Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

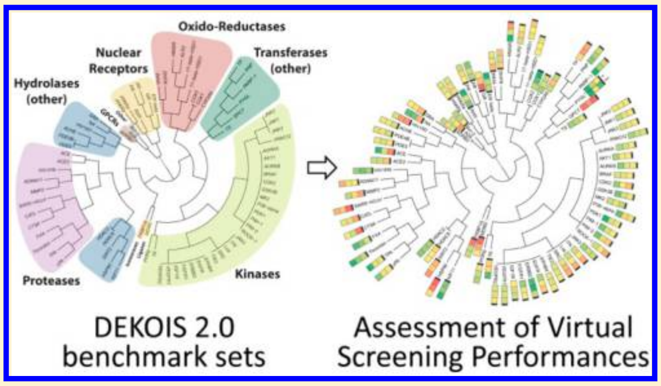
Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets

Matthias R. Bauer,[‡] Tamer M. Ibrahim,[‡] Simon M. Vogel, and Frank M. Boeckler*

Laboratory for Molecular Design and Pharmaceutical Biophysics, Department of Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Eberhard Karls University Tuebingen, Auf der Morgenstelle 8, 72076 Tuebingen, Germany

S *Supporting Information*

ABSTRACT: The application of molecular benchmarking sets helps to assess the actual performance of virtual screening (VS) workflows. To improve the efficiency of structure-based VS approaches, the selection and optimization of various parameters can be guided by benchmarking. With the DEKOIS 2.0 library, we aim to further extend and complement the collection of publicly available decoy sets. Based on BindingDB bioactivity data, we provide 81 new and structurally diverse benchmark sets for a wide variety of different target classes. To ensure a meaningful selection of ligands, we address several issues that can be found in bioactivity data. We have improved our previously introduced DEKOIS methodology with enhanced physicochemical matching, now including the consideration of molecular charges, as well as a more sophisticated elimination of latent actives in the decoy set (LADS). We evaluate the docking performance of Glide, GOLD, and AutoDock Vina with our data sets and highlight existing challenges for VS tools. All DEKOIS 2.0 benchmark sets will be made accessible at http://www.dekois.com.



■ INTRODUCTION

In silico screening methods help to select and identify promising new lead structures.[1] They have become valuable and frequently used tools in the field of drug discovery.[2] A multitude of success stories have been reported for the application of *in silico* screening tools in structure-based drug discovery.[3−6] Among these, especially protein−ligand docking tools have emerged as a useful alternative to common HTS screening.[7] However, the success of the numerous available virtual screening (VS) methods depends on the molecular target.[8,9]

In order to select the most suitable tool and to optimize the general VS workflow, methods that objectively assess the performance are essential. Standard methods to evaluate the design and screening performance of VS experiments have been proposed and discussed previously.[10−13] The purpose of these benchmarks is to evaluate the ability to rank known bioactive compounds above inactive decoy structures. To this end molecular benchmarking sets, also often referred to as decoy sets, have been composed in different manners.[14−17] It has been previously shown that benchmark sets have to satisfy certain quality criteria to avoid artificial enrichment bias.[11,12,18] Particularly, when physicochemical properties between active and decoy structures are not well matched or when the presence of potentially bioactive structures in the decoy set is not prevented (LADS = latent actives in the decoy set), such issues are commonly found to bias the actual screening performance of docking tools.[9]

For quite some time the Directory of Useful Decoys (DUD) has been the state-of-the-art benchmark set compilation comprising a total of 40 protein targets.[19] To further improve the quality of decoy sets, different methodologies and new data sets were introduced. Rohrer et al. have generated their Maximum Unbiased Validation (MUV) data sets with a spatial statistics approach using Pubchem HTS bioactivity data.[20,21] Wallach et al. presented their virtual decoy sets consisting of *in silico* generated decoy molecules.[22] With our demanding evaluation kits for objective *in silico* screening (DEKOIS) we have recently introduced an automated, fast, and reliable tool for the generation of decoy sets. This protocol provides a balanced decoy selection procedure allowing for co-optimization of active-decoy physicochemical similarity and LADS avoidance.[9] Gatica et al. introduced ligand and decoy sets for 147 GPCR targets.[23] Recently, Mysinger et al. provided an expansion of available benchmarking sets for their Directory of Useful Decoys (DUD-e).[24] Their enhanced methodology allows for chemotype diversity optimization of ligand sets, charge matching between decoys and actives, and an improved elimination of false decoys.

Received: February 19, 2013
Published: May 25, 2013

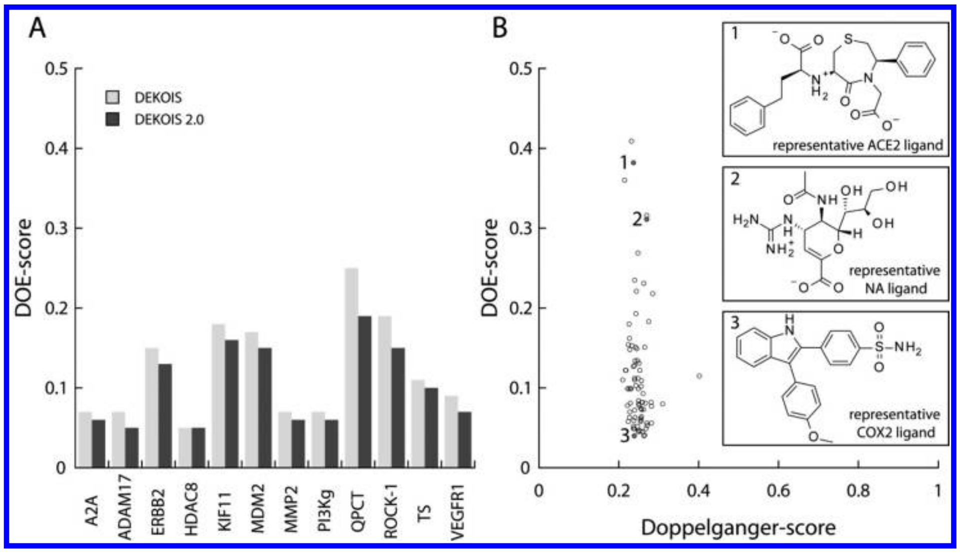


Figure 1. DEKOIS 2.0 quality assessment of physicochemical matching (DOE score)[9] and latent active avoidance (Doppelganger score).[9] (A) Comparison of DOE scores between selected DEKOIS (gray) and DEKOIS 2.0 (dark gray) data sets. (B) Scatter plot of DOE score versus Doppelganger score for all 81 DEKOIS 2.0 benchmark sets. DOE scores of below 0.1 indicate a close to optimal embedding of actives within decoys in physicochemical space, while values close to 0.5 indicate strong local clustering of actives without proper embedding with similar decoys. The Doppelganger score is based on the average of Tanimoto-based FCFP_6 similarities between mutually exclusive actives and their closest structural mimic in the decoy set. Although it cannot be assumed automatically that higher levels of shared fingerprints signify bioactivity of the decoys, the risk that molecules in the decoy set contain features that are sufficient to cause bioactivity is significantly increased. The depicted structures exemplify typical actives of the three highlighted data sets (COX2, NA, and ACE2), giving examples of the required array of physicochemical properties that suitable decoys need to match for achieving a good embedding.

In practice, the application of benchmark sets provides various benefits for the user. For instance, benchmarking can aid the development of scoring functions and optimization of screening tools.[24−26] Also the selection of multiple conformations for a target structure to improve screening performance can be guided by the application of decoy sets.[27,28] Benchmark sets can also help with the receptor optimization and validation of homology models.[29] The assessment of docking performance can ultimately help to find an efficient virtual screening workflow for a specific target. By selecting the most suitable tool, structure, or set of structures and preparation procedures, VS hit rates can be optimized. The costs of compound synthesis and testing of possibly inactive VS hits are relatively high compared to the time and effort needed to select an efficient VS approach.

So far decoy set repositories have enabled benchmarking only for a limited selection of targets. A systematic and substantial expansion of publicly available high-quality data sets is needed. The optimization of VS parameters for interesting, new targets can also help to identify new lead structures for these cases more efficiently. Eventually, by increasing the variety of binding sites and ligand classes, the predictive power and statistical significance of VS benchmarking experiments will be enhanced.

Public ligand databases like the Pubchem Bioactivity database, the BindingDB, and ChEMBL allow for the extraction and processing of bioactivity data for a wide variety of targets.[30−35] These fast growing libraries are an excellent starting point for compiling new ligand sets. However, the data have to be further processed and curated to avoid introducing potential bias, such as the following: (1) substructures that are reported to frequently cause false positive results in HTS assays, (2) reactive compounds that bind covalently to proteins, (3) neglecting diversity due to active analog bias, (4) accepting ligands with negligible bioactivity as actives, and (5) ligand types binding to different sites of the same target. All these issues should be addressed.

Our automated DEKOIS workflow, that allows for fast and reliable generation of high-quality decoys for any kind of ligand set, is an expedient tool to make use of such curated bioactivity data.[9] Moreover, we have improved the DEKOIS methodology. The optimization of physicochemical similarity between actives and decoys now includes three additional physicochemical properties: number of negative charges, number of positive charges, and number of aromatic rings. Especially neglect of charge property matching between actives and decoys was found to bias VS benchmarks.[36] The removal of LADS was further fine-tuned to eliminate potential bioactives more efficiently. To reduce analogue bias in the active set and provide a basis for chemotype enrichment experiments, we additionally optimize the structural diversity of actives.[37,38]

RESULTS AND DISCUSSION

Improved DEKOIS 2.0 Workflow. We have previously shown that an unsuitable selection of decoys introduces bias to VS benchmarks.[9] The decoy selection procedure should minimize any potential bias to obtain high-quality data sets suitable for objective *in silico* screening benchmarks. The DEKOIS methodology ensures a well-balanced matching of physicochemical properties between actives and decoys to avoid artificial enrichment and filtering of latent actives in the decoy set to prevent artificial reduction of the screening performance.

To further improve the matching of physicochemical properties between actives and decoys, we modified our workflow to consider additional complementary properties. Previously we characterized compounds by their (1) molecular weight, (2) logP, (3) number of hydrogen bond donors, (4) number of hydrogen bond acceptors, and (5) number of rotatable bonds. We now added for the physicochemical

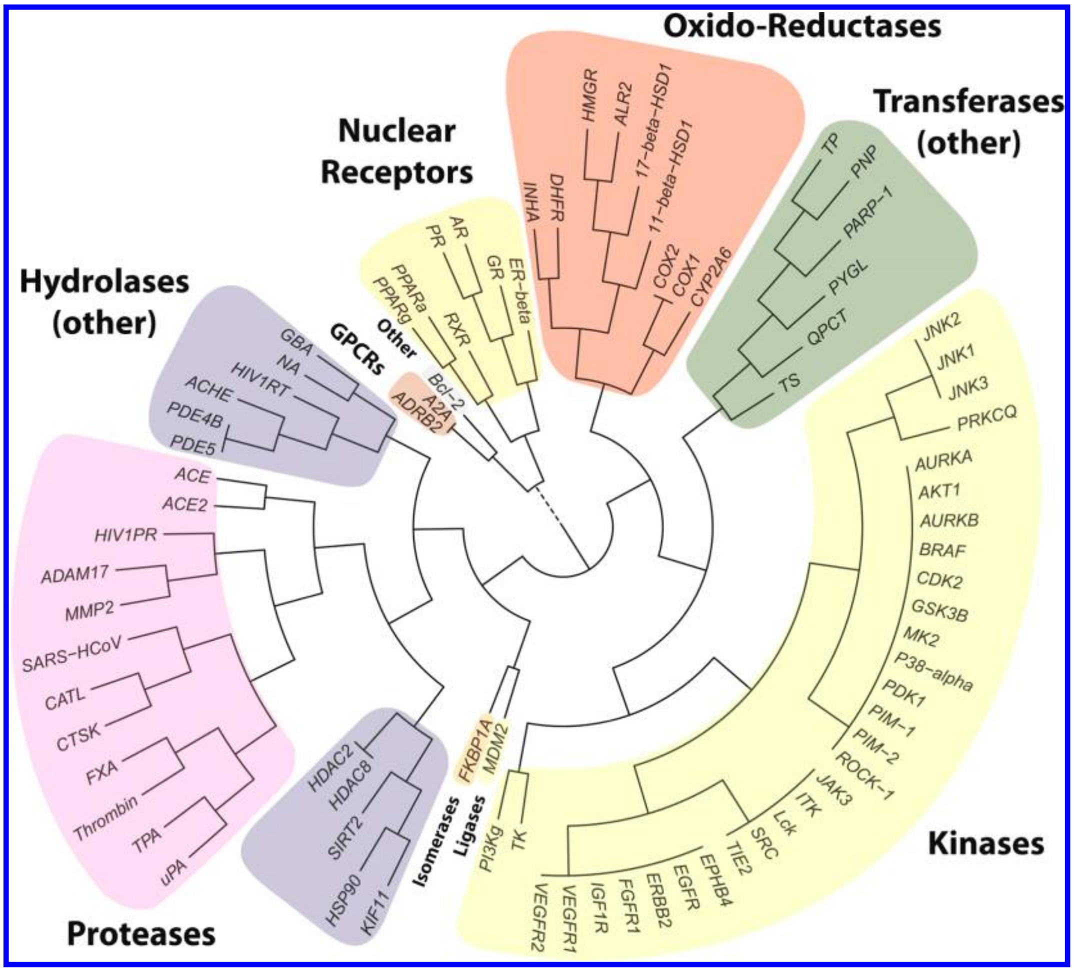


Figure 2. DEKOIS 2.0 target overview. 80 protein targets are clustered according to their enzyme commission numbers (EC) and depicted as polar dendrogram. The dashed line indicates a part of the dendrogram for which virtual EC numbers are used. Important enzyme groups are highlighted in the figure.

similarity optimization: (6) population of negatively charged states, (7) population of positively charged states, and (8) number of aromatic rings. We decided to differentiate between the occurrence of positive and negative charges in the compounds, because summation over all opposite charges would suggest that multiply charged ligands containing the same number of opposite charges could correspond to completely uncharged decoys. To consider also the ability of compounds to form diverse aromatic interactions, the number of aromatic rings was considered for the optimization as well. We had previously introduced the *deviation from optimal embedding* score (DOE score) to assess the quality of physicochemical matching between actives and decoys.[9] A detailed description of this metric can be found in the Methods section. Lower DOE scores correspond to better physicochemical matching between actives and decoys. We compared the DOE scores of the previous DEKOIS protocol and the improved DEKOIS 2.0 protocol for 12 randomly selected benchmark sets. We found that the already good physicochemical matching of the former DEKOIS protocol[9] was improved for the DEKOIS 2.0 protocol (Figure 1a).

Figure 1b shows the DOE scores of all 81 DEKOIS 2.0 data sets. Most benchmark sets possess relatively low DOE scores: We found an almost optimal decoy embedding for 57% (DOE scores <0.1) and good DOE scores for 89% of the DEKOIS 2.0 benchmark sets (DOE scores <0.2). However, for few data sets the physicochemical matching between actives and decoys is challenging. For instance, the ACE2 active set contains highly charged and rather flexible ligands (a representative ACE2 inhibitor is shown in Figure 5b). Certain combinations of features are, of course, less likely to occur frequently in the compound databases that we employ for selecting our decoys. Consequently, the available chemical space for selecting physicochemically similar decoys is considerably more restricted. Additionally, certain "bioactive substructures" should be avoided to reduce the risk of latent actives in the decoy set (LADS filter). In depth analysis of the ACE2 set revealed that particularly a strong mismatch of the number of negative charges (2.37 for the active set and 0.31 for the decoy set) caused the dramatic increase in DOE score. Still, all other features are quite reasonably matched between actives and decoys.

A different example for a challenging physicochemical property matching is the Neuraminidase (NA) benchmark set. The NA actives are unusually polar and charged (Zanamivir is shown as a representative NA ligand in Figure 1b). Again only a limited number of physicochemically similar structures are available for the decoy selection. In contrast to ACE2, where the mismatch was focused on only one property, here we found smaller deviations for logP (average values of actives: −2.37 and decoys: −0.95), number of hydrogen bond donors (actives: 4.67/decoys: 3.49), number of aromatic rings (actives:

0.35/decoys: 1.15), and the number of negative charges (actives: 1.05/decoys: 0.47).

Contrary to these problematic cases, the COX2 benchmark set is a good example for an almost ideal matching of the eight molecular properties between actives and decoys. The ligands in the COX2 active set did not possess unusual properties or molecular features (a representative COX2 ligand is shown in Figure 1b). For every COX2 ligand our decoy selection protocol could easily provide structures that are very close in physicochemical space, while avoiding bioactive substructures.

To avoid latent actives in the decoy set (LADS), we had previously employed a LADS filter within our DEKOIS methodology. In this work we modified the LADS filter to also consider frequency and size of bioactive ligand substructures that were represented by FCFP_6 (Scitegic) fingerprint bit strings. The risk of introducing potential bioactive compounds into the decoy set becomes substantially higher with an increasing overlap of substructures between actives and decoys. Additionally, the frequent occurrence of certain substructures in the active set must be considered an important factor for the assessment of their bioactivity. With lower frequencies of bioactive fragments in the ligand set, it is likely that these substructures will contribute less to bioactivity than those found more often. A comprehensive and structurally diverse selection of known ligands also facilitates this evaluation of essential substructures for bioactivity. We describe the detailed procedure of our enhanced LADS filter in the Methods section.

We previously introduced the Doppelganger score – a metric which assesses the extent of structural similarity (Tanimoto similarity with FCFP_6 fingerprints) between actives and their most structurally related decoys.[9] The Doppelganger score helps to identify problematic decoy selections within the generated benchmark sets. The methodology of this metric is described in more detail in the Methods section. In general, we found relatively low scores (0.2–0.3 Tanimoto similarity) for the DEKOIS 2.0 benchmark sets, indicating a well working LADS filter (Figure 1b). A comprehensive overview of all Doppelganger scores and DOE scores for all sets can be found in Table S1.

Target Overview and Structure Selection. There are certain requirements for providing meaningful molecular benchmarking sets for structure-based VS. First, a well curated and characterized set of ligands, also often referred to as actives, has to be compiled. Second, decoy structures have to be selected based on previously mentioned quality criteria, and, finally, one (or more) well-suited 3D structure is needed to model the ligand binding site. These essential requirements confine the eligible targets for benchmark set generation. We aimed for compiling a diverse target selection and added selected targets with a high relevance for current drug discovery efforts. Such targets of high priority were chosen based on an analysis of the most read articles of the *Journal of Medicinal Chemistry* in 2010.

The DEKOIS 2.0 library includes 81 new high quality benchmark sets for 80 protein targets (2 benchmark sets for different PYGL binding sites). Our library comprises a diverse selection of protein classes like kinases, proteases, nuclear receptors, GPCRs, oxido-reductases, transferases, hydrolases, and several others. We clustered proteins according to their enzyme commission numbers to visualize their degree of functional relatedness (Figure 2). Proteins that did not possess an EC code were clustered by "virtual EC numbers". Apart from the six main enzyme families that are described with top level codes from one to six and subsequent subclassifications, we assigned different virtual top level codes for nonclassified proteins. We describe this procedure in more detail in the Methods section.

To better understand the various advantages and disadvantages of VS approaches for specific test cases, it is important to assess and describe target characteristics. For instance, many classical targets like GPCRs, kinases, and proteases possess a well-defined and preformed concave binding site. This feature can facilitate drug discovery for these targets, as existing methods, including VS tools, and compound libraries have been successfully developed and refined for "druggable" target classes.[39−41] However, there are also examples of drug discovery efforts that focus on nonclassical binding sites. Especially targeting protein–protein interactions (PPI) has become an interesting strategy to modulate cell regulation.[42] The generally large, flat, and featureless PPIs require small molecule inhibitors that are in general larger and more lipophilic than ligands binding to classical binding pockets.[40] Such unusual target properties are also for VS approaches a new challenge. Considering the new and interesting implications this new approach can have for the development and selection of VS tools, we included benchmark sets for protein–protein interaction targets (MDM2 and BCL-2) in our DEKOIS 2.0 library. This example illustrates the importance of providing more benchmark sets for nonclassical target types to further investigate and optimize the efficiency of existing tools for these cases.

The RCSB Protein Data Bank (PDB) comprises high-quality X-ray and NMR structures for a wide variety of proteins.[43] For many targets several structures are available. It is essential to pick a suitable target structure or set of structures to conduct successful structure-based VS. Molecular benchmark sets can also aid with the selection of a well-suited structure by assessing the molecular recognition between ligands and their binding site in terms of enrichment performance.[24] Ideally, the selected structure should allow for the recognition of all ligands. The conformation of the binding pocket is essential in this context and should be inspected very carefully before performing a VS benchmark. To some extent induced-fit or conformational selection mechanisms can be observed upon ligand binding.[44] This leads often to considerable changes in binding pocket conformations. Even small changes in the atomic positions of the binding pocket can sometimes prevent the formation of important interactions or create steric clashes with docked ligands.[45] In the worst case some or all ligands of the active set cannot be recognized by a specific conformation of the target's binding site. Also the resolution of structures derived from experimental X-ray diffraction data should be taken into consideration. A better resolution, implying an enhanced accuracy of the electron density, should facilitate the correct fitting of protein residues and ligands to this electron density and reduce the chance of erroneous placement of atoms in the reported structure.

For the DEKOIS 2.0 VS benchmark we selected suitable structures for the docking of 81 benchmark sets. With respect to the issue of alternative binding sites, we ensured to select only target structures that contain the confirmed binding site for the respective active set. Structures in complex with small molecule inhibitors, that were structurally similar to the ligands in the active set, were prioritized. If several suitable structures were found, we chose the structure with the best resolution. In

Table 1. Docking Enrichment Comparison between Standard and "Large" DEKOIS 2.0 Benchmark Sets

benchmark set	GLIDE pROC AUC	GLIDE ΔpROC AUC [%]	GOLD pROC AUC	GOLD ΔpROC AUC [%]	VINA pROC AUC	VINA ΔpROC AUC [%]
A2A	0.828		0.599		0.433	
A2A large	0.888	7.3	0.692	15.5	0.435	0.4
FXA	1.643		1.305		0.831	
FXA large	1.754	6.8	1.263	3.2	0.794	4.5
VEGFR2	0.892		0.856		0.576	
VEGFR2 large	0.896	0.5	0.814	4.9	0.650	12.8

cases where target structures with small molecule ligands were not available, we also considered structures of the apo proteins. All selected structures with their respective PDB codes are reported in Table S2.

Selection of Active Ligand Data Sets. In order to compile active sets, a suitable source for bioactivity data is needed. We chose to extract this data from the BindingDB, a database that provides ligand affinities for a wide variety of protein targets, focusing mainly on candidate drug targets.[33] We compiled 81 different active sets by employing an automated ligand selection protocol in Pipeline Pilot.[46] Each generated active data set contains information about the ligand structure (SD format), the BindingDB monomer ID, binding affinity data (K_i, K_d, or IC_{50} values), the respective Uniprot target ID, and the Uniprot source organism tag. With our ligand selection protocol, we strived to address various issues that can be found in unprocessed bioactivity data.

Usually a broad range of reported ligand affinities can be found in target specific bioactivity data. However, it is difficult to decide whether all provided structures can be considered bioactive for the respective target. Some HTS screening hits may exhibit only weak and few interactions with the target binding site resulting in a low binding affinity. Compared to inhibitors with substantially higher target affinity, their bioactivity status remains questionable. It is also harder to predict correct binding modes of ligands by docking, when they possess only few and weak interactions with the binding site.[47] On the other hand, decoys exhibiting structural characteristics that typically lead to bioactivity or that form strong interaction patterns with the binding site can cause problems and should therefore be avoided. Yet, due to their similar physicochemical properties, decoys will necessarily have a tendency to contain similar "chemistry" and, thus, exhibit some interactions with the respective binding pocket. The ranking of marginally bioactive ligands above decoys by VS tools is difficult, as their weak interactions with the binding site do not clearly distinguish them from the inactive decoy structures. To facilitate a clear differentiation between actives and decoys, we decided to exclude weak binders from the selection of the ligand set. However, there are no uniformly applicable bioactivity cutoffs. Instead, relative cutoffs should be used, reflecting that binding affinity of the most potent compound classes can substantially fluctuate between targets. Consequently, we removed compounds possessing a 1000-fold weaker target affinity than the most potent inhibitor for the respective target. This should account for the variability of "bioactivity ranges" between different targets.

VS tools should not be limited to only recognize and rank specific bioactive scaffold classes above inactive decoys. The recognition of a wide variety of structurally different ligand classes is an important and desired quality and should be a good indicator for the chance to successfully identify new hits by VS. The bioactivity of structurally very similar actives can be likewise be predicted by simple 2D methods, and the additional complexity of advanced VS methods might be in vain.[38] To facilitate the evaluation of this important parameter, we optimized the structural diversity of the DEKOIS 2.0 active sets. Starting from FCFP_6 Tanimoto similarities, we clustered ligand structures into 40 structurally diverse classes. From each cluster, we selected the most potent compound. By this approach we maximize the diversity of chemotypes and reduce potential analog bias. We observed that the more homogeneous distribution of actives within the physicochemical property space, resulting from this optimization of chemotype diversity, facilitates the selection of physicochemically similar and structurally diverse decoys. However, the optimization of structural diversity depends strongly on the extent of structural diversity found in all available bioactive ligands. Especially for newer targets, often only few different scaffold classes have been reported. Employing smaller active sets can help to provide a structurally diverse selection of ligands even for these cases. Additionally, previously discussed issues in the bioactivity data can be more easily identified and avoided in this case. Based on statistical considerations, we generated active sets of identical size for all targets.

To investigate the impact of this restriction of active compounds to 40 diverse ligands for each target, we alternatively compiled larger benchmark sets for A2A, FXA, and VEGFR2, for which a large amount of bioactivity data were available. For these sets, we employed our standard ligand and decoy selection protocols, except for the selection of five ligands (or less if the cluster did not contain at least five structures) instead of one ligand (standard protocol) from each structurally diverse cluster. On average the "large active sets" contained about 4.5 times the number of ligands than the regular sets, featuring the same ratio of actives to decoys. Depending on the used docking tool, our benchmark results showed rather small changes in pROC AUC values between the standard and large benchmark sets. The total averages for the ΔpROC AUC values were found to be 4.9, 7.9, and 5.9% for Glide, GOLD, and AutoDock Vina, respectively. It should be noted, that also the heuristic nature of the docking process itself introduces fluctuations in the screening performance.[9] In addition, the profiles of the pROC curves for the standard and large data sets are quite similar (Supporting Information Figure S1). We conclude that the use of smaller-sized diverse active sets with only 40 ligands for the generation of benchmark sets yields results comparable to benchmark sets with substantially larger active sets (see Table 1).

A significant part of the reported bioactivity data, provided by academic and industrial research, is derived from high-throughput-screening (HTS) assays. However, without further confirmation of this primary assay data, interference of specific compound properties with the respective assay can lead to

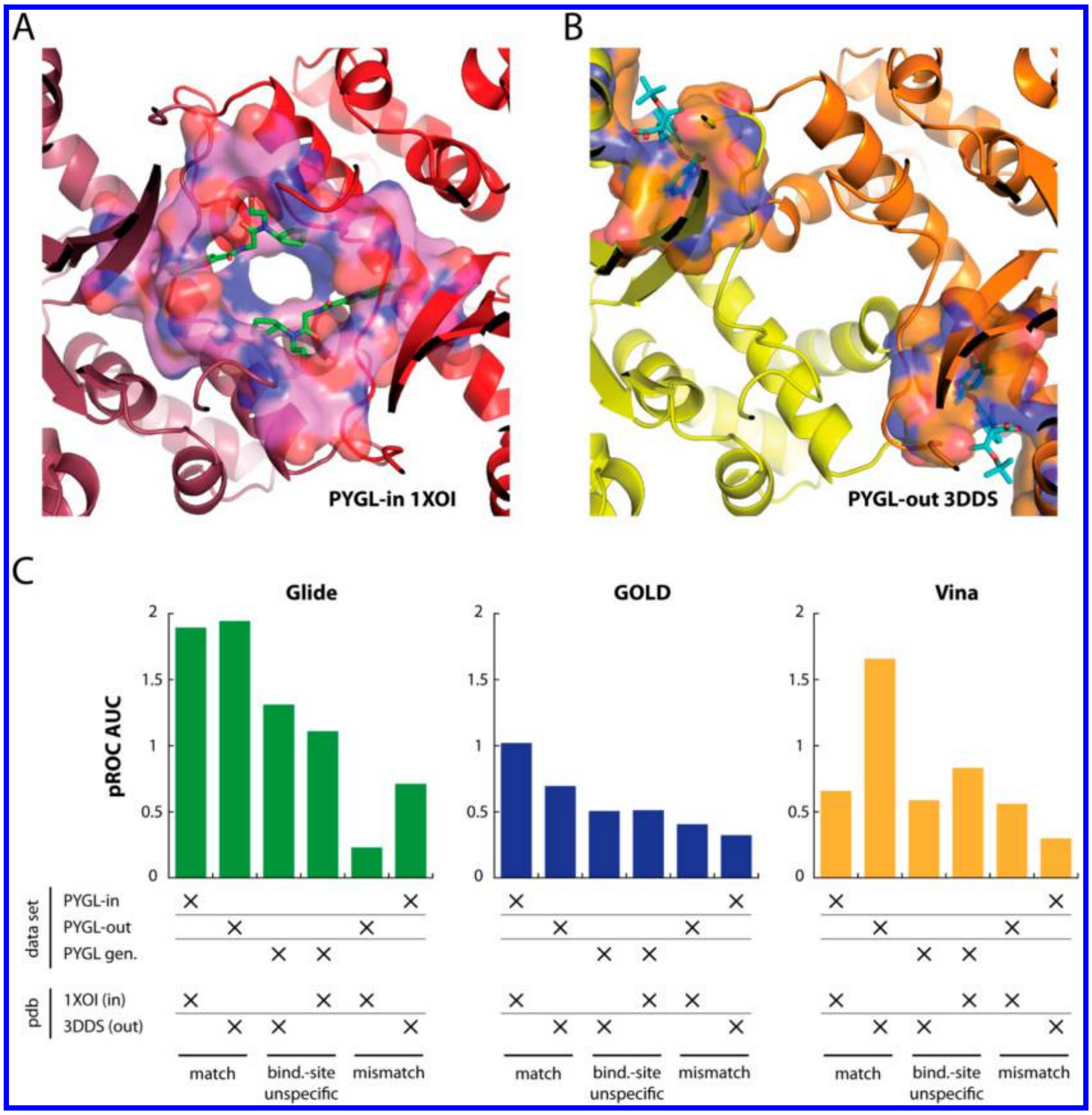


Figure 3. PYGL ligand binding sites. (A) Glycine amid inhibitors (green sticks) in complex with the PYGL homodimer (PDB code 1XOI). (B) Ligands (cyan sticks) bound to the alternative AMP binding site of the PYGL homodimer (PDB code 3DDS). The surface of binding pocket residues are shown in (A) and (B). (C) Bar charts of pROC AUC enrichment results for Glide, GOLD, and AutoDock Vina using a general or binding site specific PYGL benchmark sets.

errors in the reported bioactivities or even false positive entries. Baell et al. investigated which substructures are frequently found in so-called pan assay interference compounds (PAINS). These features can help to filter frequent false positives from biochemical HTS bioactivity data.[48] In line with these findings, we integrated a substructure filter based on PAINS in our ligand selection procedure to minimize the risk of false positives in our DEKOIS 2.0 active sets (Table S6).

A further issue in bioactivity data is the ambiguity of nonspecified stereo centers in ligand structures. For some compounds that are provided by bioactivity databases no explicit stereo configuration can be retrieved. Frequently, one stereo configuration of the ligand is significantly more bioactive. The reported bioaffinities can result from a mix of different stereoisomers or just from one compound with a certain configuration. This impairs the utility of these structures for molecular benchmarking as it is unclear if both or only one specific stereoisomer should be considered for the docking experiment. Without experimental evaluation, the inclusion of compounds with no specific stereo configuration remains a potential bias for docking benchmarks. Therefore, we discarded all structures containing nonspecified stereo centers in our ligand selection protocol.

Special Cases of Active Data Set Selection. To provide meaningful benchmark sets, it is important to ensure that all selected actives bind to the same binding site of the protein target. In case there is only one known and well-described ligand binding site for a target, the bioactivity data found in public databases should be nonambiguous. However, for some targets, several ligand binding sites are known. Apart from protein active sites, other potential binding pockets, like allosteric and protein−protein interaction sites, are sometimes targeted with small molecule inhibitors as well. As a consequence, compounds binding to different sites of the respective protein are usually not differentiated by their binding site but likely to be blended together into target-specific data sets. This ambiguous data can heavily bias the benchmark set, as ligands can only be evaluated in a meaningful way with regard to their binding pocket. Human liver glycogen phosphorylase (PYGL) is a good example for a protein with

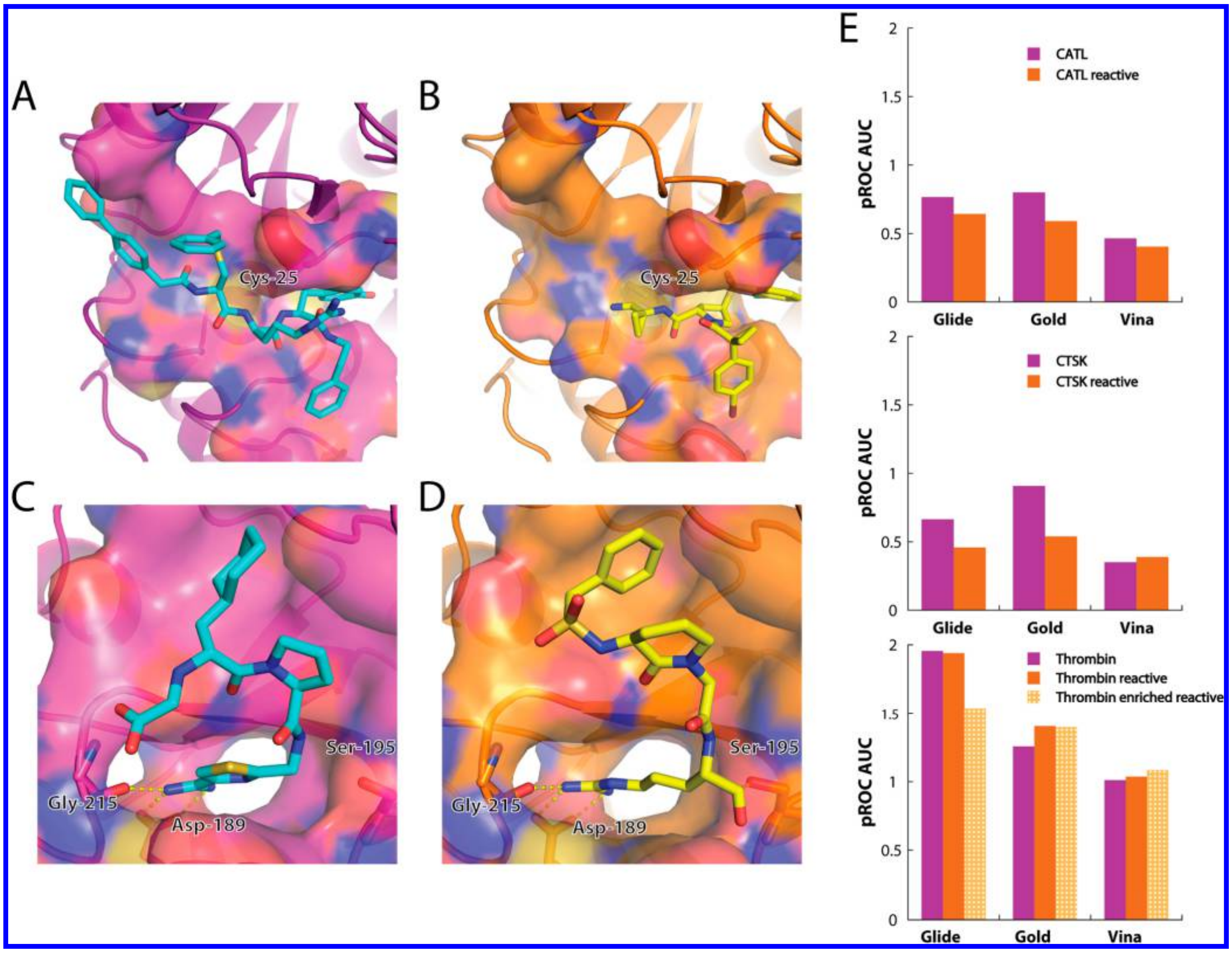


Figure 4. Binding modes of covalent and noncovalent Cathepsin and Thrombin inhibitors. Crystal structures of a noncovalent (A) and covalent (B) interaction with CATL and its catalytic residue Cys-25 are depicted (PDB codes 3BC3 and 2YJ2, respectively). The noncovalent Thrombin inhibitor with an amidine function (C) forms a strong salt bride with Asp-189 (PDB code 2FEQ). The Thrombin ligand (D) forms a covalent bond with the catalytic residue Ser-195 and additionally interacts by its guanidine function with Asp-189 (PDB code 1BA8). The bar chart (E) compares the pROC AUC docking enrichment results of benchmark sets containing no reactive ligands with sets containing reactive ligands in their active sets. Results were obtained for Glide, GOLD, and AutoDock Vina.

multiple binding sites. Glucose and other sugar derivatives bind to its active site, and other inhibitors target allosteric binding sites in order to stabilize the inactive conformation of the enzyme. In the BindingDB data, we found mainly inhibitors for the allosteric AMP site and a more recently described allosteric binding site formed by the homodimer interface (see Figure 3).[49] Based on reported literature data, we employed several substructure filters to separate bioactive structures according to their respective binding sites and compiled binding site specific ligand sets. We generated a "PYGL-in" benchmark set for the inner, allosteric binding site (PDB code 1XOI) and a "PYGL-out" benchmark set for the allosteric AMP binding pocket (PDB code 3DDS). To test the impact of binding site unspecific bioactivity data, we additionally compiled a "PYGL-general" benchmark set, which was derived from the initial, unclassified bioactivity data using our standard ligand selection protocol. We found 27 inhibitors targeting the new allosteric binding site (PYGL-in), 12 structures targeting the AMP allosteric binding pocket (PYGL-out), and 1 aminosugar targeting the enzyme active site in the ligand set of the "PYGL-general" active set.

The bar charts in Figure 3 show the VS performance of the docking tools Glide, GOLD, and AutoDock Vina for the PYGL benchmark sets. VS performance was quantified by the pROC AUC,[13] a measure of docking performance that focuses on early enrichment. As expected, the best screening results were obtained for docking the binding site specific benchmark sets against their respective binding sites. The performance for the "PYGL-general" ligand set, containing compounds binding to different allosteric sites, yielded lower enrichment performances for all three VS tools. For 1XOI, pROC AUC values were strongly impaired by 0.58 and 0.51 for Glide and GOLD, respectively, while with Vina only a marginal reduction of 0.07 was found. For 3DDS, the decrease in pROC AUC values was more substantial for Glide (0.83) and Vina (0.81) than for GOLD (0.18). Furthermore, a cross docking experiment into the mismatched binding sites using the site-specific benchmark sets resulted in even lower docking performances for all programs. These findings indicate the importance of creating

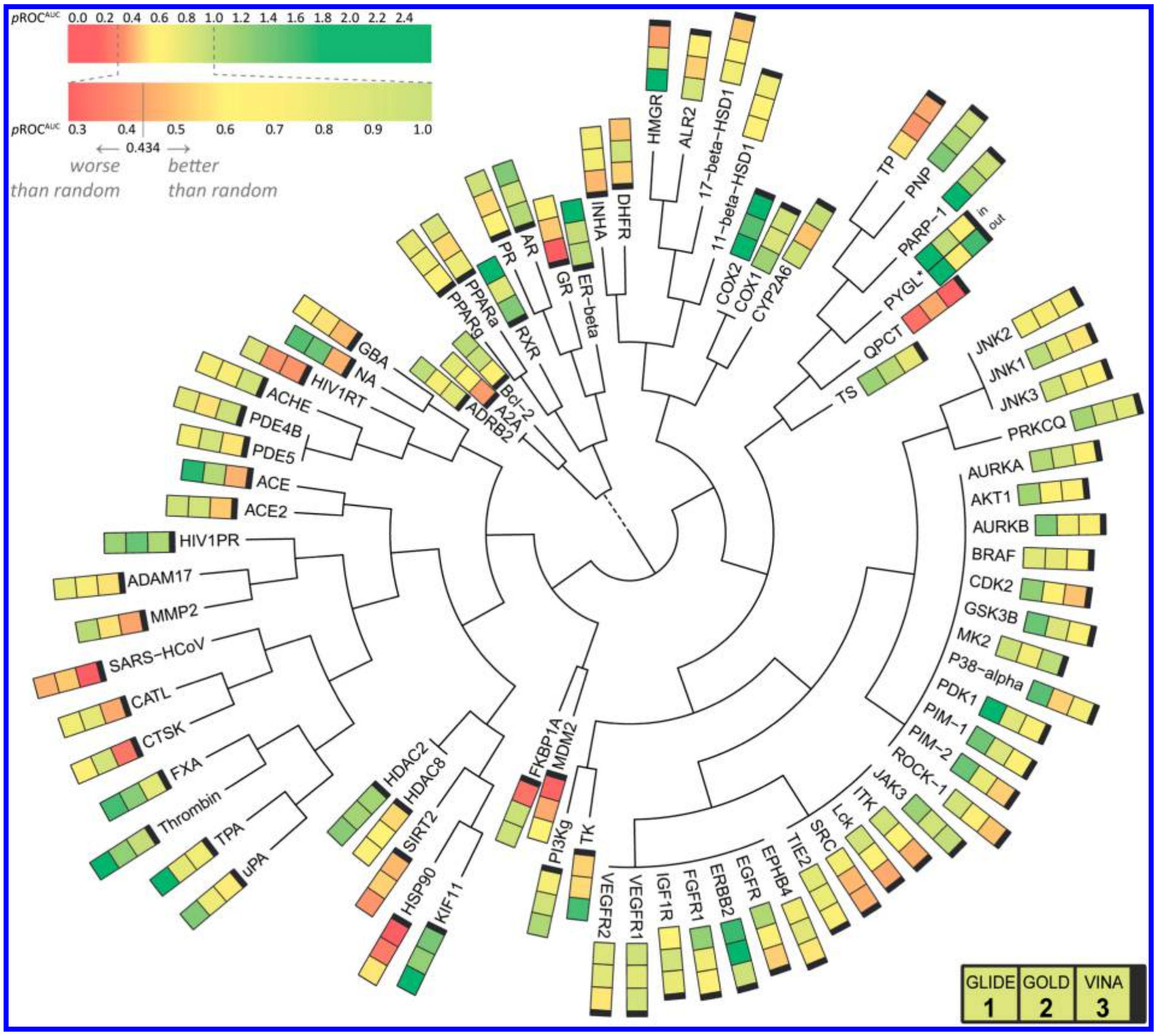


Figure 5. DEKOIS 2.0 target overview with pROC AUC enrichment results for Glide, GOLD, and AutoDock Vina. Docking calculations were conducted at default settings of the respective program. However, the computational costs for docking were substantially lower in GOLD than in Glide or Vina (Table 2). All targets are clustered by their EC codes and shown as polar dendrogram, as previously described. The calculated pROC AUC values were color-coded. A red color corresponds to a negative docking enrichment performance, orange and yellow to a random performance, and green to a reasonable or good screening performance. To facilitate the interpretation of color shades, we provide a scale for the pROC AUC values.

binding site specific ligand sets, as active sets compiled from unclassified ligand data would artificially decrease and bias the real screening performance.

Another issue arises from ligands that covalently modify the target protein or mimic reaction intermediates. Standard docking procedures cannot automatically recognize covalent binding modes resulting from the specific reactivity of ligand substructures and protein residues. Pose and scoring predictions of VS tools are usually impaired. Special program packages or protocols designed for covalent interactions have to be applied to yield meaningful predictions of poses or binding affinities.[50,51] This involves manually defining reactive groups and residues that form a covalent bond. However, comparing docking scores of covalently docked actives and noncovalently docked decoys leads to an intrinsic bias. The covalent binding mode strongly predetermines the orientation of molecules and contributes intrinsically to the docking score of actives. Meaningful benchmark results cannot be obtained.

Subsequently, we investigated the impact of covalent ligands on the screening performance of standard docking procedures. We selected three protease targets with reported covalent ligands as test cases. Especially for proteases, many covalent inhibitors can be found due to the generally higher reactivity of active site residues.[52] We compiled "reactive ligand sets" for Cathepsin L (CATL), Cathepsin K (CTSK), and Thrombin according to our standard protocol, except for disabling the reactive group filter. Table S3 gives an overview of reactive substructures known to interact covalently with their respective targets and their occurrences in the generated reactive sets. More than 75% of the CATL and CTSK ligand sets contained such substructures. Most of these reactive inhibitors possessed an electrophilic nitrile "warhead" that can form a covalent thioimidate adduct with the active side residue Cys-25 (Figure 4B). Despite the deactivated reactive group filter, the Thrombin ligand set contained only few reactive compounds. A reason for this could be the high proportion of noncovalent ligands with high affinity for Thrombin. To better estimate the possible impact of covalent binders on the screening performance, we increased the number of reactive inhibitors to compile the "enriched reactive" Thrombin ligand set. Many of these compounds contained aldehyde functions, which most probably form a transition state adduct with Ser-195 (Figure 4D).

We then compared the pROC AUC screening performance of the reactive active sets and their decoys sets with benchmark sets that did not include any covalent binders (Figure 4E). For CATL and CTSK we observed a lower docking performance for the "reactive" data sets in Glide and GOLD (pROC AUC reduction of 0.12 and 0.21 (Glide) and 0.21 and 0.37 (GOLD) for CATL and CTSK, respectively). The screening performance retrieved for reactive and nonreactive benchmark sets in Vina was, in both cases, not better than random. Pose retrieval experiments for CATL/CTSK inhibitors containing nitrile groups showed that the correct ligand binding modes could not be reproduced by docking. Based on the structure of the covalent thioimidate adduct in the crystal structure (sp^2-hybridization of the carbon atom), the linear sp-hybridized nitrile of the noncovalently linked docked actives necessarily clashes with the sulfur atom of Cys-25. This typically leads to much less favorable binding modes and as a consequence also lower docking scores for the reactive ligands.

For Thrombin, a larger change of pROC AUC was only retrieved in Glide, when comparing the standard and "enriched reactive" benchmark set. Otherwise, the Thrombin reactive benchmark set performed slightly better for GOLD or remained similar for Vina. These results were surprising as we expected for all examples a reduction in screening performance. Having a closer look at the reactive thrombin inhibitors, we found that almost all structures contained basic amidines or guanidines, which form an essential salt bridge with Asp-185 (Figure 4C and 4D). It was previously shown that the presence of such basic groups strongly correlates with ligand potency.[53] The orientation of this important amidine or guanidine function in the docked poses was usually found to be correct for both covalent and noncovalent inhibitors. Additionally, the Thrombin binding pocket exhibits enough space to accommodate different orientations of the reactive group avoiding clashes with Ser-195. Thus, the essential salt bridge and other important polar interactions are not disturbed (Figure 4D). These interactions, shared by both covalent and noncovalent Thrombin inhibitors, might explain the small difference in screening performance in our experiment.

Depending on the resulting geometry and steric requirements of the covalent bond, the size of the binding pocket, and also the presence of noncovalently interacting warhead groups, we demonstrated that the inclusion of covalent ligands can affect the screening performance. As a consequence, we decided to exclude all ligands with reactive groups from the DEKOIS 2.0 active sets. Apart from a general filter for highly reactive substructures that can covalently modify proteins, we applied a substructure filter for semireactive groups specifically for protease targets (Table S5). Covalent or transition-state binding modes of substructures were identified by in-depth literature research for all protease targets.

Docking Application. To demonstrate the utility of our DEKOIS 2.0 benchmark sets and to evaluate the impact of VS workflow parameters on the screening performance, we used three widely applied VS docking tools: Glide, GOLD, and AutoDock Vina.

The quality of the screening performance depends strongly on providing reasonably well curated input data (see aspects regarding the selection of actives and decoys, as well as the discussion about choosing suitable target structures) together with a rational setup of the VS workflow.[10,54] In general, we expect that our benchmark sets featuring optimized decoy-embedding of the actives are a demanding challenge for the used docking tools.

Figure 5 shows the individual docking performances as color-coded pROC AUC values of Glide, GOLD, and AutoDock Vina within the previously presented polar dendrogram of 80 DEKOIS 2.0 targets that were clustered by their EC codes. To quantify the individual docking performances, we used the pROC AUC, a semilogarithmic measure that emphasizes early hit recognition.[13] We prepared protein structures using the protein preparation wizard script within Maestro[55] and employed LigPrep[56] for the preparation of actives and decoys. All docking experiments were conducted with default settings of the respective program. The docking setup procedure is described in more detail in the Methods section.

In general, we obtained moderate to good screening performances for most of the DEKOIS 2.0 benchmark sets with only few exceptions. We frequently found substantial differences in the individual screening performance of the employed docking programs for the respective targets. For most kinases our benchmark data revealed similar performance profiles of the used programs. A reason for this observation could be the highly conserved ATP binding site in kinases,[57] which may result in similar compound properties of the respective actives and subsequently yields similar decoy sets. However, the kinase docking performances remained target dependent, even if the profiles of the employed VS tools did not differ as much as we observed for other target classes. This highlights the importance of separately assessing the molecular recognition of actives by the different VS tools, even for closely homologous targets and binding sites prior to virtual screening.

Most of the generated sets have very good physicochemical matching (DOE score <0.1) and show only moderate screening performance (pROC AUC of 0.84 on average). While in some data sets with suboptimal decoy embedding (e.g., HMGR, RXR, NA), some docking tools might benefit from this, in other cases we did not observe any significant difference to average performance (e.g., ACE2, PPARa, PPARg). Despite the high number of quite reasonable results for the large majority of targets, few of the targets obviously bear issues from which useful lessons could be learned. Thus, we take a closer look at benchmark results of targets that showed pROC AUC values less than 0.434 (area under the curve retrieved for "random performance"). We tried to elucidate the reasons for these poor docking results in the case of HIV1RT, HSP90, QPCT, SARS-HCoV, and SIRT2.

High Binding Site Flexibility. For the majority of bioactives, the exact binding mode and binding pocket conformation is unknown. Especially highly flexible proteins can adopt various binding site conformations that can be targeted differently by a multitude of diverse scaffolds.[58] Most likely, not all potential ligand binding modes and protein—ligand interactions can be recognized, if only one structure with a specific binding site conformation is used within the docking procedure. Hence, we expected relatively low screening performances for proteins such as HSP90 and HIV1RT. Indeed, we observed a rather strong impairment of the screening performance for these benchmark sets. The Heat

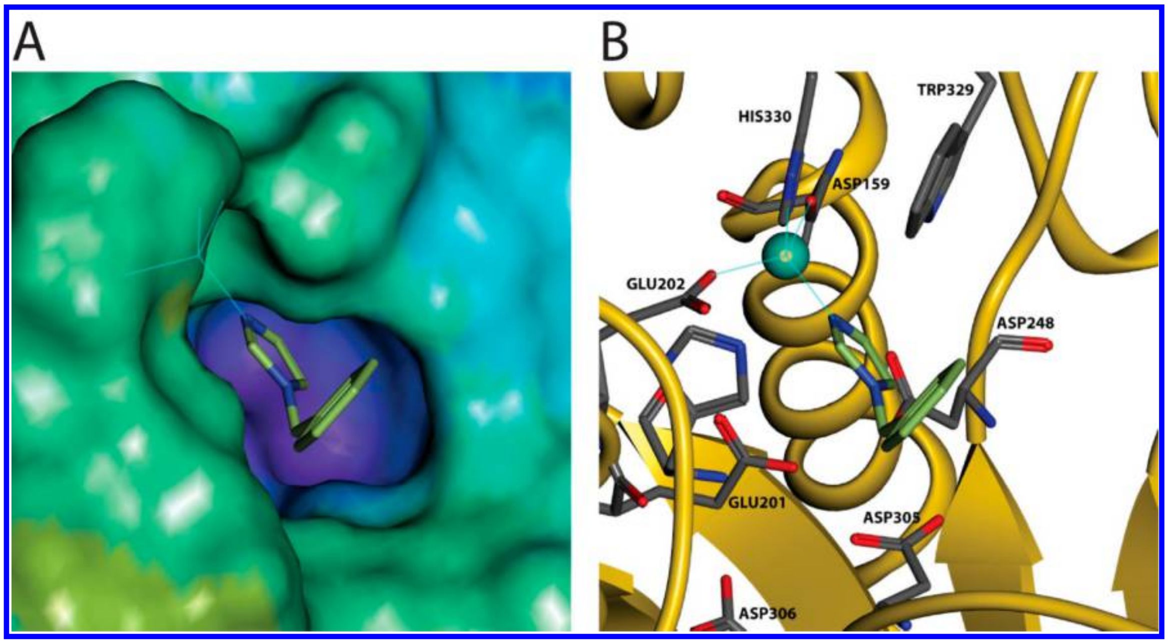


Figure 6. 1BN (1-benzyl-1*H*-imidazole) in complex with QPCT (glutaminyl-peptide cyclotransferase) showing distinct binding site features (PDB code: 2AFX). (A) Electrostatic properties of the binding site of QPCT mapped onto the molecular surface. The color gradient from blue to red indicates areas of negative to positive electrostatic potential. The blue color of the deepest part of the binding site highlights a strongly negative hotspot. (B) Binding mode of 1BN displayed with its proximal protein residues. The imidazole substructure of the ligand (capped sticks) chelates the zinc ion (small cpk sphere). Asp-159, Glu-202, and His-330 coordinate the zinc ion likewise. Glu-201, Asp-305, Asp-306, and Asp-248 form the bottom of the binding pocket and are responsible for the negative hotspot observable in (A). Both pictures were prepared with MOLCAD II.[66,67]

Shock Protein 90 (HSP90) is a chaperone protein that folds and maintains the proper conformation of other client proteins. Its druggable ATP-binding site in the N-terminal domain is reported to exhibit high flexibility and multiple ligand-binding conformations.[59] Also, the HIV-1 Reverse Transcriptase (HIV1RT) is a popular drug target for which substantial crystallographic structural data is available (over 100 structures). Its non-nucleoside reverse transcriptase inhibitor (NNRTI) binding pocket has revealed remarkable plasticity.[60] We presume that the relatively low enrichment performances for both HIV1RT and HSP90 results from high receptor flexibility. Different VS approaches that better describe the receptor flexibility, like ensemble docking, could improve the VS performance for these cases.

High Number of Low Affinity Actives. The affinities for the SARS Coronavirus 3C-like Protease (SARS-HCoV) actives range from 0.06 to 200 μM, with a median of 15 μM. Compared to the ligand potencies of other proteases, the bioactivities of these ligands are remarkably low. Fewer and weaker interactions with binding site residues are to be expected. Subsequently, the task of VS tools to enrich these moderately potent ligands from a pool of inactive decoys becomes naturally more challenging. As expected, we found quite low screening performances for all employed docking programs.

Apo Target Structure. Another example where none of the applied VS tools could obtain good benchmark results was the NAD-dependent deacetylase sirtuin-2 (SIRT2). The inhibition of SIRT2 was recently reported to have neuroprotective function.[61] Narayan et al. also showed that SIRT2 was involved in the regulation of necrosis.[62] This highlights the importance of SIRT2 as a promising target. Due to the lack of structural data for the human ligand-bound SIRT2 structures, we used a human SIRT2 apo structure (PDB: 1J8F) for benchmarking. The ligand binding site was defined as the substrate binding site and the NAD binding pocket residues. The relatively low docking performances might result from two reasons: (1) Sirtuin apo structures were reported to undergo relatively large conformational changes upon binding of the cosubstrate NAD.[63] As the crystal structure of the protein is such an apo form, it may not represent a well-suited receptor conformation for the binding of actives. (2) The reported bioactives for SIRT2 were typically weak inhibitors (IC_{50} varies from: 0.8−243.6 μM, with a median of 53.85 μM).

Accumulation of Unusual Binding Site Features. We noticed that all docking tools showed particularly bad screening performance for the Glutaminyl-peptide cyclotransferase (QPCT) benchmark set. The overlay of six QPCT structures in complex with diverse ligands revealed only small conformational changes of binding site residues, indicating no especially high receptor flexibility. Docking into the two most diverse QPCT structures only caused minor changes the screening performance. The K_i values for the best and worst QPCT active are 2.6 nM and 2.3 μM, respectively. This range of the K_i values, as well as the median of 375 nM, indicates that the awful docking performance is not attributable to insufficient potencies of the used actives. To exclude other typical issues that might impair the QPCT docking performance, we checked for correct protonation states of the privileged imidazole substructure in the active set and the absence of strong zinc chelating groups in decoy structures.

Because none of the previously discussed issues explains the docking results for QPCT, we had a closer look at the binding site requirements for recognizing actives ligands. The QPCT receptor contains a zinc ion that is coordinated by Glu-202, Asp-159, His-330 and typically an imidazole substructure within the bound actives. This deeper part of the zinc-binding pocket is strongly polar due to several negatively charged residues

(Figure 6). The imidazole is a highly privileged substructure of the reported QPCT inhibitors. Thus, despite our efforts to induce diversity in the active set, there is a strong dominance of imidazoles coordinating the zinc ion in almost all structures of the active set. For some but not all of these actives this essential part of the binding mode is correctly recognized in the pose prediction of the docking tools. Apart from this key interaction, we only observed a few additional interactions, such as the NH...π contact between Trp-329 and the aromatic system of the imidazole. It is possible that the scoring functions underestimate the quality of the zinc-imidazole interaction,[64] which is evident from the K_i value of 30 μM observed for the fragment N-methyl-imidazole.[65] In addition, other moieties in the decoy set might be able to efficiently mimic the zinc coordination of the imidazole substructure but are not recognized by our LADS filter, because there are no representatives of such ligand types in the active set. Moreover, the high density of polar and charged residues might allow for alternative ways to address this binding site, which might be an additional source for unrecognizable actives in the decoy set.

Docking Overview. For quite a few targets, Glide exhibits a preferential screening performance. It was previously reported that the Glide SP scoring function was trained and parametrized for enrichment performance as well using molecular benchmark sets.[68] In addition, our molecule and structure preparation procedures prior to docking were generally performed within the Maestro suite.[55,69] Whether this constitutes a bias toward Glide performance remains uncertain. In contrast, the scoring function ChemPLP, which we used for our GOLD docking experiments, was trained and parametrized for optimal pose prediction only.[70,71] Likewise, AutoDock Vina was also trained focusing just on optimal pose prediction.[72] This might rationalize the slightly lower performance of GOLD/ChemPLP or AutoDock Vina for some targets.

Furthermore, we observed significant differences in computational costs for the respective docking tools at their default settings. For instance, docking calculations of the complete CATL benchmark set finished significantly faster (~5 times) in GOLD than in Glide or AutoDock Vina (Table 2). The GOLD default settings are therefore directed toward fast calculations, and we expect that a more accurate GOLD setting would yield improved benchmark results.

Table 2. Overview of the Overall Docking Time for CATL Using the Same Computational conditions

	GOLD	Glide	Vina
total docking time	160.11 min	813.36 min	804.58 min
average no. of GA runs (or poses) per compound	9.7 GA runs	NA	9 poses
pROC (AUC)	0.77	0.76	0.47

Still, all the discussion about general performance trends and preferences should not prescind from the conclusion that performances can be substantially target-dependent and should be evaluated individually.

To investigate whether our data sets are also suitable for benchmarking 2D methods, we additionally tested our benchmark sets with a simple similarity-based screening technique (see the Supporting Information and Table S7).

■ CONCLUSIONS AND OUTLOOK

With DEKOIS 2.0 we provide a large library of high-quality benchmarking sets for VS tools. We have substantially extended and complemented the available target space of benchmark sets by applying our improved DEKOIS protocol and a new ligand selection protocol. Making high quality benchmark sets for structure-based VS tools requires a careful and rational selection of suitable actives and decoys. The quality of the compiled data sets was characterized using the previously established DOE score and Doppelganger score. The matching of physicochemical properties between actives and decoys was found to be improved for the DEKOIS 2.0 workflow. We generally found a low structural similarity between actives and decoys, indicating a good avoidance of false decoys (latent actives in the decoy set).

Naturally, there is an overlap of the DEKOIS 2.0 library with existing benchmark set repositories, such as the DUD-e, especially for long-established targets. The quantity and quality of structural and bioactivity data for such drug targets greatly facilitates the generation of benchmark data sets. Based on the availability of more SAR and crystallographic data also for less well-characterized targets, the target diversity will increase in the future. Considering that protocols and tools for generating benchmark sets share similarities in the general strategy but deviate in many details of the approaches, data sets from different repositories for the same target will show significant differences. We aim to continuously expand our DEKOIS 2.0 library making use of the fast and efficient automatic protocol, to further improve target diversity. Especially the technological progress in generating structural data for GPCRs or ion channels can facilitate providing new docking benchmark sets for these so far rather underrepresented target types. In addition to new target classes, new types of binding sites represent interesting challenges for future extension of DEKOIS sets. Accumulating evidence suggests that protein−protein interaction sites are major allosteric ways to modulate enzyme or receptor function. Such binding sites are rather flat and spacious and, thus, tend to favor high molecular weight ligands with particular physicochemical properties. In contrast, mutation-induced binding pockets can be rather slim and rigid but may allow for completely new therapeutic approaches. An example for such a slim mutation-induced cavity that can be targeted and stabilized with small molecules is the Y220C mutation of the core domain of the tumor suppressor p53.[3,73,74]

Based on our experience with the PYGL data sets, we propose that more attention needs to be paid to differentiating between alternative binding sites within the same target. Besides classical strategies, such as targeting active sites of enzymes, in modern drug discovery also stabilization of active or inactive conformations by allosteric modulation or disruption of protein−protein interactions have been pursued.[6] With increasing drug discovery efforts directed toward an established target, it is also more likely that such alternative ways to modulate the respective target has been discovered. Consequently, problems similar to PYGL may be encountered for such targets. For instance, the majority of HSP90 inhibitors target the ATP-binding site of the N-terminal domain. However, some inhibitors were reported to bind to the adenosine-binding pocket of the C-terminal domain or also a recently discovered, non-ATP-competitive allosteric site.[75,76] As we have demonstrated with the PYGL data sets, blending of binding site specific ligands can dramatically impair benchmark results. Therefore, it is essential to separate target-specific bioactivity data into binding site specific ligand data for the generation of meaningful benchmark sets.

We suggest that another important issue, which should be prevented, is the occurrence of covalent binders in the active set. We showed that this can decrease the actual screening performance when applying standard docking procedures. It is also nontrivial to filter covalent binders from bioactivity data. Especially for targets where reversible reaction-intermediates or irreversible covalent bonds can be formed by activated residues in the binding site, chemical functions may participate in these reactions, although in a different chemical environment, they could be regarded nonreactive. Extensive literature research has to be performed to identify ligands containing such moieties. Intensive manual inspection and curation of reported bioactives is therefore the only option for retrieving a meaningful selection from standard bioactivity databases.

In this study, we have employed a standard preparation protocol and three popular and widely used docking tools[77] to present examples of "default docking performances" retrieved for our DEKOIS 2.0 benchmark sets. Of course, these results only define a starting point for further optimization. Improved performances might be expected from tuning the molecular preparation methods and docking parameters[79,80] as well as from more refined and complex protocols involving consensus scoring,[78,79] ensemble docking, QM-polarized docking, or induced-fit/4D docking methodologies.[80] Also the utilization of new scoring functions that can recognize presently unaccounted interaction types like halogen bonding should help with better docking and scoring of certain ligand classes.[73,81−83] The actual benefit of all such efforts dedicated to enhance the quality of the screening performance can be evaluated by the present collection of DEKOIS 2.0 benchmark sets or extensions thereof, which are easily generated using our automatic workflow. All present and future data sets will be made accessible through our Web repository at http://www.dekois.com.

■ METHODS

Target Tree Clustering. The clustering and the generation of the polar dendrogram was performed with R.[84] We employed a hierarchical clustering with the ward minimum variance method. Distances were measured in Euclidean space. The polar dendrogram was generated by the phylogenetics package ape, using the fan plot type. For targets with no known EC number, we allocated virtual top level codes. The EC top level codes from one to six describe the six main enzyme groups. Subsequently, we employed further virtual top level codes for characterizing other protein groups. We used the top level code 10 for nuclear receptors and added the existing classification of nuclear receptor sequence homology as subcodes.[85] For the GPCRs and BCL-2 the top level codes of 11 and 12 were assigned, respectively.

Optimized DEKOIS Methodology. We had previously described our automated DEKOIS workflow, a decoy generation method that is integrated into a Pipeline Pilot protocol.[9] Four consecutive steps are performed during the DEKOIS protocol: (1) A 15 million ZINC compound database is classified into five, equally populated physicochemical bins. The combination of these property bins yields five-dimensional physicochemical cells that describe the compound properties. (2) Each active is assigned to its respective physicochemical property cell. (3) For each active a pool of 1500 potential decoys is selected from the respective physicochemical property cell of the active. In case the respective cell cannot provide 1500 structures, we also select compounds from adjacent neighbor cells in physicochemical space. (4) The initial pool of decoys is further optimized regarding high physicochemical similarity between actives/decoys and the avoidance of latent actives (LADS) to yield 30 structurally diverse decoys per active.

The benchmark sets presented in this work were generated with an improved version of the protocol. The physicochemical similarity between actives and decoys is now optimized with respect to eight physicochemical properties: molecular weight (MW), octanol−water partition coefficient (logP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), number of rotatable bonds (NROT), positive charge (PC), negative charge (NC), and aromatic rings (AR). All properties were calculated at a pH of 7.4 using the Chemaxon pipeline pilot components.[86]

The previously described LADS filter was further fine-tuned.[9] The LADS *score* was calculated to avoid potential bioactive structures in the decoy set. This was done within the protocol after a pool of potential decoy structures was preselected for each active from the surrounding physicochemical space. For all ligands of the active set and each preselected decoy structure FCFP_6 fingerprints (Pipeline Pilot, Scitegic)[46] were calculated. The resulting fingerprint bit strings of each individual structure from the pool of decoys were then matched with all unique FCFP_6 bit strings of the complete active set. The LADS *score* is calculated according to the following equation:

$$\text{LADS } score = \frac{\sum_{i=1}^{n} (N_{i(\text{Heavy Atoms})} \cdot f_{i(\text{FCFP 6fragment})})}{N_{\text{FCFP 6fragments}}}$$

With n being the total number of shared fingerprint strings between the structure and the active set, each substructure i was weighted by the product of its number of heavy atoms (N_i) and the frequency f_i of the occurrence of the FCFP_6 fragment i in the number of actives. The sum of the weights of shared substructures between each potential decoy and the active set was then divided by the total number of FCFP_6 fragments ($N_{\text{FCFP\_6 fragments}}$) for this decoy to yield the final LADS *score*. Using the product of N_i and f_i ensures that substructures occurring frequently in bioactive compounds exert more influence on the score, if they are larger and, thus, naturally more meaningful. Particularly small fragments such as functional groups may occur both in the active and the decoy set without indicating any bias in the quality of the molecular recognition of these decoys. A lower LADS *score* corresponds to a smaller risk for the respective structure to harbor intrinsic bioactivity. We additionally applied a filter to generally exclude preselected decoys that contain complete ligand structures as a substructure from the decoy selection procedure.

Generation of Ligand Sets. For each protein target a complete set of bioactive structures was extracted from the BindingDB. We only considered assay data measured against human proteins except for inhibitors developed against nonhuman targets. A Pipeline Pilot workflow was applied to process the Binding DB data. In general, we considered only bioactivity data reported as K_i, K_d, IC_{50}, and ΔG. Molecules that possessed a 1000-fold weaker ligand potency than the most potent bioactive in the data stream for the respective affinity measure were discarded. Due to a limited amount of bioactivity data, we had to increase this "weak binding factor" for FKBP1A, HMGR, INHA, PNP, SARS-HCOV, TK, and TPA to obtain sufficiently large ligand sets. ΔG values derived from ITC data were transformed into K_d values. Each molecule was ionized at pH 7.4 using the Chemaxon pipeline pilot standardizer component.[86] To filter ligands with undesirable physicochemical properties a filter with the following specifications was applied: 90 < Molecular Weight (MW) ≤ 900; −7 ≤ AlogP ≤ 9; Hydrogen Bond Acceptors (HBA) ≤ 18; Hydrogen Bond Donors (HBD) ≤ 18; Number of Rotatable Bonds (NROT) ≤ 18. Structures containing nonorganic atoms or unspecified stereocenters were discarded as well. A substructure filter based on the pan assay interference compounds (PAINS) by Baell et al. was employed to remove potential false positive assay hits.[48] To avoid reactive structures we used a substructure filter for following groups: acyl-halides, alkyl-halides, aldehydes, anhydrides, aziridines, disulfides, epoxides, halopyrimidines, hydrazins, isocyanates, isothiocyanates, perhaloketones, peroxides, sulfonyl-halides, and thioesters. Bioactivity data for protease targets were further filtered for substructures that were reported to form reversible reaction-intermediates or covalent modifications of the respective target. A comprehensive overview of reactive and semireactive groups used for the substructure filter can be found in Table S5. FCFP_6 fingerprints (Pipeline Pilot, Scitegic)[46] were calculated for each molecule to cluster molecules into 40 bins according to FCFP_6 Tanimoto similarities. A maximal dissimilarity partitioning method was used for the cluster process. The final ligand set was yielded by selecting the most potent structure of each bin.

Metrics. We have previously introduced the *deviation from optimal embedding score* (DOE score), a numerical metric that describes the quality of embedding ligands with physicochemical similar decoys in chemical space.[9] This metric is calculated by spatial analysis of molecular distances in a multidimensional physicochemical space. Due to the newly introduced optimization of three additional physicochemical properties (PC, NC, AR), we adapted the calculation of distances in the physicochemical space to eight dimensions. The DOE score presented in this work considers the decoy embedding quality regarding these eight physicochemical properties. To ensure an equal weight of each physicochemical property, each property was normalized.[9] A ROC curve is calculated for each active by sorting all remaining structures by distance to the respective active. Each active raises the *true positive rate* (TPR) and each decoy the *false positive rate* (FPR). The absolute value of the areas between the random distribution f(x) = x, and this ROC curve is calculated for each active a:

$$\mathrm{ABC}_{\mathrm{a}}^{\mathrm{DOE}} = \sum_{i=1}^{n} |0.5[(TPR_i + TPR_{i-1})\cdot(FPR_i - FPR_{i-1}) - (FPR_i + FPR_{i-1})\cdot(FPR_i - FPR_{i-1})]|$$

For an optimal decoy embedding of actives this area between the curves becomes zero. A complete spatial separation between actives and decoys in physicochemical space results in an $\mathrm{ABC}^{\mathrm{DOE}}$ (area between the curves) value of 0.5. The arithmetic mean of these $\mathrm{ABC}^{\mathrm{DOE}}$ values is calculated to yield the final DOE score.

The *doppelganger* score describes the extent of structural similarity between actives and their most structurally related decoys. We generated FCFP_6 fingerprints for active and decoy compounds to evaluate the structural similarity between each active and all decoys using the Tanimoto coefficient. The final *doppelganger* score is the arithmetic mean of the highest Tanimoto coefficients for mutually exclusive pairs of actives and decoys.

Preparation of Targets. Coordinates for each crystal structure in our data set were retrieved from the Protein Data Bank (PDB). A comprehensive list of PDB codes and details of the protein−ligand complexes are given in Table S2. The original PDB files were prepared by assigning bond orders, adding hydrogens, creating zero-order bonds to metals, and converting selenomethionines to methionines using the Protein Preparation Wizard in Maestro (version 9.1).[55] Identical and redundant protein chains with nonessential cofactors, ions, water molecules, and ligands were discarded. Exceptions were made for PDB codes with cofactors and cosubstrates in the ligand binding site (3tfq, 1p44, 1z11, 1ah3, and 1i00) and for structures that contained metal ions (e.g., Ca^{2+}, Zn^{2+}, and Mg^{2+}) in their ligand binding pocket (1uze, 1r4l, 3ewj, 3max, 3sff, 3k5e, 1hov, 3frg,1xp0, 2afx, and 2z94). When necessary, missing side chains in the binding site were added using Prime,[69] while metal binding states were generated at pH 7.0 by Epik.[87,88] Prepared structures were saved as MAE and PDB files. The MAE files were used for Glide[89,90] (version 5.6) docking, and the PDB files were used for GOLD[91−94] (version 5.1) docking. The PDB files were converted to PDBQT files by employing a python script (*prepare_receptor4.py*) provided by the MGLTools package (version 1.5.4) for AutoDock Vina (version 1.1.2) docking experiments.[72] The native geometry of the binding sites was preserved without in-place ligand-protein minimization.

Preparation of DEKOIS Sets. All compounds were prepared by LigPrep (version 2.4).[56] The compounds were minimized using the OPLS-2005 force field. For each molecule the protonation state at pH 7.0 was generated. The specified stereoconfiguration of actives and decoys was retained. All prepared compounds were saved as SD files for Glide and GOLD docking. The SD files were converted and split into PDB files by open-Babel[95] (version 2.3.1), which were further converted into PDBQT files by a MGLTools (version 1.5.4) python script (*prepare_ligand4.py*) for AutoDock Vina.

Docking Experiments. All docking experiments were performed with Glide, GOLD, and AutoDock Vina using default settings. The respective best-scored pose of each docked molecule was retrieved for the calculation of the pROC AUC values. For Glide (version 5.6) docking, we performed the receptor grid generation with default settings. For most targets, a grid-box with an approximate size (on average) 20 × 20 × 20 Å was retrieved. We enabled the rotation of hydroxyl groups of binding site residues. We used the standard precision docking mode (SP) and also considered Epik state penalties for metal-containing binding sites. By default, the cutoff for keeping initial docking poses was 100.0 kcal/mol (relative to the best scored initial pose). The best five final docking poses per compound were selected for postdocking minimization within Glide. For GOLD (version 5.1) docking, binding site residues were defined by specifying the crystal structure ligand coordinates and using a cutoff radius of 10 Å, with the 'detect cavity' option enabled. GOLD docking experiments were performed using the ChemPLP scoring function. The search efficiency of the genetic algorithm was kept at the standard 100% setting. The docking was terminated early when the top three solutions were within 1.5 Å RMSD. For AutoDock Vina docking, we again employed default docking parameters. The size of the docking grid was generally 20 Å × 20 Å × 20 Å, with a grid spacing of 1 Å. In cases where the ligand binding site was not completely included in the grid box, the grid dimensions were expanded accordingly. We enabled the rotation of ligand amide groups. By default, the docking was terminated when the maximum energy difference between the best scored pose and the worst one was 3 kcal/mol. Due to the different binding site definition methodologies in Glide, GOLD, and AutoDock Vina, we expect (slightly) different binding site sizes for different programs. Naturally, this can also affect the docking performance. As previously described, the calculation of pROC AUC values, enrichment factors, and generation of pROC plots was performed by pipeline pilot using R components.[9,13,84] Additionally, we tested the default docking setup for each VS tool by performing pose retrieval experiments. For each target, the cocrystallized ligand was docked into its corresponding binding site of the crystal structure. We assessed the pose prediction success by calculating the heavy atom root-mean-square deviation (RMSD) between the docked poses and the original cocrystallized ligand using Maestro (Table S4). To give an example of the differences of the computational costs (overall docking time) of the docking process, we performed a benchmark using exactly the same conditions (Table 2). In the case of CATL, GOLD performed approximately 5 times faster than the other two programs. Some docking programs might discard compounds during the docking process. To investigate the impact of this issue on pROC AUC values, we provide an adjusted pROC AUC table (Table S8). We obtained accordingly updated pROC AUC values by adding discarded compounds at the end of the score-ordered list of docked compounds. Only for the Glide results of AR, COX2, ER-beta, PIM-2, and VEGFR1, we observed larger pROC AUC changes (ΔpROC AUC > 0.05).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

We provide a short description of the 2D screening performance. Additional tables give more detailed insights into the following: (S1) target descriptions and benchmark set quality metrics, (S2) PDB codes of target structures and docking results, (S3) overview of reactive groups in reactive benchmark set examples, (S4) overview of pose retrieval docking results, (S5) filtered reactive and semireactive substructures, (S6) overview of employed PAINS substructures, (S7) screening results for a 2D Tanimoto similarity based screening method, and (S8) pROC-AUC values adjusted by integration of ligands discarded by some docking programs. Figure S1 shows pROC graphs for standard and large DEKOIS 2.0 benchmark sets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: frank.boeckler@uni-tuebingen.de.

Author Contributions

‡M.R.B. and T.M.I. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Gergana Andreeva for setting up the DEKOIS web service, Mirco Michel for his help with the R polar dendrogram script, and Ionut Onila for assistance with Molcad II. We also thank Thomas Exner for his valuable comments regarding the manuscript and Markus O. Zimmermann for proofreading the manuscript and helpful suggestions regarding halogen-bonding. T.M.I. is grateful to the German Academic Exchange Service (DAAD) for funding his Ph.D. fellowship.

■ ABBREVIATIONS

A2A, adenosine receptor A2a; AUC, area under the curve; CATL, Cathepsin L1; CTSK, Cathepsin K; DEKOIS, demanding evaluation kits for objective *in silico* screening; DOE, deviation from optimal embedding; DUD, directory of useful decoys; EC, enzyme commission; FPR, false positive rate; FXA, coagulation factor X; HBA, H-bond acceptor; HBD, H-bond donor; LADS, latent actives in the decoy set; MDM2, E3 ubiquitin-protein ligase MDM2; MW, molecular weight; NA, neuraminidase A; PAINS, pan assay interference compounds; PPI, protein−protein interaction; PYGL, glycogen phosphorylase (liver form); QPCT, glutaminyl-peptide cyclotransferase; RB, rotatable bonds; ROC, receiver operating characteristic; SIRT2, NAD-dependent protein deacetylase sirtuin-2; TPR, true positive rate; VS, virtual screening

■ REFERENCES

(1) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.

(2) McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494−502.

(3) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10360−10365.

(4) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **2012**, *14*, 133−141.

(5) Villoutreix, B. O.; Eudes, R.; Miteva, M. A. Structure-based virtual ligand screening: recent success stories. *Comb. Chem. High Throughput Screening* **2009**, *12*, 1000−1016.

(6) Vogel, S. M.; Bauer, M. R.; Joerger, A. C.; Wilcken, R.; Brandt, T.; Veprintsev, D. B.; Rutherford, T. J.; Fersht, A. R.; Boeckler, F. M. Lithocholic acid is an endogenous inhibitor of MDM4 and MDM2. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16906−16910.

(7) Tuccinardi, T. Docking-based virtual screening: recent developments. *Comb. Chem. High Throughput Screening* **2009**, *12*, 303−314.

(8) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.

(9) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 2650−2665.

(10) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.

(11) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201−212.

(12) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.

(13) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141−146.

(14) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(15) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856−5868.

(16) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.
(17) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.
(18) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.
(19) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(20) Rohrer, S. G.; Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704−718.
(21) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.
(22) Wallach, I.; Lilien, R. Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.* **2011**, *51*, 196−202.
(23) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.
(24) Mysinger, M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E) - better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(25) Pham, T. A.; Jain, A. N. Customizing scoring functions for docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 269−286.
(26) Meganathan, C.; Sakkiah, S.; Lee, Y.; Narayanan, J. V.; Lee, K. W. Discovery of potent inhibitors for interleukin-2-inducible T-cell kinase: structure-based virtual screening and molecular dynamics simulation approaches. *J. Mol. Model.* **2013**, *19*, 715−726.
(27) Spitzer, R.; Jain, A. N. Surflex-Dock: docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687−699.
(28) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 675−686.
(29) Katritch, V.; Rueda, M.; Abagyan, R. Ligand-guided receptor optimization. *Methods Mol. Biol.* **2012**, *857*, 189−205.
(30) Wassermann, A. M.; Bajorath, J. BindingDB and ChEMBL: online compound databases for drug discovery. *Expert Opin. Drug Discovery* **2011**, *6*, 683−687.
(31) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.
(32) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−633.
(33) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−201.
(34) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719−725.
(35) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **2002**, *18*, 130−139.
(36) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193−199.
(37) Good, A.; Oprea, T. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.
(38) Mackey, M. D.; Melville, J. L. Better than random? The chemotype enrichment problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154−1162.
(39) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.
(40) Jubb, H.; Higueruelo, A. P.; Winter, A.; Blundell, T. L. Structural biology and drug discovery for protein-protein interactions. *Trends Pharmacol. Sci.* **2012**, *33*, 241−248.
(41) Keller, T. H.; Pichota, A.; Yin, Z. A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* **2006**, *10*, 357−361.
(42) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450*, 1001−1009.
(43) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(44) Csermely, P.; Palotai, R.; Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **2010**, *35*, 539−546.
(45) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.
(46) *Pipeline Pilot*, 6.1.5.0 student ed.; Accelrys: San Diego, 2007.
(47) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54*, 5422−5431.
(48) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(49) Rath, V. L.; Ammirati, M.; Danley, D. E.; Ekstrom, J. L.; Gibbs, E. M.; Hynes, T. R.; Mathiowetz, A. M.; McPherson, R. K.; Olson, T. V.; Treadway, J. L.; Hoover, D. J. Human liver glycogen phosphorylase inhibitors bind at a new allosteric site. *Chem. Biol.* **2000**, *7*, 677−682.
(50) Lawandi, J.; Toumieux, S.; Seyer, V.; Campbell, P.; Thielges, S.; Juillerat-Jeanneret, L.; Moitessier, N. Constrained peptidomimetics reveal detailed geometric requirements of covalent prolyl oligopeptidase inhibitors. *J. Med. Chem.* **2009**, *52*, 6672−6684.
(51) Ouyang, X.; Zhou, S.; Su, C. T.; Ge, Z.; Li, R.; Kwoh, C. K. CovalentDock: automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constraints. *J. Comput. Chem.* **2013**, *34*, 326−336.
(52) Powers, J. C.; Asgian, J. L.; Ekici, O. D.; James, K. E. Irreversible inhibitors of serine, cysteine, and threonine proteases. *Chem. Rev.* **2002**, *102*, 4639−4750.
(53) Costanzo, M. J.; Almond, H. R., Jr.; Hecker, L. R.; Schott, M. R.; Yabut, S. C.; Zhang, H. C.; Andrade-Gordon, P.; Corcoran, T. W.; Giardino, E. C.; Kauffman, J. A.; Lewis, J. M.; de Garavilla, L.; Haertlein, B. J.; Maryanoff, B. E. In-depth study of tripeptide-based alpha-ketoheterocycles as inhibitors of thrombin. Effective utilization of the S1' subsite and its implications to structure-based drug design. *J. Med. Chem.* **2005**, *48*, 1984−2008.
(54) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: a critical review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.
(55) *Suite 2012: Protein Preparation Wizard*; Schrödinger, LLC: New York, NY, 2012.
(56) *Ligprep*, 2.4; Schrödinger, LLC: New York, NY, 2010.
(57) Hanks, S. K.; Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **1995**, *9*, 576−596.
(58) Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **2002**, *11*, 184−197.
(59) Biamonte, M. A.; Van de Water, R.; Arndt, J. W.; Scannevin, R. H.; Perret, D.; Lee, W. C. Heat shock protein 90: inhibitors in clinical trials. *J. Med. Chem.* **2010**, *53*, 3−17.
(60) Ivetac, A.; McCammon, J. A. Molecular recognition in the case of flexible targets. *Curr. Pharm. Des.* **2011**, *17*, 1663−1671.

(61) Luthi-Carter, R.; Taylor, D. M.; Pallos, J.; Lambert, E.; Amore, A.; Parker, A.; Moffitt, H.; Smith, D. L.; Runne, H.; Gokce, O.; Kuhn, A.; Xiang, Z.; Maxwell, M. M.; Reeves, S. A.; Bates, G. P.; Neri, C.; Thompson, L. M.; Marsh, J. L.; Kazantsev, A. G. SIRT2 inhibition achieves neuroprotection by decreasing sterol biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 7927−7932.

(62) Narayan, N.; Lee, I. H.; Borenstein, R.; Sun, J.; Wong, R.; Tong, G.; Fergusson, M. M.; Liu, J.; Rovira, I. I.; Cheng, H. L.; Wang, G.; Gucek, M.; Lombard, D.; Alt, F. W.; Sack, M. N.; Murphy, E.; Cao, L.; Finkel, T. The NAD-dependent deacetylase SIRT2 is required for programmed necrosis. *Nature* **2012**, *492*, 199−204.

(63) Huhtiniemi, T.; Salo, H. S.; Suuronen, T.; Poso, A.; Salminen, A.; Leppanen, J.; Jarho, E.; Lahtela-Kakkonen, M. Structure-based design of pseudopeptidic inhibitors for SIRT1 and SIRT2. *J. Med. Chem.* **2011**, *54*, 6456−6468.

(64) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.

(65) Buchholz, M.; Heiser, U.; Schilling, S.; Niestroj, A. J.; Zunkel, K.; Demuth, H. U. The first potent inhibitors for human glutaminyl cyclase: synthesis and structure-activity relationship. *J. Med. Chem.* **2006**, *49*, 664−677.

(66) Brickmann, J.; Exner, T. E.; Gimmler, J.; Lautenschläger, P.; Heiden, W.; Moeckel, G.; Zahn, D. *MOLCAD II*; MOLCAD GmbH: Darmstadt, 2000.

(67) Brickmann, J.; Exner, T. E.; Keil, M.; Marhofer, R. J. Molecular graphics - Trends and perspectives. *J. Mol. Model.* **2000**, *6*, 328−340.

(68) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.

(69) *Prime*, 2.2; Schrödinger, LLC: New York, NY, 2010.

(70) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.

(71) Korb, O.; Ten Brink, T.; Victor Paul Raj, F. R.; Keil, M.; Exner, T. E. Are predefined decoy sets of ligand poses able to quantify scoring function accuracy? *J. Comput.-Aided Mol. Des.* **2012**, *26*, 185−197.

(72) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(73) Wilcken, R.; Liu, X.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-enriched fragment libraries as leads for drug rescue of mutant p53. *J. Am. Chem. Soc.* **2012**, *134*, 6810−6818.

(74) Wilcken, R.; Wang, G.; Boeckler, F. M.; Fersht, A. R. Kinetic mechanism of p53 oncogenic mutant aggregation and its inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 13584−13589.

(75) Taldone, T.; Sun, W.; Chiosis, G. Discovery and development of heat shock protein 90 inhibitors. *Bioorg. Med. Chem.* **2009**, *17*, 2225−2235.

(76) Chang, D. J.; An, H.; Kim, K. S.; Kim, H. H.; Jung, J.; Lee, J. M.; Kim, N. J.; Han, Y. T.; Yun, H.; Lee, S.; Lee, G.; Lee, J. S.; Cha, J. H.; Park, J. H.; Park, J. W.; Lee, S. C.; Kim, S. G.; Kim, J. H.; Lee, H. Y.; Kim, K. W.; Suh, Y. G. Design, synthesis, and biological evaluation of novel deguelin-based heat shock protein 90 (HSP90) inhibitors targeting proliferation and angiogenesis. *J. Med. Chem.* **2012**, *55*, 10863−10884.

(77) Mihasan, M. What in silico molecular docking can do for the 'bench-working biologists'. *J. Biosci.* **2012**, *37*, 1089−1095.

(78) ten Brink, T.; Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J. Chem. Inf. Model.* **2009**, *49*, 1535−1546.

(79) ten Brink, T.; Exner, T. E. pK(a) based protonation states and microspecies for protein-ligand docking. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 935−942.

(80) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149−164.

(81) Kuhn, B.; Fuchs, J. E.; Reutlinger, M.; Stahl, M.; Taylor, N. R. Rationalizing tight ligand binding through cooperative interaction networks. *J. Chem. Inf. Model.* **2011**, *51*, 3180−3198.

(82) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and applications of halogen bonding in medicinal chemistry and chemical biology. *J. Med. Chem.* **2013**, *56*, 1363−1388.

(83) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: a systematic evaluation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 935−945.

(84) Team, R. D. C. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(85) Zhang, Z.; Burch, P. E.; Cooney, A. J.; Lanz, R. B.; Pereira, F. A.; Wu, J.; Gibbs, R. A.; Weinstock, G.; Wheeler, D. A. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res.* **2004**, *14*, 580−590.

(86) *Chemaxon pipeline pilot components*, 5.4.1.1; Chemaxon: Budapest, 2011.

(87) *Epik*, 2.1; Schrödinger, LLC: New York, NY, 2010.

(88) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.

(89) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(90) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(91) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(92) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(93) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(94) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(95) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.