JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Facing the Challenges of Structure-Based Target Prediction by Inverse Virtual Screening

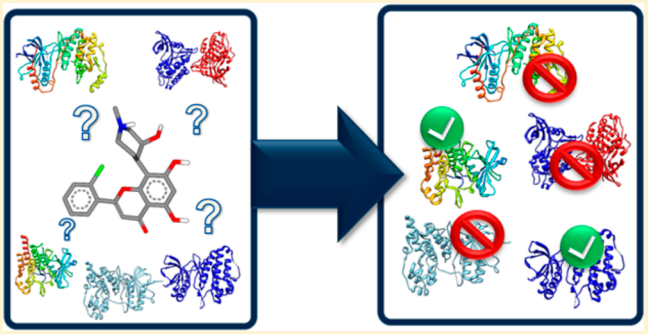
Karen T. Schomburg,[†] Stefan Bietz,[†] Hans Briem,[‡] Angela M. Henzler,[†] Sascha Urbaczek,[†] and Matthias Rarey*[,†]

[†]Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Global Drug Discovery, Medicinal Chemistry, Bayer Pharma AG, 13353 Berlin, Germany

Supporting Information

ABSTRACT: Computational target prediction for bioactive compounds is a promising field in assessing off-target effects. Structure-based methods not only predict off-targets, but, simultaneously, binding modes, which are essential for understanding the mode of action and rationally designing selective compounds. Here, we highlight the current open challenges of computational target prediction methods based on protein structures and show why inverse screening rather than sequential pairwise protein−ligand docking methods are needed. A new inverse screening method based on triangle descriptors is introduced: *i*RAISE (*inverse* Rapid Index-based Screening Engine). A Scoring Cascade considering the reference ligand as well as the ligand and active site coverage is applied to overcome interprotein scoring noise of common protein−ligand scoring functions. Furthermore, a statistical evaluation of a score cutoff for each individual protein pocket is used. The ranking and binding mode prediction capabilities are evaluated on different datasets and compared to inverse docking and pharmacophore-based methods. On the Astex Diverse Set, *i*RAISE ranks more than 35% of the targets to the first position and predicts more than 80% of the binding modes with a root-mean-square deviation (RMSD) accuracy of <2.0 Å. With a median computing time of 5 s per protein, large amounts of protein structures can be screened rapidly. On a test set with 7915 protein structures and 117 query ligands, *i*RAISE predicts the first true positive in a ranked list among the top eight ranks (median), i.e., among 0.28% of the targets.



INTRODUCTION

Controlling the protein selectivity of a lead compound in drug discovery is crucial for avoiding adverse effects and, thus, lowering the high attrition rates of drugs during the past decade.[1−3] For rational protein selectivity enhancement, the uttermost goal is the complete target profile for a compound on human targets. Furthermore, protein target predictions may reveal hidden opportunities in drug repurposing projects,[4−6] support the difficult but promising design process of multitarget drugs,[7−9] and reveal targets of drugs with so far unknown mechanisms-of-action (in the DrugBank,[10] more than 80 drug entities are registered with unknown mechanisms-of-action). Other scientific fields, such as biotechnology, are profiting from target prediction methods, e.g., for the design of in vitro synthetic reaction pathways.[11]

Strategies to predict targets for a small molecule are either computational or experimental.[12,13] So far, the use of experimental activity assays for a broad range of targets still dominates in drug development processes. However, computational methods can complement or reduce—and even substitute—some costly and time-consuming experimental methods. In contrast to high-throughput screening of thousands of molecules for one target, no such time- and cost-efficient experimental methods exist for screening thousands of proteins.

Depending on the available data, computational target prediction methods can be classified as ligand-based, network-based, side-effect-based, or protein-structure-based.

Ligand-based methods couple ligand similarity measurements with experimental data.[14−18] Network-based methods exploit available data on ligand and target interactions for compiling networks and deduce thereof new predictions.[19−22] Side-effect-based methods derive target predictions from phenotypic (adverse) effects of drugs.[23] Protein-structure-based methods use docking, pharmacophore searching, binding site comparison or protein−ligand interaction fingerprints to predict new targets.[24]

Ligand-based, network-based, and side-effect-based methods show good results, if the molecules with available data are similar enough to those for which predictions should be made, following the paradigm of "*If something has been observed, knowledge can be deduced for similar things*". However, these methods fail to predict effects that are outside the compound

Received: February 28, 2014

   dx.doi.org/10.1021/ci500130e | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

domain used to generate the respective model. Protein-structure-based methods are dependent on three-dimensional (3D) protein structures. Furthermore, pharmacophore searches, binding-site comparisons, and interaction fingerprints need at least one starting co-crystallized complex as input. Docking-based target prediction is the only method that is independent of such preliminary information, needing only the 3D protein structure and the active site location, e.g., identified by any co-crystallized ligand or a pocket identification algorithm. The amount of available 3D structures of proteins grows rapidly, promising increasing importance for this method in future.

In the following, we will focus on docking-based target prediction methods. These approaches have one further major advantage: simultaneously with predicting a target, the binding mode of a ligand to a protein is predicted. However, compared to classic protein−ligand docking, the reverse setup has different requirements. Four main challenges must be addressed in the development of structure-based target prediction methods:

(1) Preprocessing and handling many protein structures: In classical screening, a single protein is used and the active site preparation is rather complex and time-consuming. For inverse screening, the method must be able to deal with at least 10^4 structures, calling for completely automated time-efficient processes.
(2) Efficient and consistent handling of structural data: Protein structure data is storage-demanding defining a need for new approaches to handle large amounts of protein structures consistently and efficiently.
(3) Ranking of targets: As has been stated and observed previously,[25] scoring functions that were developed for assessing protein−ligand complexes in classic docking are problematic when applied to intertarget ranking. Measures accounting for the diversity of protein pockets concerning shape and properties[26] must be included.
(4) Significant evaluation methods: Prospective evaluation is expensive and not feasible for intermethod comparison. Therefore, reliable datasets for retrospective studies are needed. For inverse docking/screening, no standard evaluation datasets exist yet on which new methods can be evaluated and compared among each other, such as, e.g., the DUD[27] for classic docking and virtual screening. The main problem is the categorization of targets as true negatives for small molecules. Unfortunately, literature data rarely reveal negative results, i.e., if a molecule does not interact with a protein. In summary, a dataset is needed that contains a sufficient number of molecules and proteins with a reliable assignment of targets and nontargets.

So far, the available docking-based target prediction methods only barely account for the challenging requirements of the reverse scenario.

Invdock, which is the first published docking-based target prediction method, uses the DOCK docking algorithm.[28,29] A threshold score is applied to avoid the high computing time of classic docking approaches for the reverse setup. Once any pose in a cavity is found with a score better than the threshold, the exhaustive search of the best pose is aborted. For evaluation, a redocking study on nine proteins,[28] as well as a screening of eight compounds against the TTD (therapeutic target database[30]) of, at that time, 1040 structures of 38 proteins related to side effects were conducted. Of the 43 experimentally documented protein−ligand interactions, Invdock finds 38.[29]

TarFisDock, which is another inverse docking approach,was published in form of a web-service of inverse docking based on DOCK against the PDTD (Potential Drug Target Database[31]).[32] For evaluation, Li and co-workers screened the PDTD of, at that time, 698 structures of 371 targets with two compounds. For vitamin E, 50% of reported targets were ranked among the first 10% of the targets and for 4H-tamoxifen among the first 5%.

Another inverse screening application utilizes DOCK for building a chemical-protein interactome for deriving relevant genes of adverse drug reactions, in particular for the identification of risk-alleles.[33,34]

The sc-PDB is a subset of the protein−ligand complexes contained in the Protein Data Bank[35] relevant to drug design.[36,37] Paul et al. inversely screened an early version of the database with 2148 binding sites of 1045 different proteins with five chemically diverse ligands with GOLD.[38] Of these, for four compounds, enrichment factors at 1% of the dataset between 26 and 102 are reported and, for one compound (AMP = adenosine monophosphate), poor performance was reported, leading the authors to recommend to use the inverse docking approach for selective ligands.

In an application study of Muller, the sc-PDB was also screened using GOLD with five combinatorial molecules sharing a 1,3,5-triazepan-2,6-dione scaffold.[39] Of five experimentally tested targets, one was confirmed as true target. Later, Kellenberger evaluated a combination of GOLD docking scores and an interaction fingerprint for the ranking of true targets for the same four ligands[25] on the sc-PDB consisting at that time of 4300 protein ligand binding sites of 1550 different proteins. For the four compounds, targets were predicted with AUCs between 0.7 and 0.95 for the GoldScore and AUCs between 0.45 and 0.9 for the interaction fingerprint scoring.

An evaluation of Glide in an inverse docking scenario on the Astex Diverse Set shows limitations in its intertarget ranking capability and coins the term "interprotein scoring noise".[40] It was found that a correction of the standard Glide scoring function, considering protein properties, improved target predictions.

In the following, we introduce a new structure-based target prediction method, which is based on protein−ligand screening and applies measures to account for the requirements of the reverse setup. Special care is taken regarding the preparation and handling of protein structure data. In addition, the method is a true inverse screening approach, substantially reducing the computing time compared to the application of a classic docking method. Finally, to address the intertarget ranking issue, several special scoring measures are applied, taking into account the diversity of protein pockets.

METHODS

The overall screening process is divided into two parts: a preliminary registration procedure and the actual screening procedure (see Figure 1). The registration procedure enables fast screening by performing precalculations and data setup only once for a protein dataset. The screening procedure can then be performed recurrently on the prepared dataset. The basis of the screening technology is a descriptor-based bitmap search, called RAISE technology (where RAISE represents RApid Index-based Screening Engine). Since the RAISE

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

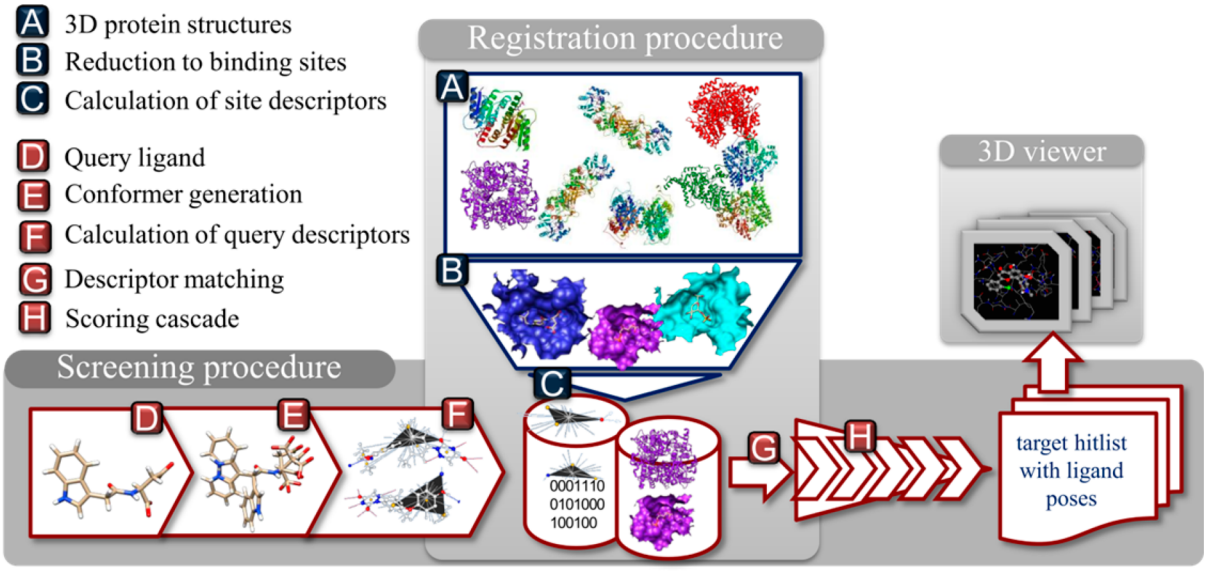


Figure 1. Workflow of the *i*RAISE inverse screening algorithm. Steps A–C of the registration procedure must be done only once for a dataset of protein structures. Steps D–H are part of the screening procedure.

technology, in this context, is applied in an inverse protein–ligand scenario, the tool is thus called *i*RAISE (*inverse*-RAISE).

Registration Procedure. Starting with a set of 3D protein structures (Figure 1A), first, the active sites are determined with a radius of 6.5 Å around a reference ligand (Figure 1B). The reference ligand can either be supplied by the user or identified in an automatic mode. In automatic mode, all ligands in a protein structure are used to build active sites except for co-factors, ions, crystallization agents, solution buffer agents, and ligands with covalently bound metals. The exclusion list is compiled by joining our own list with those from other publications[37,41,42] and contains, in total, 1207 PDB HET codes (see the Supporting Information). Next, descriptors are calculated for all active sites (Figure 1C, see the section entitled "Triangle Descriptor"). Finally, the active sites and the protein structures are stored in a relational database (see the section entitled "Protein Structure Database"), enabling efficient and consistent data handling.

Screening Procedure. Initially, conformations for the query ligand are sampled with the CONFECT conformation generator[43] (Figure 1E). Triangle descriptors are calculated for each conformation (Figure 1F). Then, these descriptors are matched against the protein descriptors (Figure 1G). A descriptor match corresponds to one protein–ligand pose with at least three matching interactions and a rough shape fit. Each found pose is then scored by the Scoring Cascade (Figure 1H; see the section entitled "Scoring Cascade"). The result of the screening procedure is a ranked list of targets for the query ligand as well as poses of the ligand for all hit targets are the results of the screening procedure.

Triangle Descriptor. In order to obtain a rapid screening procedure, ligand–protein matching is abstracted by a descriptor representing pharmacophoric and steric features. With the descriptor, the time-consuming multiple sequential placing of each ligand into each active site is circumvented. In *i*RAISE, the same triangle descriptor that was published for the virtual screening tool TrixX is used.[44,45] In brief, the descriptor has the following properties. A triangle descriptor consists of interaction spots of type hydrogen bond acceptor, donor, or hydrophobic at the corners. Each hydrophilic interaction spot is annotated with one or several interaction directions. The triangle side lengths encode the distance between the interaction spots. The shape of the ligand and, accordingly, the space of the active site around a triangle descriptor is encoded by 80 bulk rays originating from the center of the triangle limited by the surface of the ligand or the protein, respectively.

A novel feature of the triangle descriptor in *i*RAISE is the integration of flexibility of hydrophilic rotatable terminal groups (such as hydroxyl groups) of the active site and the query molecule. Rotatable groups are handled by interaction spots with multiple directions (for acceptors) or multiple possible interaction spots (for donors).

Unifying Query Descriptors. In molecular conformations where, e.g., only a terminal part of the ligand changes, many identical descriptors are generated. Therefore, a clustering procedure is applied reducing the total descriptor set to unique descriptors. Since the triangle descriptor contains only binned values, the clustering of identical descriptors is performed rapidly. The clustering reduces the number of descriptors significantly with which the index will then be queried: For an average of 200 query ligand conformations, the unique descriptor clustering reduces the amount to 35% of all descriptors.

Storing and Matching. The triangle descriptors of the active sites are stored in a FastBit compressed bitmap index[44,46] In *i*RAISE, this index is extended by storing the coordinates of the triangle corners next to the triangle descriptors, enabling immediate superposition of descriptors. This modification is mandatory for inverse screening due to the time-consuming calculation of active site descriptors, in contrast to ligand descriptors in a standard virtual screening setup.

The triangle descriptors of the query ligand are matched complementary concerning interaction spots, interaction directions, shape, and triangle side lengths to those of the active sites in the bitmap index. Once a match is found, the transformation needed to superpose the triangles is applied to the respective conformation(s) of the query molecule, producing the pose(s) in the protein active site.

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

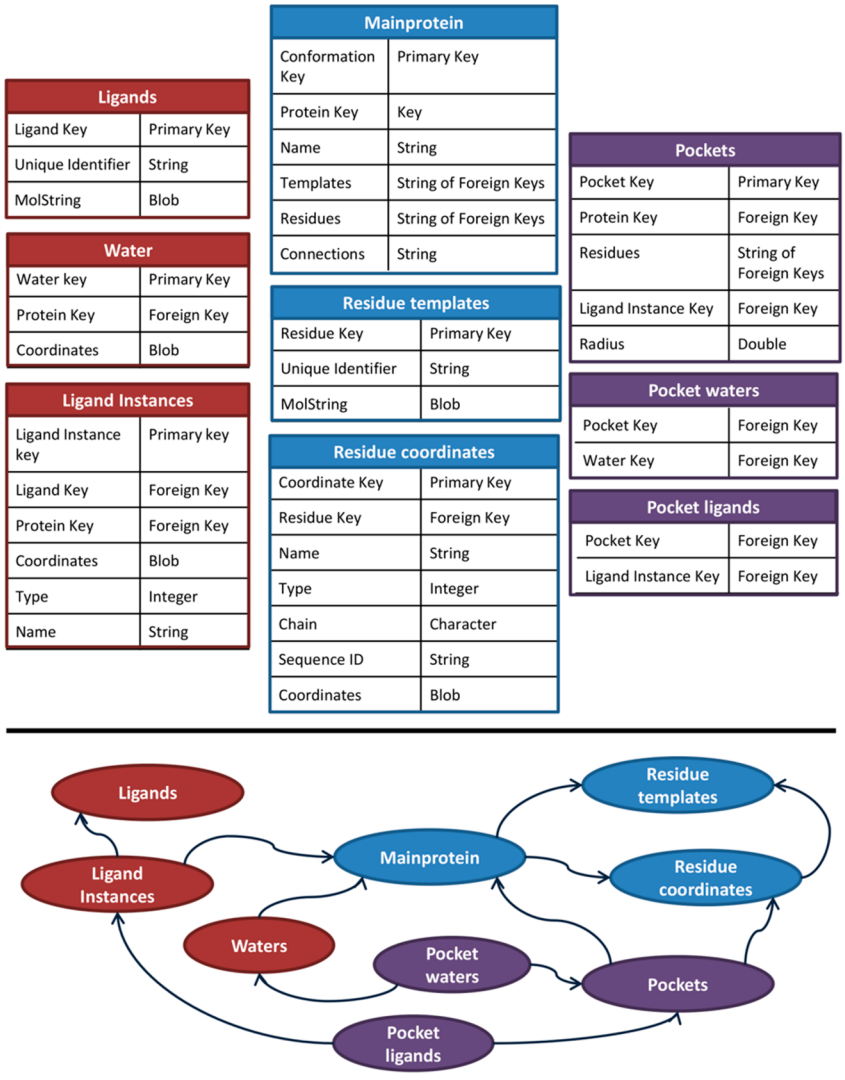


Figure 2. Database scheme of the protein structure database. Blue tables code the information on the protein. Red tables contain information on small molecules such as ligands, co-factors, metal ions, and water molecules. Purple tables code the information on the active site.

Protein Structure Database. For a storage- and time-efficient handling of the protein structure data as well as consistent representation of active sites calculated in the preparation procedure, a SQLite (www.sqlite.org) database is used. In Figure 2, the scheme of the database is sketched. The database consists of tables for protein data (blue tables); for data of ligands, water molecules, and metal ions (red tables); and for active site data (purple tables).

The protein data is represented in three tables: the *Mainprotein* table, containing general information, such as the protein name; and the *Residue templates* and *Residue coordinates* tables, containing amino acids. The *Mainprotein* table has two keys: a protein key and a conformation key, enabling the storage of protein ensembles. The *Residue templates* table contains each topological distinct amino acid of all proteins once. The USMILES[47] is used as unique identifier while the MolString contains the information on atoms, bonds, and valence states needed for reinitialization.[48,49] While each amino acid is added only to the *Residue templates* table if a topologically identical one previously has not been registered there, its coordinates, name, type, chain, and sequence index are written to the *Residue coordinates* table. Each *Residue coordinates* entry is mapped with a key to a *Residue templates* entry. With this setup of storing amino acids, the repeated information on the chemical composition of amino acids is stored only once in the database.

For storing ligands, the same concept of templates is used. The *Ligands* table contains a unique identifier in form of a USMILES and the MolString, coding the topology of a ligand. In this table, only a new entry is added, if the table yet does not contain the USMILES of the ligand. The coordinates of the ligands and distinct data such as the name and the corresponding protein key are stored in the *Ligand Instances* table. Water molecules are handled separately in the *Water* table containing the coordinates, a water key, and the protein key.

Active sites are stored in the *Pockets* table, which contains the key of the corresponding protein, the radius, the key of the ligand used as reference and the keys of the residues belonging to the active site. The keys of further ligands or metal ions contained in the active site are stored in the *Pocket ligands* table and keys of active site water molecules in the *Pocket water* table.

Storing the protein−ligand complexes in the database takes only 60% of the size needed to store the raw PDB files: For storing protein−ligand complexes of 100 random files from the PDB, the database size is 36MB (in comparison to 63MB for

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

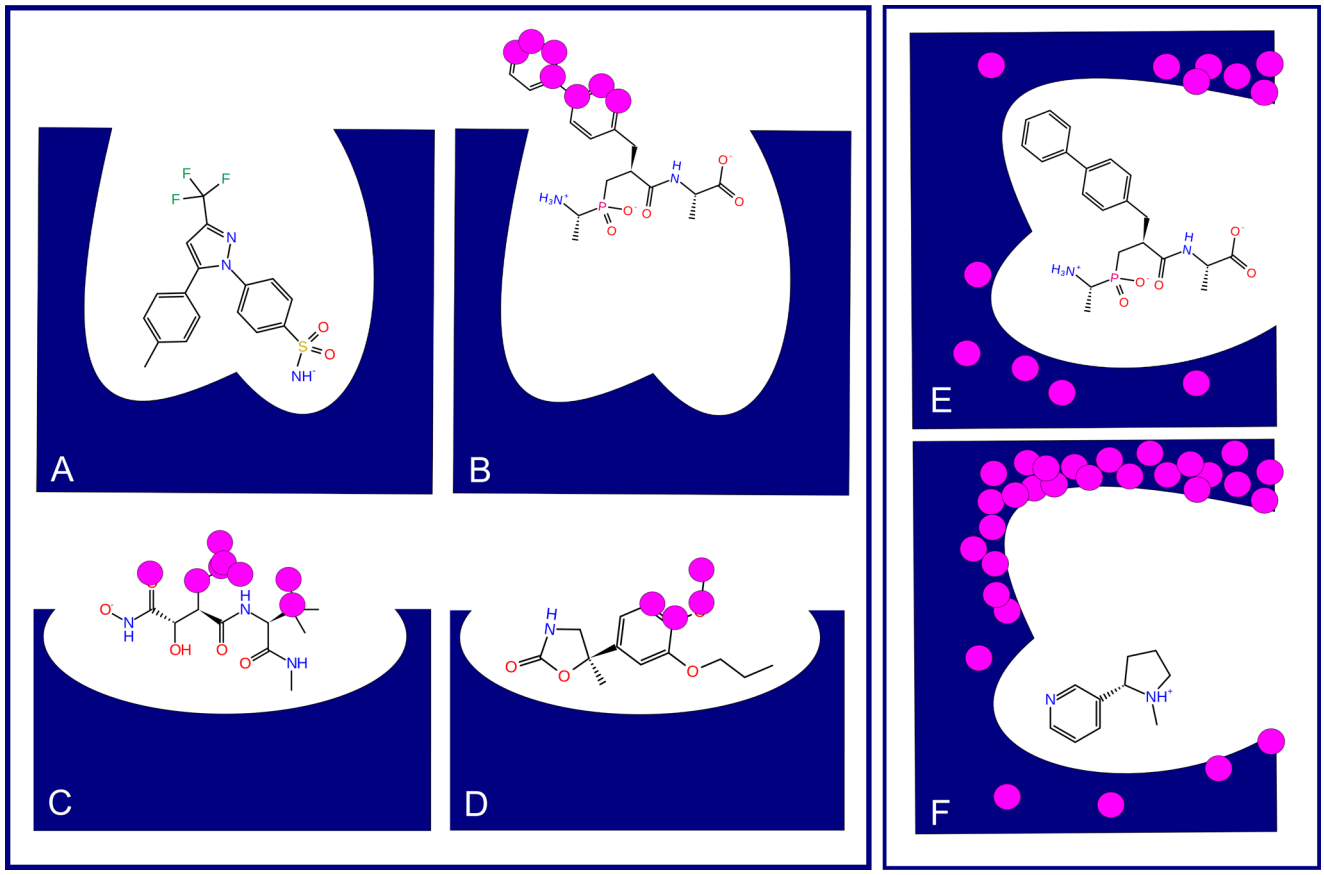


Figure 3. Schematic representation of the ligand coverage and pocket coverage in the Scoring Cascade. Panels (A–D) represent ligand coverage: (A) a reference ligand in a buried pocket with all atoms covered; (B) a ligand pose protruding into the solution with some noncovered atoms, highlighted by pink circles (this pose would be discarded due to insufficient ligand coverage); (C) a shallow pocket with a reference ligand with some noncovered atoms; (D) a docking pose in the shallow pocket with fewer noncovered atoms (thus, this ligand would not be discarded). Panels (E and F) represent pocket coverage: (E) a pocket with a reference ligand, which occupies most of the pocket and therefore has only a few noncovered pocket atoms; and (F) a small ligand in a large binding pocket with many noncovered atoms in the pocket, the score of which would be weighted down, because of insufficient pocket coverage.

the raw files), 333MB for 1000 random PDB files (in comparison to 558MB for the raw files), and 3.4GB for 10.000 random PDB protein files (in comparison to 5.7GB for the raw files). Reading a protein and the annotated active site from the database on average takes only 60 ms, independent from the database size as long as the database fits into main memory.

Scoring Cascade. A Scoring Cascade of five steps, accounting for active site diversity, is used to overcome interprotein scoring noise. The Scoring Cascade starts with all ligand poses obtained from the descriptor matches for one protein. It applies five steps to discard irrelevant poses and obtain a score comparable among proteins with diverse features:

1. Clash Test. The clash test discards clashing poses rapidly with a grid representation of the active site. This step already discards two-thirds of all poses from the descriptor matching.

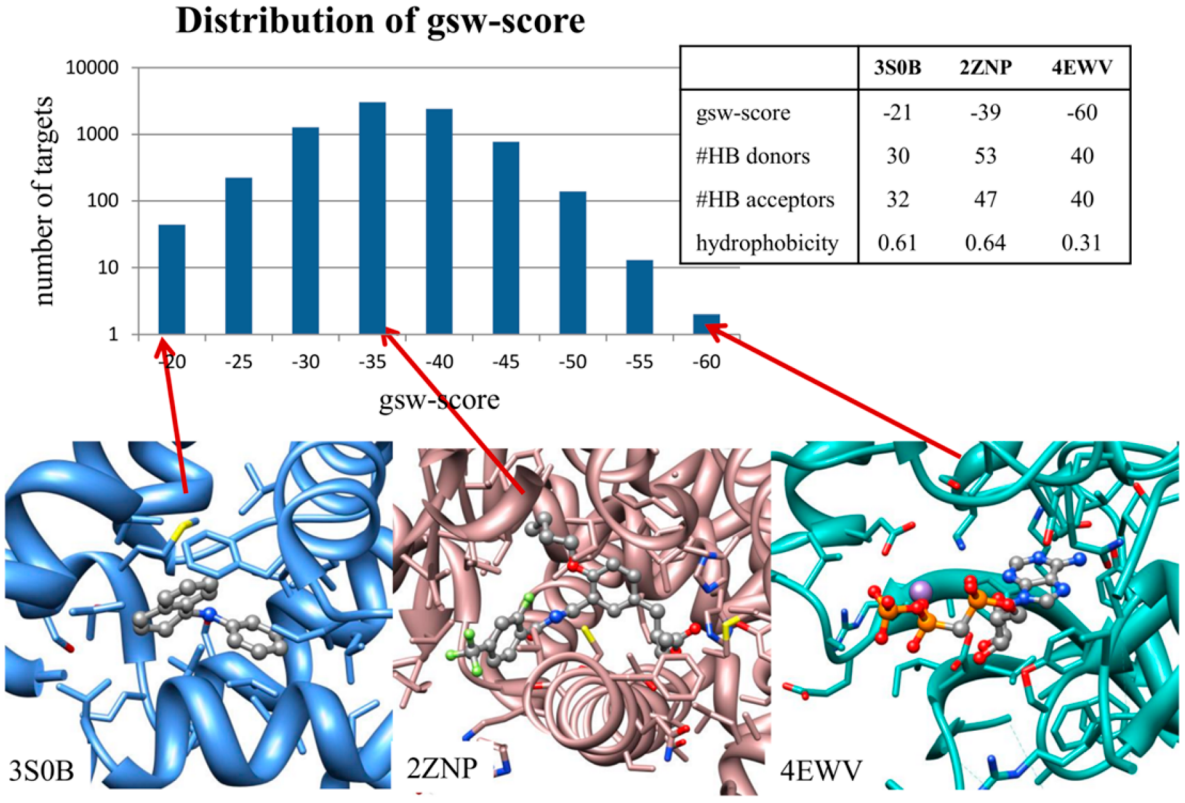
2. Interaction Score. The second step is the scoring of each pose with a simple interaction score based on Lennard-Jones potentials for hydrophilic interactions, metal interactions, hydrophobic contacts, and hydrophobic–hydrophilic mismatches (see Supporting Information for Lennard-Jones parameters). Beforehand, for each pose, the best hydrogen bond network in the active site is calculated with Protoss.[50] Also, the reference ligands that were used to determine the active sites are scored with this simple interaction score.

3. Reference Score Cutoff. For each pose, its interaction score is compared to the interaction score of the reference ligand of that active site. If the score of the pose is less than 75% of the score of the reference ligand, then this pose is discarded. This step discards, on average, 50% of the target matches for a query ligand.

Taking the reference ligand into account in scoring renders the method dependent on reliable crystallized protein–ligand complexes. In addition, not each ligand of each co-crystallized complex binds with high binding energy. However, since this step is used only as a cutoff, the binding affinity variations are not problematic: A reference ligand with low binding energy is scored only lowly by the energy-based scoring function and, thus, fewer ligands are discarded by this cutoff step.

4. Ligand Coverage Score. This score measures how well a ligand is buried in a pocket and is used to discard poses that protrude with a large part into solvent. Consulting the reference ligand enables comparing pockets with different shapes, e.g., comparing scores of shallow to buried pockets. If the coverage of a ligand pose multiplied by a factor of 1.2 is less than the coverage of the reference ligand, or if <10% of all ligand atoms are covered in total, the pose is discarded.

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX



	3S0B	2ZNP	4EWV
gsw-score	-21	-39	-60
#HB donors	30	53	40
#HB acceptors	32	47	40
hydrophobicity	0.61	0.64	0.31

Figure 4. Distribution of target-specific gsw-scores. The complex 3S0B with a minor gsw-score of −21 is highly hydrophobic. The complex 2ZNP with an average gsw-score of −39 is a larger pocket containing many hydrogen bond partners. The complex 4EWV with a high gsw-score of −60 is hydrophilic and contains a metal ion. (#HB donors = number of hydrogen bond donors, #HB acceptors = number of hydrogen bond acceptors, hydrophobicity = number of hydrophobic amino acids of active site divided by total number of amino acids of active site.)

The *ligand coverage* is the average ligand atom coverage (where A is the set of heavy atoms of a ligand):

$$\text{ligand coverage} = \frac{1}{|A|}\sum_{a\in A}\text{coverage}(a)$$

with the coverage of an atom a given as

$$\text{coverage}(a) = \begin{cases} 1 & \text{if close receptor atoms} + \dfrac{1}{|N(a)|}\displaystyle\sum_{b\in N(a)}\text{coverage}(b) > 3 \\ 0 & \text{otherwise} \end{cases}$$

Close receptor atoms are all atoms of the active site, which are in a radius of 4.5 Å of the ligand atom a. Furthermore, the average coverage of the covalently bound atoms $N(a)$ is added. As shown exemplarily in Figure 3, the ligand coverage is able to differentiate between binding scenarios to pockets with different shapes.

5. Pocket Coverage Score. The fifth and final step of the Scoring Cascade is the pocket coverage score, which addresses how well a ligand fills a pocket. Poses that produce insufficient pocket coverage in comparison to the pocket coverage produced by the reference ligand (<80% of the reference pocket coverage) are weighted down with a factor of 0.8. The pocket coverage is calculated as follows:

$$\text{pocket coverage} = \frac{1}{|P|}\sum_{a\in P}\text{coverage}(a)$$

where P is the set of pocket atoms.

Thus, the *pocket coverage* is the number of covered active site protein atoms divided by the total number of active site protein atoms. The *coverage* of a receptor atom a is calculated using the following formula:

$$\text{coverage}(a) = \begin{cases} 1 & \text{if distance to any ligand atom} < 4.5\ \text{Å} \\ 0 & \text{otherwise} \end{cases}$$

Since all pockets are determined with a cutoff distance of 6.5 Å around the ligand, reference ligands have a pocket coverage of ∼5%−40%. In Figures 3E and 3F, a schematic representation of the pocket coverage demonstrates how query ligands are scored higher in pockets that they fill in a similar way as a reference ligand. This step is only used as a weight to the score instead of as a cutoff.

Gauss Cutoff Score. After the Scoring Cascade, the score of each pose is used to rank the proteins as targets for the query ligand. A final step is applied to further tune the ranking capability of *i*RAISE. A Gaussian score weight (gsw-score) is applied to the score of the Scoring Cascade (sc-score) to be able to decide if a score is statistically significant for a protein pocket. Taking the sc-score in relation to the gsw-score results in the final *i*RAISE score, which is statistically significant, if it is >1. The gsw-score is a characteristic score for each protein pocket. It is obtained using the following steps:

(1) The complete protein pocket library of an *i*RAISE project is screened with the 84 chemically diverse ligands of the Astex Diverse Set.[51]

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(2) For each protein pocket, the sc-score is calculated for those of the 84 ligands, which could be placed into the pocket. The reference score cutoff-step (Step 3) of the Scoring Cascade is omitted to obtain a full spectrum of scores.
(3) The gsw-score of each protein pocket is defined as the average score of all 84 ligand scores. Parameterization showed that using a score of the average plus twice the standard deviation as assessment of statistical significance was too strict.
(4) All protein pockets with less than 20 of the 84 ligand scores are discarded, since the average score would be calculated from too few data points.

In Figure 4, the score distribution of the scores on the ~8000 targets of the sc-PDB dataset[37] screened with all 84 Astex ligands is shown. The scores are binned at units of 5. The gsw-score ranges between −21 and −60, and the median average score of all targets is −39. Between 0 and all of the 84 ligands can be docked to structures. The median is 67 ligands per target. Only 77 of the more than 8000 protein structures of the data set are scored with less than 20 ligands, leaving totally 7915 protein pockets. In Figure 4, a hydrophobic pocket with a minor gsw-score, a pocket with an average score and a hydrophilic pocket containing a metal ion are shown.

RESULTS

The evaluation of inverse screening tools still poses a huge challenge since no standard evaluation datasets or standard statistical metrics are established. Therefore, we focus on quantitative statistical evaluation, in comparison to other methods and evaluation of the gain of the individual steps of the Scoring Cascade.

In total, five evaluation experiments were performed:

(1) Binding mode prediction evaluation: Redocking study with RMSD values
(2) Evaluation of the ranking capacity of the Scoring Cascade
(3) Comparison of the ranking capability of *i*RAISE with FlexX/Hyde and Glide
(4) Comparison of ranking capabilities with a pharmacophore-based method on a large dataset
(5) Computing-time analysis

Software and Data Sets. *FlexX/Hyde.* FlexX[52] was applied as integrated in the LeadIT software suite (version 2.1, www.biosolveit.de) and for scoring the Hyde scoring function[53] was used. Default parameters were used. The docking was started with a conformation generated by Corina.[54]

Glide. For Glide docking, the XGlide script (Version 3.3, provided by Schrodinger, Inc.) was used. XGlide automates the protein preparation and Glide grid generation step, based on the native X-ray ligand complex. Subsequently, XGlide performs Glide SP (Version 6.1) docking runs with default parameters. Starting conformations of the input ligands were generated by Corina.[54]

Pharmacophore Search. For comparison of the ranking capability of *i*RAISE with a pharmacophore-based search strategy, we used the results published by Meslamani et al.[55]

Astex Diverse Set. For experiments 1−3, the Astex Diverse Set[51] was used. This dataset consists of 85 manually curated high quality diverse protein−ligand complexes. In the screening experiments, all 84 ligands were screened against all 85 protein structures (one ligand is present twice in the dataset). The objective of this experiment is to predict the co-crystallized target for each of the 85 query ligands as a true target and to rank the true target to the first positions of the list of all targets. This experiment is suited for simple evaluation and for comparison to other methods, but some caution is necessary when interpreting the results: Redocking the ligands into the target with which they have been co-crystallized is an artificial use case and is only useful for proof-of-concept evaluation. For estimating practical applicability, experiments with structures not crystallized with the query ligand are necessary. Furthermore, for the Astex Diverse Set, it is not known, whether ligands bind to multiple of the 85 targets.

sc-PDB Diverse Set. For the fourth experiment, the sc-PDB Diverse Set[55] consisting of 157 diverse ligands and the sc-PDB protein structure data set was used. The sc-PDB is a subset of the Protein Data Bank filtered with quality and druggability criteria. Meslamani used the 2010 version of the sc-PDB for the pharmacophore searches. This version is no longer available; therefore, we used the 2012 version of the sc-PDB.[56] We downloaded the original PDB files from the PDB instead of using the preprocessed files contained in the sc-PDB. Of originally 8077 structures, we used 7992, since, of these, 51 were discarded due to several errors during initialization of the reference ligand or the protein, 25 due to a mismatch of the reference ligand provided in sc-PDB, and 9 due to obsolete PDB codes. Annotation of the 7992 structures with the gsw-score further reduces the number to 7915. The sc-PDB Diverse Set of ligands consists of 157 ligands, which are co-crystallized with targets of the sc-PDB. Of these, we took a subset of 117 ligands of which the co-crystallized PDB structure reported by Meslamani was also present in the sc-PDB 2012 version. The 7915 structures of the sc-PDB 2012 were clustered by UniProtPK ID, as provided with the sc-PDB 2012 version, resulting in 2879 different proteins. True positive structures for the 117 ligands were assigned by two protocols: First, proteins with the same UniProtPK ID as the co-crystallized protein are considered true positives only.[55] Second, also structures with the same EC number as the co-crystallized protein were considered as true positives.

1. Redocking Study. As an initial study, we evaluated the ability of *i*RAISE to predict binding modes comparing the poses generated by *i*RAISE with the crystal structures. *i*RAISE was started with a Corina-generated conformation of the Astex ligands and up to 200 conformations were sampled. In Figure 5, the bars show the sum of ligands that can be predicted with RMSDs lower than the value indicated on the abscissa. In the 30 best-scored poses of each ligand, a solution with RMSD

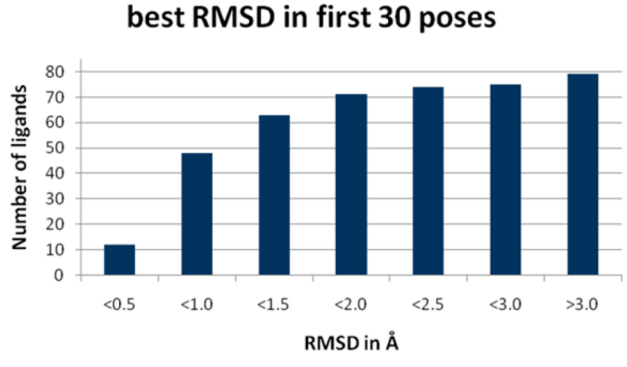


Figure 5. RMSDs for redocking each Astex ligand into its true target with *i*RAISE.

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

values of <2.0 A were observed in ∼84% of the cases. This value is slightly below the performance of optimized protein−ligand docking methods. One has to keep in mind, that *i*RAISE is a fully automated procedure enabling large throughput and therefore does not perform a post-optimization of poses. In such a scenario, the redocking performance is comparable to the state of the art.

2. Evaluation of the Ranking Capacity of the Scoring Cascade. In order to evaluate the effect of the Scoring Cascade, the rank of the true target for each Astex ligand was compared if only the simple interaction score, as a ranking measure, was used opposed to the full sc-score, based on the ligand poses obtained from descriptor matching (see Figure 6). This

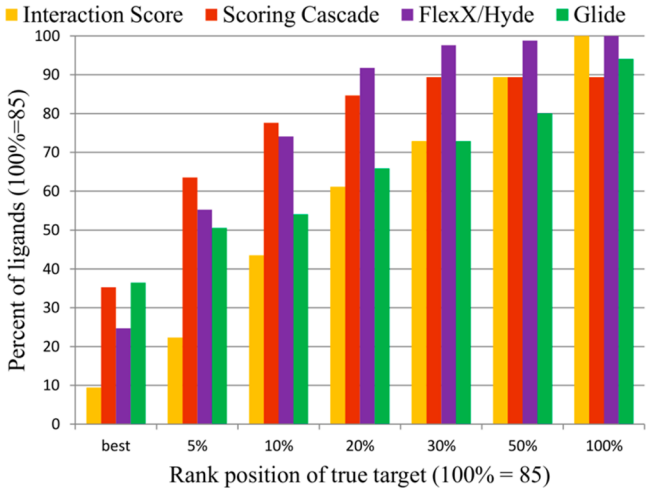


Figure 6. Ranking of true target for each of the 85 ligands of the Astex Diverse Set. The yellow bars show the ranks of the *i*RAISE poses scored only with the simple interaction score, the red bars show the ranking of the *i*RAISE poses scored with the full Scoring Cascade. The purple bars show the ranking if the FlexX docking with Hyde scoring is used and the green bars show the ranking of Glide. On the *y*-axis, the percentage of ligands is annotated, on the *x*-axis, the rank at which the true target was found. "Best" means the true target was found at rank 1 and the percentages show among which percent of the score-ordered list of targets the true target was ranked.

evaluation highlights two important points. First of all, the Scoring Cascade indeed ranks the true targets much better than the simple interaction score. With the Scoring Cascade, ∼35% of all ligand queries result in a ranking of the true target at position one, and for more than two-thirds of all ligands the rank of the true target in the score-ordered list is among the best 5%, i.e. among the top four ranks. Second, since the Scoring Cascade tunes the selectivity, for some ligands the true target is not found in the target list. For these examples, the scoring is too strict, as can be seen in the diagram in Figure 6 in the fact that only 90% of the ligands get a rank for their true target.

3. Comparison of the Ranking Capability of iRAISE with FlexX/Hyde and Glide. As a third test on the Astex Diverse Set, we compared the ranking results to standard docking approaches to see how *i*RAISE performs in comparison (see Figure 6). The diagram shows that Glide and *i*RAISE predict almost the same number of targets at the first position (∼35%). Ranking the true target in the first 5% of the target is accomplished by *i*RAISE for ∼63% of the ligands, which is superior to standard docking. At 10%, *i*RAISE's performance is still marginally superior to the docking programs, while at higher percentages FlexX/Hyde shows better performance. This is due to the fact that *i*RAISE, as discussed in the previous section, does not generate a pose for each ligand in its true target, because of the selectivity enhancement of the Scoring Cascade. However, on large datasets, enrichment at the first percentages is important for the choice of targets to test experimentally. In summary, the diagram shows that taking the co-crystallized reference ligand into account, as in the Scoring Cascade, substantially improves inverse screening performance.

4. Comparison to a Pharmacophore-Based Method. Following the evaluation protocol by Meslamani,[55] we screened 117 ligands of the sc-PDB Diverse Set against the 7915 protein structures of the sc-PDB 2012 with *i*RAISE. The rank of the first true positive target in a score-ordered list of all targets is consulted as a measure of success. Therefore, Table 1 contains the median rank (median of the 117 ligands) of the first true positive target identified by the methods in absolute number, as well as in percentage of the dataset. In Table 1, the results of *i*RAISE, as well as the data extracted from the supporting information of Meslamani's publication, are shown. The medians of four different pharmacophore-based methods (rigid1, rigid2, flex1, flex2), and of Surflex[57] and Plants[58] docking were extracted for 117 ligands from Meslamani[55] (the names of the methods are adopted). Since the number of protein clusters (UniProtKB ID clusters) differs in the sc-PDB version used by Meslamani and the one screened by *i*RAISE (2556 vs 2879 different proteins), we calculated the percentage of the first rank on all clustered proteins. For screening with *i*RAISE, two protocols for ligand preparation were used: Initially, 200 conformations of the ligand were used without the co-crystallized ligand (called "*i*RAISE flex" in Table 1). Then, the co-crystallized ligand structure was used as input for *i*RAISE screening without generating conformations (called "*i*RAISE crystal" in Table 1). Clearly, the second experiment is artificial and much "easier" for the method, since the correct conformation already is used. We performed this experiment to be able to compare the results to the pharmacophore-based methods, since those deduce the pharmacophore from the co-crystallized complex, which, then again, is a true positive in the dataset.

For the experiment with the crystal ligand, *i*RAISE performs better than the pharmacophore methods, independent of the way how true positives are annotated, following Meslamani by UniProtKB ID or by the EC number. The first true positive is found at 0.07% of the database, while the best pharmacophore-based method ranks the first true positive at 0.16% of the proteins. For the *i*RAISE screening with conformations, the first true positive is ranked at 1.15% of the protein structures, following the assignment of true positives by UniProtKB ID, compared to 2.5% of Surflex and 4.4% of Plants. If the EC number is considered during assignment of true positives, *i*RAISE ranks the first true positives at the median at 0.28%. In contrast to the UniProtKB ID, the EC number is not organism-specific, but, nevertheless, does classify the same protein justifying the usage of EC numbers in this case. The comparison of both assignment methods shows that the targets ranked toward the beginning of the list, which are not true positives after the UniProtKB ID, are nevertheless frequently correct predictions. The analysis shows, in total, that the ranking of *i*RAISE of the first true positive is comparable to the pharmacophore-based method and clearly outperforms both docking-based methods Surflex1 and Plants1. The docking

dx.doi.org/10.1021/ci500130e | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

Table 1. Median Ranking of First True Positive Identified by Pharmacophore-Based Methods (rigid1, rigid2, flex1, flex2), Two Docking Methods (Surflex1 and Plants1), Two Docking Plus Interaction Fingerprint-Based Methods (Surflex2 and Plants2), for *i*RAISE with Conformations (*i*RAISE flex), and for *i*RAISE with the Crystallized Ligand (*i*RAISE crystal)[a]

method	median rank of first TP on 2556 proteins	median rank in percent of proteins	median rank first TP on 2879 proteins	median rank in percent of proteins	median rank first TP on 2879 proteins with EC-TP[b]	median rank in percent of proteins with EC-TP[b]
rigid1 (pharm)	4	**0.16**				
rigid2 (pharm)	4	**0.16**				
flex1 (pharm)	6	**0.23**				
flex2 (pharm)	4	**0.16**				
Surflex1	65	**2.5**				
Surflex2	11	**0.43**				
Plants1	113	**4.4**				
Plants2	29	**1.13**				
*i*RAISE flex			33	**1.15**	8	**0.28**
*i*RAISE crystal			2	**0.07**	2	**0.07**

[a]All results except those from *i*RAISE are extracted from the supporting information given in the Meslamani work.[55] Boldface font indicates numbers that are comparable and should be used for interpretation; other numbers are given for completeness. [b]EC-TP = annotation of true positives (TP) with EC numbers.

methods Surflex1 and Plants1 can only be compared to the *i*RAISE flex method, since the results of starting them with the co-crystallized ligand are not available. However, the results of Surflex2 and Plants2 can be compared to the *i*RAISE crystal results, since they take into account interaction fingerprints derived from co-crystallized complexes, which also helps to select the correct conformation. The ranks of the first true positives of all 117 ligands are listed in the Supporting Information.

5. Computing-Time Analysis. To evaluate the computing time of *i*RAISE, its two steps—the registration procedure and the screening procedure—are evaluated separately. The registration procedure, including the triangle descriptor generation and the protein database generation, takes, on average, ∼7 s per target (all time measurements on a workstation with Intel Core i5/3570 CPU@3.4 GHz, 4 cores and 8GB RAM, single-threaded). The screening step requires, on average, 7 s per target (a median of 5 s per target) for a query ligand with an average conformation ensemble size of 200. However, the screening is highly dependent on the structure of the query ligand, ranging from, e.g., 1 s per target for a ligand with few triangle descriptors such as indirubin-3′-monoxime up to 38 s per target for a small hydrophilic ligand with many hits during the descriptor matching step such as pantoate (examples from the Astex Diverse Set). The *i*RAISE procedure is easily parallelizable with an automatic data partition during precalculation of data chunks of ∼100 proteins. Therefore, with a small computing cluster of ∼128 cores, a nonredundant PDB protein set with ∼50 000 proteins can be screened within ∼1−2 h.

CONCLUSION

With *i*RAISE, we introduce the first structure-based inverse screening method, which deviates from classic reverse docking approaches by applying several measures for facing the challenges of the reverse setup. An abstraction of protein−ligand matching with a triangle descriptor breaks the sequential screening course and saves computing time. To handle huge amounts of protein structures efficiently and consistently, a protein structure database was introduced. By precalculating and storing descriptors and active sites only once for a set of protein structures, screening with a query ligand can be performed rapidly. The problem of interprotein scoring noise of common docking scoring functions is addressed by a five-step Scoring Cascade, which substantially increases selectivity of the target ranking. To assess the statistical significance of a score for a protein structure, we introduced a Gaussian-based weighting score. Weighting the *i*RAISE score with it, the ranking of proteins is further improved. The resulting score can be used as a cutoff to decide up to which ranks proteins should be tested experimentally. Such a dynamic approach is better suited than a fixed cutoff at, e.g., 10% of the ranking list, since experimentally testing many targets for a ligand is much more complex than screening the same amount of ligands for one target. Therefore, in target prediction, false positives have a worse effect than in ligand prediction. Adding selectivity in the true positive assignment led to missing some true target structures, because of the strict scoring scheme. Therefore, the balance of selectivity versus sensitivity is an area of improvement in *i*RAISE.

*i*RAISE has been evaluated thoroughly, concerning its binding mode prediction and ranking capabilities. On the Astex Diverse Set with 85 diverse high-quality protein−ligand complexes, it has been shown that the Scoring Cascade boosts the ranking of the true target at the first position from 9% to 35%. Furthermore, the ability of *i*RAISE to predict the correct binding mode was evaluated by root-mean-square deviations (RMSDs) on the Astex Diverse Set. Of the 85 complexes, 74 were redocked with a RMSD value of <2.0 Å. The comparison to classic docking methods shows that *i*RAISE outperforms these in ranking, because of its measures accounting for protein pocket diversity. Finally, we evaluated the performance of

dx.doi.org/10.1021/ci500130e | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

*i*RAISE on a large data set of 7915 protein structures and 117 diverse ligands. The first true positive was ranked at 0.28% of the dataset, i.e., it is found among the first 8 ranks (median). In comparison to four pharmacophore-based protocols and two docking-based methods, *i*RAISE performs comparably and even better, if the same amount of preinformation is incorporated.

So far, *i*RAISE has only been evaluated on retrospective experiments. Prospective evaluation would be the next step to prove its usability. The *i*RAISE software is available for Linux operating systems (www.zbh.uni-hamburg.de/raise).

■ ASSOCIATED CONTENT

Supporting Information

List of PDB HET codes excluded for pocket detection. Parameters of simple interaction scoring function. Ranks of first true positives of sc-PDB Diverse Set. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 004940428387350. E-mail: rarey@zbh.uni-hamburg.de.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This project is part of the Excellence Cluster in Excellence Initiative by the State of Hamburg "Fundamentals of Synthetic Biological Systems (SynBio)" (www.tu-harburg.de/synbio).

Notes

The authors have made the *i*RAISE software available at www.zbh.uni-hamburg.de/raise.

■ ACKNOWLEDGMENTS

The authors thank the BioSolveIT GmbH for the opportunity to use the HYDE scoring function and the LeadIT software suite, as well as the NAOMI framework underlying *i*RAISE. Furthermore, the authors thank Lara Kuhnke from Bayer Pharma AG for technical assistance with the XGlide script.

■ ABBREVIATIONS

TP, true positives; RMSD, root-mean-square deviation

■ REFERENCES

(1) Khanna, I. Drug discovery in pharmaceutical industry: Productivity challenges and trends. *Drug Discovery Today* **2012**, *17*, 1088−102.

(2) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem.* **2007**, *2* (6), 874−880.

(3) Huggins, D. J.; Sherman, W.; Tidor, B. Rational approaches to improving selectivity in drug design. *J. Med. Chem.* **2012**, *55* (4), 1424−1444.

(4) Ashburn, T. T.; Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discovery* **2004**, *3* (8), 673−683.

(5) Liu, Z.; Fang, H.; Reagan, K.; Xu, X.; Mendrick, D. L.; Slikker, W., Jr; Tong, W. *In silico* drug repositioning−what we need to know. *Drug Discovery Today* **2013**, *18* (3), 110−115.

(6) Ekins, S.; Williams, A. J.; Krasowski, M. D.; Freundlich, J. S. *In silico* repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today* **2011**, *16* (7), 298−310.

(7) Roth, B. L.; Sheer, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3* (4), 353−359.

(8) Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today* **2013**, *18*, 495−501.

(9) Bottegoni, G.; Favia, A. D.; Recanatini, M.; Cavalli, A. The role of fragment-based and computational methods in polypharmacology. *Drug Discovery Today* **2012**, *17* (1), 23−34.

(10) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Drugbank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672 (www.drugbank.ca, accessed December 2013).

(11) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **2009**, *27* (2), 157−167.

(12) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: Applications to targets and beyond. *Br. J. Pharmacol.* **2007**, *152*, 21−37.

(13) Jenwitheesuk, E.; Horst, J. A.; Rivas, K. L.; Van Voorhis, W. C.; Samudrala, R. Novel paradigms for drug discovery: Computational multitarget screening. *Trends Pharmacol. Sci.* **2008**, *29* (2), 62−71.

(14) Niijima, S.; Yabuuchi, H.; Okuno, Y. Cross-target view to feature selection: Identification of molecular interaction features in ligand-target space. *J. Chem. Inf. Model.* **2010**, *51* (1), 15−24.

(15) Keiser, M. J.; Roth, B. L.; Armbruster, B. L.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197−206.

(16) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring polypharmacology using a ROCS-based target fishing approach. *J. Chem. Inf. Model.* **2012**, *52* (2), 492−505.

(17) Kinnings, S. L.; Jackson, R. M. ReverseScreen3D: A structure-based ligand matching method to identify protein targets. *J. Chem. Inf. Model.* **2011**, *51* (3), 624−634.

(18) Nettles, J. H.; Jenkins, A.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802−6810.

(19) Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4* (11), 682−690.

(20) Gregori-Puigjane, E.; Mestres, J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High. Throughput Screening* **2008**, *11* (8), 669−676.

(21) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. *Mol. BioSyst.* **2009**, *5* (9), 1051−1057.

(22) Nonell-Canals, A.; Mestres, J. In silico target profiling of one billion molecules. *Mol. Inf.* **2011**, *30* (5), 405−409.

(23) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263−266.

(24) Rognan, D. Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **2010**, *29* (3), 176−187.

(25) Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: Methods and problems. *J. Chem. Inf. Model.* **2008**, *48* (5), 1014−1025.

(26) Gowthaman, R.; Deeds, E. J.; Karanicolas, J. Structural Properties or Non-Traditional Drug Targets Present New Challenges for Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53* (8), 2073−2081.

(27) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789−6801.

(28) Chen, Y. Z.; Zhi, D. G. Ligand−protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43* (2), 21−226.

(29) Chen, Y. Z.; Ung, C. Y. Prediction of potential toxicity and side effect protein targets of a small molecules by a ligand−protein inverse docking approach. *J. Mol. Graph. Model.* **2001**, *20* (3), 199−218.

(30) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2002**, *30* (1), 412−415.
(31) Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H. PDTD: A web-accessible protein database for drug target identification. *BMC Bioinf.* **2008**, *9* (1), 104.
(32) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: A web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **2006**, *34*, W219−W224.
(33) Yang, L.; Chen, J.; He, L. Harvesting Candidate Genes Responsible for Serious Adverse Drug Reactions from a Chemical-Protein Interactome. *PLoS Comput. Biol.* **2009**, *5*, e1000441.
(34) Yang, L.; Wang, K.; Chen, J.; Jegga, A. G.; Luo, H.; Shi, L.; Wan, C.; Guo, X.; Qin, S.; He, G.; Feng, G.; He, L. Exploring Off-Targets and Off-Systems for Adverse Drug Reactiongs via Chemical-Protein Interactome−Clopazine-Induced Agranulocytosis as a Case Study. *PLoS Comput. Biol.* **2011**, *7*, e1002016.
(35) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, B. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(36) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An annotated database of druggable binding sites from the protein databank. *J. Chem. Inf. Model.* **2006**, *46* (2), 717−727.
(37) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: A database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **2011**, *27* (9), 1324−1326.
(38) Paul, N.; Kellenberger, E.; Bret, G.; Müller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **2004**, *54* (4), 671−680.
(39) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In Silico-Guided Target Identification of a Scaffold-Focused Library: 1,3,5-Triazepan-2,6-diones as Novel Phospholipase A2 Inhibitors. *J. Med. Chem.* **2006**, *49*, 6768−6778.
(40) Wang, W.; Zhou, X.; He, W.; Fan, Y.; Chen, Y.; Chen, X. The interprotein scoring noises in glide docking scores. *Proteins* **2011**, *80* (1), 169−183.
(41) Strömbergsson, H.; Kleywegt, G. J. A chemogenomics view on protein−ligand spaces. *BMC Bioinf.* **2009**, *10* (6), S13.
(42) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49* (23), 6716−6725.
(43) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem.* **2013**, *8* (10), 1690−1700.
(44) Schlosser, J.; Rarey, M. Beyond the virtual screening paradigm: Structure-based searching for new lead compounds. *J. Chem. Inf. Model.* **2009**, *49*, 800−809.
(45) Schellhammer, I.; Rarey, M. TrixX: Structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput. Aided Mol. Des.* **2007**, *21* (5), 223−238.
(46) Wu, K. FastBit: An efficient indexing technology for accelerating data-intensive science. *J. Phys.: Conf. Ser.* **2005**, *16*, 556.
(47) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(48) Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA—Interactive manipulation of molecule collections. *J. Cheminf.* **2013**, *5*, 38.
(49) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *J. Chem. Inf. Model.* **2011**, *51* (1), 3199−3207.
(50) Bietz, S.; Urbazek, S.; Rarey, M. Protoss: A holistic approach to predict tautomers and protonation states in protein−ligand complexes. *J. Cheminf.* **2014**, DOI: 10.1186/1758-2946-6-12.
(51) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50* (4), 726−741.
(52) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470−489.
(53) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claußen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided Mol. Des.* **2012**, *26*, 701−723.
(54) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(55) Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H.-O.; Rognan, D. Protein−ligand-based pharmacophores: Generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **2012**, *52* (4), 943−955.
(56) http://cheminfo.u-strasbg.fr, accessed January 2013.
(57) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **2007**, *21* (5), 281−306.
(58) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced protein−ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84−96.