J Comput Aided Mol Des (2014) 28:927–939
DOI 10.1007/s10822-014-9769-4

An integrated approach to knowledge-driven structure-based virtual screening

Angela M. Henzler · Sascha Urbaczek · Matthias Hilbig · Matthias Rarey

Received: 3 April 2014 / Accepted: 23 June 2014 / Published online: 4 July 2014

Abstract In many practical applications of structure-based virtual screening (VS) ligands are already known. This circumstance requires that the obtained hits need to satisfy initial made expectations i.e., they have to fulfill a predefined binding pattern and/or lie within a predefined physico-chemical property range. Based on the RApid Index-based Screening Engine (RAISE) approach, we introduce cRAISE—a user-controllable structure-based VS method. It efficiently realizes pharmacophore-guided protein-ligand docking to assess the library content but thereby concentrates only on molecules that have a chance to fulfill the given binding pattern. In order to focus only on hits satisfying given molecular properties, library profiles can be utilized to simultaneously filter compounds. cRAISE was evaluated on a range of strict to rather relaxed hypotheses with respect to its capability to guide binding-mode predictions and VS runs. The results reveal insights into a guided VS process. If a pharmacophore model is chosen appropriately, a binding mode below 2 Å is successfully reproduced for 85 % of well-prepared structures, enrichment is increased up to median AUC of 73 %, and the selectivity of the screening process is significantly enhanced leading up to seven times accelerated runtimes. In general, cRAISE supports a versatile structure-based VS approach allowing to assess hypotheses about putative ligands on a large scale.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9769-4) contains supplementary material, which is available to authorized users.

A. M. Henzler · S. Urbaczek · M. Hilbig · M. Rarey (✉)
Center for Bioinformatics (ZBH), University of Hamburg,
Bundesstraße 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

Keywords Structure-based virtual screening · Protein-ligand docking · Pharmacophore · Molecular properties · RAISE

Introduction

Virtual screening (VS) assists researchers in picking a few candidates from a vast amount of compounds giving a hint which chemical class of substances might be worth for optimization and further experimental testing. There exist various VS strategies [1]. Which strategy is deployed depends on the kind of information given in advance. Structure-based methods basically require a protein structure. Docking calculations predict the binding mode of a ligand that is assessed by scoring its protein-ligand interactions. In contrast to other VS approaches, the rather thorough assessment of compounds is at the expense of efficiency. Moreover, confronted with the well-known scoring problem, protein-ligand docking occasionally fails to predict the native binding mode [2] particularly, when protein flexibility is involved [3]. Pharmacophore-based strategies require a pharmacophore hypothesis given in advance. Meanwhile often applied in VS scenarios, the widespread feature-based models can be established from already known bioactive compounds, apoproteins, or protein-ligand complexes [4–8]. If a pharmacophore model compiles a few essential features representing commonly established protein-ligand interactions, the feature matching approach of pharmacophore-based VS is expedient to support fast compound selection. The scoring generally relies on geometric criteria assessing the alignment of the queried and matched features. Structure-based pharmacophore modeling offers the possibility to state excluded volume spheres and thereby to define a steric imprint of the targeted protein. Since they geometrically limit the search space, VS gets more restrictive. However, in contrast to the atomic protein representation of classical structure-based methods, excluded volume spheres are generally porous and untyped i.e., they allow to roughly assess the shape but miss essential atom type information which is required to assess the electrostatic propensity of the screened compounds to bind to the target. If a pharmacophore hypothesis and a protein structure are both available, an integrated approach is motivated by observations made in several studies. It has been shown that combining docking with pharmacophore filtering improves binding mode predictions and the enrichment of actives [9–11]. Pharmacophore-based docking may therefore serve as an attractive alternative to substitute consecutive or parallel pharmacophore filtering and docking phases in screening projects. There already exist docking approaches that allow the propagation of pharmacophore hypotheses. Methods such as Gemdock [12], SP-Dock [13], and Gold [14] use the additional information to adapt their underlying scoring function. Additional terms examine the similarity of a posed ligand to the pharmacophore hypothesis giving rather similar poses a greater weight. As demonstrated by FlexX-Pharm [15], a pharmacophore hypothesis can also reduce the underlying search space. Incremental construction algorithms like FlexX [16] can discard partial solutions as soon as the given hypothesis cannot be fulfilled anymore. As a result, poses obeying the pharmacophore emerge and the guided approach can be applied in VS more efficiently.

Besides the observed synergetic effects with respect to prediction quality and efficiency, together with a pharmacophore-based docking engine, the highly interactive process of pharmacophore modeling can pave the way towards a user-directed VS process. With the development of cRAISE our main concern was to provide an externally controllable platform for structure-based VS. Herein we describe the methodology of cRAISE which is a completely redesigned adaptation of the TrixX approach [17, 18]. cRAISE is now based on the Naomi framework [19], a robust chemical model which is designed to appropriately describe organic molecules relevant in the context of drug discovery. Nevertheless, cRAISE still captures the core idea of TrixX which postulates that a VS compares to a search that is only realized efficiently under the support of indexing-techniques. Essential search attributes, such as pharmacophore-like descriptors, are precalculated and stored in a way that allows to directly access relevant and omit irrelevant data during the search. The indexing requires costs and its benefit becomes apparent if multiple searches are performed. Moreover, an index requires that the prepared data remains unchanged throughout its complete lifetime. We assume that a typical large library, e. g. collections of external vendor catalogs or in-house collections, hardly changes its content but is frequently queried with diverse target proteins—a screening scenario for that our approach is designed. Under this premise computational effort can be shifted to a preparative process that enables efficient, succeeding VS runs. The most probable conformations of compounds can be computed in advance, stored, and accessed later without traversing through the conformation space again. However, the aim of cRAISE to intervene in VS applications seems to be limited by the necessity of the TrixX approach to utilize a static compound library. Various screening projects may demand that the library content satisfies project-specific requirements, e. g. omit compounds that later will lead to experimental artifacts. Moreover, VS is often performed in iterations learning from and following-up on first round results. Once a screening result is obtained, analyzed, and the molecules retrieved show properties not corresponding with the expectations, it arises the need to adapt initially made hypotheses. Opposed to the former implementation cRAISE now offers a broad range of search possibilities in order to avoid a recalculation of the index with a restricted library in such situations. It enforces guided docking runs when a pharmacophore hypothesis is stated. The additional information tailors the search space as soon and as much as possible. Moreover, molecular library profiles about constitutional or topological ligand features can be stated and utilized to gain further external control over the VS process.

The results of a method requiring external knowledge strongly depend on the provided information. Nevertheless, in order to reveal insights how cRAISE can be controlled and how it reacts on the given information, we automatically derived pharmacophore models covering a range of strict to rather relaxed model definitions. Utilizing this data, we evaluated our method to demonstrate the directionality of the pharmacophore-driven approach i.e., its capability to suggest solutions that meet the externally made expectations, which was our major design goal. Within our study we could also observe the synergetic effect of pharmacophore-guided docking and thus confirm the results that have been already stated by others. The pharmacophore models were derived from standardized forms of the Astex Diverse [20] and the DUD [21] datasets that have been previously used to comparatively assess the most popular docking algorithms. A complete issue of this journal addresses the competition to which our results can be directly compared. [22–29] On the given datasets the predictions of those methods strongly depended on the data preparation and the utilized docking protocol, thus, mean AUC values ranging between 59 and 80 % were reported. Our guided redocking and enrichment studies on this data show that if a pharmacophore model is chosen appropriately, a binding mode

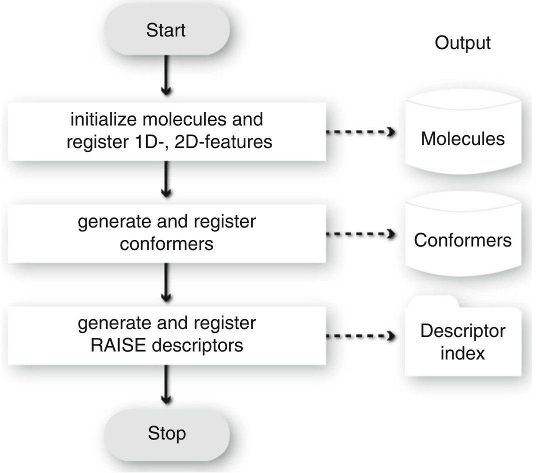
Start
Output
initialize molecules and register 1D-, 2D-features
Molecules
generate and register conformers
Conformers
generate and register RAISE descriptors
Descriptor index
Stop

Fig. 1 Library preparation workflow: precalculation of constitutional and topological molecule features, conformations, and RAISE molecule descriptor indexing. The output is reused throughout succeeding VS runs

below 2 Å is successfully reproduced for 85 % of well-prepared structures. Compared to unguided predictions, we were able to increase enrichment with our hybrid screening approach resulting to a median AUC of 73 % with automatically derived pharmacophore models and there is still room to enhance the enrichment further with more sophisticated models. Benchmark studies on subsets of the ZINC database [30, 31] show that external knowledge in form of pharmacophore models and molecular profiles enhance the selectivity of the screening process leading up to seven times accelerated runtimes. All in all, our method provides a versatile tool to intervene in structure-based VS by means of pharmacophore hypotheses and library profiles. Thereby it allows to encounter the generally conflicting aims of structure-based VS that requires choosing a trade-off between accuracy and efficiency when utilizing large-scaled molecular libraries.

Methods

Overview

cRAISE is a two-tiered procedure. In the preparatory phase, molecular feature detection, conformational sampling [32], and descriptor generation for the given compound library is realized. The features and conformations are stored in a database, the descriptors in a bit-compressed index both remaining static throughout subsequent VS runs (see also Fig. 1).

As illustrated in Fig. 2, the screening phase derives combined spatial and physico-chemical RAISE descriptors from a given protein active site that are translated to SQL-like queries. Using the compressed bitmap index structure, the queries detect matching molecule descriptors. The conformers of the matches are fetched from the database, placed into the active site by descriptor superimposition, and scored keeping only the best pose of a compound for the final hit list of the screening run. An optional pharmacophore hypothesis is used to restrict the set of queries and furthermore, to early reject poses before they are actually scored. A library profile is directly encoded in the query such that violating molecules are never fetched.

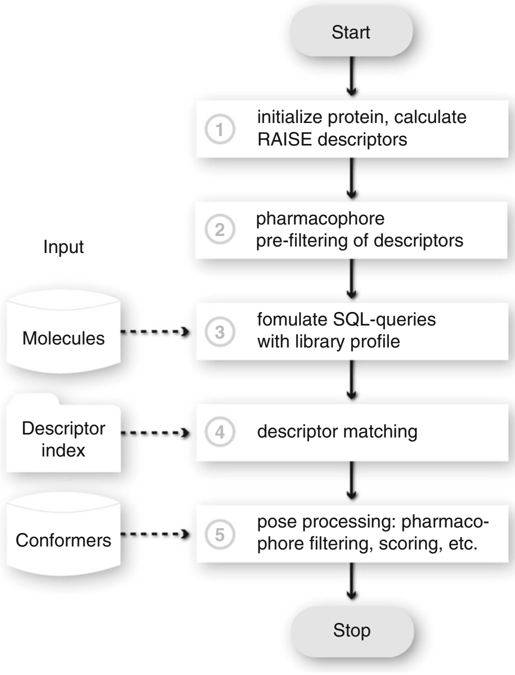
Start
Input
1 initialize protein, calculate RAISE descriptors
2 pharmacophore pre-filtering of descriptors
Molecules
3 fomulate SQL-queries with library profile
Descriptor index
4 descriptor matching
Conformers
5 pose processing: pharmacophore filtering, scoring, etc.
Stop

Fig. 2 Screening workflow: RAISE protein descriptor calculation, SQL-query generation, descriptor matching, pose generation, scoring. A pharmacophore hypothesis affects step 1, 2, and 5, a library profile step 3 and 4 of the workflow

The descriptor and concepts behind the indexing and matching phase have been described before [17]. Here we focus on the processing of pharmacophore hypotheses and library profiles to support externally guided VS.

cRAISE docking and virtual screening

cRAISE places ligands into a targeted protein active site by aligning complementary interaction sites. The determination of interaction sites is now based on the Naomi model [19]: for hydrogen-bond donor or positively charged atoms, donor sites are placed at a distance of an idealized hydrogen bond (2.8 Å) away from the heavy atom center into the directions of their protons. Acceptor sites remain on the center of a hydrogen-bond acceptor or a negatively charged heavy atom. Donor and acceptor sites possess attached directions indicating the orientation of protons or lone pairs and thus, possible interaction directions. Hydrophobic sites are undirected and reside on aliphatic and aromatic regions of small molecules. They are placed at carbons of acyclic aliphatic chains, at halogens, and on centers of aliphatic and aromatic rings. For adjacent carbon atoms, only a single site is created in the middle of the bond. In branched regions where a tertiary or quaternary central carbon is bound only to carbons, a single site is created in the center of the branch. The hydrophobic protein active site counterparts reside in volumes with a mostly hydrophobic environment. They are calculated by probing the active site volume with methyl-like representatives that are assessed by a classical Lennard–Jones (12, 6) potential. Surrounding hydrophilic atoms contribute to repulsion, but do not contribute to the attractive part of the potential. The top-scored representatives are selected and converted into hydrophobic interaction sites. Basically, cRAISE identifies an interaction in a protein-ligand complex if complementary interaction sites properly align in three-dimensional space i.e., if a donor site covers the site of an acceptor and if both sites posses roughly opposite interaction directions. Hydrophobic interaction sites do not have to fulfill the direction criterion.

Each triplet of interaction sites forms the corners of a triangle, the basis of the RAISE descriptor. A corner encodes the type (donor, acceptor, or hydrophobic) and obtains the associated interaction direction(s) of hydrophilic interaction sites. The descriptor additionally stores the lengths of the triangle sides. Some constitutional and geometric criteria ensure triangle angles being not too acute. Furthermore, each triplet must contain at least one hydrophilic corner. A special feature of the RAISE descriptor is that it encodes molecular shape relative to pharmacophoric features in a transformation invariant fashion. This is achieved with the lengths of 80 steric bulk rays that originate from the center of the triangle. The rays locally describe the van-der-Waals volume of a molecule or the interior volume of an active site. In order to decide whether a molecule fits into the active site, descriptor features can be simply compared. A descriptor match is recognized if complementary triangle corner types, opposite interaction directions, similar triangle side lengths, and an inclusion of all of the 80 ligand bulk rays in their respective active site descriptor counter-parts is detected. Then cRAISE accesses the molecule of origin designated by the descriptors compound/conformation ID from the molecule database. The coordinates of the triangle corners are used to calculate an affine transformation that superposes a pair of matching molecule and active site triangles. The transformation is applied to the molecule producing the actual pose. The basic idea of the RAISE screening procedure is to avoid evaluating each molecule descriptor. This is achieved with an efficient bitmap indexing and compression system [33, 34]. Essentially, cRAISE performs rigid body docking. Molecular flexibility is introduced with conformers generated with an integrated conformer sampling method based on CONFECT [32]. CONFECT was reparameterized for the cRAISE docking methodology: for rather small and rigid compounds slight suboptimal conformers are generated in order to increase the chance of a shape fit. However, to keep a large-scale application tractable it is necessary to provide an upper bound for the number of generated conformers [35]. Thus, the conformation set is restricted to at most 250 conformers per compound. For rather large and flexible compounds, a k-medoid cluster algorithm using the TFD [36] as a distance measure, samples rotatable bonds rather granular and selects diverse representatives out of the conformation space.

Table 1 Supported library profile features

Type	Features
Range[a]	Total charge, molecular weight, volume, topological polar surface area (TPSA), calculated octanol/water partition coefficient (logP), number of heavy atoms, hetero atoms, hydrogen-bond donors, hydrogen-bond acceptors, aromatic atoms, halogenic atoms, total number of bonds, rotatable bonds, maximum number of continuous rotatable bonds, number of ring systems, individual rings, aromatic rings, maximal ring size, maximal ring system size, number of stereo centers
Existence[b]	Chemical elements of the periodic table, any predefined molecular pattern (SMARTS), common functional groups (alcohol, ether, ketone, aldehyde, ester, amine, amide, amidine, guanidine, azide, nitrile, pyrrole, furan, thiophene, phenyl, pyridine)

[a] Features registered and evaluated on a value range

[b] Features registered and examined for existence

Integration of library profiles

The molecular feature handling of cRAISE is based on functionalities of MONA [37], a tool for visualization and statistical analysis of molecular libraries. Constitutional and topological features of compounds are calculated during the library preparation step and stored in the molecule database. Basically, the registered features can be categorized according to the kind of supported query. Table 1 summarizes all supported molecular features.

In order to support a guided VS run, a library profile can be defined by an arbitrary combination of feature range and existence conditions. As soon as a profile is given, cRAISE determines the IDs of compounds that are in accordance with the conditions prior to descriptor matching. The IDs constrain the SQL-queries and enforce a fetching of

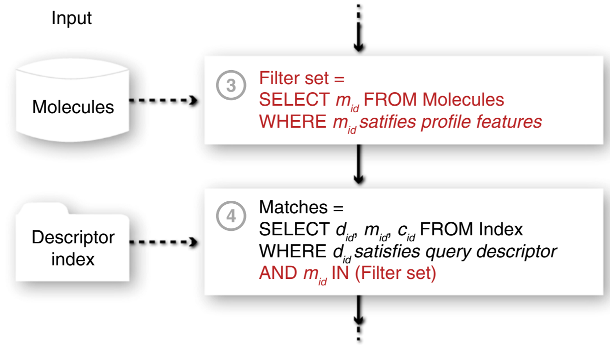
Input
Molecules
3
Filter set =
SELECT m_id FROM Molecules
WHERE m_id satifies profile features
Descriptor index
4
Matches =
SELECT d_id, m_id, c_id FROM Index
WHERE d_id satisfies query descriptor
AND m_id IN (Filter set)

Fig. 3 For each protein descriptor, a library profile reformulates its SQL-query to select respective matches from the descriptor index (step 3 and 4 of the screening workflow). d_{id}, m_{id}, c_{id} denote descriptor, molecule, and conformer IDs, respectively. They enable proper compound/conformer selection for pose initialization

Table 2 Supported pharmacophore features and their interpretation during pose sampling

Type	Feature interpretation
Donor inclusion[a]	Place only an H-bond donor/cation center with proper proton direction here
Acceptor inclusion[a]	Place only an H-bond acceptor/anion center with proper lone pair direction here
Hydrophobic inclusion[b]	Place only a hydrophobic group here
Hydrophilic inclusion[a]	Place an H-bond donor/acceptor/cation/anion center with proper proton/lone pair direction here
Any inclusion[b]	Place any atom center here
Donor exclusion[b]	Do neither place H-bond donor nor cation atom centers here
Acceptor exclusion[b]	Do neither place H-bond acceptor nor anion atom centers here
Hydrophobic exclusion[b]	Do not place hydrophobic atom centers here
Hydrophilic exclusion[b]	Do neither place H-bond donor, acceptor, cation, nor anion atom centers here
Any exclusion[b]	Do not place atom centers here

[a] Directed feature

[b] Undirected feature

appropriate descriptor matches from the index. Figure 3 visualizes this process.

Integration of pharmacophore hypotheses

cRAISE supports the specification of pharmacophore-type inclusion and exclusion features directly influencing its pose sampling stage. An inclusion feature is a constraint defining a region in the protein active site where a ligand atom has to reside. Exclusion features define forbidden regions of the active site. Each feature has either the type *donor*, *acceptor*, *hydrophobic*, *hydrophilic*, or *any* and is represented by a sphere defined by center and tolerance radius. Some features are directed to further constrain the location of a ligand atom with an appropriate proton or lone pair orientation. Table 2 summarizes all kinds of supported feature types and describes which constraints are enforced during the cRAISE pose sampling stage. A pharmacophore hypothesis can be defined by an arbitrary set of inclusion and exclusion features. Additionally, the number of essential inclusion features N_e states how many inclusion features have to be fulfilled simultaneously by a placed ligand.

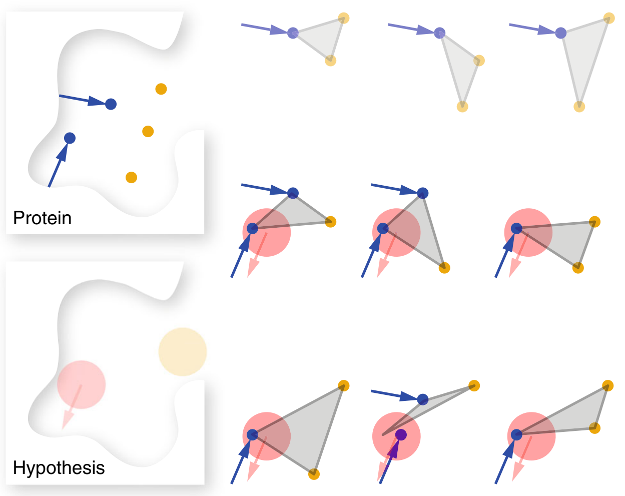

Fig. 4 A pharmacophore hypothesis is locally tested on query construction (step 2 of the screening workflow): an acceptor feature (*red*) restricts the calculation of query descriptors. Only active site descriptors with a complementary donor (*blue*) corner in the feature sphere and an opposite direction are built. Descriptors that would be built without the hypothesis are omitted (*grayed out*). Hydrophobic (*yellow*) features are not evaluated at this stage

Predefined pharmacophore features are used at two stages of the screening process: (1) prior to descriptor matching to reject pharmacophore violating query descriptors and (2) during the post-matching phase to reject pharmacophore-violating poses. Figure 4 demonstrates the effect of a pharmacophore hypothesis on query construction. Only triangles with at least one corner contained in a *donor*, *acceptor*, *hydrophilic*, or *any* inclusion feature are built. Since RAISE descriptors cover molecules only locally, false negative predictions could occur if one enforces more than one inclusion feature at this stage. Hydrophobic features do not restrict query descriptors but are evaluated in the post-matching phase. In consequence, hypotheses stating only hydrophobic features do not influence the query construction at all.

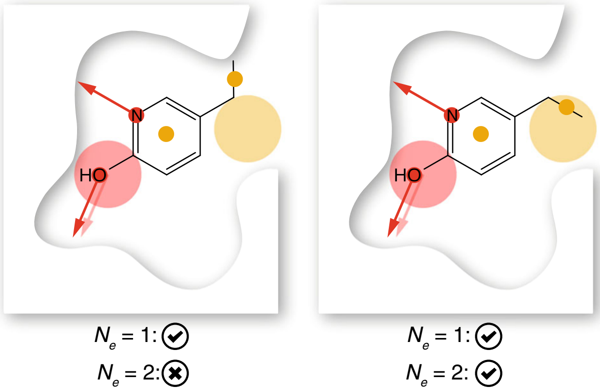

Fig. 5 A pharmacophore hypothesis is globally tested for poses (during step 5 of the screening workflow). *Left* the pose satisfies a single feature but violates the hypothesis if it requires two features. *Right* another pose satisfies the hypothesis

Figure 5 shows how a hypothesis is tested in the post-matching phase. All inclusion features are tested and if at least N_e features are fulfilled, the pose globally satisfies the hypothesis. This simple approach realizes the following task: If N_e equals the number of inclusion features defined by the user any of the generated poses need to obey the complete pharmacophore model, otherwise, for each pose all possible feature combinations of the given size are tested until either a single or none of them is fulfilled.

Hierarchical pose filtering and scoring scheme

A docking engine applied for large-scale VS produces a large amount of poses. The direct use of elaborate scoring functions too early without prior pose reduction hinders the throughput and eliminates the advantage in speed gained by the non-sequential screening paradigm of RAISE. The hierarchical pose-filtering scheme introduced here is intended to efficiently eliminate poses with sparse contacts, clashes, and pharmacophore violating poses as much and as soon as possible and to rapidly assess the quality of fit for succeeding poses. Initiating a VS run, cRAISE calculates information relevant for pose evaluation in advance. It determines an active site volume by computing the convex hull [38] from the active site atoms. A fine granular clash grid for hydrophilic and hydrophobic probes (0.25 Å voxel spacing) detects a clash for a grid point if the probe sphere contains any atom center. Moreover, a probe is assessed with its surrounding protein atoms and individual score contributions, namely possible protein counter interaction directions and Lennard–Jones-like potential values, are annotated at the grid. If a pharmacophore hypothesis states an exclusion feature, it is quasi seen as a protein atom sphere and grid points therein are flagged as clashes of the respective feature type.

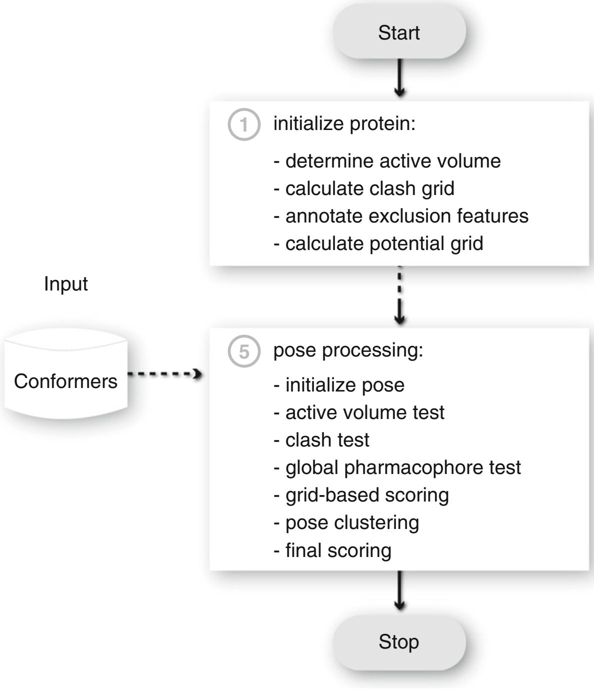

Fig. 6 Step 1 and 5 of the screening workflow: Relevant information is precalculated (*step 1*) and accessed later for pose evaluation (*step 5*)

The actual pose processing starts with two initial placement tests: to avoid sparse contacts with the protein, the majority of pose atoms has to reside in the precalculated convex hull after transformation. Ligand atoms clashing into the protein or intruding into exclusion volumes are rapidly identified using the clash grid. Then, each inclusion feature is tested until the number of fulfilled inclusion features N_e is achieved. The initial scoring stage accesses individual potentials for each atom and evaluates opposed interaction directions. These contributions compile the complete score for a ligand. It estimates the quality of fit by an empirical scoring function that accounts for hydrogen bonds, metal interactions, lipophilic contacts, and the loss of torsional entropy of the ligand. Essentially, the function is the Boehm scoring function [39], which was recalibrated on the Iridium Highly Trustworthy dataset [40] (v1.1) for which K_i/K_d values are published. Instead of piecewise linear penalty functions, Lennard–Jones-like potentials honor good and penalize close atom contacts (see Supplementary Material). The cosine assesses the angle deviations from the ideal geometry of opposed interaction directions. After ranking the poses by this score, similar poses closer than 0.5 Å RMSD to higher ranked ones are eliminated. Eventually, the pose-processing phase captures the scoring discrepancies that might occur due to grid mapping. It re-ranks the poses according to the cRAISE scoring function but this time evaluated on exact pose atom coordinates. The best pose for each compound contributes to the final hit list of the screening run. Figure 6 summarizes the individual preparation and pose evaluation steps.

Results and discussion

Datasets

For evaluating the pharmacophore-guided binding mode predictions and screening performance, data sets provided by the organizers of the ACS docking symposium 2011 were used. Several docking tools and scoring functions have already been evaluated with these standardized sets [22–29]. In the following, we refer to the datasets as $\text{Astex}_{\text{ACS}}$ and DUD_{ACS} respectively. The $\text{Astex}_{\text{ACS}}$ set comprises crystal structures of 85 protein targets of the Astex Diverse Set [20] with rerefined protein heavy atom, hydrogen atom, and ligand coordinates. For monomeric structures only a single ligand is provided as reference for active site definition, while for multimeric structures all ligands are supplied. Taking protein atoms with a heavy atom distance of 6.5 Å from any ligand atom center into account, all in all 146 well-resolved active site definitions can be obtained from the dataset. The symposium organizers provided a non-crystallographic structure for each ligand as starting point for docking calculations in order to support an objective, comparable evaluation. We will report values that take all ($n = 146$) and only a single, namely the qualitatively best site ($n = 85$) into account. Since the quality of the multiple active site copies differs, the values will provide a hint at what the precision of our method is and moreover, will provide comparability to the other, already published methods that haven been evaluated with this dataset and reported the values, as well. The DUD_{ACS} set covers active and decoy ligands for the 40 different targets of the DUD dataset [21]. The protein structures have been rerefined by the symposium organizers as well but opposed to the targets of the $\text{Astex}_{\text{ACS}}$ set, the targets of the DUD_{ACS} set retained their key crystal waters. The Supplementary Material summarizes further corrections made to the datasets. Our large-scale studies were carried out with subsets of the ZINC database [30, 31]. From the ZINC clean leads subset [41] comprising 4,230,832 compounds at access time, we randomly selected one, two, and three millions of unique compounds. We further refer to these libraries as $\text{ZINC}_{\text{CL1M}}$, $\text{ZINC}_{\text{CL2M}}$, and $\text{ZINC}_{\text{CL3M}}$ set, respectively.

Definition of pharmacophore hypotheses

Pharmacophore models introduce a strong bias in docking calculations. Although this is intended in practical applications, it causes major deficiencies in validation studies with respect to objectivity and reproducibility. We therefore decided to derive them automatically from individual protein-ligand complexes. The method initially places inclusion features centered at sites complementary to protein interaction sites. Tolerance spheres are scaled to a 1.7 Å radius, a common default value for structure-based models. To identify features that propagate binding, then only those that are complementary to the given ligand with respect to interaction type and geometry are selected. In order to create a scenario close to practice, feature sphere centers are not positioned on ligand atoms. Instead, they are derived relatively to protein atom coordinates. We generated various versions of each model linearly decreasing the number of inclusion features N_e being essential, i.e. are expected to be fulfilled. A relaxation of N_e has two practicable applications: First, a structure-based model derived from protein atom coordinates can result in feature constellations that require tensed or even unrealistic ligand conformations for entire fulfillment. Second, usually not all features are obeyed by all binding ligands in a VS scenario. Relaxed models allow to implicitly account for both situations and to explore only subsets of feature combinations of size N_e during the pharmacophore matching phase.

Pharmacophore-guided binding mode predictions

With the $\text{Astex}_{\text{ACS}}$ set and automatically derived models for these structures we performed pharmacophore-guided binding mode predictions. Figure 7 exemplarily depicts the guided top predictions found for four complexes of the $\text{Astex}_{\text{ACS}}$ set (green). They were predicted on lower ranks if the placement procedure was not guided by any feature. cRAISE does not explicitly change the score of a pose based on pharmacophore information but discards poses that contradict the given features. As a result, guided predictions let pharmacophore fulfilling poses emerge on higher ranks if the unguided prediction does not already rank a fulfilling pose on top. These observations show how guided binding mode predictions implicitly exert leverage on pose ranking.

Sometimes the automatically derived, structure-based models require extremely tensed conformations for optimal fulfillment. We explicitly neglect to put features on reference ligand coordinates, a procedure that would capture such situations. Instead, we implicitly relieve some tension during pharmacophore matching by relaxing the number of demanded features N_e. If a few feature definitions geometrically contradict each other within the model, the relaxation enables a recovery of near native poses that would require unrealistic ligand coordinates for matching all features.

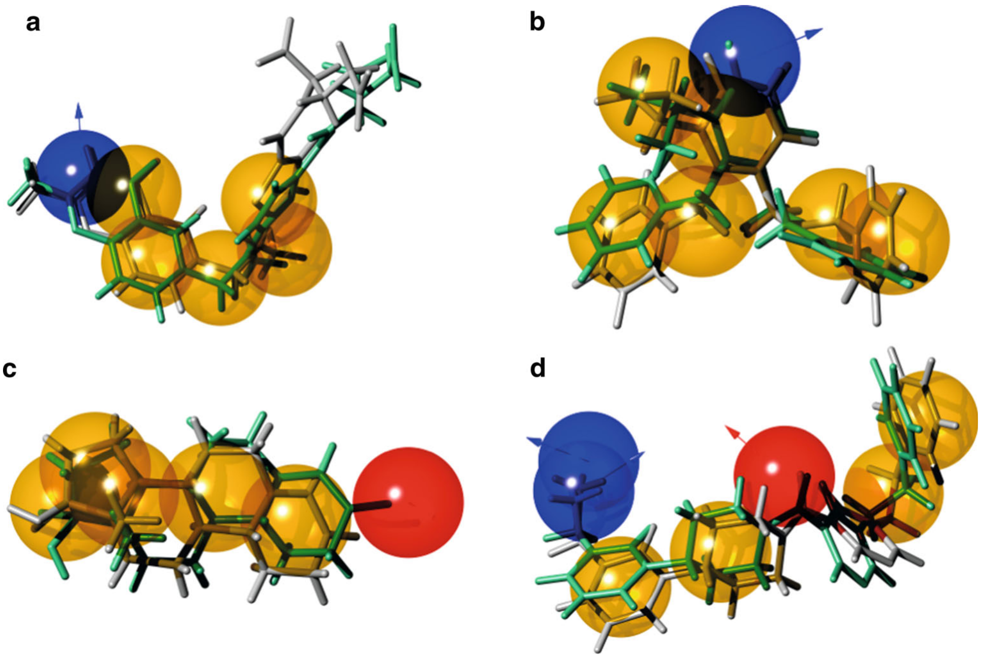

Fig. 7 Pharmacophore-guided binding mode predictions (*green*) are identified close to native binding modes (*gray*) on higher ranks. Donor inclusion (*blue*), acceptor inclusion (*red*), hydrophobic inclusion (*gold*). **a** 1hvy_3 at rank 1 (unguided 147), **b** 1jla_1 at rank 1 (unguided 5), **c** 1sqn_1 at rank 1 (unguided 5), **d** 2bm2_2 at rank 1 (unguided 143)

In order to quantify the effect of pharmacophore-driven binding mode prediction for the entire Astex$_{ACS}$ set, we characterized a successful prediction as a reproduction of a pose with a root mean square deviation (RMSD) of less than 2.0 Å to the respective reference ligand. A partially predicted pose was characterized by an RMSD above 2.0 Å, but below 3.0 Å. The total success rate was defined as the percentage of successful reproductions on all complexes. Figure 8 plots the success rates of the top sampled cRAISE poses for linearly increasing the number of demanded inclusion features. A tendency to guide binding mode predictions is observed if the poses have to satisfy up to 80 % of the demanded inclusion features (blue bars). Partially predicted poses and docking failures are reduced (green bars). However, for some structures the models were already too strict to successfully recover a pose at all (red bars) and the trend to direct top predictions further by demanding more features being fulfilled is reversed. Those failures can be captured by appropriate N_e relaxation which is part of the pharmacophore elucidation process before a guided docking with cRAISE can be accomplished. In order to show what our method can potentially achieve if this task is realized properly, we selected a good model for each target i.e., N_e^g which minimizes the RMSD of the top prediction (N_e^g-bars). On the Astex$_{ACS}$ set, the number of features was typically relaxed by 5–25 % to allow the poses to fulfill tensed models at least partially. Table 3 summarizes the guided success rates on the Astex$_{ACS}$ set with appropriately relaxed models and compares it with unguided predictions. A paired *t*-test was used to assess the significance of the comparisons. Therefore, we compared

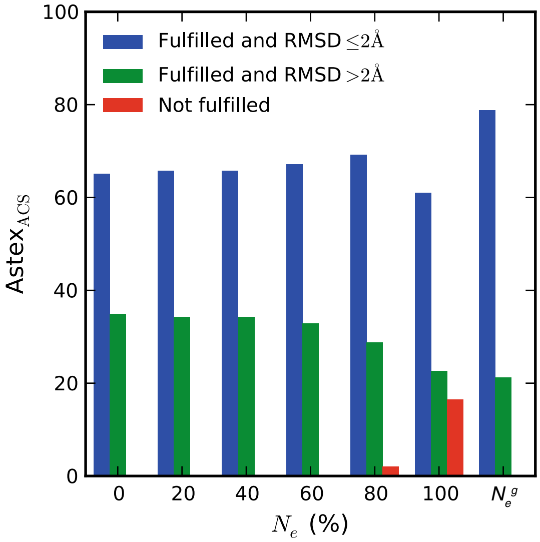

Fig. 8 Success rates of top predictions linearly increasing the number of demanded inclusion features N_e (*blue bars*). Less top predictions deviate from the native binding mode if they have to satisfy more pharmacophore features (*green bars*). Too strict model definitions lead to docking failures if the features cannot be satisfied by any pose (*red bars*). A relaxation of N_e recovers those failures (N_e^g*-bars*)

the paired RMSD differences within various rank cutoffs (for all protein-ligand complexes docked without any and with the use of the N_e^g-model). Under the assumption that the paired differences are independent and identically normally distributed, the probability *p* that the difference between the guided versus the unguided top predictions is random is far below 0.001. With increased rank cutoff the value successively increases since the chance of finding an identical pose within those ranks by the unguided docking

Table 3 Pharmacophore-guided and unguided pose sampling success rates (%) of the Astex$_{ACS}$ $n = 85$ ($n = 146$ in braces)

Rank	≤1.0 Å	≤2.0 Å	≤3.0 Å	p
Guided				
1	35 (32)	85 (80)	97 (95)	<0.001
5	41 (40)	91 (87)	97 (95)	0.002
20	48 (45)	93 (91)	99 (99)	0.151
32	51 (47)	93 (93)	100 (99)	0.227
All	52 (49)	95 (95)	100 (100)	0.562
Unguided				
1	29 (25)	71 (64)	84 (82)	–
5	38 (36)	86 (81)	94 (95)	–
20	46 (44)	87 (84)	97 (98)	–
32	47 (45)	91 (88)	99 (98)	–
All	55 (51)	97 (97)	99 (98)	–

Paired *t*-test *p* was determined for the complete dataset

runs increases as well. Moreover, the success rates reveal that the pharmacophore fulfilling poses with an RMSD >2 Å (green bars) mostly correspond to partially docked ligands. They are the result of pharmacophore models guiding the prediction by features covering poses only locally (compare e. g. the model of 1hvy in Fig. 7a). As a result, these models allow the remaining flexible ligand portion to freely explore unconstrained regions of the binding site. In general, our observations suggest that pharmacophore-guided binding mode prediction directs pose sampling and appropriately influences pose ranking.

Pharmacophore-guided enrichment studies

To assess the enrichment performance under pharmacophore type constraints we automatically derived models from the initially given 40 protein-ligand complexes of the DUD$_{ACS}$ set as described above. We performed pharmacophore-guided VS runs on the libraries consisting of the respective actives and decoys sets. Thereby, the enriched hits had to share at least a linearly increased amount of common features. The total area under (AUC) the receiver operating characteristic (ROC) curve served as a measure for the discriminative power of our method to separate actives from decoys. We additionally determined the true positive rate at a false positive rate of 1 and 2 % of the ROC showing the ability of our method to enrich actives early. The total enrichment performance was determined by averaging the AUC values of all 40 screening runs.

Since our models were derived from individual protein-ligand complexes i.e., determined from a single active, their features depict a superset of the real pharmacophore which is a particular feature combination therein. A

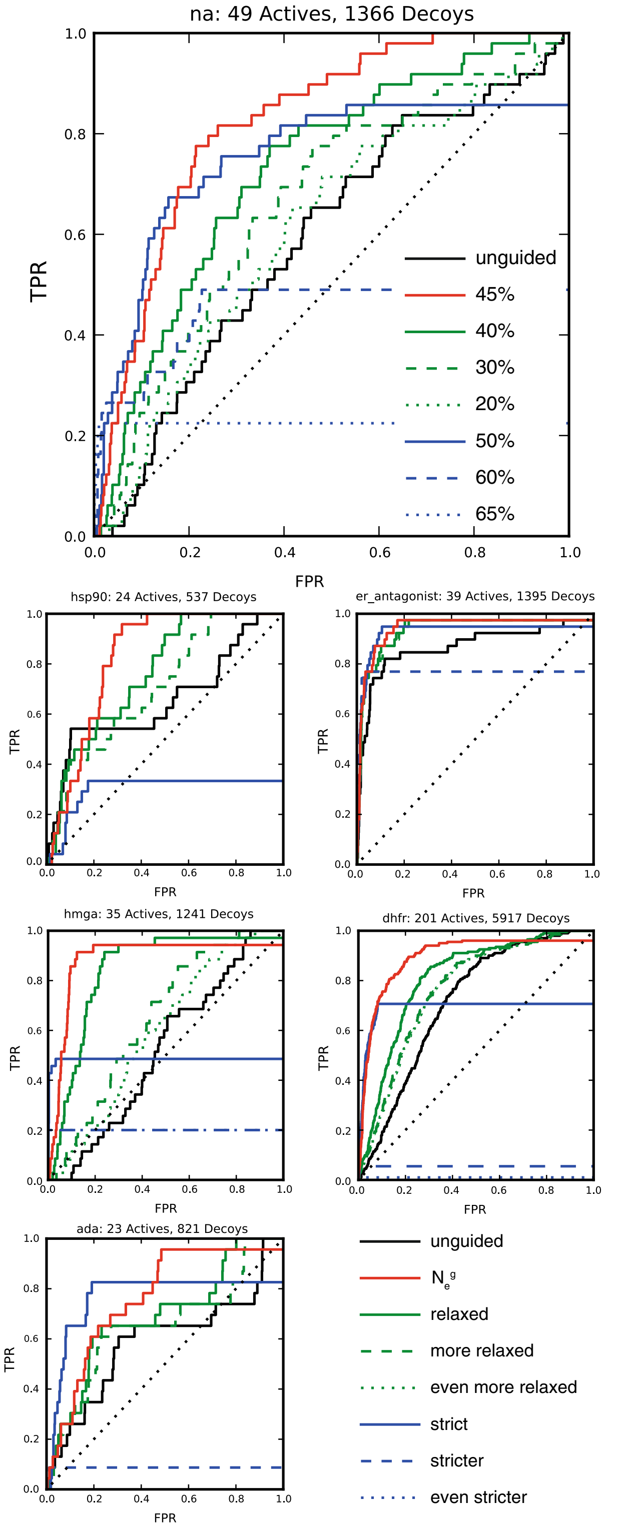

Fig. 9 Impact of strict and relaxed model definitions on enrichment behavior of the neuraminidase (na, $N_e^g = 45\,\%$), the human heat shock protein 90 kinase (hsp90, $N_e^g = 75\,\%$), the estrogen receptor antagonist (er_antagonist, $N_e^g = 65\,\%$), the hydroxymethylglutaryl-CoA reductase (hmga, $N_e^g = 65\,\%$), the dihydrofolate reductase (dhfr, $N_e^g = 55\,\%$), and the adenosine deaminase (ada, $N_e^g = 75\,\%$)

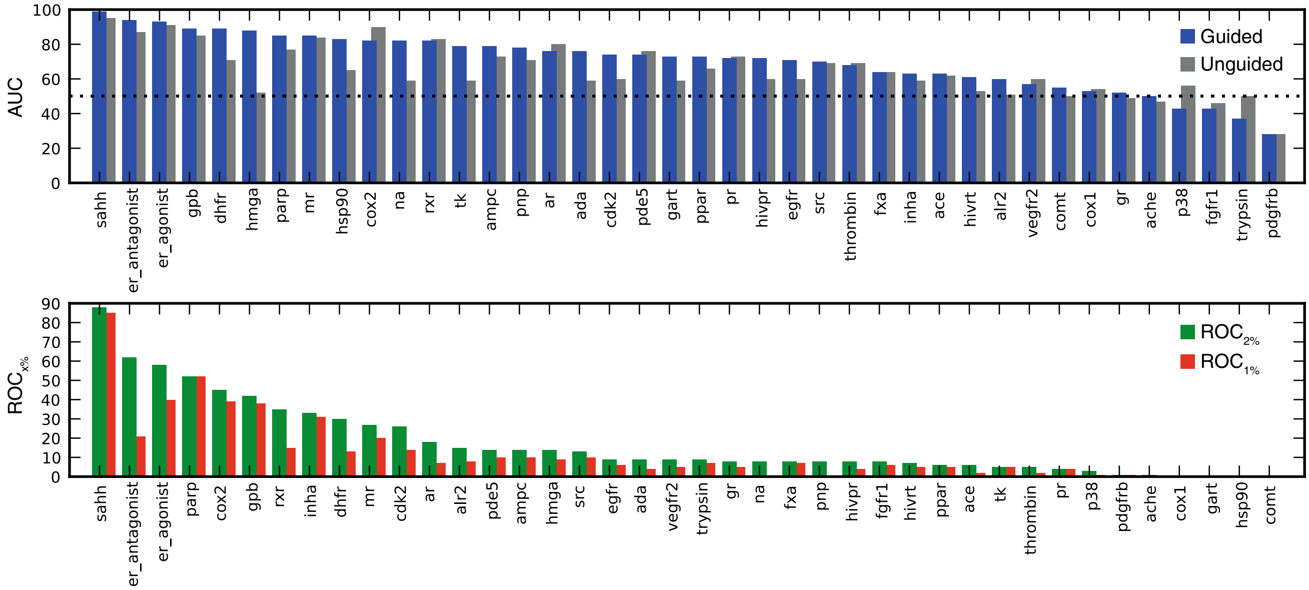

Fig. 10 ROC$_{1\%}$, ROC$_{2\%}$, and AUC for the N_e^g-models of the individual DUD$_{\mathrm{ACS}}$ sets

stringent model demanding all features being fulfilled identifies solely actives with that specific binding pattern. A relaxed model allows to introduce some tolerance and to explore various feature combinations during the pharmacophore matching phase. The degree of relaxations influences the enrichment of actives. This is demonstrated in Fig. 9 by example of six DUD$_{\mathrm{ACS}}$ targets. Demanding more features being fulfilled, the global enrichment with respect to the AUC metric is improved (green curves). In contrast, strict models enrich only subsets of actives—however, most often earlier (blue curves). All in all, these results show that cRAISE VS can be externally controlled by utilizing pharmacophore hypotheses.

It was not our goal to provide a tool for pharmacophore elucidation but for externally guided VS if well-prepared pharmacophore hypotheses have been already stated. In order to show what our screening method can potentially achieve in this case, in analogy to the above described guided docking experiments, we selected a good model for each target i.e., N_e^g, but this time with respect to maximize the global enrichment (Fig. 9, red curves). Basically, N_e^g represents an upper bound for the size of the actual pharmacophore. In our experiments this value ranged between 10 % and 95 % and there is still room to further improve the models, namely, if one determines the perfect N_e^g-combination that is able to recognize all actives. The following results represent realistic key figures what to expect from pharmacophore-guided VS. In Fig. 10 enrichment values for VS runs on the DUD$_{\mathrm{ACS}}$ sets guided by these models are plotted (blue bars). Most often guided VS significantly improves the early as well as the global enrichment of actives with respect to unguided predictions (gray bars). In case of hmga the pharmacophore model turns screening towards a highly directive process resulting in an AUC that is maximally improved by 36 %, while unguided VS enriches actives only close to random. A few models decreased the AUC, most often, if our model generation was confronted with problematic complexes that lead to ambiguous feature definitions. Table 4 summarizes the performance on the complete dataset and shows the statistical information on the enrichment metrics including 95 % confidence limits on the mean metrics.

Table 4 Pharmacophore-guided and unguided enrichment performance on the DUD$_{\mathrm{ACS}}$ sets

	ROC$_{1\%}$	ROC$_{2\%}$	AUC
Guided			
Mean	0.123 (±0.055)	0.177 (±0.064)	0.704 (±0.052)
SD	0.173	0.201	0.164
Median	0.060	0.090	0.730
Min	0.000	0.000	0.280
Max	0.850	0.880	0.990
Unguided			
Mean	0.087 (±0.041)	0.142 (±0.059)	0.651 (±0.047)
SD	0.129	0.185	0.145
Median	0.050	0.090	0.610
Min	0.000	0.000	0.280
Max	0.520	0.700	0.950

Error ranges represent 95 % confidence limits

Table 5 Number of pharmacophore-guided and unguided query descriptors

	Guided	Unguided
sahh	5,678	10,677
gpb	15,433	37,579
hsp90	2,934	19,637
fxa	3,996	31,510
er_agonist	7,756	13,042
dhfr	7,491	36,943

Large-scale studies

We conclude with guided and unguided large-scale VS studies and runtime evaluations on subsets of the the ZINC comprising one, two, and three million compounds, respectively. For these lead-like libraries on average 239 conformations per compound and 108 descriptors per conformation were generated and stored in the $\mathrm{ZINC_{CL1M}}$, $\mathrm{ZINC_{CL2M}}$, and $\mathrm{ZINC_{CL3M}}$ indices. Since it affects the runtime of cRAISE, we chose representative targets of the $\mathrm{DUD_{ACS}}$ set reflecting lower and upper bounds with respect to the number of query descriptors. Pharmacophore models showing promising capabilities in the above enrichment studies were used to guide the following VS runs. Table 5 summarizes the target data employed in the experiments. Inclusion features of type donor, acceptor, or hydrophilic restrict the number of query descriptors of the targeted protein active site. Our models contain on average four of these features which reduce the ordinary queries (on average 30,000) by around three quarters. Basically, N_e does not affect the number of queries and increasing this number does not restrict the search space further. To verify how many inclusions are satisfied, the poses have to be actually built. However, increasing N_e reduces the kept poses forwarded to the scoring stage. Thus, a stricter pharmacophore model can save expensive scoring calculations.

All computations were performed in a parallel screening setting on a high performance cluster of 25 Intel Xeon CPU E5630 dual quad core nodes with 2.53 GHz. Each process consumed maximally 8 GByte of main memory. The $\mathrm{ZINC_{CL1M}}, \mathrm{ZINC_{CL2M}}$, and $\mathrm{ZINC_{CL3M}}$ indices were split into packages of 2,500 compounds (á 6.9 GByte) each and were distributed over the local hard drives of each cluster node in order to reduce the network load during a VS run. Our measured runtimes are given in form of wall clock times. The processing time t_c for a single compound is given by $t_c = t_{\mathrm{total}}/N$, where N is the number of given compounds and t_{total} is the total VS time i.e. the sum of the wall clock times of all distributed jobs. t_c allows to estimate the cRAISE screening time independent of the parallel setup and the size of the employed compound library. The processing time t_m for a single conformation is given by $t_m = t_{\mathrm{total}}/M$, where M is the number of generated conformations. It allows to estimate the cRAISE screening time if externally provided conformers are processed. The parallel run time t_{p} is variable due to the current availability of compute nodes. Thus, it was estimated by the average of the VS time of individual jobs on basis of an optimal availability of 200 cores. Then, t_p reflects the best possible run time in a parallel setting of 25 freely available dual quad core nodes.

The observed timings of our large-scale experiments are summarized in Table 6. We achieved an up to seven times accelerated run time with pharmacophore hypothesis guiding the screening process. Basically, the runtime varies from target to target. The cause is found in the selectivity of the query: If an index contains n descriptors, a single query descriptor has the potential to extract all of them. If a target possesses m query descriptors, in the worst case, a screening produces $n \times m$ descriptor matches. Even if this worst-case scenario never occurs, the run time depends on how many poses are actually processed in the post-matching phase i.e. on the amount of extracted index descriptors. Table 6 shows the observed selectivity values $\sigma = \#\mathrm{matches}/\#\mathrm{index\,descriptors}$ for the targets employed in the VS runs. A selectivity of 1 indicates that the whole index is extracted. These values correlate with the observed runtimes. Guiding a VS by a pharmacophore hypothesis generates queries that are more selective and explains the accelerated run time behavior of cRAISE.

Molecular profiles—an example

The definition of molecular profiles allows to further guide the VS process with respect to retrieve only hits satisfying user-defined molecular properties. For the sake of completeness we demonstrate here a simple screening scenario: The $\mathrm{ZINC_{CL3M}}$ library contains three millions of lead-like compounds. We defined a molecular profile with Mona restricting this library to molecules with a molecular weight of at most 300, maximally 5 rotatable bonds, and a logP of at most 3.5 (provided in the Supplementary Material). It was utilized to determine the runtime if the $\mathrm{ZINC_{CL3M}}$ library is simultaneously filtered during a VS run of er_agonist. The retrieved timings were as expected. The profile indirectly reduced the library by 75 to 707,770 % compounds and the timings ($t_c = 2.06$ s, $t_p = 9.25$ h) were reduced by approximately the same amount. This experiment verifies that library profiles can be used ad hoc during a VS run without the necessity to rebuild the static index for a restricted library.

Table 6 Timings on the $\text{ZINC}_{\text{CL1M/2M/3M}}$ sets

Average per conformation t_m, per compound t_c, parallel runtime t_p, selectivity σ

	Guided				Unguided			
	t_m (s)	t_c (s)	t_p (h) (1M/2M/3M)	σ	t_m(s)	t_c (s)	t_p (h) (1M/2M/3M)	σ
sahh	0.01	1.20	1.66/3.32/5.00	0.07	0.01	2.75	3.65/7.13/10.40	0.13
gpb	0.01	2.80	3.73/7.73/11.70	0.35	0.04	9.73	12.58/26.22/38.78	0.87
hsp90	0.01	1.23	1.66/3.43/5.15	0.21	0.04	8.89	13.05/25.75/37.03	1.38
fxa	0.01	3.32	4.38/8.93/13.87	0.75	0.09	21.24	30.37/59.00/89.37	5.40
er_agonist	0.02	5.68	7.62/15.10/24.05	1.28	0.03	8.23	10.83/23.22/34.30	1.54
dhfr	0.02	4.49	5.98/12.15/18.70	1.06	0.10	24.81	34.12/63.32/103.37	6.10

Conclusion

We have described cRAISE, a VS tool that propagates additional knowledge to support pharmacophore-driven pose sampling and library profiling in structure-based VS. This is particularly useful if hypotheses about desired key interactions and/or physico-chemical features of the compounds are known beforehand. In such situations cRAISE allows to focus on predictions that are of major interest. Our pharmacophore-guided approach leads to an effective search space reduction and as a result, it reduces computational demand. Opposed to many other pharmacophore-guided docking approaches, it thereby provides a screening platform that allows the testing of hypotheses on a large scale.

Our results demonstrate that pharmacophores allow to externally direct the docking process. The implemented search space reduction does not lead to a loss of quality. To the contrary, if the models are prepared properly, they offer the chance to improve binding mode predictions. Poses that violate the given feature definitions are either not generated at all or rejected before they are actually scored. The procedure lets pharmacophore fulfilling poses emerge without the need to adapt the underlying scoring function. The presented enrichment studies reveal that early as well as global enrichment can be enhanced by this mean. Relaxed models offer the possibility to simultaneously evaluate different feature combinations. They enforce only some pharmacophoric commonality on the retrieved screening results. Nevertheless, by the use of strict model definitions it is possible to focus on compounds that reveal a specific binding pattern.

Confronted with millions of compounds, the once prepared cRAISE descriptor index basically enables fast information retrieval in succeeding VS runs. In order to benefit from our index-based VS technique the precalculated information needs to be permanently stored and the content needs to remain unchanged throughout its complete lifetime. The methods of cRAISE introduced here provide a versatile interface to support flexible queries on this static compound library for different screening projects. Given pharmacophore definitions are utilized to guide cRAISE to extract only information of molecules with an improved chance to result in a pharmacophore-obeying pose. We showed that pharmacophore definitions can drastically accelerate the screening process. Moreover, cRAISE allows to state library profiles by constitutional and topological ligand conditions. The additional constraints restrict the index-based search further and filter out compounds without any loss of efficiency simultaneously during a VS run.

Our introduced hybrid method demonstrates how to gain external control over structure-based VS. Essentially, taking the best out of both worlds, it is a first step towards an integrated, synergetic VS platform combining structure- and ligand-based techniques. Relevant for any three-dimensional VS strategy is the consideration of tautomers and ionization states of query and library molecules. cRAISE provides the option to account for these degrees of freedom during pharmacophore guided and unguided VS. This extension accompanied with the respective results will be published separately. The cRAISE software is available for Linux operating systems (http://www.zbh.uni-hamburg.de/raise).

Acknowledgments We would like to thank Nadine Schneider for the fruitful discussions about protein-ligand interactions and scoring functions, moreover, Christin Schärfer for her commitments to conformer generation. This work was financially supported by the BMWI-ZIM Project KF2563701.

References

1. Sotriffer C (2011) Virtual screening. WILEY-VCH Verlag GmbH & Co, KGaA, Weinheim
2. Sotriffer C, Matter H (2011) The challenge of affinity prediction: scoring functions for structure-based virtual screening. In: Sotriffer C (ed) Virtual screening. Wiley-VCH Verlag GmbH & Co, KGaA, Weinheim, pp 177–221

3. Henzler AM, Rarey M (2010) Mol Inform 29:164–173
4. Sanders MP, McGuire R, Roumen L, de Esch IJ, de Vlieg J, Klomp JP, de Graaf C (2012) Med Chem Commun 3:28–38
5. Wallach I (2011) Drug Dev Res 72:17–25
6. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) J Med Chem 53:539–558
7. Yang SY (2010) Drug Discov Today 15:444–450
8. Dror O, Shulman-Peleg A, Nussinov R, Wolfson H (2004) Curr Med Chem 11:71–90
9. Tintori C, Corradi V, Magnani M, Manetti F, Botta M (2008) J Chem Inf Model 48:2166–2179
10. Muthas D, Sabnis YA, Lundborg M, Karln A (2008) J Mol Graph Model 26:1237–1251
11. Peach ML, Nicklaus MC (2009) J Chem Inform 1:6
12. Yang JM, Shen TW (2005) Proteins 59:205–220
13. Fradera X, Knegtel RM, Mestres J (2000) Proteins 40:623–636
14. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P (2004) J Chem Inf Comp Sci 44:793–806
15. Hindle SA, Rarey M, Buning C, Lengauer T (2002) J Comput Aided Mol Des 16:129–149
16. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470–489
17. Schlosser J, Rarey M (2009) J Chem Inf Model 49:800–809
18. Schellhammer I, Rarey M (2007) J Comput Aided Mol Des 21:223–238
19. Urbaczek S, Kolodzik A, Fischer JR, Lippert T, Heuser S, Groth I, Schulz-Gasch T, Rarey M (2011) J Chem Inf Model 51:3199–3207
20. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) J Med Chem 50:726–741
21. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789–6801
22. Neves MAC, Totrov M, Abagyan R (2012) J Comput Aided Mol Des 26:675–686
23. Spitzer R, Jain AN (2012) J Comput Aided Mol Des 26:687–699
24. Schneider N, Hindle S, Lange G, Klein R, Albrecht J, Briem H, Beyer K, Claußen H, Gastreich M, Lemmen C, Rarey M (2012) J Comput Aided Mol Des 26:701–723
25. Novikov FN, Stroylov VS, Zeifman AA, Stroganov OV, Kulkov V, Chilov GG (2012) J Comput Aided Mol Des 26:725–735
26. Liebeschuetz JW, Cole JC, Korb O (2012) J Comput Aided Mol Des 26:737–748
27. Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC (2012) J Comput Aided Mol Des 26:749–773
28. Corbeil CR, Williams CI, Labute P (2012) J Comput Aided Mol Des 26:775–786
29. Repasky MP, Murphy RB, Banks JL, Greenwood JR, Tubert-Brohman I, Bhat S, Friesner RA (2012) J Comput Aided Mol Des 26:787–799
30. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45:177–182
31. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) J Chem Inf Model 52:1757–1768
32. Schärfer C, Schulz-Gasch T, Hert J, Heinzerling L, Schulz B, Inhester T, Stahl M, Rarey M (2013) Chem Med Chem 8:1690–1700
33. Wu K (2005) J Phys: Conf Ser 16:556–560
34. Wu K, Ahern S, Bethel EW, Chen J, Childs H, Cormier-michel E, Geddes C, Gu J, Hagen H, Hamann B, Koegler W, Lauret J, Meredith J, Messmer P, Otoo E, Perevoztchikov V, Poskanzer A, Rübel O, Shoshani A, Sim E, Stockinger K, Weber G, Zhang Wming (2009) J Phys Conf Ser 180:1
35. Kirchmair J, Ristic S, Eder K, Markt P, Wolber G, Laggner C, Langer T (2007) J Chem Inf Model 47:2182–2196
36. Schulz-Gasch T, Schärfer C, Guba W, Rarey M (2012) J Chem Inf Model 52:1499–1512
37. Hilbig M, Urbaczek S, Groth I, Heuser S, Rarey M (2013) J Chem Inform 5:38
38. Barber CB, Dobkin DP, Huhdanpaa H (1996) ACM Trans Math Softw 22:469–483
39. Böhm HJ (1994) J Comput Aided Mol Des 8:243–256
40. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD (2012) Drug Discov Today 17:1270–1281
41. Zinc clean leads (2012) UCSF Univerity of California, San Francisco. http://zinc.docking.org/subsets/clean-leads. Accessed 7 Dec 2012