J Comput Aided Mol Des (2012) 26:701–723
DOI 10.1007/s10822-011-9531-0

Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function

Nadine Schneider · Sally Hindle · Gudrun Lange · Robert Klein · Jürgen Albrecht · Hans Briem · Kristin Beyer · Holger Claußen · Marcus Gastreich · Christian Lemmen · Matthias Rarey

Received: 31 August 2011 / Accepted: 12 December 2011 / Published online: 27 December 2011

Abstract The HYDE scoring function consistently describes hydrogen bonding, the hydrophobic effect and desolvation. It relies on HYdration and DEsolvation terms which are calibrated using octanol/water partition coefficients of small molecules. We do not use affinity data for calibration, therefore HYDE is generally applicable to all protein targets. HYDE reflects the Gibbs free energy of binding while only considering the essential interactions of protein–ligand complexes. The greatest benefit of HYDE is that it yields a very intuitive atom-based score, which can be mapped onto the ligand and protein atoms. This allows the direct visualization of the score and consequently facilitates analysis of protein–ligand complexes during the lead optimization process. In this study, we validated our new scoring function by applying it in large-scale docking experiments. We could successfully predict the correct binding mode in 93% of complexes in redocking calculations on the Astex diverse set, while our performance in virtual screening experiments using the DUD dataset showed significant enrichment values with a mean AUC of 0.77 across all protein targets with little or no structural defects. As part of these studies, we also carried out a very detailed analysis of the data that revealed interesting pitfalls, which we highlight here and which should be addressed in future benchmark datasets.

Keywords Virtual screening · Docking · Protein–ligand interactions · Binding affinity · Scoring · HYDE

Introduction

The major hurdle in a structure-based virtual screening process is still the classification of screened compounds into strong binders, weak binders and non-binders. Considering the huge amount of compounds which are tested during a virtual screening run, the challenge is to determine in milliseconds, or a few seconds at the very most, if a compound will bind to a certain protein target. Achieving both is the balancing act required of the scoring function incorporated into the virtual screening software. With the only data to hand being the three-dimensional coordinates of the atoms of the protein and pose of the respective compound, the scoring function must be able to classify the compounds as binders or non-binders. To do this, the scoring function has to model the physico-chemical interactions between the two molecules. Molecular interactions have been investigated experimentally and theoretically for more than 50 years, but nevertheless they are still not

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9531-0) contains supplementary material, which is available to authorized users.

N. Schneider · M. Rarey (✉)
Center for Bioinformatics, University of Hamburg,
Bundesstr. 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

S. Hindle · H. Claußen · M. Gastreich · C. Lemmen
BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin,
Germany

G. Lange · R. Klein · J. Albrecht
Bayer CropScience AG, Industriepark Hoechst, G836,
65926 Frankfurt am Main, Germany

H. Briem
Bayer Pharma AG, Global Drug Discovery,
Müllerstr. 178, 13353 Berlin, Germany

K. Beyer
Bayer Pharma AG, Global Drug Discovery, Aprather Weg 18A,
42113 Wuppertal, Germany

understood in every detail [1, 2]. Moreover, oftentimes not all factors that may finally contribute to the binding affinity of the compound are available to the scoring function—one current topic of interest is the role of water, not only during binding but also in the unbound state of both the protein and ligand [3].

A plethora of different scoring functions [4] have been developed over the last 20 years with the aim of obtaining a high correlation to experimental binding affinity. Most of these functions model molecular interactions using empirical data concerning, for example, hydrogen bonds, metal interactions, electrostatic attraction and the hydrophobic effect. The calibration of the scoring function is done mostly using protein–ligand crystal structures (see for instance [5, 6]) and experimentally measured affinities (see for instance [7–9]). This leads to the largest drawback of these types of scoring functions: They only rely on stabilizing interactions. It is known, however, that destabilizing contributions (e.g. desolvation of polar atoms) also play a major role in molecular interactions [10–12]. Due to the lack of experimental results (crystal structures and binding affinity data) of non-binders and weak binders, new approximated terms are then added to the scoring functions to counterbalance this deficiency, which unfortunately often leads to complicating the model instead of improving the results. On the contrary, Pearlman et al. recently developed a new paradigm for the creation of scoring functions [13]. They postulated the following criteria for a scoring function: First of all, keep it simple and intuitive; secondly, it should require no postparameterization tweaking and should be applicable to multiple target systems without reparameterization; finally, it should be rapidly evaluated for any potential ligand. The same criteria were important to us during the development of our new scoring function HYDE. We even extended the criteria to disallow any calibration of the scoring function on protein–ligand complexes and experimental affinity data. Our reasoning for this is that, as well as the absence of information about destabilizing contributions mentioned above, with such data, a lot of noise caused by experimental inaccuracies is introduced into the model.

The HYDE scoring function depends on a HYdration term and a DEsolvation term which are calibrated using octanol/water partition data (logP) of small molecules. In our model we include all essential interactions and therefore consistently describe hydrogen bonding, the hydrophobic effect and desolvation. Consequently, our method not only considers the favorable molecular interactions arising during the binding process but also the unfavorable contribution to binding affinity which stems from the desolvation of hydrophilic atoms in the molecular interface. In our experience, considering the desolvation penalty of hydrophilic groups definitely aids the detection of false positives during a virtual screening experiment.

In this paper we have accomplished a large-scale validation study of the HYDE scoring function by testing its performance in redocking and virtual screening experiments. This study was carried out in the course of the 241st ACS National Meeting where a Docking and Scoring Symposium was organized by Gregory Warren, Neysa Nevins and Georgia McGaughey. The goal of this Symposium was to evaluate the optimal performance of current docking and scoring methods on public data as well as the suitability of these benchmark datasets for reliable validation of computational methods. In this context, we used the Astex diverse set [14] to assess the ability of the HYDE scoring function to find the bioactive conformation of a ligand from a pool of poses produced by the docking algorithm of FlexX [15, 16]. The virtual screening performance is tested using all 40 targets of the Directory of Useful Decoys (DUD) [17] as well as the WOMBAT ligands published by Good and Oprea [18]. We show from the results of the redocking experiments that the HYDE scoring function is perfectly suited to application in lead optimization during a drug discovery process. We also demonstrate how HYDE can be successfully deployed for hit identification in virtual screening.

The paper is organized as follows. Firstly, we introduce our recently developed HYDE scoring function [19] and illustrate how useful it can be in lead optimization with a detailed example. Secondly, we explain the different methods used in this evaluation study. These methods are only summarized here, since detailed descriptions can be found elsewhere [15, 19–21]. The next section gives a detailed analysis of our redocking and virtual screening results. For both, we also discuss the problems and pitfalls which we primarily detected in the input data during our validation. Finally, we summarize the performance of the HYDE scoring function in large scale applications and our findings concerning the data contained in the two benchmark datasets.

Methods

The HYDE scoring function

The HYDE scoring function [19] is a recently developed scoring function based on HYdration and DEsolvation terms. It consistently describes the energetically favorable contributions of hydrogen bonding and the hydrophobic effect, as well as the energetically unfavorable contribution of polar desolvation to the binding affinity. The overall intention is to estimate the change in desolvation and saturation between the unbound and the bound molecules in a protein–ligand complex. The HYDE scoring function takes following form:

$$\Delta G_{HYDE} = \sum_{atoms\ i} \Delta G^{i}_{desolvation} + \Delta G^{i}_{saturation} \quad (1)$$

Both the desolvation and saturation terms were derived from the Gibbs–Helmholtz equation. The two atom-based terms take the following form:

$$\Delta G^{i}_{desolvation} = -2.3RT \cdot \left(acc^{i}_{unbound} - acc^{i}_{bound}\right) \cdot p\log P^{i} \quad (2)$$

$$\Delta G^{i}_{saturation} = \frac{2.3RT}{f_{sat}} \cdot \left(sat^{i}_{bound} - sat^{i}_{unbound}\right) \cdot p\log P^{i} \quad (3)$$

The $p\log P^{i}$ factor is the partial logP of an atom i, which we have calibrated on experimental logP values of small and simple molecules taken from the Starlist [22]. The constant factor in both equations arises from converting the natural logarithm of the Gibbs-Helmholtz equation to the common logarithm used in the partial logP factor. In the desolvation term, the *acc* value is the solvent accessible area of an atom i in the *unbound* or in *bound* state respectively. In the saturation term, which is very similar to the desolvation term, a saturation factor f_{sat} [23] is included. The saturation factor describes the incomplete saturation of the hydrogen bond network in bulk water; it is temperature dependent. At a temperature of 273 K the saturation factor is $f_{sat} = 0.89$ while at 310 K it is only about $f_{sat} = 0.84$. We also calculate the change in saturation for an atom i by considering the number of intermolecular and intramolecular interactions sat^{i} in the *bound* and in the *unbound* state. To model these interactions, HYDE uses sections of spherical surfaces—known as interaction surfaces—based on the FlexX interaction model [15] (similar to that found in LUDI [24]). Here, interactions between the ligand and protein are modeled by overlapping interaction surfaces. A perfect overlap describes a geometrically perfect hydrogen bond. Interactions that deviate from the perfect geometry are scaled according to a penalty factor until a certain threshold at which HYDE considers the hydrogen bond to be no longer made.

We do not use measured affinities to train our scoring function; the only calibrated values we use are the partial logP parameters. For this reason HYDE is a very general scoring function, whose good performance is not restricted to target systems on which it is trained, like most other scoring functions. We achieve a correct prediction of the hydrophobic effect (≈ -110 J/Å^2) [25], hydrogen bonds in vacuo (≈ -16 kJ/mol) [26], hydrogen bonds in water (-2 to -6 kJ/mol) [27] and also the affinity loss due to unsatisfied hydrogen bond functions (≈ 6 kJ/mol) [28, 29]. The HYDE score estimates the binding affinity of a ligand and protein, since the single terms are derived from the Gibbs–Helmholtz equation as described above. The finer details of the HYDE scoring function implementation will be described in a subsequent publication.

Since the HYDE score is atom-based, we can introduce a very intuitive atom coloring scheme for visualization of the HYDE score, which is outlined in Fig. 1. The atom-based contributions of the HYDE scoring function are mapped to the ligand atoms in the protein–ligand complex. We use three different colors to represent the energetically contribution of an atom to the protein–ligand complex. Red atoms contribute energetically unfavorably to the overall binding energy, while green atoms have a favorable ΔG_{HYDE} contribution. White atoms are energetically negligible. The schematic shown in Fig. 1 highlights the properties of the HYDE scoring function: The red ligand atom (O8) is an aromatic oxygen, which is located in the region of a protein carbonyl oxygen. Both atoms desolvate each other and, because in this case two hydrogen bond acceptors are facing each other, no hydrogen bond can be formed to compensate the desolvation penalty of both atoms. In this case, we get a desolvation cost of $\Delta G_{HYDE} = 10.6$ kJ/mol, where the main contribution arises from the carbonyl oxygen of the receptor, while the aromatic oxygen itself contributes less since it is only a weak acceptor. The whole contribution is mapped to the aromatic oxygen of the ligand. The green atom on the right (C6) is an aromatic carbon which is situated in a hydrophobic pocket of the protein. It is desolvated by the surrounding atoms and desolvates them also. Here, we achieve a favorable ΔG_{HYDE} contribution through the hydrophobic effect of -3.6 kJ/mol, which is also mapped to the aromatic carbon of the ligand. A third case which may occur is shown on the left of Fig. 1: A hydrogen bond donor interacting with a hydrogen bond acceptor (N7 on the ligand). A prerequisite for the hydrogen bond is the desolvation of the hydrogen bond donor and acceptor—here there is an unfavorable ΔG_{HYDE} contribution. This unfavorable energy is overcompensated by the geometrically perfect hydrogen bond. The sum of these contributions is $\Delta G_{HYDE} = -2.2$ kJ/mol, which is the energy of the hydrogen bond and which is mapped onto the aromatic nitrogen of the ligand. The reason for mapping the protein contributions on ligand atoms is to facilitate the simple identification of potential optimization sites at the ligand.

Datasets and data preparation

We used two public available benchmark datasets in our study, the Astex diverse set for redocking experiments and the DUD dataset for virtual screening runs. Both datasets were carefully revised by the organizers of the Docking and Scoring Symposium and then supplied to attendees. We summarize shortly the main revision steps accomplished by the organizers:

All ligands of the *Astex diverse set* [14] were extracted from the complexes and converted into *sd* files. It was

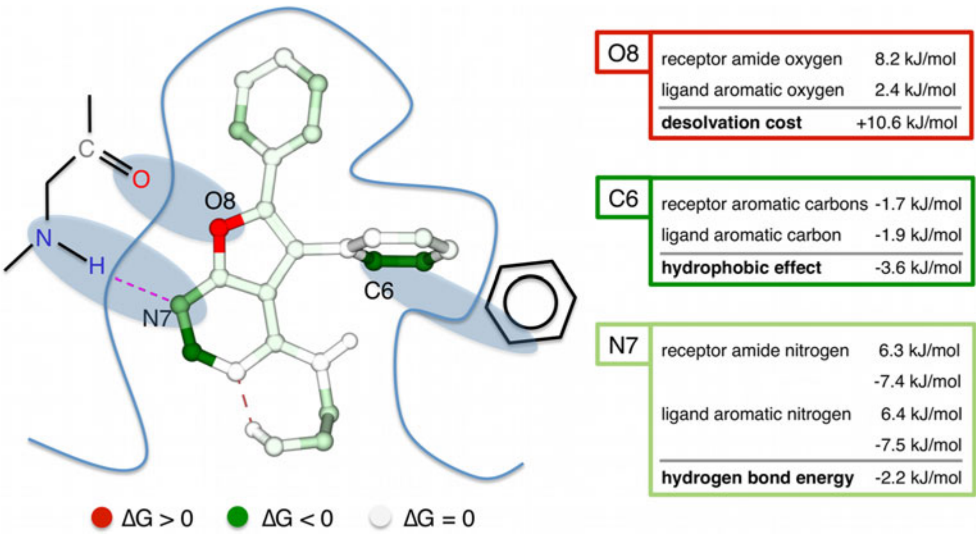
C
O
N
H
O8
N7
C6
O8
receptor amide oxygen 8.2 kJ/mol
ligand aromatic oxygen 2.4 kJ/mol
desolvation cost +10.6 kJ/mol
C6
receptor aromatic carbons -1.7 kJ/mol
ligand aromatic carbon -1.9 kJ/mol
hydrophobic effect -3.6 kJ/mol
N7
receptor amide nitrogen 6.3 kJ/mol
-7.4 kJ/mol
ligand aromatic nitrogen 6.4 kJ/mol
-7.5 kJ/mol
hydrogen bond energy -2.2 kJ/mol
● ΔG > 0
● ΔG < 0
○ ΔG = 0

Fig. 1 HYDE coloring scheme showing the energetically favorable contributions of hydrogen bonding and the hydrophobic effect, as well as the energetically unfavorable contribution of polar desolvation to the binding affinity

ensured that all atom types and bond orders are consistent to primary literature. Hydrogen atoms were added to the receptors only using the program REDUCE (v 3.13) [30, 31], then all atoms were re-refined using phenix.refine [32]. In ten of the proteins, a covalently modified amino acid is found. In this case no automated method can be used to re-refine all atom positions. Hence, for these ten receptors, only the hydrogen positions were optimized using the MMFF94S force field in the OpenEye software package (SZYBKI) [33]. All crystallographic waters were removed from the receptors and finally the revised structures were supplied as PDB files. The extracted ligands were provided to serve as a reference to define the active sites in the protein. In contrast to the original version of the Astex diverse set, all ligands available in a crystal structure (i.e. in multimeric proteins) were provided, hence the number of pockets increased from the original 85 pockets to 151 pockets. Additionally, a non-crystallographic conformation of each ligand was supplied as starting point for the docking.

The receptors of the *DUD* dataset [17] were processed in a similar way to the receptors of the Astex diverse set. Electron density maps were not available for all targets, hence the DUD receptors were not newly refined. The added hydrogen atoms were also minimized using again OpenEye's implementation of the MMFF94S force field (SKYBKI). Crystal waters specifically mentioned by the DUD authors were retained. Duplicate active and decoy compounds were removed from the DUD dataset. Strong acids/bases were deprotonated/protonated using the MOE DBV "wash" functionality in the MOE 2009.10 database version [34], if not already done so in the original DUD ligand tautomer/protomer. The 3D structures were generated with Schrodinger's (Maestro v9.1) LigPrep (FF: OPLS_2005, Ionization: Retain original State, Desalt and Generate tautomers deselected, Stereoisomers: Determine chiralities from 3D structure) [35].

We decided that, in addition to carrying out calculations using *original data*, we would also apply our own active site preparation method and compare the two sets of results. We refer from now on to two datasets: One where we leave the data in its original form as requested by the organizers—called *original data,* and one where we allowed our own software to prepare the data as a user would during standard application—called *standardized data.* The latter differs slightly from the former in terms of hydrogen bond network optimization of the protein structures and some manual corrections to some of the data. The hydrogen bond network optimization is carried out automatically with ProToss [21] during the receptor preparation in the LeadIT software [16] (for more detailed information see the Experimental Details section).

Preparation of active sites

Pre-processing of some of the PDB data was necessary despite rectifications and updates from the organizers. See Data Preparation in the Supplementary Material for more information about the changes made.

LeadIT was used to automatically prepare the protein ready for docking—and for rescoring later—from the PDB files provided. The active site was selected by taking all amino acids, cofactors, waters and ions lying within 6.5 Å of any reference ligand heavy atoms. The protonation state of the protein depended on the data being used (*original* or *standardized*—see above). LeadIT then assigns coordination geometries to any metal ions [36]. This is important in order to determine whether there are any coordination sites free for the ligand during docking. In fact, if this is the case, the coordination sites are automatically set by LeadIT as pharmacophore type constraints [20] to be applied during docking. The automatically determined coordinations (and pharmacophore constraints) were left untouched for the *original data* calculations. For the calculations on the *standardized data*, the coordinations were sometimes changed, or the pharmacophore constraints deleted—see Supplementary Material for detailed information on interactive changes made. Finally, the protein was saved to a so-called project file ready for input into the docking and rescoring software.

Preparation of ligands

Although the ligands were already prepared by the organizers—and some issues already addressed in the Data Guidelines and ensuing communications—some fixes and alterations were still necessary. Changes were made only to the Astex ligands and to the reference ligands in the DUD datasets. The active and decoy ligand sets for WOMBAT and DUD were unaltered. The changes were made for the *standardized data* but not for the *original data*. See the Supplementary Material for further information. In addition, all conjugated functional groups, such as carboxylates or amidines, were described as delocalized systems with "delocalized" formal charges during docking as this is necessary for FlexX. All ligands were converted from *sd* format into *mol2* format using Naomi [37].

Structural target analysis

Additionally, we classified the 40 DUD targets according to structural quality criteria. An experienced crystallographer carried out a detailed visual inspection whereby the electron density maps of the structures were also consulted. The classification can be found in column 2 of Table 3 in the Results section. We defined four different quality criteria; Q1 structures have no or only marginal structural deficiencies, structures with deficiencies on the ligand and with waters required to obtain reasonable results were classified as Q2. The next two quality levels, Q3 and Q4, show poor electron density and crystal packing effects (a detailed listing can be found in the Supplementary Material Table S8).

Workflow

In both redocking and virtual screening experiments, the calculations were basically carried out in two stages: A docking calculation followed by rescoring of the docking poses. The docking engine used was the latest version of FlexX [15] (in the LeadIT software suite 2.0.1 [16]), while rescoring was carried out using the HYDE scoring function in a stand-alone tool. Information about the protein was shared between the two pieces of software using LeadIT project files ('*fxx*' files), while the docking solutions were saved and then loaded into the HYDE scoring tool using the *mol2* file format.

Docking with FlexX

FlexX is an established docking tool [38] that docks ligands into the active site using a robust incremental construction algorithm. The ligand is decomposed into several components and then, guided by interactions with the protein, is reconstructed in the active site using various placement strategies to fully explore the conformational space. As in HYDE, interactions between the ligand and protein are modeled by overlapping interaction surfaces which are based on the interaction model in LUDI [24]. These interaction surfaces define regions around a functional group that are available for interactions with a partner functional group. The placement strategies, incremental build up procedure and scoring function (based on the LUDI scoring function [7]) are fast, making FlexX also suited to large scale screening applications. Recently, a new placement algorithm called the Single Interaction Scan (SIS) has been implemented in FlexX, which was originally designed for docking small fragments that make few directed interactions, but which has also proved very successful for docking of hydrophobic molecules and for docking into hydrophobic pockets. In both the redocking and screening experiments, poses resulting from the standard placement algorithm and the SIS algorithm were combined to give the best chance of having maximum diversity in the docking poses. More information about certain parameter settings etc. used during the calculations is given in the Supplementary Material.

Rescoring with HYDE

The HYDE scoring function includes no terms to estimate the steric arrangement of a protein–ligand complex. Therefore, to obtain good hydrogen bond geometries, to avoid clashes between the ligand and the protein as well as within the ligand, and also to relax the strain energy of the ligand, two optimization procedures were available prior to scoring. In the first, the hydrogen bond network within the protein and between the protein and ligand is optimized using ProToss [21]. The second is an optimization/minimization of the ligand in the active site using a numerical optimization algorithm. The implementation of the HYDE scoring function together with the two optimization procedures in a stand-alone HYDE tool is explained in more detail in the Experimental Details section.

Validation measures

In the redocking experiments, our performance was measured by calculating the Root Mean Square Deviation (RMSD) between the coordinates of the heavy atoms of the crystallographically determined ligand structure and the docked pose. We used the MOE software (v 2010.10) [39] to do this after the optimization of the pose within the HYDE tool. A good metric for the ability of a method to find the natural binding mode of a ligand in the pool of decoy poses is the RMSD value of the best scored pose. Finding the correct binding mode considering only the best scored pose is a challenging task, therefore, often the best RMSD among the *x* top scoring poses are considered. In this study, we consider the first 32 top scoring poses to find the best RMSD, as requested by the organizers of the Symposium.

We estimated the overall virtual screening performance of our methods by plotting Receiver Operator Characteristic (ROC) curves and calculating the Area Under Curve (AUC) for all targets. The optimal value here is AUC = 1.0, while a method that randomly selects active molecules from a set of decoy molecules will achieve an AUC value of 0.5. The AUC value is estimated using the trapezoidal rule [40].

To obtain a ROC curve, the sensitivity (SE) is plotted against inverse specificity (SP) (1 − specificity) of a method. In other words, the signal is plotted versus the noise; in a virtual screening context this takes the form of the true positive ratio versus the false positive ratio. These are as follows [41] (TP = true positives, TN = true negatives, FP = false positives, FN = false negatives):

$$\mathrm{SE} = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$$\mathrm{SP} = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{5}$$

Looking at the true positive rates when 0.1, 1 and 2% of the decoys have been found gives us a good estimate of the performance of the method within the first few percent of the ranked dataset, i.e. a good descriptor for early enrichment. For comparison, the random selection of active compounds from a pool decoys will achieve true positive rates of 0.001, 0.01 and 0.02 respectively at these thresholds. It is worth noting that the true positive rate at 0.1% is not meaningful for datasets containing less than 1,000 decoy molecules.

An additional way of evaluating the performance of a method in structure-based virtual screening is the *null hypothesis*. The focus here is to determine if the structure-based method includes the information which it can only get from the protein (interactions formed between ligand and protein) to discriminate active from decoy compounds. To carry out null hypothesis testing, the active and decoy molecules of a certain target are docked into another target of the dataset. The “wrong” target should be related in some way to the “right” to get a reasonable evaluation. If a method still achieves a similar enrichment on the wrong target, two possibilities exist that could cause this effect. Firstly, the right and the wrong target may be too related to draw conclusions on the performance of the method itself. Secondly, the actives and the decoys may differ in some way which is identified by the scoring function without regarding the protein structure. This is not necessarily an error or a bias in the scoring function, but it is undesirable in the drug development process as it may lead to highly unspecific binders.

Results and discussion

Redocking studies on Astex diverse set

In the redocking studies we used the FlexX docking module of the LeadIT software for generating docking poses for all defined pockets of the Astex dataset targets. These generated docking poses were optimized, then rescored and reranked using the HYDE scoring function. The detailed set-up is described in the Methods section as well as in the Experimental Details. Both the *original data* and *standardized data* were used as input. We calculated heavy atom RMSD values for all optimized poses to assess the similarity to the bioactive conformation of the ligand.

Overview of performance in redocking accuracy

The overall results for both versions of the dataset can be found in Table 1. If we consider only poses with RMSD values below 2 Å as being a correct prediction of the binding mode, we find the correct pose for 93% of the pockets in the *standardized data* and for 89% in the *original data* within the top 32 scored poses. If we restrict the definition of a correct binding mode prediction to be a pose with an RMSD below 1 Å, we achieve success in 69% of the pockets for the *original data* and even 77% for the *standardized data*. The difference between the two versions of the dataset becomes even more obvious when we consider only the top ranked pose. For the *original data*, the RMSD of the top ranked pose is below 2 Å for 66% of the pockets whereas for the *standardized data*, the RMSD of the top ranked pose is below 2 Å for 75% of the pockets.

Table 2 shows the statistics across all pockets for both datasets. Considering the 32 best-scored poses for the *standardized data*, the lowest RMSD value we find in the whole dataset is 0.16 Å, while the median RMSD is 0.53 Å.

Table 1 Percent of docking poses with RMSD values equal or better than 0.5, 1.0, 1.5 and 2.0 Å

	Original data (151 pockets)				Standardized data (147 pockets)			
Rank	0.5	1	1.5	2	0.5	1	1.5	2
1	9.9	35.1	51.7	66.2	16.3	**45.6**	66.7	**74.8**
≤2	12.6	42.4	62.3	71.5	21.1	55.8	74.1	80.3
≤3	13.9	46.4	64.9	76.8	25.9	63.9	78.9	84.4
≤4	14.6	48.3	65.6	77.5	27.2	65.3	79.6	85.0
≤5	17.9	51.0	68.9	78.8	27.9	68.0	80.3	86.4
≤20	32.5	64.2	80.8	86.8	41.5	76.2	86.4	91.2
≤32	39.1	68.9	84.8	88.7	46.3	**76.9**	88.4	**92.5**

All results are given in percent of total pockets

Table 2 RMSD statistics on the original and the standardized data

	Original data		Standardized data	
Rank	1	≤32	1	≤32
Mean	2.15	0.91	1.89	0.78
STD	1.97	0.85	2.06	0.74
Median	1.45	0.59	1.06	0.53
Min	0.27	0.19	0.22	0.16
Max	9.57	6.49	9.65	5.80

All results are given in Å

The following more detailed results only deal with the *standardized data*. In Fig. 2a, the redocking results of the top scoring pose of all pockets of all 85 protein–ligand complexes are plotted. There are some fluctuations amongst the results for pockets of one protein–ligand complex—these arise from slightly different binding modes amongst the pockets, different side chain conformations and the inclusion of non-biological pockets. Some of these cases are discussed below. In Fig. 2b, the results of the top scoring pose considering only the best performing pocket of a complex are shown. In 82% of the complexes we achieve RMSD values below 2 Å. There are only a few outliers—a more detailed discussion of three of the outliers (marked in dark blue, Fig. 2b) is given in the following section. Finally, we present the detailed results considering the best 32 poses for all pockets (see Fig. 3a)—there are only a few complexes showing fluctuations amongst the pockets and in 93% of the complexes we found a RMSD value below 2 Å. Again, when taking only the best pocket into account, we get RMSD values below 2 Å in 96% of the complexes with even 82% of the complexes having RMSD values below 1 Å (see Fig. 3b).

Detailed discussion of the results for redocking

In this section we consider only the *top scoring pose* for a complex in the *standardized data*. As mentioned above, for some complexes of the dataset there are fluctuations in the redocking performance for the different pockets. These are discussed below. Following on from there, the three outliers highlighted in Fig. 2b will be analyzed in more detail and finally we highlight three examples where the redocking performance was excellent.

Fluctuations amongst pockets of one complex

In *1sq5*, a small, hydrophilic and very flexible ligand is situated in a large hydrophilic pocket. Here, the binding mode amongst the four pockets is slightly different and at least four waters can be found in the active site. One of the water molecules is a bridging water molecule that interacts with two hydrogen bond functions of the ligand and one of the active site. A second water molecule is bound very tightly in a small hydrophilic pocket of the active site, forming three hydrogen bonds with the protein and which can probably, therefore, not be displaced by the ligand. If these two important waters were present as part of the input, a more consistent binding mode could have been predicted at the top rank. A similar case can be found in *1hvy*, where the very flexible tail of the ligand is found at the exit of the binding pocket stabilized by at least two water molecules. Different side chain conformations occur in the two pockets of *1ia1*. MET 25 A and MET 25 B, situated at the bottom of the pocket, have different orientations and therefore influence the shape of the pocket. In *1hp0*, the fluctuations in the redocking performance of the two pockets arise from the metal interactions with the calcium ion, whose coordination geometry is very distorted. Poor electron density is another reason causing different performance assessment in the different pockets of a complex. In *1j3j*, for example, one pocket is well-resolved while the other pocket suffers from very poor electron density and high temperature factors. Also in the dataset, non-biological binding sites (crystal artifacts), e.g.

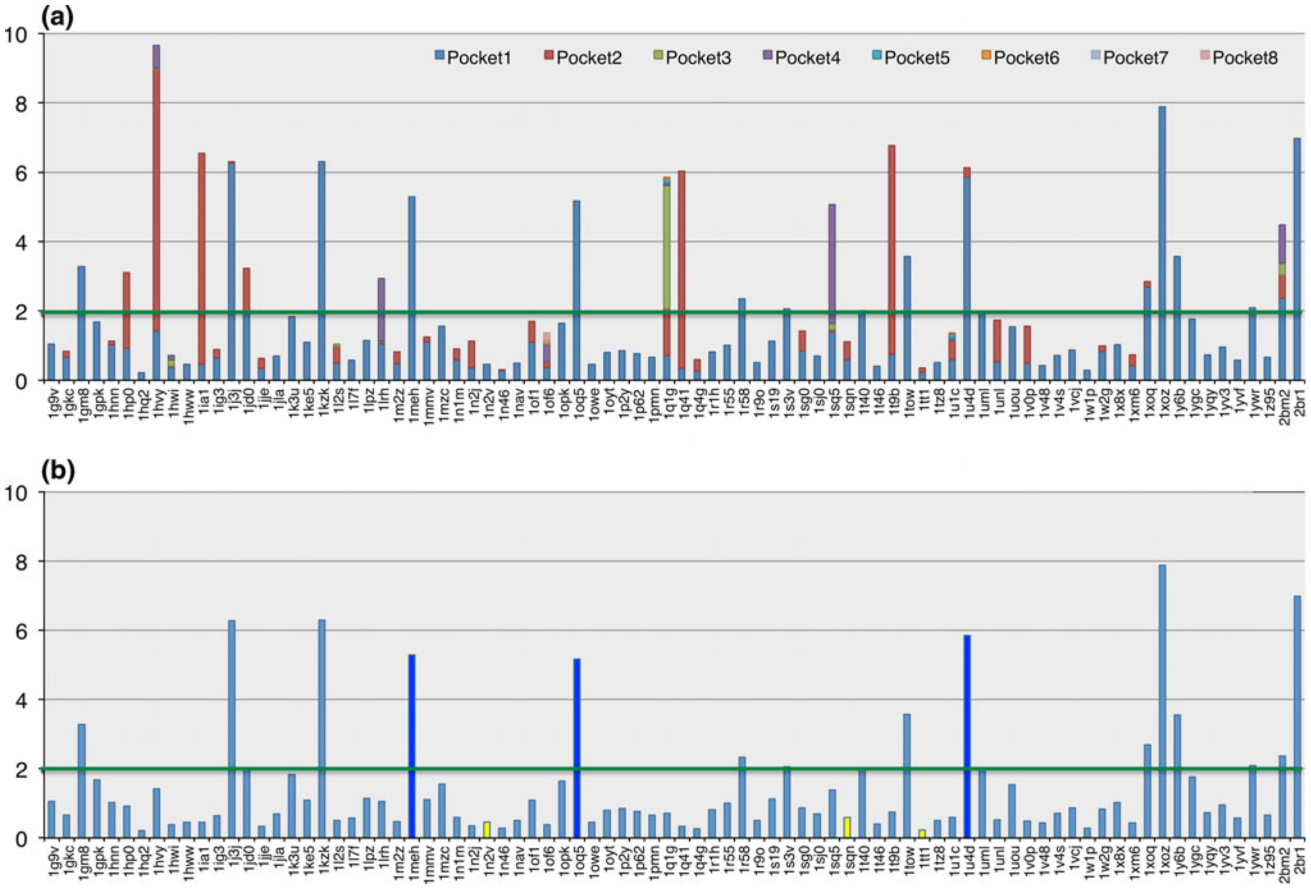

Fig. 2 **a** RSMD of the *best scoring* pose for *all pockets* of all 85 Astex diverse set complexes (*standardized data*). **b** RSMD of the *best scoring* pose of the *best pocket* of all 85 Astex diverse set complexes respectively (*standardized data*). *Dark blue*: Outliers discussed in detail further below. *Yellow*: Good performing examples which are discussed in more detail further below

in *1l2s*, or incomplete binding sites were included. For example, in *1tz8* four of the five pockets were incomplete, i.e. only one half of each pocket was included in the PDB file provided. In addition, those four pockets were essentially only two pockets, since the reference ligands, which were extracted automatically from the PDB file, comprise of two copies of each ligand that represent only alternate conformations for the ligand. For 1tz8, the only real way to achieve correct redocking would be to use a biological subunit, which is also deposited in the PDB Databank and contains only two complete pockets. We removed the four incomplete pockets in our *standardized data*.

Three outlying complexes with respect to RMSD

1. *1meh: An example of a soaked structure with few directed interactions*: In 1meh only two hydrogen bonds were formed between the protein and the carboxylate group of the flexible tail of the ligand. An aromatic ligand oxygen can be found whose lone pairs point towards a backbone amide nitrogen of the protein (N GLY 314 A), but the expected hydrogen bond interaction cannot be established within the hydrogen bond parameters of HYDE (Fig. 4b). The distance between the acceptor and the donor, as well as the angles, are out of range, which might be an artifact of soaking. The HYDE score of the crystal structure is quite poor with a value of −14 kJ/mol compared to experimental affinity of −46 kJ/mol ($K_i = 9nM$) [42]. The best scored pose achieves a HYDE score of −35 kJ/mol with a RMSD value of 5.3 Å. Here, the ligand is flipped in the binding pocket and the flexible tail interacts with a small hydrophilic pocket where three well-defined water molecules can be found in the crystal structure (Fig. 4a). The pose with the best RMSD value of 1.2 Å achieves a HYDE score which is only 4 kJ/mol weaker ($\Delta G_{HYDE} = -31$ kJ/mol). Here, the pose is shifted about 1 Å compared to the crystal structure, so that the missing hydrogen bond mentioned before between the aromatic oxygen of the ligand and the backbone nitrogen of the protein can be formed. An additional hydrogen bond can be established by a hydroxyl group of the ligand. Hence, this binding mode seems to be more reasonable than the

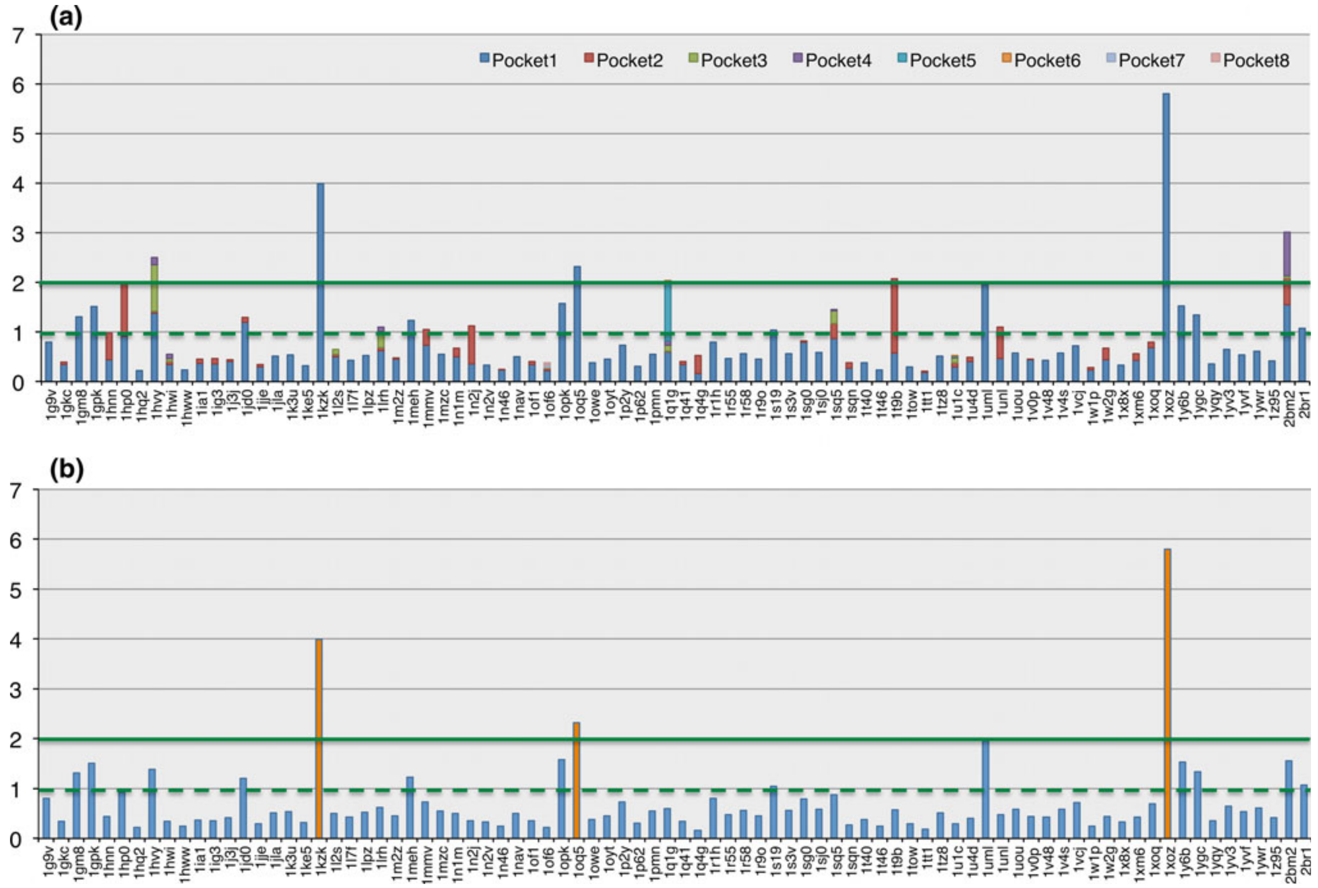

Fig. 3 **a** RSMD of the *best-of-32* pose of *all pockets* of all 85 Astex diverse set complexes (*standardized data*). **b** RSMD of the *best-of-32* pose of the *best pocket* of all 85 Astex diverse set complexes respectively (*standardized data*). *Orange*: No better pose was generated by the docking tool, see example 1oq5 below

crystal structure to us. Both poses are shown in the HYDE coloring scheme in Fig. 4c, d.

2. *1oq5: No good pose found by the docking engine*: In 1oq5 we found the interaction of the sulfonamide group of the ligand with the zinc ion of the protein to be the only directed interaction. Also, this interaction is not clearly defined since in the original crystal structure from PDB, the deprotonated nitrogen of the sulfonamide group coordinates the zinc ion [43], while in the supplied crystal structure (*original data*) one of the oxygen atoms coordinates the zinc. We corrected the sulfonamide in the reference ligand of the supplied structure to the conformation of the original crystal structure, since an interaction between a deprotonated nitrogen with a metal ion is more common than an interaction between the sulfonamide oxygen and an ion. The best scored pose achieves an RMSD value of 5.2 Å with a HYDE score of −28 kJ/mol which is quite low compared to the measured affinity of −49 kJ/mol (IC50 2.1 nM). The pose is again flipped in the binding pocket and the metal ion is coordinated by the sulfonamide oxygen, so that the sulfonamide nitrogen interacts with the protein (Fig. 5a). The best pose by RMSD with a value of 2.3 Å only gets a HYDE score of −20 kJ/mol (Fig. 5b). The interaction with the metal ion is formed perfectly in this pose, but the toluene is shifted so that an amide group of an asparagine side chain (N ASN A 67) is desolvated. For this complex we did not find a pose close to the crystallographic binding mode, the pose mentioned above achieves the best RMSD value across all generated poses. Fig. 5c shows the assessment of the crystallographic binding mode with HYDE. Here, the predicted affinity is $\Delta G_{\mathrm{HYDE}} = -44$ kJ/mol, which is in good agreement with the experimental affinity of −49 kJ/mol. Hence, if a good pose had been generated, HYDE would be able to detect it.
3. *1u4d: Poor electron density in the binding pocket*: In 1u4d we found very poor electron density for large parts of the binding pocket, as well as high temperature factors (Fig. 6a, b). In this complex, only the well-resolved part of the binding pocket reveals geometrically good hydrogen bonds with the crystal structure ligand. Therefore, the best pose by RMSD with a good

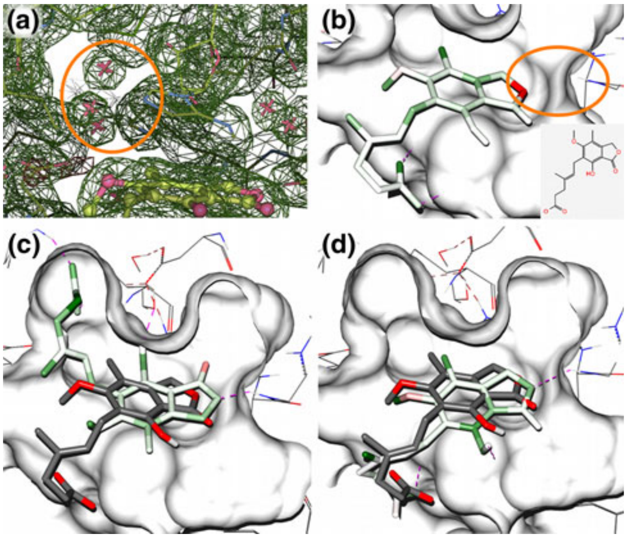

Fig. 4 Redocking results for complex 1meh. The crystal ligand is shown in atom color. **a** Three well-resolved waters in a small hydrophilic pocket of the complex. **b** Crystal ligand in HYDE coloring. The aromatic ligand oxygen (*red atom*) points towards a backbone amide nitrogen of the protein, but no hydrogen bond is formed since distance between the acceptor and the donor as well as the angles are out of range of our hydrogen bond parameters. **c** Best scoring pose (HYDE coloring): RMSD = 5.3 Å and ΔG_{HYDE} = −35 kJ/mol. **d** Best pose by RMSD: RMSD = 1.2 Å and ΔG_{HYDE} = −31 kJ/mol

RMSD of 0.4 Å only achieved a poor HYDE score of −21 kJ/mol (Fig. 6d). The best scoring pose with an RMSD value of 5.9 Å and again a quite low HYDE score of −27 kJ/mol is flipped and shifted compared to the reference ligand. This happens to avoid the desolvation costs of burying hydrophilic atoms in the poorly resolved loop of the binding pocket without making interactions to them (Fig. 6c).

Three examples where HYDE excelled

1. *1n2v: Rejection of an artificial binding mode by HYDE*: An interesting complex is *tRNA-guanine transglycosylase bound with compound 6* (1n2v), where two binding modes of the ligand seem to be possible (Fig. 7a, b). In the original crystal structure, a conserved water (HOH 1404 A) can be found interacting with three amino acids (GLN 107 A, TYR 106 A, GLY 105 A) on the right side of the pocket as shown in Fig. 9a. As already mentioned, all waters were removed from the PDB files in the supplied dataset. Despite this, HYDE is still able to identify the correct binding mode at the first rank. The RMSD of this pose is 0.5 Å and it achieves a HYDE score of −23 kJ/mol, which again is in perfect agreement with the experimental value of −23 kJ/mol. In Fig. 7c, d, both possible binding modes are shown in the HYDE coloring scheme. (Note that this coloring scheme is different to the one used above: Here, the scores for the protein atoms are represented at the protein atoms and are not mapped onto the respective ligand atoms as before.) At first glance it is already obvious that the artificial binding mode on the right is unfavorable for the protein atoms mentioned above, which should interact in fact with the conserved water. The ligand carbonyl group, which points to these atoms, is not able to interact adequately, leaving some protein hydrogen bond functions unsaturated. This is detected by the HYDE scoring function and penalized with a poor score of only −12 kJ/mol.
2. *1sqn: Strong hydrophobic binder*: For the *progesterone receptor* complexed with *norethindrone*, which is quite a hydrophobic complex, HYDE achieves a perfect prediction. Again, the best scoring pose in each of the two binding pockets of the complex have reasonably good RMSD values of 1.1 and 0.6 Å (Fig. 8) and HYDE scores of −50 kJ/mol and −55 kJ/mol respectively, which are in excellent agreement with the measured affinity of −54 kJ/mol.
3. *1tt1: Strong hydrophilic binder*: 1tt1 contains the very hydrophilic binder *kainate* that forms eight geometrically perfect hydrogen bonds with the protein. In both pockets of this complex we achieve very good RMSD values of 0.2 and 0.4 Å for the best scoring poses

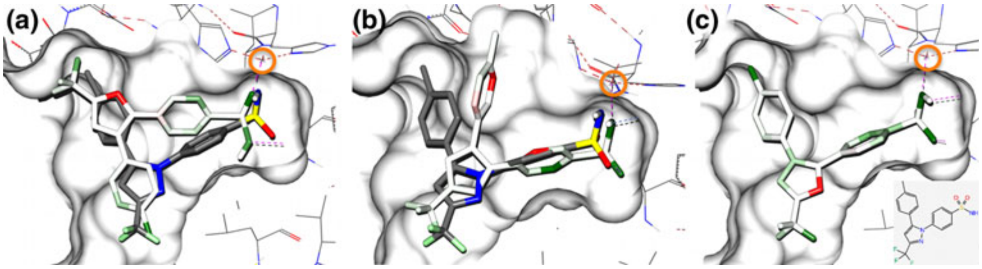

Fig. 5 Redocking results for complex 1oq5. The crystal ligand is shown in *atom color*. Zinc ion is highlighted in *orange circle*. **a** Best scoring pose (HYDE coloring): RMSD = 5.2 Å and ΔG_{HYDE} = −28 kJ/mol. **b** Best pose by RMSD (HYDE coloring): RMSD = 2.3 Å and ΔG_{HYDE} = −20 kJ/mol. **c** Crystal ligand in HYDE coloring: $\Delta G_{experimental}$ = −49 kJ/mol and ΔG_{HYDE} = −44 kJ/mol

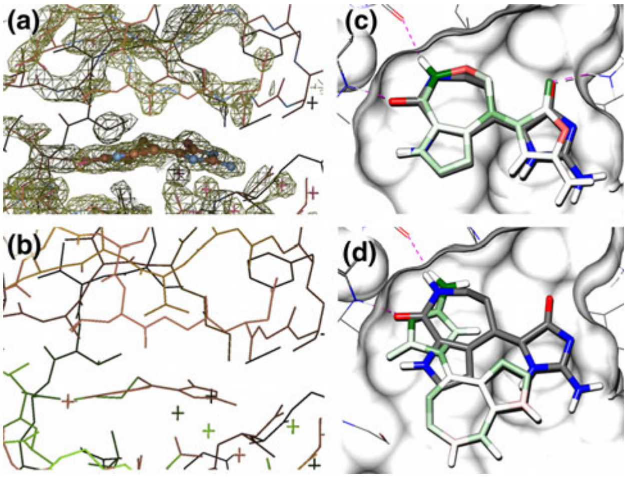

Fig. 6 Redocking results for complex 1u4d. The crystal ligand is shown in *atom color*. **a** Poor electron density of the binding pocket. **b** High temperature factors (B-factors) of the residues in the binding pocket (*red*/*brown*: high B-factor, *green*/*blue*: low B-factor). **c** Best-scoring pose (HYDE coloring): RMSD = 5.9 Å and ΔG_{HYDE} = −27 kJ/mol. **d** Best pose by RMSD (HYDE coloring): RMSD = 0.4 Å and ΔG_{HYDE} = −21 kJ/mol

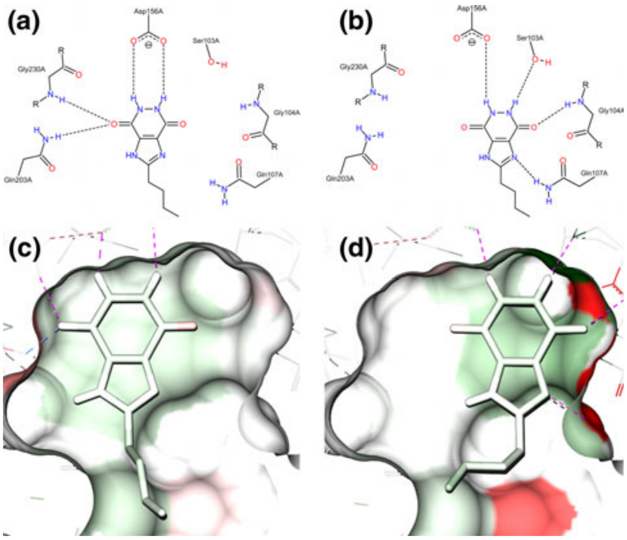

Fig. 7 Redocking results for complex 1n2v ($\Delta G_{experimental}$ = −23 kJ/mol). Two binding modes are possible here. Atoms are colored according to their HYDE score contribution—note that the contributions of the protein atoms are not mapped to the ligand atoms in this illustration. **a** Natural binding mode. **b** Artificial binding mode. **c** Best scoring pose: RMSD = 0.2 Å and ΔG_{HYDE} = −23 kJ/mol. **d** A pose in the artificial binding mode: RMSD = 2.3 Å and ΔG_{HYDE} = −12 kJ/mol. Unfavorable desolvation of protein atoms is directly obvious

(Fig. 9). In addition, the HYDE scores for these poses, which are −44 kJ/mol and −45 kJ/mol respectively, reflect perfectly the experimental binding affinity of −43 kJ/mol.

Screening experiments on the DUD

Again, we used the LeadIT software for generating the docking poses. These generated docking poses were optimized using both optimization procedures, then rescored and reranked using the HYDE scoring function. The detailed set-up is described in the Methods section as well as in the Experimental Details. Both the *original data* and *standardized data* were used as input. We calculated AUC values for all 40 targets in the dataset to assess the overall screening performance. We also calculated the true positive rates at 0.1, 1 and 2% to measure performance amongst the first few percent of the ranked dataset (a more detailed description of the validation measures can be found in the Methods section).

Overview of performance in virtual screening

Table 3 summarizes our performance in virtual screening on the *standardized data*. In comparison to these results Table 4 shows the overall statistics for the *original data*; a corresponding table for detailed results on the *original data* can be found in the Supplementary Material (Table S10). The results were sorted according to the second column. This column contains a quality assessment of the 40 targets; the structural quality is decreasing from Q1 to Q4 targets. Detailed guidelines of this classification were outlined in the Methods section as well as in Supplementary Material (Table S8). We achieve a mean AUC value of 0.72 considering the whole dataset, while, if we limit our analysis to only the Q1 structures, the mean AUC value is enhanced to 0.77. The true positive rates are even more affected by the quality of the input structure—see columns 3–5. Compared to the performance with *original data* (Table 4), the screening results are slightly better using our prepared protein structures. The main difference in performance is shown in the comparison of the true positive rates.

Detailed results of virtual screening performance

In this section we discuss deficiencies found in the data that led to a moderate performance on three targets: p38 MAP kinase, Thymidine kinase and COMT. This analysis is followed by three examples where we have achieved very successful results for the targets estrogen receptor, HMG-CoA reductase and neuraminidase.

1. *p38 MAP kinase (1kv2) Different structural classes of inhibitors which do not all bind to this protein conformation*: This target exhibits a very flexible active site, so that large conformational changes may occur during the binding of different ligands. The degree of

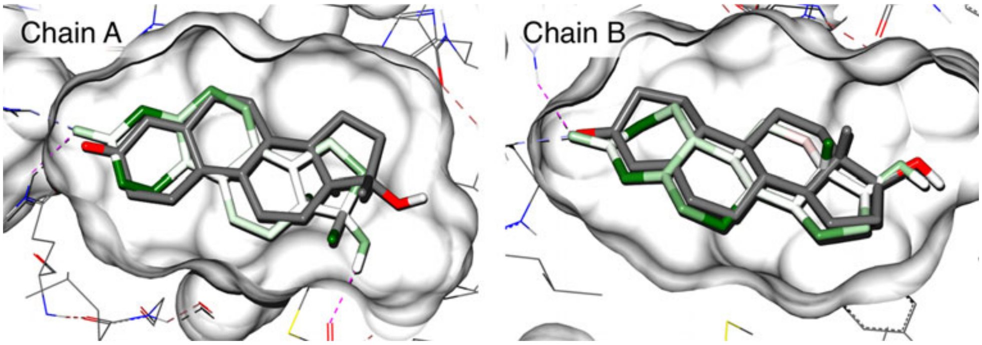

Fig. 8 Redocking results of complex 1sqn ($\Delta G_{experimental} = -54$ kJ/mol). Both pockets of the complex are shown. The crystal ligand is shown in *atom color*. *Left*: Best-scoring pose of 1sqn chain A (HYDE coloring): RMSD = 1.1 Å and $\Delta G_{HYDE} = -50$ kJ/mol. *Right*: Best-scoring pose of 1sqn chain B (HYDE coloring): RMSD = 0.6 Å and $\Delta G_{HYDE} = -55$ kJ/mol

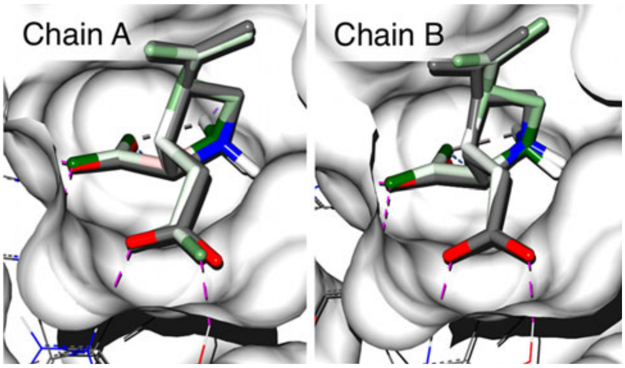

Fig. 9 Redocking results of complex 1tt1 ($\Delta G_{experimental} = -43$ kJ/mol). Both pockets of the complex are shown. The crystal ligand is shown in *atom color*. *Left*: Best-scoring pose of 1tt1 chain A (HYDE coloring): RMSD = 0.2 Å and $\Delta G_{HYDE} = -44$ kJ/mol. *Right*: Best-scoring pose of 1tt1 chain B (HYDE coloring): RMSD = 0.4 Å and $\Delta G_{HYDE} = -45$ kJ/mol

conformational change becomes apparent in Fig. 10a, which shows two superposed crystal structures of p38 MAP kinase, one in DFG-out conformation (1kv2) and the other in DFG-in conformation (1a9u). The crystal structure 1kv2 with the DFG-out conformation is found in DUD. This conformation sterically interferes with ATP binding and consequently with the inhibitors binding to the ATP pocket as well. The collision of the inhibitor from 1a9u with the DFG loop of 1kv2 is directly obvious. In addition, the binding mode of the inhibitors in both structures is completely different, as shown in Fig. 10b. We originally achieved an AUC value of 0.51 for p38 MAP kinase which means there was no differentiation between the actives and the decoys. As a consequence, we analyzed the active compounds in the dataset and found they comprised three different structural classes of inhibitors [19]. The dataset contains 22 pyridinyl-imidazoles (PI), 190 triarylpyrroles (TA) and 44 diaryl-urea derivatives (DU) (Fig. 10c). Only the diaryl-urea derivatives are type-II kinase inhibitors which are able bind to the DFG-out conformation of p38 MAP kinase. After excluding the two other structural classes, PI and TA, from the set of active compounds and rerunning the virtual screening for p38 MAP kinase, we observed a tremendous increase of the AUC value from 0.51 to 0.92 (Fig. 10d).

2. *Thymidine Kinase (TK) (1kim) Actives and decoys topological highly similar—true decoys?* TK is known to be a difficult virtual screening target. The challenges are receptor flexibility, water-bridged interactions and the low affinity of most known ligands [45]. The enrichment we achieve for this target is only moderate with an AUC value of 0.71 and primarily no good true positive rates (Fig. 11d). The analysis of the highest ranked compounds shows that the binding modes for the actives as well as the false positives are reasonable, and that the chemical and topological differences between these are only marginal. This is also illustrated in Fig. 11, which presents the native ligand of thymidine kinase (Fig. 11a–c), two of the active (Fig. 11e, g) and two of the decoy compounds (Fig. 11f, h). In fact, one of these decoys (Fig. 11h) emerged to be an agonist of Thymidine Kinase 2 [46]. According to this dataset it is unfeasible to expect good true positive rates.
3. *Catechol O-Methyltransferase (COMT) (1h1d) substrate-specific recognition group is contained in 5% of the decoys*: COMT belongs to the class of metalloenzymes. The active site is a small and shallow cavity which is totally exposed; only the catechol head group of the ligand that interacts with a magnesium ion is buried inside the cavity. The binding mode of the native ligand is influenced by crystal packing (Fig. 12c) and almost the whole affinity is induced by the catechol interaction with the magnesium ion. The experimental binding affinity is about $\Delta G_{experimental} = -47$ kJ/mol, while we predict a HYDE score

Table 3 Results of virtual screening on the *standardized data*

Standardized data								
Target	Quality	Virtual screening performance				Null hypothesis performance		
		tAUC[a]	0.1%[b]	1%[b]	2%[b]	Non-target	ntAUC[c]	tAUC- ntAUC[d]
ada	Q 1	0.59	0.00	4.35	4.35	ace	0.60	−0.01
ar	Q 1	0.66	0.00	16.22	21.62	rxr	0.41	0.26
cox1	Q 1	0.62	4.00	1.00	1.00	sahh	0.67	−0.05
er_agonist	Q 1	0.91	2.99	23.88	38.81	mr	0.72	0.20
er_antagonist	Q 1	0.94	3.58	23.08	41.03	ppar	0.71	0.23
fxa	Q 1	0.92	2.82	21.83	33.80	thrombin	0.83	0.09
gpb	Q 1	0.77	0.00	5.77	9.62	hivrt	0.39	0.38
gr	Q 1	0.50	0.00	0.00	2.56	pr	0.51	−0.02
hivpr	Q 1	0.79	5.66	24.53	26.42	trypsin	0.49	0.30
hivrt	Q 1	0.60	10.00	12.88	17.50	gpb	0.46	0.14
hmga	Q 1	0.88	37.14	40.00	42.86	ache	0.62	0.26
hsp90	Q 1	0.74	0.00	4.17	8.33	egfr	0.63	0.12
mr	Q 1	0.77	40.00	53.33	53.33	er_agonist	0.66	0.11
na	Q 1	0.95	12.99	40.12	59.18	cox2	0.38	0.57
p38	Q 1	0.51	3.12	6.25	7.81	vegfr2	0.65	−0.14
parp	Q 1	0.94	3.03	31.21	51.52	inha	0.37	0.57
ppar	Q 1	0.92	7.41	45.80	60.49	er_antagonist	0.72	0.20
pr	Q 1	0.52	0.00	3.70	3.70	gr	0.63	−0.10
rxr	Q 1	0.94	0.00	50.00	55.80	ar	0.50	0.43
sahh	Q 1	0.93	6.06	54.55	57.58	cox1	0.80	0.13
ace	Q 2	0.69	18.37	26.53	30.61	ada	0.33	0.37
ache	Q 2	0.50	0.00	0.00	0.30	hmga	0.32	0.19
alr2	Q 2	0.52	3.85	19.23	19.23	ampc	0.53	0.00
pde5	Q 2	0.64	1.96	9.80	15.69	comt	0.57	0.07
ampc	Q 3	0.84	0.00	19.05	23.81	alr2	0.67	0.16
cdk2	Q 3	0.67	8.00	1.00	1.00	pdgfrb	0.38	0.29
cox2	Q 3	0.81	0.99	11.47	18.82	na	0.85	−0.03
dhfr	Q 3	0.85	0.50	7.46	13.43	gart	0.34	0.50
gart	Q 3	0.94	0.00	14.29	19.05	dhfr	0.73	0.21
pnp	Q 3	0.53	8.00	8.00	8.00	tk	0.50	0.04
src	Q 3	0.79	3.87	18.06	23.87	fgfr1	0.57	0.22
tk	Q 3	0.71	0.00	0.00	0.00	pnp	0.55	0.16
trypsin	Q 3	0.63	0.00	0.00	2.27	hivpr	0.79	−0.16
comt	Q 4	0.89	0.00	0.00	0.00	pde5	0.44	0.45
egfr	Q 4	0.65	1.80	6.98	10.95	hsp90	0.56	0.09
fgfr1	Q 4	0.51	0.17	2.54	4.24	src	0.78	−0.27
inha	Q 4	0.60	5.88	9.41	13.06	parp	0.65	−0.05
pdgfrb	Q 4	0.26	0.64	1.91	3.18	cdk2	0.40	−0.14
thrombin	Q 4	0.81	0.00	6.15	10.77	fxa	0.83	-0.03
vegfr2	Q 4	0.70	0.00	4.05	8.43	p38	0.35	0.34
		Overall performance						
Mean		0.72	4.82	15.72	20.60		0.57	0.15
STD		0.17	8.84	15.99	19.07		0.16	0.20
Median		0.73	1.88	9.61	14.56		0.57	0.15
Min		0.26	0.00	0.00	0.00		0.32	−0.27
Max		0.95	40.00	54.55	60.49		0.85	0.57

Table 3 continued

Standardized data								
Target	Quality	Virtual screening performance				Null hypothesis performance		
		tAUC[a]	0.1%[b]	1%[b]	2%[b]	Non-target	ntAUC[c]	tAUC- ntAUC[d]
		Only considering Q1 targets						
Mean		0.77	6.94	23.13	29.87		0.59	0.18
STD		0.17	11.41	18.68	21.95		0.14	0.20
Median		0.78	3.08	22.46	30.11		0.62	0.17
Min		0.50	0.00	0.00	1.00		0.33	−0.14
Max		0.95	40.00	54.55	60.49		0.83	0.57

[a] AUC of the target with its own ligand/decoy set. For the purpose of comparison the random value is 0.5

[b] True positive rates when 0.1, 1 and 2% of the decoys have been found. For the purpose of comparison the random values are 0.001, 0.01 and 0.02 respectively

[c] AUC of the corresponding non-target (wrong target) with the ligand/decoy set of the target

[d] Difference of the target AUC and the non-target AUC value

Table 4 Overall performance of virtual screening on the *original data*

	Original data Overall performance					
	tAUC[a]	0.1%[b]	1%[b]	2%[b]	ntAUC[c]	tAUC-ntAUC[d]
Mean	0.71	3.69	14.34	20.59	0.58	0.13
STD	0.18	6.21	15.23	18.80	0.16	0.18
Median	0.70	0.65	8.06	14.15	0.58	0.11
Min	0.25	0.00	0.00	0.00	0.20	0.28
Max	0.95	28.57	53.06	63.27	0.89	0.54
	Only considering Q1 targets					
Mean	0.75	5.14	21.46	29.74	0.59	0.16
STD	0.19	7.65	17.92	21.78	0.17	0.19
Median	0.79	2.43	18.25	30.94	0.58	0.15
Min	0.38	0.00	0.00	3.70	0.20	0.14
Max	0.94	28.57	53.06	63.27	0.89	0.54

[a] AUC of the target with its own ligand/decoy set. For the purpose of comparison the random value is 0.5

[b] True positive rates when 0.1, 1 and 2% of the decoys have been found. For the purpose of comparison the random values are 0.001, 0.01 and 0.02 respectively

[c] AUC of the corresponding non-target (wrong target) with the ligand/decoy set of the target

[d] Difference of the target AUC and the non-target AUC value

of $\Delta G_{HYDE} = -54$ kJ/mol where the catechol interaction itself accounts for $\Delta G_{HYDE} = -42$ kJ/mol (Fig. 12a, b). We achieve a good AUC value of 0.89 for COMT though the true positive rates are low (Fig. 12d). The reason for this is the composition of the decoy set in the DUD. We found 5% of the decoys containing a catechol group and nearly half of which were sterically able to bind to the magnesium ion (for example, see Fig. 12e). This is illustrated in Fig. 12f, g, where we show one of the active compounds ($\Delta G_{HYDE} = -46$ kJ/mol) and a decoy structure ($\Delta G_{HYDE} = -54$ kJ/mol) which both coordinate the magnesium ion. Since the decoys were not experimentally validated, compounds containing catechol should not be included in the decoy set of catechol O-methyltransferase.

Examples for successful screens on three Q1 targets

1. *Estrogen Receptor with agonists (ER agonist)*: For the relatively hydrophobic estrogen receptor, we achieve an AUC value of 0.91 and very high true positive rates TP(0.1%) = 2.99, TP(1%) = 23.88 and TP(2%)

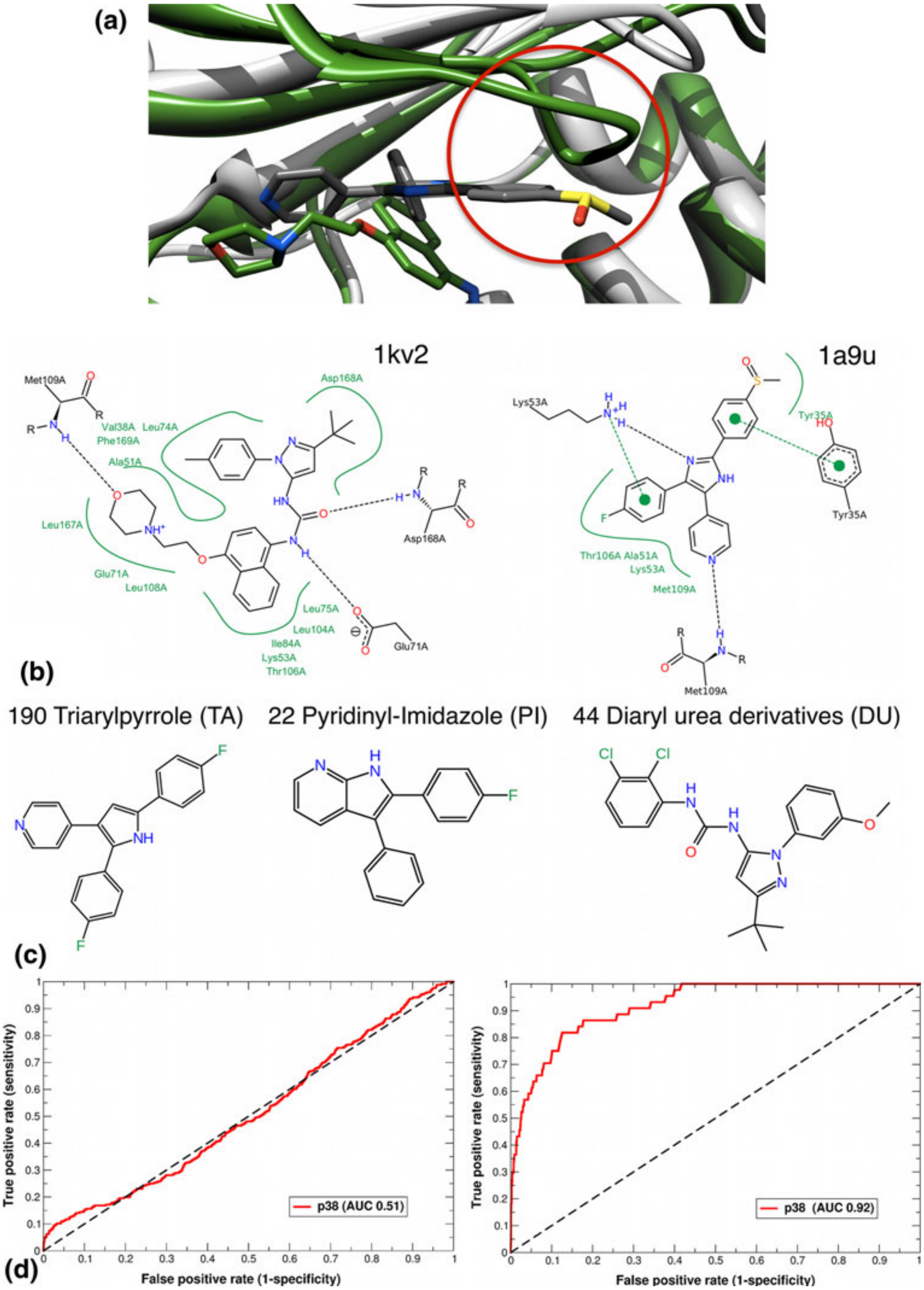

Fig. 10 Virtual screening results and analysis of target p38. **a** Superposition of two p38 MAP kinase structures. *Green*: 1kv2 in DFG-out conformation, *grey*: 1a9u in DFG-in conformation. Ligand of 1a9u interferes with the DFG loop of 1kv2. **b** Different binding modes of the ligands in 1kv2 and in 1a9u [44]. **c** Examples of the three different structural classes found in the active compounds: zinc00006854, zinc03832150 and zinc03833973. **d** ROC plots of p38 MAP kinase. *Left*: ROC plot for p38 Map kinase including all three structural classes of actives; the first two classes interfere with the DFG loop of the crystal structure. *Right*: ROC plot including only the "true" actives (DU compounds)

= 38.81 for screening the agonist actives and decoys. The binding mode of the best scoring compound (Diethylstilbestrol) is illustrated in Fig. 13a. The predicted HYDE score of $\Delta G_{HYDE} = -60$ kJ/mol is in good agreement with the experimentally measured affinity ($\Delta G_{experimental} = -53$ to -56 kJ/mol) found in a similar complex (3erd) of the PDB.

2. *HMG-CoA Reductase (HMGA)*: For HMG-CoA reductase with a very large binding pocket and relatively flexible ligands, our AUC value of 0.88 is marginally lower than that of the estrogen receptor but the true positive rates TP(0.1%) = 37.14, TP(1%) = 40.0 and TP(2%) = 42.86 are very high. Amongst the highest twelve ranked compounds, we find only actives, including well-known inhibitors like Fluvastatin or Simvastatin, which are nanomolar binders. The HYDE scores of these range from −69 to −49 kJ/mol corresponding to nanomolar binding affinity. In Fig. 13b the binding mode of the best scoring compound is shown, which is analog to the crystal ligand (Compactin,

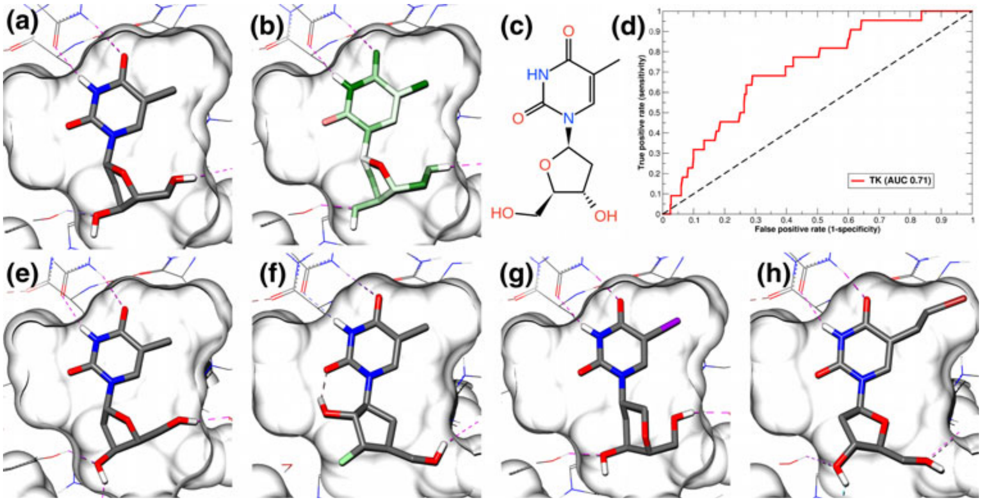
(a)
(b)
(c)
O
HN
O
N
O
HO
OH
(d)
True positive rate (sensitivity)
False positive rate (1-specificity)
TK (AUC 0.71)
(e)
(f)
(g)
(h)

Fig. 11 Virtual screening results and analysis of actives and decoys for target TK. At the bottom, two actives and two decoys are illustrated. The similarity is obvious and many decoys contain the thymin-like group which is important for recognition. **a** Crystal ligand (thymidine) in atom color (PDB: 1kim). **b** Crystal ligand in HYDE coloring scheme (($\Delta G_{experimental} = -30$ kJ/mol, $\Delta G_{HYDE} = -31$ kJ/mol). **c** Thymidine. **d** ROC plot of TK. **e** Best pose of active compound zinc00025672. **f** Best pose of decoy compound zinc00010140. **g** Best pose of active compound zinc03805724. **h** Best pose of decoy compound zinc00001043—this emerged to be an agonist of Thymidine Kinase 2

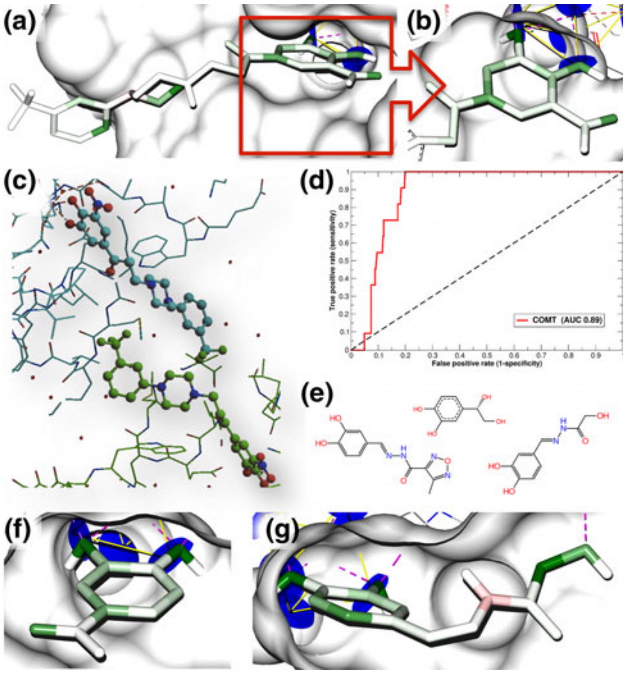
(a)
(b)
(c)
(d)
True positive rate (sensitivity)
False positive rate (1-specificity)
COMT (AUC 0.89)
(e)
(f)
(g)

Fig. 12 Virtual screening results and analysis of target COMT. **a** Crystal ligand (PDB: 1h1d) in HYDE coloring scheme ($\Delta G_{HYDE} = -54$ kJ/mol). **b** Interaction of the catechol group of the crystal ligand with the magnesium ion in detail ($\Delta G_{HYDE} = -42$ kJ/mol). **c** Crystal packing in the structure of 1h1d. The ligand binding mode is influenced by crystal packing with a symmetry related molecule. **d** ROC plot of COMT. **e** Examples of decoy molecules. **f** Pose of an active compound (zinc03814484) bound to COMT in HYDE coloring scheme ($\Delta G_{HYDE} = -46$ kJ/mol). **g** Pose of a decoy compound (zinc00366295) bound to COMT in HYDE coloring scheme ($\Delta G_{HYDE} = -54$ kJ/mol)

$\Delta G_{experimental} = -44$ kJ/mol). In the ROC plot, a plateau can be observed after about 60% of the actives have been identified. After investigating the active compounds, it emerged that all but two of 35 compounds share a HMG-like moiety but that 18 were presented in the inactive lactone form. This inactive form is not able to bind to the HMG binding pocket: In vivo the inactive form is enzymatically hydrolyzed to the active hydroxyl-acid form [47]. Hence, these inactive forms should be excluded from the active dataset as they are already present in their active form.

3. *Neuraminidase (NA)*: For the very hydrophilic neuraminidase we get a nearly perfect overall virtual screening performance with an AUC value of 0.95. Figure 13c again shows the best scoring compound in HYDE coloring and the crystal ligand (Zanamivir) in atom color. Here, we achieve a good early enrichment of actives, considering the first ten ranks we find seven actives and only three decoys. The HYDE score of these is in the range of −51 to −44 kJ/mol, this is in good agreement with experimental affinity of well-known binders like Zanamivir ($K_i = 0.1$ nM, $\Delta G_{experimental} = -55$ kJ/mol) or Oseltamivir (IC50 = 1–7 nM, $\Delta G_{experimental} = -49$ kJ/mol).

Null hypothesis performance

An additional validation concerning virtual screening performance is the null hypothesis test, which we have described in the validation part of the Methods section. For

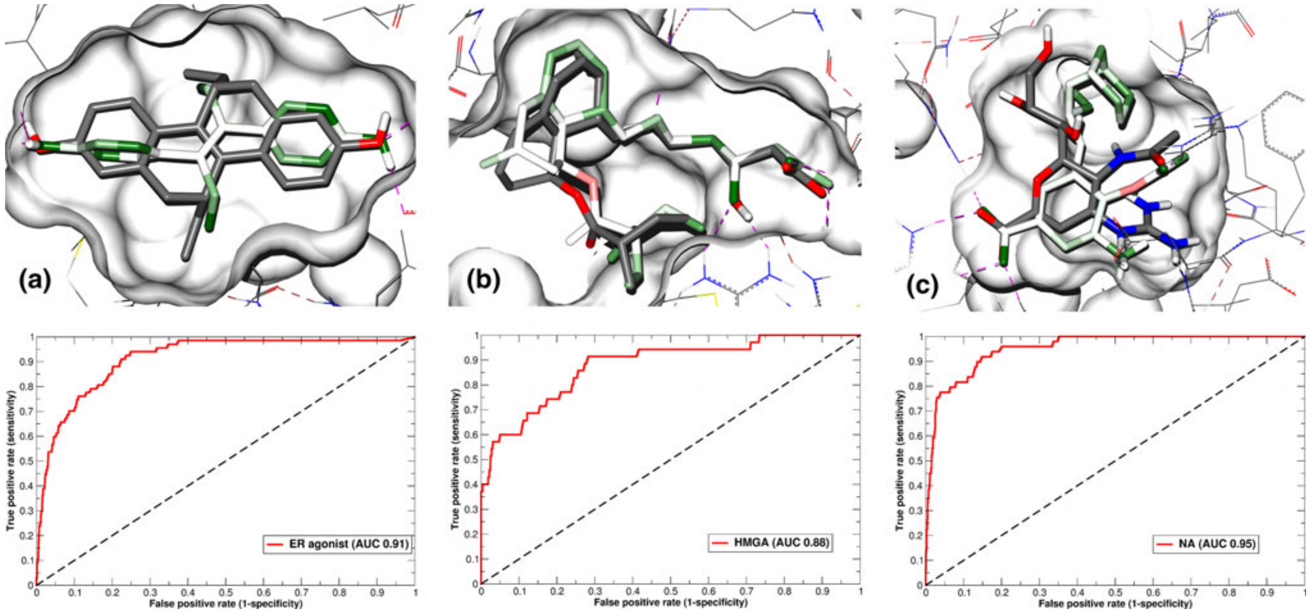

Fig. 13 Virtual screening results for estrogen receptor agonist (ER agonist), HMG-CoA reductase (HMGA) and neuramindase (NA). At the top, the best scoring compound (active) of each target is shown in the HYDE coloring scheme. The crystal ligands are represented in *atom colors*. At the bottom the ROC plots were shown. **a** Best scoring compound (active) ER agonist: zinc03861549 ($\Delta G_{HYDE} = -60$ kJ/mol) **b** Best scoring compound (active) HGMA: zinc04134476 ($\Delta G_{HYDE} = -69$ kJ/mol). **c** Best scoring compound (active) NA: zinc04134491 ($\Delta G_{HYDE} = -51$ kJ/mol)

this evaluation, we received pairings of the DUD targets from the organizers of the Symposium. The results of this test can be found in columns 7 and 9 of Table 3.

The DUD pairings partly comprised very closely related targets (see column 1 and column 8 of Table 3). For example, factor Xa is paired with thrombin. The target AUC value of factor Xa is 0.92—we achieve a very good enrichment for this target. Exchanging the protein structure of factor Xa with the structure of thrombin also results in a very high non-target AUC value of 0.83. As mentioned before, the goal of this analysis is to measure the true signal of the method. Here, it first seems that our performance is quite poor in the null hypothesis test. But if we consider again the factor Xa thrombin pairing, we can find an explanation for this result. Formerly, when the crystallization of factor Xa was not possible, medicinal chemists often used thrombin as a surrogate protein for the design of new factor Xa inhibitors [48]. Consequently, it is not so surprising that the virtual screening performance of our method is independent of whether factor Xa or thrombin is used as the target structure. The same is true for kinase structures. Here, it was also common to use different kinase structures for the development of inhibitors when the original kinase structure could not be obtained by crystallization [49, 50].

Another problem of null hypothesis testing is the pairing of different nuclear receptors with each other. These targets all contain a very large and hydrophobic binding pocket with few specific interaction sites. Often, the active compounds of these nuclear receptors are also very similar and differ only in size and hence in the number of hydrophobic contacts they may establish. Therefore, docking smaller hydrophobic compounds into large hydrophobic cavities will always produce reasonably scored poses and so can result in high non-target AUC values.

Screening experiments with WOMBAT ligands

In addition to virtual screening using the DUD ligands we carried out a second virtual screening experiment using the WOMBAT dataset of ligands—also provided by the organizers of the Symposium. This dataset includes ligands for eleven of the DUD targets which are designed to represent a broader range of chemical diversity for these targets than the DUD ligands themselves. The motivation of the authors was to avoid the problem of potential analog bias found with DUD.

In Table 5, the results for virtual screening taking the WOMBAT ligands as the actives instead of the DUD ligands are shown on the *standardized data*. Results on the *original data* can be found in the Supplementary Material (Table S11). The supplied WOMBAT dataset contains enumerated stereoisomers for some of the compounds, however it is not clear which of the stereoisomers is the active one. The original WOMBAT data does not include stereo-chemistry information for the compounds. One possibility would be to include the rankings of all stereoisomers in the screening statistics. This could lead to bias in the results if one active compound is ranked highly/lowly several times due to the different isomers. Therefore, to avoid this bias the results presented here only include the best scoring stereoisomer for each compound. Results including all stereoisomers can be found in the Supplementary Material (Table S12 and S13).

We see a sizable decrease in the mean AUC across nearly all targets. In fact, we consider such a simple comparison almost meaningless for the following reasons. It is important to note that only the active compounds of the targets are provided in the WOMBAT dataset, while the same DUD decoy sets had to be used for the experiments. The decoy sets in DUD were designed by deriving decoy compounds from the active compounds. Introducing new chemotypes into the active set will lead to a bias if the decoy set is not updated simultaneously. In addition, the challenge arising with the inclusion of more chemotypes is obvious—very diverse scaffolds require a flexible protein structure to achieve a reasonable binding mode. In our virtual screening workflow, the heavy atom protein structure is rigid. Consequently, enrichment studies using active compounds with very diverse scaffolds will result in poorer enrichment rates as some of the actives will be missed. Also, with the WOMBAT dataset, the problems with the p38 MAP kinase target which undergoes a large conformational change upon binding of different compound classes (as discussed in the Result section of DUD) become more pronounced. This is because from the set of WOMBAT actives for p38, only seven compounds are able to bind to the DFG-out conformation of the protein structure present in this target.

Table 5 Results of virtual screening using the WOMBAT ligands (actives) compared to using the DUD actives (last column)

Target	Quality	Standardized data				DUD actives
		AUC	0.1%[a]	1%[a]	2%[a]	AUC
ar	Q 1	0.56	0.00	0.00	0.00	0.66
cox2	Q 1	0.67	0.00	1.32	2.63	0.81
er_antagonist	Q 1	0.64	0.00	4.69	10.86	0.94
fxa	Q 1	0.85	1.87	8.13	15.71	0.92
hivrt	Q 1	0.51	3.26	6.06	9.09	0.60
p38	Q 1	0.47	0.42	5.08	6.78	0.51
ppar	Q 1	0.69	0.00	0.00	0.00	0.92
alr2	Q 2	0.55	0.00	2.38	2.38	0.52
pde5	Q 2	0.52	0.81	5.68	6.03	0.64
cdk2	Q 3	0.68	1.32	11.84	16.45	0.67
egfr	Q 4	0.66	0.00	4.05	9.46	0.65
		Overall performance				
Mean		0.62	0.70	4.48	7.22	0.71
STD		0.11	1.06	3.56	5.74	0.16
Median		0.64	0.00	4.69	6.78	0.66
Min		0.47	0.00	0.00	0.00	0.51
Max		0.85	3.26	11.84	16.45	0.94

DUD decoys were used in both cases. Only the best scoring stereoisomer is considered

[a] True positive rates when 0.1, 1 and 2% of the decoys have been found. For the purpose of comparison the random values are 0.001, 0.01 and 0.02 respectively

In the following we analyze the two targets for which the screening results with DUD and with WOMBAT active ligands differ the most (Figs. 14c, d and 15c, d).

1. *Peroxysome Proliferator-activated Receptor gamma (PPARg)*: This target is noted for two characteristics: It is very flexible and ligands bind very unspecifically in the large hydrophobic pocket. A large part of the binding affinity is induced by the hydrophobic effect caused by the burial of large hydrophobic ligands in the hydrophobic binding site. The only specific interaction is formed by a glitazone or a carboxylate group on the ligand with the catalytic residues—two histidines, a serine and a tyrosine (OG SER 289, NE2 HIS 323, OH TYR 473, NE2 HIS 449)—deep in the binding pocket. The binding site is able to almost double its volume by conformational change of amino acid side chains near the catalytic center (e.g. compare 1fm6 to 1fm9). The crystal structure found in DUD (1fm9) is presented in this "open" conformation. Here, the geometrical arrangement of the catalytic residues is slightly different than in the "closed" conformation (1fm6). Two classes of actives bind to the different conformations: Compounds containing a branched carboxylate bind to the open conformation, while compounds containing a glitazone group prefer the closed conformation. The WOMBAT ligands comprise more glitazones than the DUD ligands, and only about five branched carboxylate compounds can be found in the WOMBAT set. On the other hand, the DUD actives consist mostly of branched carboxylate compounds. The branched carboxylate compounds are also larger, meaning the derived DUD decoys have a higher molecular weight (Fig. 14b), while overall the WOMBAT ligands are smaller. The distribution of the molecular weight of the WOMBAT and the DUD ligands is illustrated in Fig. 14a. As mentioned above, the binding site of this target is rather unspecific. This means the larger hydrophobic decoys found in DUD can bind in the active site with a better HYDE score than the smaller WOMBAT actives. This theory is

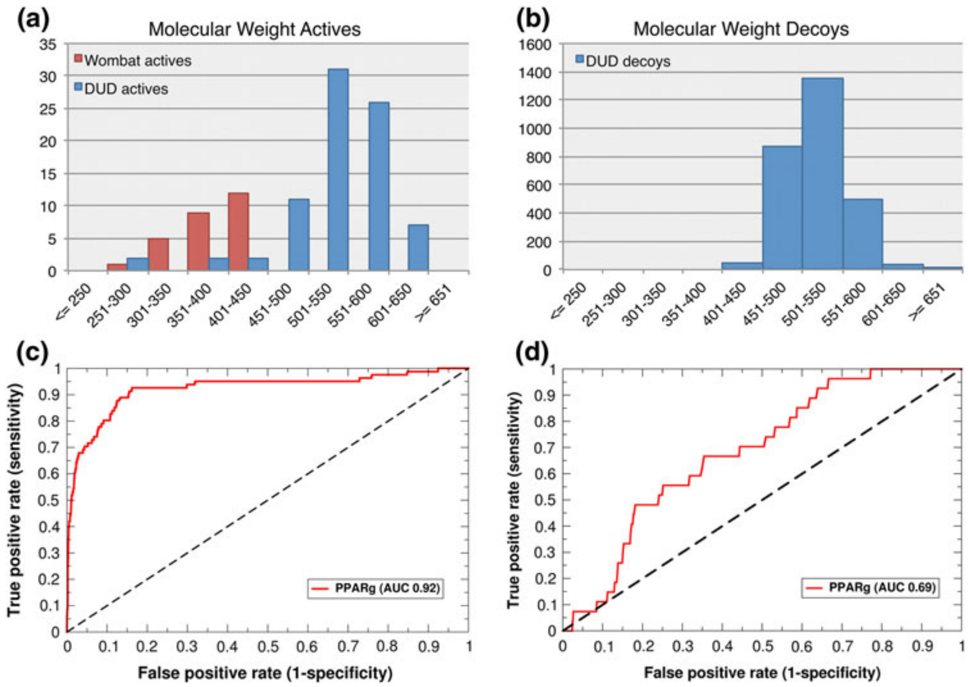

Fig. 14 Comparison of the molecular weight of the WOMBAT ligands with the DUD actives of PPARg and the corresponding ROC plots. **a** Histogram of the molecular weight of the WOMBAT ligands and the DUD actives. **b** Histogram of the molecular weight of the DUD decoys. **c** ROC plot of virtual screening using the DUD actives and decoys. **d** ROC plot of virtual screening using the WOMBAT ligands and the DUD decoys

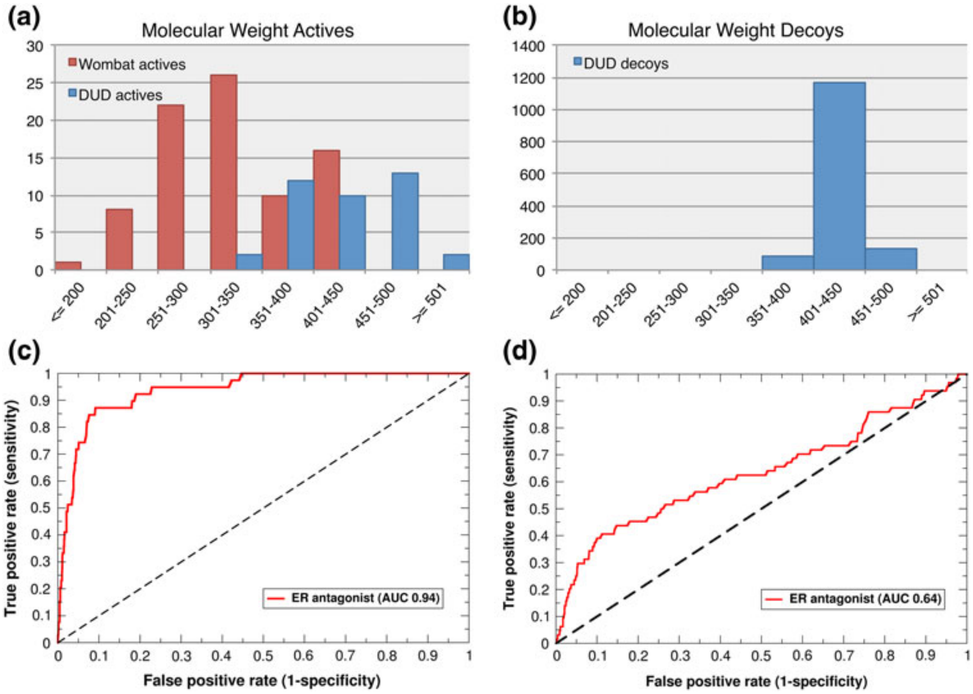

Fig. 15 Comparison of the molecular weight of the WOMBAT ligands with the DUD actives of ER antagonist and the corresponding ROC plots. **a** Histogram of the molecular weight of the WOMBAT ligands and the DUD actives. **b** Histogram of the molecular weight of the DUD decoys. **c** ROC plot of virtual screening using the DUD actives and decoys. **d** ROC plot of virtual screening using the WOMBAT ligands and the DUD decoys

confirmed in practice, as all of the high scoring decoys contain the carboxylate group necessary for the specific interaction in the catalytic center of the target. For these reasons it is not surprising that the enrichment performance is poorer for this target using the WOMBAT ligands. Furthermore the decoys for PPARg were not experimentally validated, therefore it would not be unusual if they would turn out to be binders in this unspecific binding site.

2. *Estrogen receptor (antagonist) (ER antagonist)*: In the WOMBAT dataset, agonists and antagonists were not separated for this target. However, the authors of WOMBAT pointed out they only want to keep molecules designated as antagonists [18]. For this reason we have chosen the larger cavity of the ER antagonist from DUD as the target structure. ER is, like PPARg, a flexible target with a large hydrophobic pocket. Two specific interaction sites can be found

deeply buried in the binding pocket. In this case, the WOMBAT dataset contains more ligands (64) than the DUD dataset (39). Again, we have a disparity in the distribution of molecular weight in the ligand sets—Fig. 15a shows a comparison of the molecular weight distributions. The weight of the WOMBAT ligands is more scattered, ranging from very small compounds (one below 200 Da, possibly agonists) to average sized compounds (400–450 Da) with an average of weight 326 Da. In comparison, the DUD ligands tend to be larger—and hence also the DUD decoys (Fig. 15b), which have a mean molecular weight of 425 Da. So we find here the same bias as already discussed for PPARg. Moreover, half of the WOMBAT ligands are only weak binders with a micromolar experimental binding affinity. The identification of these is known to be a challenge in virtual screening experiments.

Considering the bias in molecular weight between the WOMBAT ligands and the DUD decoys we decided to normalize the HYDE score by the number of heavy atoms of the ligand and rerank the actives and decoys. This result is shown in Table 6 (corresponding results using the *original data* can be found in the Supplementary Material in Table S14). For nearly all targets we now achieve better AUC values, with a remarkable increase of the mean AUC value from 0.62 to 0.71. The two targets discussed above—the estrogen receptor and the Peroxysome Proliferator-activated Receptor gamma—achieve very good AUC values of 0.93 and 0.92, respectively. This shows that especially for the hydrophobic targets it is necessary to have similarly sized actives and decoys to allow a conclusion about the virtual screening performance of a method, since the most affinity is achieved for these targets by unspecific hydrophobic contacts.

Conclusion

In this paper we described our recently developed scoring function HYDE and carried out a large-scale validation of this scoring function in redocking and virtual screening experiments.

HYDE proved to be an effective tool for application in virtual screening, and most advantageous for the detailed analysis of single protein–ligand complexes. When the 32 best scored poses were considered, HYDE was able to find the bioactive conformation in 93% of cases. HYDE scored the bioactive conformation of a ligand in 75% of cases at the top rank. In virtual screening, we achieved a mean AUC value of 0.72 considering all DUD targets. If we disregarded those targets with structural deficiencies, the mean AUC value increased to 0.77.

Table 6 Results for virtual screening with normalized HYDE score using the WOMBAT ligands (actives) compared to using the DUD actives (last column)

Target	Quality	Standardized data				DUD actives
		AUC	0.1%[a]	1%[a]	2%[a]	AUC
ar	Q 1	0.61	2.78	6.38	12.82	0.66
cox2	Q 1	0.71	2.74	6.45	9.85	0.81
er_antagonist	Q 1	0.93	18.89	40.59	53.66	0.94
fxa	Q 1	0.90	9.00	25.55	36.06	0.92
hivrt	Q 1	0.51	0.00	0.00	0.96	0.60
p38	Q 1	0.54	0.00	0.00	1.35	0.51
ppar	Q 1	0.92	8.16	36.03	48.15	0.92
alr2	Q 2	0.51	0.00	4.76	4.76	0.52
pde5	Q 2	0.74	5.12	16.05	22.29	0.64
cdk2	Q 3	0.73	3.80	12.52	15.95	0.67
egfr	Q 4	0.67	0.00	3.23	5.83	0.65
		Overall performance				
Mean		0.71	4.59	13.78	19.24	0.71
STD		0.16	5.73	14.28	18.70	0.16
Median		0.71	2.78	6.45	12.82	0.66
Min		0.51	0.00	0.00	0.96	0.51
Max		0.93	18.89	40.59	53.66	0.94

DUD decoys were used in both cases. Only the best scoring stereoisomer is considered

[a] True positive rates when 0.1, 1 and 2% of the decoys have been found. For the purpose of comparison the random values are 0.001, 0.01 and 0.02 respectively

A clear advantage of using the HYDE scoring function is the very simple and above all intuitive interpretation of the results. The HYDE score directly reflects the free binding energy. It is also an atom-based score which means the user can detect which parts of the ligand and protein contribute favorably to the binding energy or which parts are detrimental to binding. By applying a red/green color scheme to the atom scores, this information can be visualized directly, as we have shown here in many detailed examples.

The main intention of the organizers of the Docking and Scoring Symposium was to allow participants to demonstrate the performance of current docking methods on publicly available datasets. Despite the fact that these datasets were revised and refined by the organizers in advance, many problems still remain to be fixed. We have shown on multiple examples how pitfalls in the data lead to problems when using public data blindly for the validation of computational tools. A lot of work and thoughtful research is necessary to prepare a dataset for validation purposes. The argument that in a real-life scenario errors in the data will also occur is correct, but the goal of developers of new methods is to validate their new method and not the dataset they use for validation. Having a high quality reliable dataset would be a huge benefit in improving existing methods, as well as for developing new refined methods. For the users of these methods, it would be also an advantage—they can choose which method will perform best for their application instead of having to choose which method will struggle best to overcome problems in the data.

During the evaluation we detected many problems within the different datasets, from poor electron density and missing important structural waters through to wrongly annotated decoys and the necessity to consider the protein conformation with respect to classes of active compounds. Most frustration arises when examining the work of others and realizing they have already detected the same problems in the data. It proved to a long, meticulous and painstaking procedure to accomplish this detailed analysis of the data and results, and therefore it would be appropriate to include our findings—and certainly those of others from the Symposium—in a new release of these datasets.

Experimental details

The HYDE tool

The stand-alone HYDE tool was implemented for this project to include the version of the HYDE scoring function available in the LeadIT release version 2.0.1. The HYDE tool can read LeadIT project files and ligand files and has the ProToss functionality built in along with a numerical optimizer. The HYDE scoring algorithm uses an interaction model based on the FlexX interaction model similar to that found in LUDI. The main difference between the two interaction models is that in HYDE the interaction surfaces are often smaller (or even use multiple surfaces to describe available interacting regions rather than just the one used in FlexX). The reason is obvious: HYDE needs a much more accurate model for assessing good versus weak interactions, while FlexX requires a certain amount of flexibility for forming interactions during the docking process. Because the docking pose is placed using those more generously sized interaction surfaces, the final pose may not form interactions that correspond exactly to the surfaces used in HYDE. Therefore, we allow HYDE to optimize the pose in the active site before the final score is calculated.

Firstly, before the numerical optimization, HYDE can use ProToss to optimize the hydrogen bond network to the docking pose. We use this functionality when scoring screening poses, where ligands other than the native ligand are docked into the active site. Secondly, a numerical optimizer is applied to fine-tune the pose by optimizing the HYDE score while ensuring the ligand conformation and clash are not compromised. For more details about the numerical optimizer implementation, see the Supplementary Material.

The HYDE tool outputs detailed information about the score for the pose, along with the optimized pose conformation.

ProToss

ProToss is a method that can extremely rapidly resolve common crystallographic ambiguities in the protein, such as flipped amide groups and histidine residues, as well as determine hydrogen positions and protonation states. Hydrogen bond networks are evaluated in ProToss by applying an empirical scoring function to possible intra-protein and protein–ligand hydrogen bond geometries using a high-speed dynamic programming approach. The hydrogen bond geometries are described using the interaction surfaces found in FlexX or HYDE depending on the current application, i.e. active site preparation in LeadIT, or hydrogen bond network optimization for a pose in the HYDE tool. Note that ProToss in its current version does not change any of the heavy atom positions on the ligand but may change its hydrogen positions.

Workflow details

The redocking and screening calculations were run in a virtual screening environment, which controlled the workflow, distributed the calculations across a large computing cluster (consisting of 10 processors yielding 120 compute nodes via hyperthreading) and collected the data in the forms described in the Validation Measures in the Methods section.

The workflow was implemented as follows for the redocking calculations on the Astex dataset:

1. FlexX docking: Output the best 200 poses each from the standard placement algorithm and from the SIS algorithm according to the FlexX score
2. HYDE rescoring: Numerical optimization, output HYDE scores and optimized poses
3. Results: Calculate the heavy atom RMSD of the optimized poses to the given reference ligands, summarize the RMSD for the best scored pose and for the best 32 scored poses

The workflow for the screening calculations on the DUD and WOMBAT datasets was as follows:

1. FlexX docking: For all actives and decoys, output the best 20 poses each from the standard placement algorithm and from the SIS algorithm according to the FlexX score

2. HYDE rescoring: ProToss optimization of the hydrogen bond network followed by numerical optimization, output HYDE scores and optimized poses
3. Results: Calculate ROC plots, AUC values and true positive values

Acknowledgments We thank the organizers of the Docking and Scoring Symposium for the arrangement of the very interesting session at the 241st ACS National Meeting. The HYDE project was funded by Bayer CropScience AG and Bayer Pharma AG. We also thank the whole team at the BioSolveIT GmbH for carrying out the elaborate computations for this study.

References

1. Kubinyi H (2001) In: Testa B, van de Waterbeemd H, Folkers G, Guy R (eds) Pharmacokinetic optimization in drug research, 1st edn. Wiley-VCH, Weinheim, Germany
2. Bissantz C, Kuhn B, Stahl M (2010) A medicinal chemist's guide to molecular interactions. J Med Chem 53(14):5061–5084
3. Mobley DL, Dill KA (2009) Binding of small-molecule ligands to proteins: "What You See" is not always "What You Get". Structure 17:489–498
4. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 153:7–26
5. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol 295:337–356
6. Muegge I (2006) PMF scoring revisited. J Med Chem 49: 5895–5902
7. Böhm HJ (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. J Comput Aided Mol Design 8:243–256
8. 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
9. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA et al (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J Med Chem 49:6177–6196
10. Savage HJ, Elliott CJ, Freeman CM, Finney JL (1993) Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. J Chem Soc Faraday Trans 89:2609–2617
11. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793
12. Fleming PJ, Rose DG (2005) Do all backbone polar groups in proteins form hydrogen bonds? Protein Sci 14:1911–1917
13. Pearlman DA, Rao BG, Charifson P (2008) FURSMASA: a new approach to rapid scoring functions that uses a MD-averaged potential energy grid and a solvent- accessible surface area term with parameters GA fit to experimental data. Proteins 71: 1519–1538
14. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortonson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. J Med Chem 50(4):726–741
15. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
16. LeadIT (version 2.0.1) BioSolveIT GmbH, Sankt Augustin. http://www.biosolveit.de/leadit/. Accessed 25 Aug 2011
17. Huang N, Shoichet B, Irwin JJ (2006) Benchmarking set for molecular docking. J Med Chem 49(23):6789–6801
18. 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
19. Reulecke I, Lange G, Albrecht J, Klein R, Rarey M (2008) Towards an integrated description of hydrogen bonding and dehydration: reducing false positives in virtual screening using the HYDE scoring function. ChemMedChem 3(6):885–897
20. Hindle S, Rarey M, Buning C, Lengauer T (2002) Flexible docking under pharmacophore type constraints. J Comput Aided Mol Des 16(2):129–149
21. Lippert T, Rarey M (2009) Fast automated placement of polar hydrogen atoms in protein-ligand complexes. J Cheminformatics. doi:10.1186/1758-2946-1-13
22. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR. Hydrophobic, electronic, and steric constants. American Chemical Society, Washington, DC
23. Lange G, Klein R, Albrecht J, Rarey M, Reulecke I (2010) European patent specification EP2084520
24. Böhm HJ (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. J Comput Aided Mol Design 6:593–606
25. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203
26. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106(3):765–784
27. Fersht AR, Shi JP, Knill-Jones J, Lowe DM, Wilkinson AJ, Blow DM, Brick P, Carter P, Waye MMY, Winter G (1985) Hydrogen bonding and biological specificity analysed by protein engineering. Nature 314:235–238
28. McComas CC, Crowley BM, Boger DL (2003) Partitioning the loss in vancomycin binding affinity for D-Ala-D-Lac into lost H-bond and repulsive lone pair contributions. J Am Chem Soc 125:9314–9315
29. Foloppe N, Fisher LM, Howes R, Kierstan P, Potter A, Robertson AGS, Surgenor AE (2005) Structure-based design of novel Chk1 inhibitors:# insights into hydrogen bonding and protein−ligand affinity. J Med Chem 48(13):4332–4345
30. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285: 1735–1747
31. REDUCE (version 3.13) Richardson Laboratory, Duke University, North Carolina. http://kinemage.biochem.duke.edu/software/reduce.php. Accessed 25 Aug 2011
32. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D 66:213–221
33. SZYBKI, OpenEye Scientific Software, Santa Fe. http://www.eyesopen.com/szybki. Accessed 25 Aug 2011
34. MOE Molecular Operating Environment (database version 2009.10) Chemical Computing Group, Montreal. http://www.chemcomp.com/software-chem.htm. Accessed 25 Aug 2011
35. LigPrep (Maestro v9.1) Schrödinger, New York. http://www.schrodinger.com/products/14/10/. Accessed 25 Aug 2011

36. Seebeck B, Reulecke I, Kämper A, Rarey M (2008) Modeling of metal interaction geometries for protein-ligand docking. Proteins Struct Funct Bioinform 71:1237–1254
37. Urbaczek S, Kolodzik A, Fischer JR, Lippert T, Heuser S, Groth I, Schulz-Gasch T, Rarey M (2011) NAOMI—on the almost trivial task of reading molecules from different file formats. J Chem Inf Mod. doi:10.1021/ci200324e
38. Kubinyi H (2006) In: Ekins S (ed) Computer applications in pharmaceutical research and development, 1st edn. Wiley, New York
39. MOE Molecular Operating Environment (version 2010.10) Chemical Computing Group, Montreal. http://www.chemcomp.com/software-moe2010.htm. Accessed 25 Aug 2011
40. Yeh KC, Kwan KC (1978) A comparison of numerical integrating algorithms by trapezoidal, lagrange, and spline approximation. J Pharmacokinet Phar 6:79–98
41. 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
42. Sintchak MD, Fleming MA, Futer O, Raybuck SA, Chambers SP, Caron PR, Murcko MA, Wilson KP (1996) Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic acid. Cell 85:921–930
43. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem 47:550–557
44. Stierand K, Rarey M (2010) Drawing the PDB—protein-ligand complexes in two dimensions. J Med Chem Lett 1:540–545
45. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
46. Pérez-Pérez MJ, Priego EM, Hernández AI, Familiar O, Camarasa MJ, Negri A, Gago F, Balzarini J (2008) Structure, physiological role, and specific inhibitors of human thymidine kinase 2 (TK2): present and future. Med Res Rev 28:797–820
47. Corsini A, Maggi FM, Catapano AL (1995) Pharmacology of competitive inhibitors of HMG-CoA reductase. Pharmacol Res 31(1):9–27
48. Wiley MR, Weir LC, Briggs S, Bryan NA, Buben J, Campbell C, Chirgadze NY, Conrad RC, Craft TJ, Ficorilli JV, Franciskovich JB, Froelich LL, Gifford-Moore DS, Goodson T, Herron DK, Klimkowski VJ, Kurz KD, Kyle JA, Masters JJ, Ratz AM, Milot G, Shuman RT, Smith T, Smith GF, Tebbe AL, Tinsley JM, Towner RD, Wilson A, Yee YK (2000) Structure-based design of potent, amidine-derived inhibitors of factor Xa: evaluation of selectivity, anticoagulant activity, and antithrombotic activity. J Med Chem 43(5):883–889
49. Breault GA, Ellston RPA, Green S, James SR, Jewsbury PJ, Midgley CJ, Pauptit RA, Minshull CA, Tucker JA, Pease JE (2003) Cyclin-dependent kinase 4 inhibitors as a treatment for cancer. Part 2: identification and optimisation of substituted 2, 4-Bis Anilino Pyrimidines. Bioorg Med Chem Lett 13:2961–2969
50. Gassel M, Breitenlechner CB, König N, Huber R, Engh RA, Bossemeyer D (2004) The protein kinase C inhibitor Bisindolyl Maleimide 2 binds with reversed orientations to different conformations of protein kinase A. J Biol Chem 279:23679–23690