-fly

mechanism, which determines the most favorable state for each of the generated binding modes.

eHiTS offers six different levels of accuracy that differ with respect to the number of poses carried over from the initial stage of the exhaustive rigid fragment docking to higher levels of refinement. The accuracy levels allow the user to choose the preferred balance between precision and computational cost. The tuning process is done at eHiTS' default accuracy level, 3, and the docking results reported hereafter in the context of validation were done at accuracy 6, i.e. the highest.

Virtual screening

The DUD set [38] is a valuable benchmark for virtual screening applications. It covers a diverse set of 40 biological targets, for each of which a set of known active molecules is provided, as well as a set of decoys, assumed to be inactive and yet structurally similar to the actives. Receptors, actives and decoy compounds were all used as provided in the downloadable set with no preparation steps. In the case of NA, active ligands were recovered from the PDB, and a set of 28 molecules has been found with no overlap with the DUD NA actives. Screening runs were carried out at accuracy 3.

Binding affinity

Assessments of binding affinities were done using the refined set of the PDBBind benchmark [39, 40]. The PDBBind database offers a wide selection of protein–ligand complexes from the PDB that have reliable measured values of the binding affinity, as Kd, Ki and IC50. The refined set consists of complexes that were solved at a resolution higher than 2.5 Å, pertain to pharmaceutically relevant small molecules bound to binding pockets composed of naturally occurring amino acids, and for which Ki and Kd values were published. The PDBBind database and the refined set are updated annually to reflect new published data as well as enhancements in the refining protocols of the dataset. The PDBBind data is provided as presplit receptor and ligand files. Assessment of binding affinity was done by carrying out docking runs at the default accuracy. eHiTS-Energy values, rather than eHiTS-score values were used for the analysis.

All docking runs were performed with eHiTS 2009.1. The results reported below strictly refer to test sets that are independent of the tuning sets.

Results and discussion

Rank tuning

BACE1

As a pivotal protein in the amyloid-β pathway, BACE1 has become a main target for drug discovery related to Alzheimer's disease [41]. BACE1's large binding pocket poses a challenge for the design of potent inhibitors that can cross the blood–brain barrier. From a docking perspective, docking relatively small molecules into a large binding site is a considerable challenge for the scoring function, as the conformations space of the ligand is vast, and is likely to include many regions of little integrity. Score tuning, in such circumstances, can potentially promote certain type of interactions and energy terms that will advance the more accurate poses in ranking. We randomly selected 29 complexes to be used in tuning for this target, and the remaining 44 complexes from the BACE1 set were used as a test set. The results for the test set are shown in Table 1.

From the column associated with the full test set it is evident that the tuned weight set offers a significant improvement over the global weight set in terms of top rank pose RMSD and a slight improvement in terms of docking success rate, defined as percentage of top-rank poses under 2.0 Å RMSD. The average closest pose is shown as a reference. As mentioned above, the closest conformation marks the theoretical limit of the rank tuning method. While the tuned results remain far from that limit, overall the tuned score is successful in promoting the better poses, and a 26% improvement is observed for the top-rank RMSD.

The distribution of RMSD values of the top rank poses, shown in Fig. 2, demonstrates that the effect of tuning is to shift the entire distribution to the left, and to reduce the

Table 1 Rank tuning results for BACE1 test-set

	Entire BACE1 test-set	2QK5 family	1TQF family
Top-rank pose—global weight set	2.06 (80%)	1.16 (100%)	2.43 (71%)
Top rank pose—tuned weight set	1.52 (86%)	1.11 (100%)	1.69 (81%)
Closest pose	0.78	0.69	0.81

Listed are: average top-rank pose RMSD, and docking success rate in parentheses

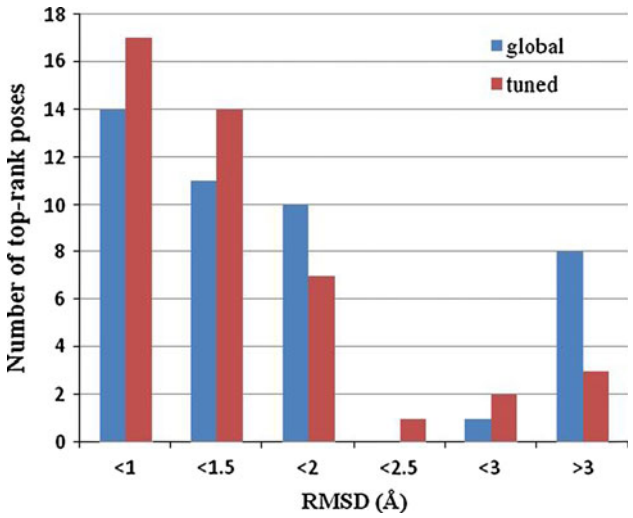

Fig. 2 RMSD distribution of the top-rank poses of the 44 BACE1 complexes

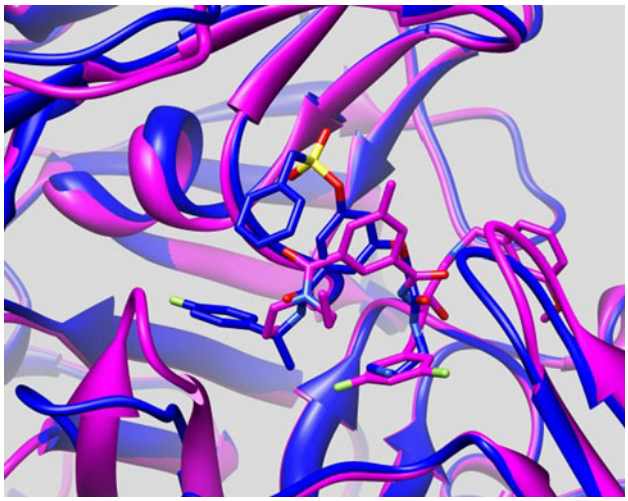

Fig. 3 PDB complexes 2QK5 (*magenta*) and 1TQF (*blue*) superimposed. Main differences observed in loop regions in the binding pocket are typical to the two geometric families classified by eHiTS

number of high RMSD outliers. The median for the top rank pose drops from 1.4 Å for the global weight set to 1.08 for the tuned set. The number of poses with RMSD greater than 4 Å is 7 for the global set and 2 for the tuned set. From the 44 complexes, in 22 cases the tuned weight set promotes lower RMSD pose to the top rank, compared to 12 cases where higher RMSD poses are put at the top by the tuned set. In 10 cases the pose selected as top-rank is the same for the global and tuned weight-sets.

A closer inspection of the rank-tuning results of this BACE1 set reveals that eHiTS clustered the structures into two geometric families, denoted arbitrarily 2QK5 and 1TQF after the first PDB code classified in each family. The training sets included 10 and 19 complexes, and the test sets included 13 and 31 complexes for the 2QK5 and 1TQF families, respectively. Comparison of the structures of members of each family shows differences primarily in loop regions in the binding pocket, as shown in Fig. 3. Table 1 shows the effect of tuning on the individual families.

It is interesting to note the marked difference in performance between the families, which may be due to a more confined pocket in the case of the 2QK5 family. Clearly, very little is achieved through tuning when the global weight set is as well performing as it is for 2QK5. The main contribution to the tuning effect comes, therefore, from the 1TQF family which shows 30% improvement in top rank RMSD on average. For the 1TQF family, the median RMSD drops from 1.50 to 1.22 Å using the tuned set.

CDK2

The binding pocket geometry in the case of the CDK2 complexes is very well conserved, and eHiTS clusters the

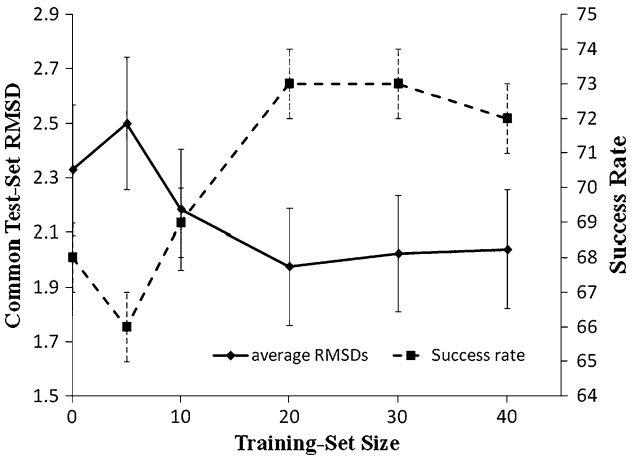

Fig. 4 Average top rank RMSD and docking success rate for the common test set achieved with tuning sets of various sizes for CDK2. Average values over 5 tuning iterations are shown for each tuning set

entire set into one geometric family. The large set is very useful for testing the effect of the size and composition of the training set on the quality of rank tuning. Five training sets of various sizes were selected at random such that each set was a subset of the bigger training sets. This was done in order to keep a large common test set, and to avoid skewing of the results due to variances in the test sets. For each training set, the tuning process was repeated five times in order to study the scattering of the results due to the stochastic nature of the tuning process. Figure 4 shows the average top-rank RMSD and docking success rate for the common test-set of 102 complexes using score weight sets obtained with tuning sets of increasing size. The training set of size zero represents eHiTS' global weight set. The error bars for the global set reflect the

standard error for the test set computed using bootstrapping. For the other data points, the error bars were computed as the standard error of the sample of five averages (from five tuning runs), where for each average the error was computed using bootstrapping.

Figure 4 indicates that a training set of size 5 is not sufficient for rank tuning, and it does not show, in most tuning repetitions, any improvement compared to the global weight set. The best convergence to improved results is seen with a set size of 20 complexes or more. Although the expected trend would be of improved results with increasing size of the training set, an increased variance of the results is observed for the set size of 40 complexes. This is due to the fact that the tuning utility has a limit on the number of regression iterations, which is insufficient to handle the large number of complexes. Overall, however, with only one exception, all tuning repetitions using training sets of 10 or more complexes improve the docking performance on the test set, with roughly 10% improvement in the top-rank RMSD, and 5% in the docking success rate.

Enrichment tuning

We evaluated the effect of enrichment tuning on eHiTS' screening performance using the DUD set. The protocol of Cross et al. [42] in terms of system setup and data analysis was followed to enable comparison with other docking results presented in that article. Ligands and decoys were rank-ordered based on the score using eHiTS' global weight set versus eHiTS' pre-tuned weight sets when available. In all docking jobs combined, roughly 200 compounds, i.e. less than 0.2% of the cases, failed to dock and they were appended at random order at the bottom of the respective ranked lists. For compounds with several tautomeric or protonation states, only the best scoring tautomer was included in the ranked list.

The pre-tuned weight sets combine rank-tuning with enrichment tuning, i.e. with ligand-based score scaling. Those weight-sets are trained with PDB structures and ligands only. Clearly there could be some overlap between the DUD actives and PDB ligands of the targets. To demonstrate the effect of tuning while guaranteeing the independence of the test set, we extracted NA ligands from 90 PDB complexes and filtered out from the library ligands that are part of the DUD set. Since, as shown previously, the NA complexes span many geometric families, we performed pure enrichment tuning, skipping rank tuning, with 28 PDB ligands that were left after filtering. The 28 PDB ligands were also used to train a separate LASSO filter. The ROC curves [43] for the NA screening are shown in Fig. 5. The Receiver Operating Characteristic (ROC) curves are particularly suitable for the evaluation of

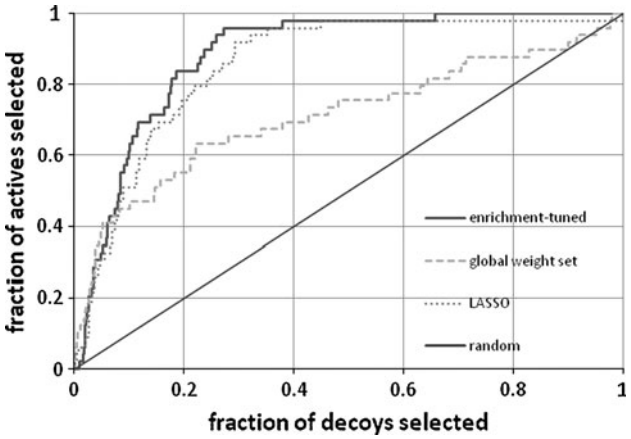

Fig. 5 ROC curves for the neuraminidase target in the DUD set screened using eHiTS with the enrichment-tuned scoring (*solid*) and the global score weight-set (*dashed*) and using the ligand based tool, LASSO (*dotted*)

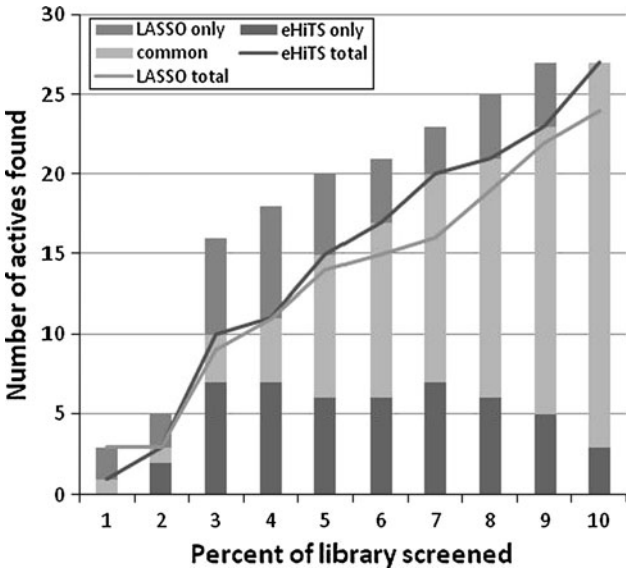

Fig. 6 The number of actives recovered by tuned eHiTS and LASSO and the level of overlap between them as a function of the percentage of DUD's NA library screened

screening performance since they allow for effective differentiation between two populations, such as active and inactive, and they do not suffer from dependence on the composition of the dataset in the same manner as enrichment curves are. In addition, the area under the curve (AUC) provides a normalized measure for the screening performance, ranging from 0 for inverted performance (all actives recovered at the bottom of the list), 0.5 for random selection of compounds, and 1 for perfect performance.

The LASSO and the global weight-set curves in Fig. 5 manifest the infamous superiority of ligand based methods over docking techniques for screening purposes. However,

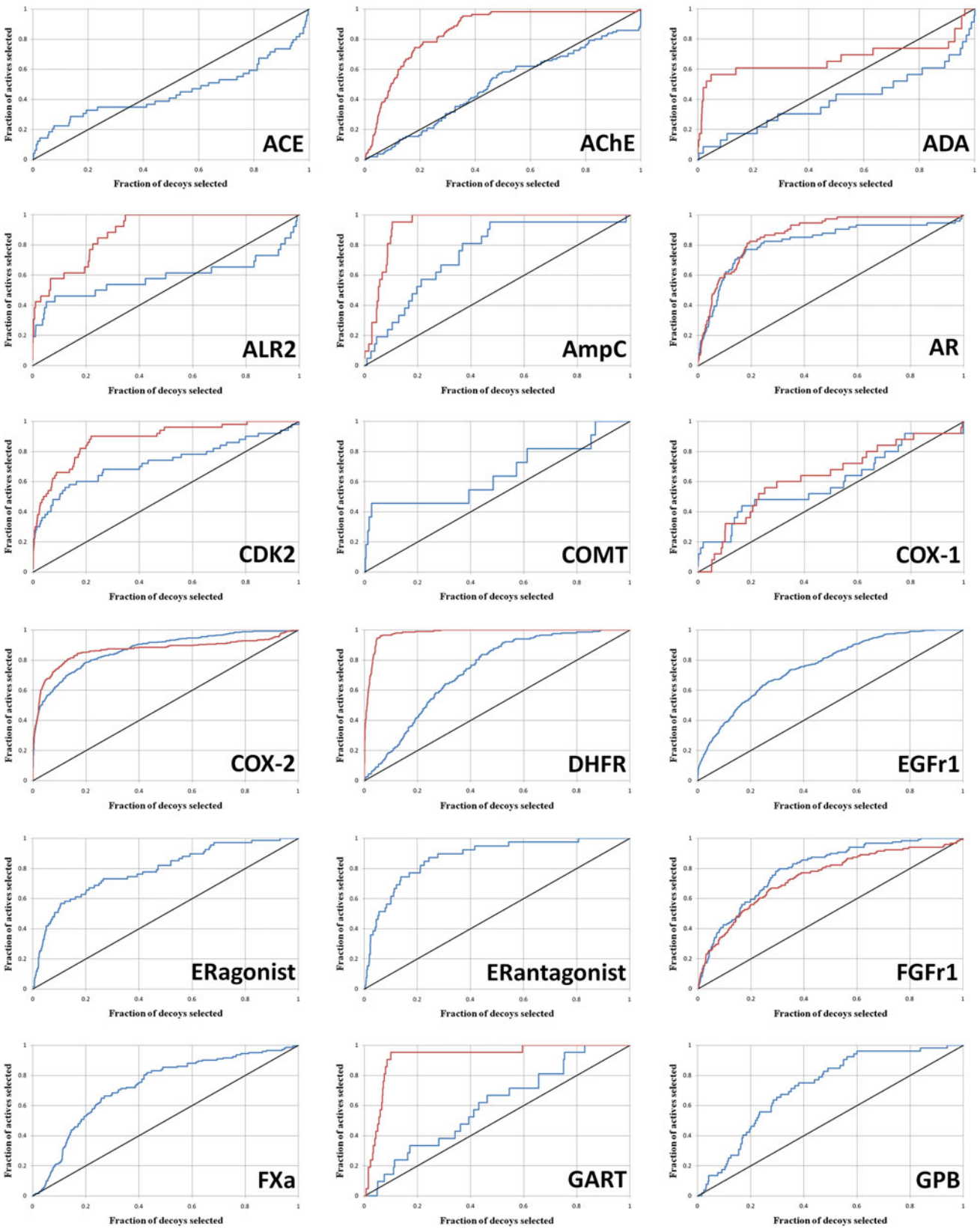

Fig. 7 ROC curves for the 40 targets in the DUD set. Results obtained with the global weight sets are shown in *blue*. For 25 cases that matched pre-tuned sets, the tuned ROC curve is shown in *red*. Random behavior is shown as a diagonal for reference

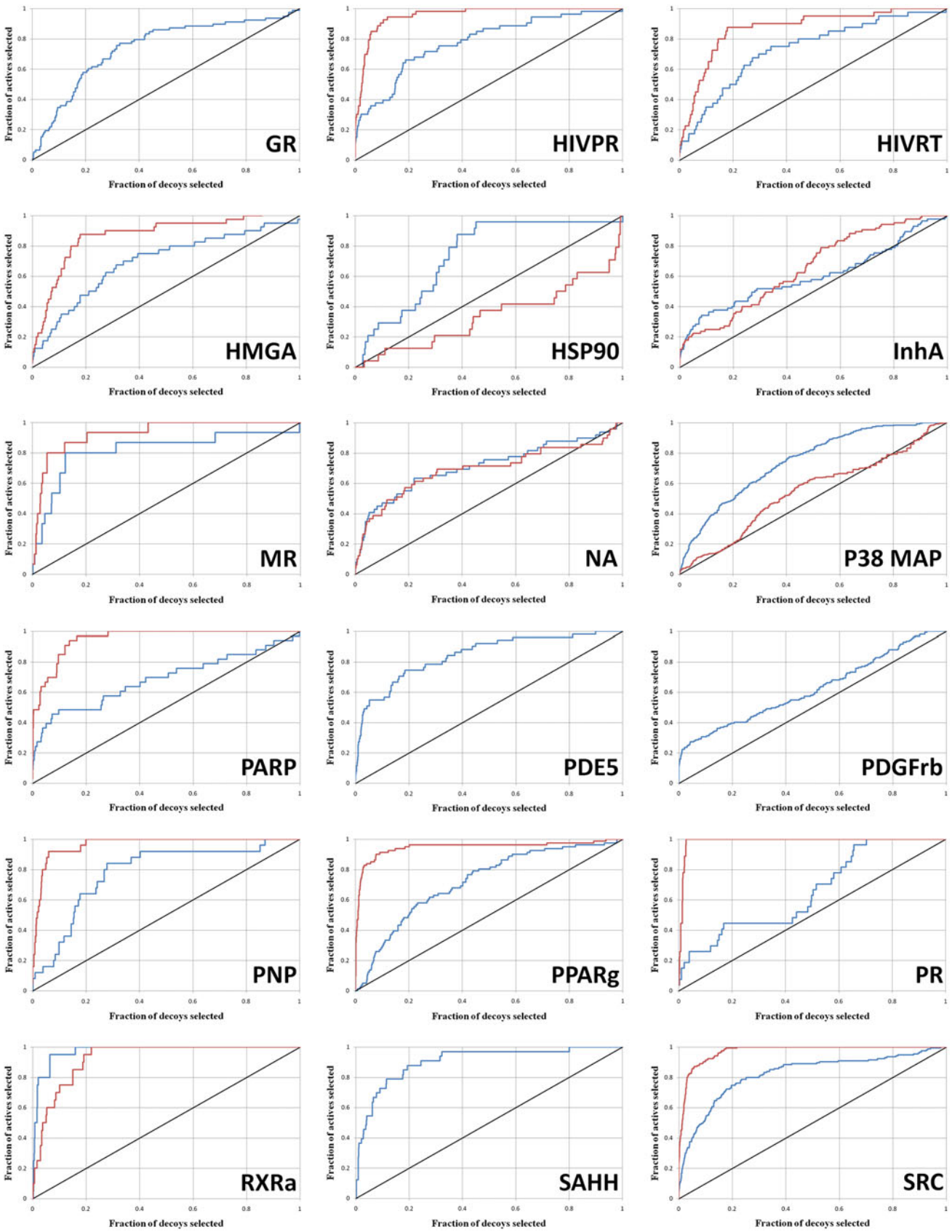

Fig. 7 continued

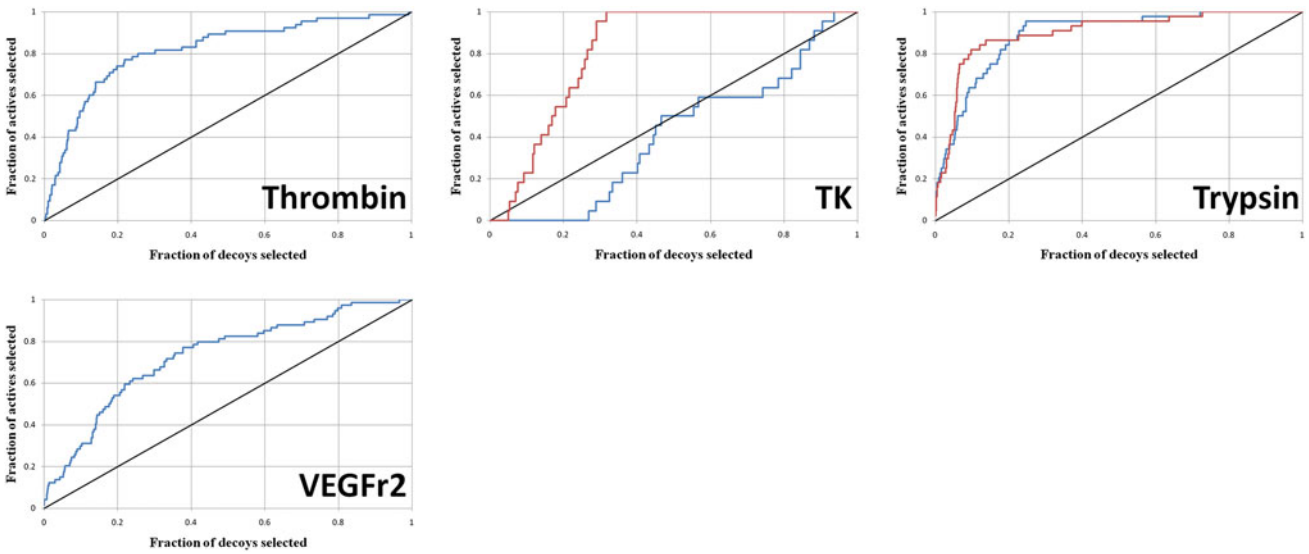

Fig. 7 continued

Table 2 Number of PDB codes used to tune the weight sets matched to 23 of the DUD targets, and the number of distinct DUD actives associated with the respective targets

Target	PDB codes used for tuning	DUD actives	Target	PDB codes used for tuning	DUD actives
AChE	24	105	HMGA	13	35
ADA	17	23	InhA	17	85
ALR2	74	26	MR	11	15
AmpC	41	21	PARP	5	33
AR	41	74	PNP	27	25
CDK2	105	50	PPAR	27	81
COX-1	9	25	PR	6	27
COX-2	9	348	RXR	7	20
DHFR	45	201	SRC	13	155
GART	13	21	TK	19	22
HIVPR	196	53	Trypsin	110	44
HIVRT	44	40			

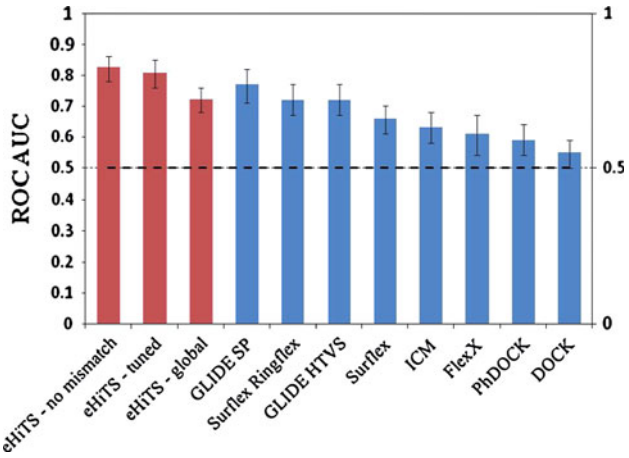

Fig. 8 Mean ROC-AUC values. *Error bars* indicate 95% confidence interval. Values for programs other than eHiTS were reproduced from Cross et al

the enrichment-tuned docking screening, i.e. the convolution of docking and ligand similarity, is comparable to the LASSO curve, demonstrating the synergy between the methods. The ROC-AUC [44] associated with the global scoring, the LASSO screening, and the tuned-eHiTS is 0.71, 0.86 and 0.89, respectively.

A relatively small gain in this particular case in enrichment using docking compared to ligand based screening is achieved at a considerable computational cost. The cost-effectiveness of docking has to be estimated by finding out whether it is able to identify different actives than the ligand-based method alone does. Figure 6 shows the level of overlap between the sets of compounds recovered from the total 49 DUD actives using eHiTS and LASSO as a function of the percentage of library screened. It is interesting to note that for the lower percentages of screened compounds, the overlap is quite limited, and the two methods expose, for the most part, different actives. Except at 1% of the library, docking outperforms ligand-based screening, and offers additional actives that would not be selected by LASSO. In drug discovery, where every hit may have a profound impact on the discovery effort, the performance shown in Fig. 6 justifies the utilization of the more expensive docking method, and suggests that a combination of the methods may have a bigger effect still.

Table 3 ROC-AUC and ROC enrichment values for the entire DUD set and biological families within the set

	ROC-AUC		ROC_{En} (0.5%)		ROC_{En} (1%)		ROC_{En} (2%)		ROC_{En} (5%)		ROC_{En} (10%)	
	Glob	Tuned	Glob	Tuned	Glob	Tuned	Glob	Tuned	Glob	Tuned	Glob	Tuned
Entire DUD (40)	0.72	0.83	16.5	29.3	12.8	19.8	8.9	14.3	5.6	8.9	3.9	5.8
NHRs (8)	0.80	0.88	12.8	23.8	15.4	20.7	11.1	17.9	7.4	11.1	5.1	6.9
Kinases (9)	0.71	0.79	15.4	21.6	11.1	14.4	7.3	9.8	4.8	6.1	3.4	4.3
Serine proteases (3)	0.81	0.81	11.4	12.9	9.5	8.8	6.9	6.6	5.0	5.5	4.6	5.2
Metallo enzymes (4)	0.59	0.75	19.9	24.3	14.0	17.3	11.4	16.4	5.9	8.3	3.4	4.5
Folate enzymes (2)	0.67	0.95	1.5	39.0	1.3	27.0	1.1	21.2	1.4	13.7	1.7	9.6

For the purpose of comparing the eHiTS performance to that of other docking applications, as given in Cross et al., the ROC curves for the full DUD set are shown in Fig. 7. 25 out of the 40 targets in the set were matched to one of the pre-tuned weight sets available in the commercial version of eHiTS. For those targets, both the curve associated with the global weight-set and the curve associated with the enrichment-tuned scoring are shown. For the other 15 cases, where the targets were not matched to any of eHiTS' pre-tuned geometric families, only the ROC curves obtained with the global weight set are shown. Two cases out of the 25 were mismatched with certainty: P38 MAP was matched to a myoglobin family, and HSP90 was matched to a family that included NADH oxidase, thioredoxin reductase and other proteins. In those cases the global weight set is expectedly performing better in screening, and the mismatch can be identified prospectively.

Several of the cases plotted in Fig. 7, among them AChE, AmpC, DHFR, GART, HIVPR, PR, PPARg and PNP, show a significant improvement in the enrichment performance upon using the tuned set. Others, like MR and HIVRT, demonstrate a more moderate improvement, while a few cases show a negative effect. To explain the differences in the effects, we recall that since the family clustering is based on geometric considerations, it is inevitable that on occasion receptors from different biological families will be clustered together. This is illustrated in the cases of the COX-1 and COX-2 targets which were matched to the same family. The scoring of this family was tuned using four complexes of each of the COX receptors as well as prostaglandin H2 synthase. The marginal improvement in the ROC curves using the tuned scoring in these two cases may be a result of this "contamination" of the tuning set, a contamination which clearly can be avoided when a user makes conscious decision with regard to the composition of the tuning set.

Table 2 lists the number of distinct DUD actives and the number of PDB codes in the eHiTS family associated with each of the 23 targets that were matched to relevant families, i.e. excluding P38 MAP and HSP90. In most cases, the DUD actives outnumber the PDB complexes. The PDB codes often have a significant redundancy in terms of the ligands, i.e. the number of PDB codes is merely a higher bound for the number of ligands that are used for training. In the MR case, for example, the 11 codes offer only 5 distinct ligands. This redundancy, along with possible training set contamination by PDB codes from other families, and along with the limited overlap between the DUD actives and the PDB ligands, suggests that while the test set is not independent, the results are indicative of the capabilities of the method.

The DUD screening results are summarized in Fig. 8 as average ROC-AUC. Results referring to other docking programs are presented by courtesy of Cross et al. [42]. The error bars represent the 95% confidence interval computed using bootstrapping. Screening with eHiTS' global weight set yields results that are on a par with GLIDE-HTVS and Surflex Ringflex. The other two eHiTS columns represent the out-of-the-box performance on all 40 targets using the family-tuned sets when applicable, and the global set otherwise. The no-mismatch value was computed with the same data as for the tuned case, except that the global weight-set ROC-AUC values were used for the incorrectly matched HSP90 and P38 MAP instead of the tuned ones. The AUC increased from 0.72 for the global set to 0.83 for the combination of 17 global and 23 tuned sets. The considerable increase is of statistical significance.

For most targets the ROC curves are appreciably better than random behavior, and demonstrate an improved performance for the tuned scoring. Table 3 illustrates that the enhanced performance is manifested not only in the overall ROC curve, but also in early retrieval of actives. The table lists the ROC enrichment values [45] for the DUD set with a break-up of the data to the main target types represented in the DUD set, and shows, within the confines of the limited statistical power, that the general trends are reflected in most target types individually. Blowouts of the ROC curves reflecting early enrichments, and ROC enrichment values for DUD target are provided in the supporting materials, online resource 1.

Good and Oprea [46] curated the actives from the DUD set and eliminated compounds that do not meet basic

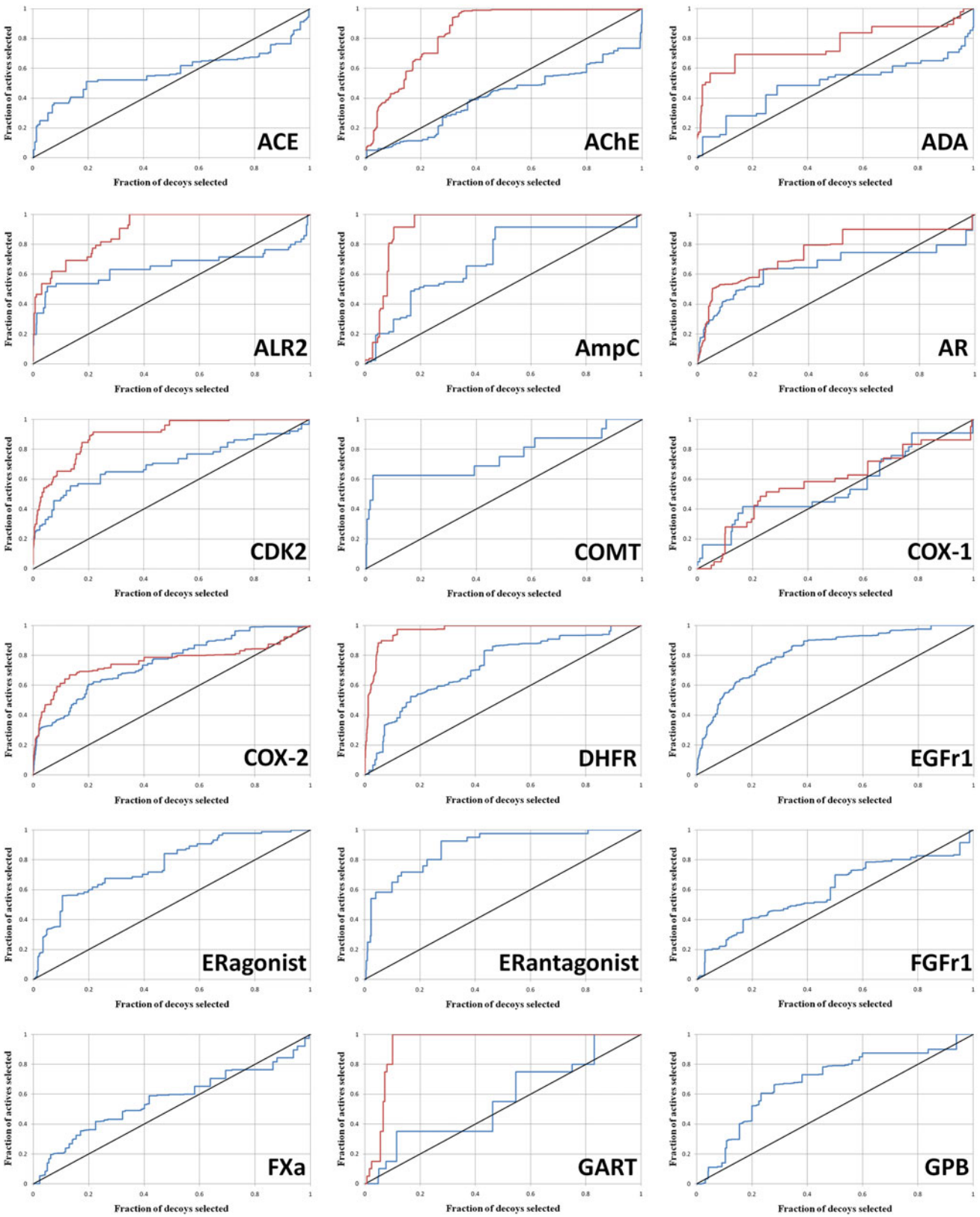

Fig. 9 ROC curves for the 40 targets in the DUD set accounting for scaffold redundancy. Results obtained with the global weight sets are shown in *blue*, results with the pre-tuned weight sets are in *red*, when available

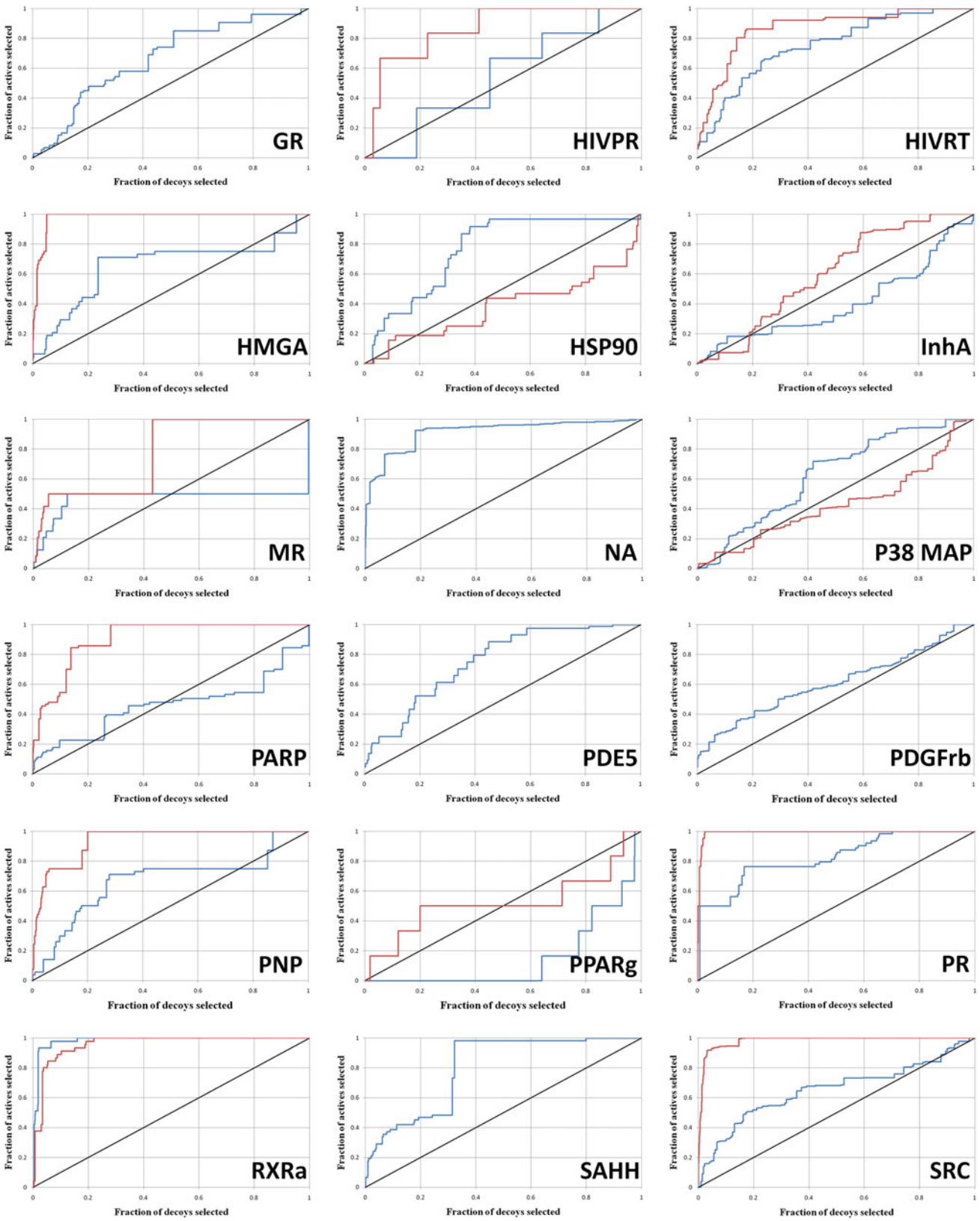

Fig. 9 continued

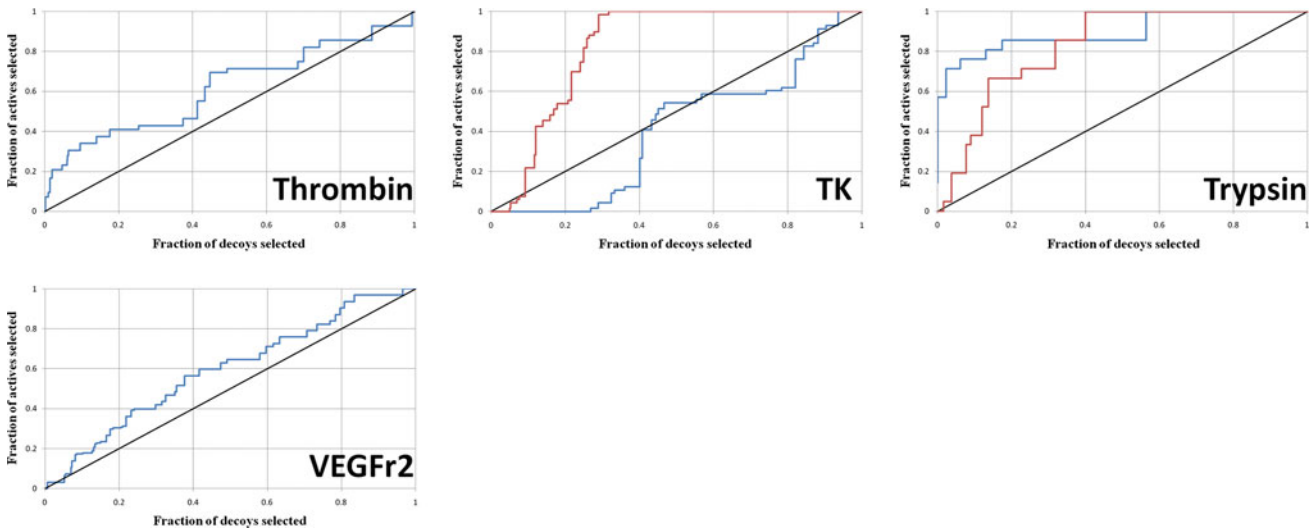

Fig. 9 continued

criteria of lead-likeness in terms of molecular weight and logP. They further clustered the remaining actives by their chemotype, or scaffold similarity. Their work, however, did not address the decoys set, but see Wallach and Lilien's for discussion about its composition [47]. Scaffold redundancy is particularly problematic when ligand-based methods are considered. Since the enrichment-tuning includes a ligand-based element, accounting for active similarity is crucial for estimating the effect of the method. Figure 9 repeats the ROC curves of Fig. 7 but considering the Good and Oprea clusters. The curves were computed by assigning to each active a weight equal to the inverse of the number of actives in its corresponding cluster. Thus, for each target, the weights of all the actives add up to the number of clusters found for this target.

The ROC curves for the DUD clusters are generally lacking compared to those in Fig. 7, suggesting that the docking and scoring may be more successful in retrieving specific chemotypes. However, use of the tuned scoring shows improved enrichment over the global set in the majority of cases. The exceptions are again the mismatched HSP90 and P38 MAP.

The average ROC-AUC values for the eHiTS screening accounting for the chemotype clusters are shown in Fig. 10. The average AUC increases from 0.65 using the global set to 0.77–0.78 using the tuned scoring for the entire set. Comparing only those targets for which tuned scoring exists, the average AUC improves from 0.61 to 0.84.

Early enrichment values for the DUD set, considering scaffold redundancy demonstrate the superiority of the tuned scoring, and can be found in the supporting materials.

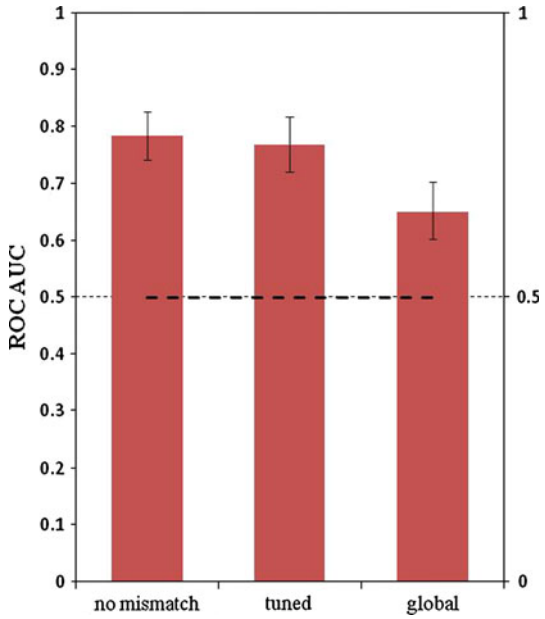

Fig. 10 Mean ROC-AUC values for the DUD clusters. *Error bars* indicate 95% confidence interval

Affinity tuning

In the case of binding affinity, the tuning process does not benefit from abundance of experimental data like the cases of rank- and enrichment-tuning. Until recently the PDBBind had been virtually the sole attempt to collect and digest experimental values from the literature and associate them with structural information into a coherent benchmark. A more recent effort by Heather Carlson and coworkers is aiming to develop a similar knowledge-base as a resource for comparison and development of docking methods [48].

Table 4 The correlation between eHiTS' binding score and experimental log(Kd) for the PDBBind

Scored structures	R_P	SD	ME
All docked ligands	0.74	1.18	0.93
Docked ligands w RMSD <2 Å	0.76	1.12	0.88
All crystallographic poses	0.59	1.36	1.08

Pearson correlation coefficient, standard deviation and mean error are shown for different data

Ideally, the customization of the scoring function for binding affinity estimation should be done in a similar manner to the rank-tuning process, i.e. on a family-based manner. However, for the vast majority of the geometric families clustered by eHiTS there are only few or no affinity values. Consequently, a single, universal, score weight-set has been developed for binding affinity estimation. The special purpose scoring scheme can be used, for example, to rank-order a set of assumed actives and prioritize them for lab experiments.

The linear regression results for the correlation between eHiTS binding score and the pKd values of the PDBBind complexes are given in Table 4. The data shows a moderate correlation between the eHiTS score and the experimental binding affinities for the 1,370 complexes from the refined set of the PDBBind that docked successfully, and for a subset of complexes that their top-rank docking pose was under 2 Å. In addition, the correlation is shown for the crystallographic poses with no docking or optimization. The correlation obtained with the X-ray modes is limited compared to the one achieved with the docked poses. This is expected since the tuning of the binding energy weight set is done on conformations that are on the eHiTS potential surface. For the docked ligands the correlation with experimental affinities is sound, and the eHiTS results in general are in good standing compared to those achieved with other tools [2, 49].

Conclusions

Knowledge-based approaches have become a mainstream in the molecular docking paradigm, and they benefit from an ever expanding base of experimental data and from continuous progress in computational methods. The concept of score tuning and customization is a natural extension of the fundamental knowledge-oriented methodology. The main impediment to this approach is the risk of over training, and therefore developers as well as users should be conscious to the statistical limitations of the techniques.

We showed that the three tuning methods are very effective in addressing some of the fundamental shortcomings of scoring in the context of molecular docking. The rank-, enrichment- and affinity-tuning approaches presented in this study are general, and can be implemented in other docking programs and scoring functions. However, they are particularly easy to implement in eHiTS given the structure of its native scoring function (see the work of Kinnings et al. [30] where the eHiTS score was tuned using machine learning techniques for specific targets). eHiTS includes pre-tuned scoring schemes for several hundred geometric families clustered from the PDB, as well as a global, universal, scoring scheme that is used for receptors that do not match any of the pre-tuned families. During docking runs, eHiTS selects the appropriate customized scoring by comparing the binding pocket geometries of the target receptors to the characteristic distance matrices of the tuned geometric families. A supplementary utility allows the users to perform the tuning procedures on their own data. In addition, a universal customized scoring scheme is available for binding affinity estimation.

The rank tuning process has been designed as a final stage refinement in the sense that it does not alter the generated poses, but rather only changes their rank-ordering. It is possible to perform deeper tuning that will generate binding modes which are closer to the crystallographic conformations, and a procedure of this sort had been used in previous versions of eHiTS in the past. Conceptually, this process identified binding patterns and indirectly imposed automatic constraints. We have found, however, that deep tuning is much more costly, and requires more experimental data to achieve the desired stability, and therefore it has not been included in eHiTS 2009 version. Work is ongoing to advance this concept, as well as to improve the convergence of the current stochastic rank-tuning given the considerable variance in results across different iterations, and different training sets.

The convolution of the docking score with a ligand similarity measure is manifestly a powerful means to improve enrichment in screening scenarios beyond the levels of each approach separately. We have not experimented with the relative weights given to each factor, but clearly this may offer an avenue for further improvement. We have shown that a mismatch of the family (e.g. HSP90), or "contamination" of the training set with actives of a different biological family (e.g. COX-1) may have a devastating effect on the results of the screening. In the case of the eHiTS pre-tuned sets, one can verify that a set is correctly matched by inspecting the PDB codes that were used to generate it. Furthermore, the occasional contamination is strictly a result of the automation of the clustering process, and is altogether avoided when a user consciously chooses the complexes and actives to be used in tuning.

Binding affinity tuning is currently limited in scope given the shortage of benchmarks in the area. We hope that

current efforts in the community, as well as our own data collection will enable us to expand this capability in the future.

Acknowledgments The authors thank Bashir Sadjad for his diligent coding during the work on this project. We also thank Dan Harris for his application development of a previous version of the eHiTS tuning utility and Tony Cook for reviewing an earlier version of this manuscript. We acknowledge Jason Cross and coauthors for permission to reproduce data from their paper.

References

1. Englebienne P, Moitessier N (2009) Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? J Chem Inf Model 49:1568–1580
2. Wang R, Lu Y, Fang X, Wang S (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800. J Chem Inf Comput Sci 44:2114–2125
3. 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 47:1504–1519
4. 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, Peishof CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
5. Stahl M, Rarey M (2001) Detailed analysis of scoring functions for virtual screening. J Med Chem 44:1035–1042
6. Schulz-Gasch T, Stahl M (2003) Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 9:47–57
7. Kontoyianni M, McClellan LM, Sokol GS (2004) Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 47:558–565
8. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL III (2004) Assessing scoring functions for protein–ligand interactions. J Med Chem 47:3032–3047
9. Plewczynski D, Łaźniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32:742–755
10. Oda A, Tsuchida K, Takakura T, Yamaotsu N, Hirono S (2006) Comparison of consensus scoring strategies for evaluating computational models of protein–ligand complexes. J Chem Inf Model 46:380–391
11. Cornell WD (2006) Recent evaluations of high throughput docking methods for pharmaceutical lead finding—consensus and caveats. (ed) David C. Spellmeyer. Annu Rep Comput Chem 2:297–323
12. Jain AN (2006) Scoring functions for protein-ligand docking. J Comput Aided Mol Des 7:407–420
13. Muegge I, Martin YC (1999) A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 42:791–804
14. Gohlke H, Hendlich M, Klebe G (1999) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295:337–356
15. Tøndel K, Anderssen E, Drabløs F (2006) Protein Alpha Shape (PAS) Dock: a new gaussian-based score function suitable for docking in homology modelled protein structures. J Comput Aided Mol Des 20:131–144
16. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11:425–445
17. Wang R, Lai L, Wang S (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des 16:11–26
18. 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 47:1750–1759
19. Rarey M, Kramer B, Langauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
20. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW (2000) Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 14:731–751
21. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M (2005) LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model 23:395–407
22. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
23. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202
24. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shaw DE, Shelley M, Perry JK, 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 47:1739–1749
25. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comp Chem 13:505–524
26. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2007) eHiTS: a new fast, exhaustive flexible ligand docking system. J Mol Graph Model 26:198–212
27. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2006) eHiTS: an innovative approach to the docking and scoring function problems. Curr Protein Pept Sci 7:421–435
28. Fradera X, Knegtel RMA, Mestres J (2000) Similarity-driven flexible ligand docking. Proteins 40:623–636
29. Mooij WTM, Verdonk ML (2005) General and targetes statistical potentials for protein-ligand interactions. Proteins 61:272–287
30. Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, Bourne PE (2011) A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 51:408–419
31. Amini A, Shrimpton PJ, Muggleton SH, Sternberg MJE (2007) A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming. Proteins 69:823–831
32. Vriend G (1996) WHAT_CHECK. [Online] Radboud University, Nijmegen Medical Centre. http://swift.cmbi.ru.nl/gv/whatcheck/. Accessed 27 April 2011
33. Kleywegt GJ, Harris MR, Zou J, Taylor TC, Wählby A, Jones AT (2004) The Uppsala electron-density server. Acta Cryst D 60:2240–2249
34. Moscona A (2005) Neuraminidase inhibitors for influenza. N Engl J Med 353:1363–1373
35. Powell MJD (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput J 7:155–162

36. Reid D, Sadjad BS, Zsoldos Z, Simon A (2008) LASSO—ligand activity by surface similarity order: a new tool for ligand based virtual screening. J Comput Aided Mol Des 22:479–487
37. Sutherland JJ (2007) Lessons in molecular recognition 2: assessing and improving cross-docking accuracy. J Chem Inf Model 47:2293–2302
38. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
39. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. J Med Chem 47:2977–2980
40. Wang R, Fang X, Lu Y, Yang CY, Wang S (2005) The PDBbind database: methodologies and updates. J Med Chem 48:4111–4119
41. Cole SL, Vassar R (2007) The Alzheimer's disease β-secretase enzyme, BACE1. Mol Neurodegener 2:22–46
42. Cross JB, Thompson DC, Rai BK, Baber JC, Yi Fan K, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49:1455–1474
43. 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 48:2534–2547
44. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
45. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:239–255
46. Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22:169–178
47. Wallach I, Lilien R (2011) Virtual decoy sets for molecular docking benchmarks. J Chem Inf Model 51:196–202
48. Carlson HA, Dunbar JB Jr, Gestwicki JE, Stuckey J, Showalter HD, Wang S (2009) CSAR—Community Structure-Activity Resource. [Online] University of Michigan. http://www.csardock.org/. Accessed 27 April 2011
49. Raub S, Steffen A, Kamper A, Marian CM (2008) AIScore—chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. J Chem Inf Model 48:1492–1510