Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds

Jochen Schlosser and Matthias Rarey*

Center for Bioinformatics, Research Group for Computational Molecular Design, University of Hamburg,
Bundesstrasse 43, 20146 Hamburg, Germany

Received January 15, 2009

The standard approach to structure-based high-throughput virtual screening is a sequential procedure: Each molecule of a given library is screened against the target protein, eventually generating a ranked list of molecules. In this paper a new paradigm avoiding the sequential screening pipeline is presented. Based on a novel descriptor, compounds can be directly accessed by their chemical and shape complementarity to a given protein active site. The docking calculation is performed inherently during the search process since each search result automatically implies a ligand pose in the active site. The new method named TrixX BMI is ideally suited for application scenarios in which medicinal chemists request a certain pharmacophore interaction pattern to a protein. By using an innovative indexing technology, sublinear runtimes in the number of ligands can be achieved. Redocking experiments show that TrixX BMI correctly predicts the pose of the bioactive conformation within an rmsd of less than 2.0 Å of the cocrystallized ligand in 80% of 85 protein−ligand complexes of the Astex Diverse Set. In addition to that several comparative enrichment experiments show that TrixX BMI is on a competitive basis to established virtual screening technology, while the observed runtimes are clearly below one second per compound.

INTRODUCTION

The concept of sequential screening is widely used for lead identification in pharmaceutical research. In experimental as well as in virtual lead identification, the method of choice is frequently the individual testing of large compound collections.[1] In an experimental setup, (high-throughput) screening is an expensive endeavor with known weaknesses like high error rates[2,3] and low ligand efficiency.[4,5] In the past years fragment-based lead identification strategies were developed that are successfully applied today.[6] In a computational setup, the fragment-based approach can also be followed. Virtual screening is, however, a much more flexible and adaptable process. For example, compounds can be created at no costs, preselections can be easily made, and, most importantly, the space of compounds to be screened can be preorganized in a multitude of ways. Therefore, the question remains, whether there are alternative routes for virtual lead identification besides classical virtual screening and fragment-based strategies.

For structure-based computational molecular design, the prediction of protein−ligand complex geometries and their respective binding affinities are at the center of attention. A variety of methods including frequently applied docking engines like Glide,[7] Gold,[8] Icm,[9] Autodock,[10] and FlexX[11] have been developed in the past decade (see refs 12, 1, and 13 for reviews). The deficiencies of docking methodology addressing this prediction are well-known.[13] Especially imperfect scoring functions,[14,15] inadequate modeling of protein flexibility,[16,17] and low coverage of the large space of possible configurations result in only weak correlations between predicted and experimentally measured binding affinities. At the same time, a large variety of compounds have to be considered resulting in limited compute resources per compound, even in setups on large compute clusters.

Standard docking tools differ in their underlying docking method, e.g. incremental construction, random search, multiconformer docking, or hierarchical combinations of these, but they share one basic algorithmic concept: All compounds in a given collection are subject to docking, although many of the compounds processed during a virtual screening run do not fit into the protein active site of interest. In this paper, we present a computational method which allows the preorganization of compounds in a way that only those compounds with reasonable chemical and steric complementarity are considered in the actual docking calculation. Such a method does not follow the classical screening concept anymore since only relevant parts of the compound collection will be used during the calculation. This is possible by merging techniques similar to three-dimensional pharmacophore matching,[18,19] with docking and indexing technology. Besides its novel methodology, TrixX BMI offers a substantial speed-up of the whole docking process which leaves room to consider the promising candidates in much more detail. Although not considered in this paper, due to the separate preprocessing of compound libraries and target proteins, new routes for addressing protein flexibility can be envisioned.

So far only few approaches deviate from the iterative protocol speeding up the overall docking time significantly. PhDock[20] is a cluster-based approach which uses a three-dimensional pharmacophore to describe the spatial arrangement of functional groups within a ligand. After conformationally sampling all ligands and clustering the resulting

* Corresponding author e-mail: rarey@zbh.uni-hamburg.de.

10.1021/ci9000212 CCC: $40.75 © 2009 American Chemical Society
Published on Web 04/09/2009

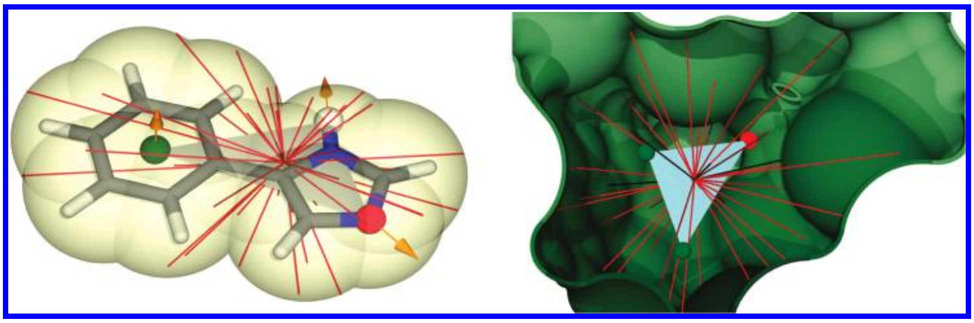


Figure 1. TRIXX descriptor for compounds (left) and protein active site (right). Three interaction centers (red, green, and white spheres), their interaction direction (orange arrows), and 40 of the 80 rays representing shape relative to the triangle (red rays).

pharmacophores with regard to similar spatial arrangements, a modified version of DOCK 4.0[21,22] is used to place the pharmacophores into the binding site. Once a pharmacophore has been docked successfully the procedure expands and scores all associated ligand conformers by superposing them with the pose prediction of the corresponding pharmacophore. This approach scores only those ligands that are linked to a correctly docked pharmacophore.

Three-dimensional pharmacophore applications like DISCOVERY STUDIO,[23] LIGAND SCOUT,[24] PHASE,[25] and UNITY[26] preselect compounds from annotated compound databases by automatically generating a pharmacophore description of the active site using inclusion and exclusion volumes to describe steric constraints. Compound features are not clustered, and therefore all compounds in the database are used within a search. A variant of these tools is SHAPE4[27] which uses a negative image of the active site as input to a modified search based on the OE SHAPE TOOLKIT.[28] This concept incorporates shape into the pharmacophore removing the computationally inefficient usage of volumes as steric description but still needs to scan the complete database of annotated ligands.

TRIXX,[29] recently developed in our group, applies the idea of a target-driven search algorithm in contrast to the iterative docking procedure. TRIXX decouples the analysis of the compound library from the actual screening phase by preprocessing all compounds once. This first step splits each ligand into fragments and stores triangle based descriptors in the compound index, which is realized using a relational database system. During screening, query descriptors for the target active site are used to extract initial matches from the index satisfying the given query constraints. The final step consists of linking compatible fragment matches which, if possible, generates pose predictions of library compounds.

In this paper, we present the second generation of the TRIXX technology named TRIXX BMI. The basic approach of TRIXX BMI is similar but removes the risk of noncompatible fragment placements by removing the linking in favor of incremental construction. Furthermore, we increase the fragment size considerably. Most notably, we developed a novel description of steric properties which now is part of the index and allows clash predictions solely on descriptor level. This results in the need of a different indexing system. The upcoming sections describe this new approach in more detail and discuss its performance concerning runtime, redocking accuracy, and enrichment rates.

METHODS

1) TRIXX Descriptor. The most prominent concept used in TRIXX BMI is the TRIXX descriptor. In its basic form it resembles a three-point pharmacophore of interactions between functional groups of a ligand[29] and a protein active site.[30] A pharmacophore is supposed to represent electronic and steric features necessary to trigger or block a compounds biological response.[31] We use interaction types and directions to describe electronic features and an eighty-dimensional distance vector for steric properties (see Figure 1). In contrast to exclusion or inclusion volumes used by other approaches, our steric description is aligned to each descriptor via a local coordinate system based on the remaining descriptor properties. The novel usage of steric properties already on the abstract descriptor level thus yields a fully automatically derived structure based steric pharmacophore filter which is able to preselect compounds for a given protein active site with a matching pharmacophore and a reasonable steric fit simultaneously. Most importantly, this selection mechanism can be executed via the use of modern indexing technology thus speeding up the process enormously.

2) Descriptor Matching. Since the TRIXX descriptors are used to capture complementary site and compound pharmacophores, some properties need to be calculated accordingly. Figure 2 illustrates matching compound and site descriptors (Figure 2a) and highlights the similarities and differences of the two. The main difference concerns the calculation of steric bulk description. A compound descriptor needs to identify the atoms farthest away from its center thus describing the compounds extension (exit distance). A site descriptor needs to find the closest receptor atoms thus describing the cavity relative to the descriptor (clash distance). In both cases TRIXX uses a distance vector to store eighty different directions which are aligned according to a descriptor specific coordinate system. In addition to differences in the shape description, the hydrophilic interaction types of a site descriptor are inverted during the query, such that hydrogen bond donors (white) are matched with hydrogen bond acceptors (red) and vice versa. If a descriptor does not have unique interaction types as in the example, there is more than one possible superposition. In Figure 2b the two possible superpositions of the example descriptors and their corresponding clash predictions are shown. A look at Figure 2c reveals that the descriptor correctly predicts the nonclashing pose on the left and discards the clashing pose on the right side. The filter decision is already made on the descriptor level by simply comparing corresponding exit and clash distances. In a real VHTS scenario this comparison is done

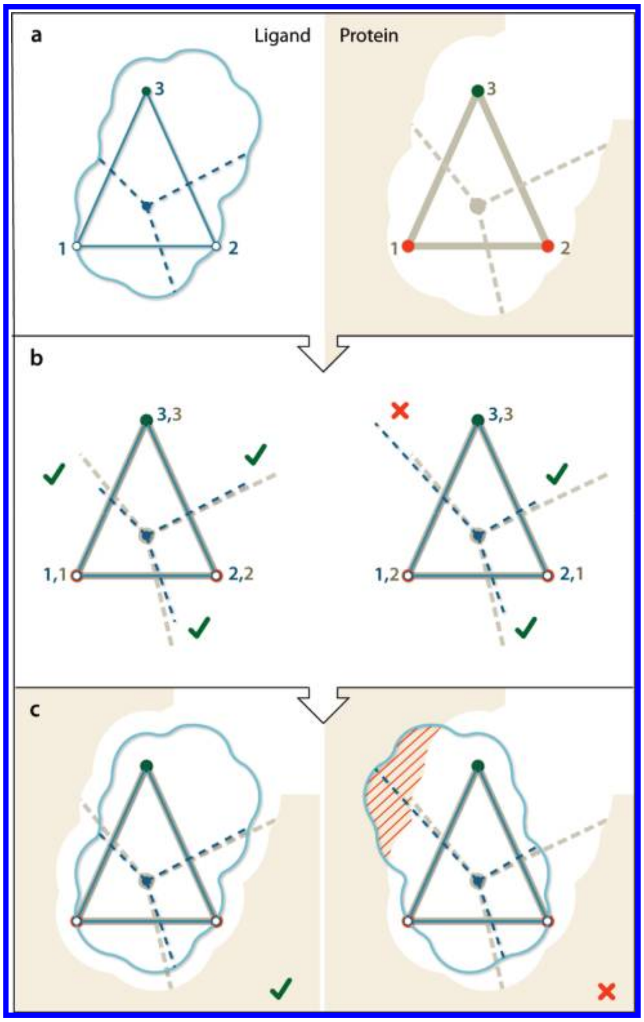


Figure 2. Multiple descriptor alignments and steric clash detection using the TRIXX BMI descriptor. a) Triangle descriptor and three selected bulk dimensions for ligand (left) and site (right) descriptor. b) The two possible alignments and the TRIXX BMI steric bulk comparison on the abstract descriptor level. c) The corresponding poses demonstrate the correctness of the clash predicted by comparing the descriptor's steric bulk description above.

for all compound descriptors of the index by posing one query for each site descriptor. Individual compounds are not necessary which separates TRIXX BMI from other docking methodologies. To cover for some flexibility and uncertainty of the active site, TRIXX BMI deploys error ranges for steric bulk distances and side lengths.

3) Descriptor Size. Since a large number of descriptors are necessary to represent a single compound, descriptor size is an important factor concerning disk space in general and I/O load during virtual screening. The size of the new descriptor compared to the original TRIXX descriptor only differs in the requirements of the steric bulk vector: There are 4 bytes for the interaction types, 36 bytes for interaction coordinates, and 12 bytes concerning the triangle side lengths. Our new 80 dimensional bulk vector needs 320 bytes which leads to 372 bytes storage requirements for each uncompressed descriptor. Therefore a compound library of 1 million ligands, having 10 conformations on average and about 100 descriptors each, needs about 350 GByte to store the raw descriptor data. The real on-disk requirements are further reduced by index compression.

4) Workflow. The complete workflow of TRIXX BMI is presented in Figure 3. First a preprocessing step analyzes all library compounds, identifies their pharmacophore features, and stores TRIXX ligand descriptors in a compressed indexing system optimized for high-speed data retrieval. The problem of conformational flexibility is addressed via usage of Corina[32] ring conformations and an in-house torsion driver to cover the search space adequately. Only ligands violating user-defined flexibility thresholds are split into large fragments which are also subject to conformational sampling. The information content of the TRIXX descriptors of each ligand, stored in the indexing system, is exploited during actual virtual screening runs using different targets in the next part of the workflow.

At the start of each virtual screening experiment the target protein is analyzed and the active site is selected manually. Favorable interaction spots are identified automatically, and TrixX descriptors are calculated for the active site. The steric descriptor properties are calculated using all atoms within a radius of 7.5 Å around the geometric center of the descriptor thus covering 15 Å in diameter. Since the average depth of a drug-binding cavity lies in the range between 6.8 and 11.4 Å[33] and the maximum depth is in between 13.0 and 22.9 Å, the bulk descriptor covers the steric properties of both ligand and active site adequately. From here on TRIXX BMI differs significantly from other available methods: Each TRIXX site descriptor can be used as a search key to an indexing system holding the ligand descriptors. This allows the identification of matching compounds by formulating SQL-like queries and thus breaking the iterative screening paradigm. TRIXX BMI does not look at individual compound descriptors or molecules but uses the precalculated index to identify initial pose predictions and to discard incompatible library compounds. Since each match already describes a reasonable pose, the postprocessing only needs to perform a clash-test and to score. Only compounds that were subject to fragmentation within compound indexing need to be incrementally constructed using the FlexX algorithm with the initially placed fragment as base placement.

5) User-Defined Pharmacophore Constraints. A typical aim in structure-based virtual screening is to find lead structures forming interactions to specific active site atoms. Most docking tools therefore allow modeling of protein-based pharmacophore constraints which have to be obeyed by all retrieved hits. The concept of pharmacophore constraints can easily be integrated into the TRIXX methodology. The constraints are used to automatically filter the generated TRIXX site descriptors. The pharmacophore constraints guide the search by using only those site descriptors that fulfill parts of the desired interaction pattern. This ensures that all matches returned show a high probability of obeying the pharmacophore constraints thus reducing the number of false positives and improving the runtime significantly. Each potential site descriptor is used as a query to the index if at least one of the edges of the corresponding triangle connects two receptor spots which are part of the pharmacophore. All other descriptors are discarded. For a typical number of pharmacophore constraints between two and five the number of queries is reduced by about 1 order of magnitude compared to the unconstraint search. Furthermore, pharmacophore definitions are often buried deeply within the active site which increases the overall selectivity of the queries due

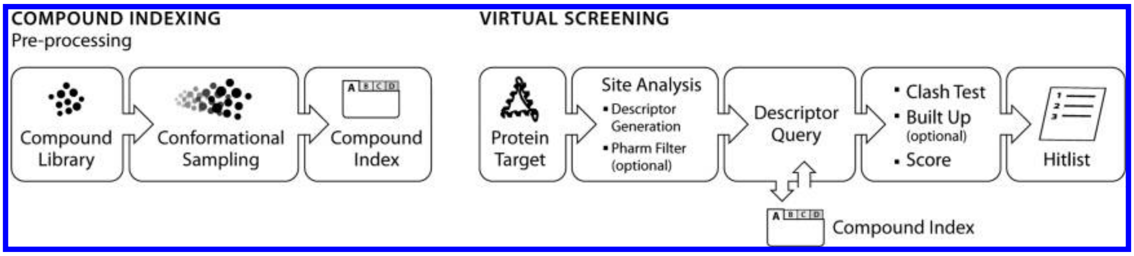
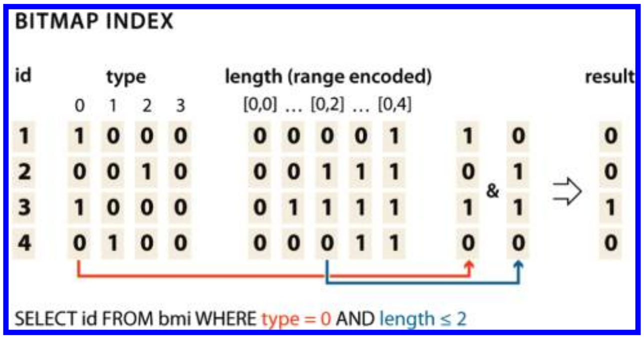


Figure 3. TrixX BMI workflow, with the preprocessing phase on the left and the virtual screening part on the right.



id	type 0	type 1	type 2	type 3	length (range encoded) [0,0]	…	[0,2]	…	[0,4]			result
1	1	0	0	0	0	0	0	0	1	1	0	0
2	0	0	1	0	0	0	1	1	1	0	1	0
3	1	0	0	0	0	1	1	1	1	1	1	1
4	0	1	0	0	0	0	0	1	1	0	0	0

SELECT id FROM bmi WHERE type = 0 AND length ≤ 2

Figure 4. Example query on a bitmap index.

Table 1. Redocking Results[a]

	rmsd [Å] ≤			
	1.0	1.5	2.0	2.5
GOLD (av best20)	N.A.	N.A.	64 [75]	N.A.
FlexX (best20)	38 [45]	53 [62]	61 [72]	62 [73]
TrixX BMI (best20)	24 [28]	45 [53]	63 [74]	67 [79]
TrixX BMI (best200)	26 [31]	49 [58]	68 [80]	74 [86]

[a] Number ([%]) of poses found (within the best n ranks) out of 85 protein/ligand complexes using conformational ensembles based on Corina structures (TrixX BMI) respectively Corina structures (Gold, FlexX) as input.

to more stringent steric constraints. This leads to a further reduction of hits returned by the indexing system. Note that the same compound index is used in unconstraint as well as constraint screening scenarios.

6) Indexing Technology. Due to the high dimensionality of the query an indexing system which is able to handle these kinds of queries efficiently is necessary. The TrixX BMI compound index is mostly static, in the sense that it is not altered during virtual screening. This allows the usage of a stand-alone indexing tool to avoid the overhead of transaction processing within database management systems. These criteria are fully met by the Fastbit[34] indexing technology. It is based on compressed bitmap indices which are especially well suited for high dimensional queries. Bitmap indexing systems build one index for each descriptor attribute A. Each index consists of $|A|$ columns depending on the cardinality of the attribute. A set bit in row i in column j of the index then represents the value of $descriptor_i[A] = j$. These so-called vertical partitions can be efficiently combined using logical AND/OR operations to construct the final result of a query (see Figure 4). Furthermore, it is possible to adjust the bit encoding to different types of queries. For example range encoding, where a set bit in row i, column j represents $descriptor_i[A] \leq j$. Because TrixX always uses exactly one kind of query on each attribute, the different encoding schemes can be utilized to reduce the system's I/O load and the computational cost of combining the results.

Another important aspect of Fastbit is its ability to compress the raw descriptor data by about a factor of 2 while still allowing efficient logical operations without the need to decompress the data. For high cardinality attributes, like floating point values, a binning scheme is used. Since a bitmap index already suffices to answer queries which use exact bin boundaries only the compressed bitmap indices need to be stored, the raw data can be deleted. This can be done because TrixX BMI uses the same binning scheme for site as well as compound descriptors. Thus, only exact bin boundaries are part of the queries, and the system never needs to revert to the original descriptor data to check intermediate values. Once again, the I/O load is reduced which also holds for the storage requirements on hard disk.

The query itself consists of almost 100 different constraints (interaction type, bulk distances, interaction directions, and side lengths) describing a potential descriptor match. Here, the canonical ordering scheme of the triangle geometries is vital for the direct comparison of the different attributes. This schema not only defines an order of comparison for side lengths and interaction types but also supplies a local coordinate system for directional and bulk constraints.

The most important difference to the original TrixX descriptor is the steric bulk compatibility-check. The full descriptor can now be exploited already during the query. TrixX descriptors suffice to predict a receptor−ligand clash; no immediate postprocessing of the initial matches is necessary, and each match returned by the index yields an initial pose prediction.

RESULTS

The performance of TrixX BMI is measured with respect to different criteria.

1. First, the overall *redocking performance* using the Astex Diverse Set[35] is measured in terms of rmsd to the cocrystallized ligand comparing TrixX BMI to FlexX and GOLD.

2. In the next section TrixX BMI's capability to *enrich* known actives of four different target proteins from the DUD database[36] is compared to FlexX and Dock. The DUD database is chosen since it is freely available, and results for DOCK have been published. The targets are chosen since their corresponding ligands represent different fractions of chemical space and detailed pharmacophores have been published (see Table 2).

3. Here, we use the available pharmacophore information to assess the enrichment performance of TrixX BMI compared to FlexX-Pharm.[37]

4. *Runtime experiments* compare TrixX BMI to FlexX and FlexX-Pharm. Average runtimes over 285 targets (AFX_{285}) using a random catalog of 2000 (Z_2) leadlike

Table 2. Pharmacophore Constraints for the Chosen Targets from the DUD

target	type	detail[a]	e/o[b]	P_{min}, P_{max}[c]	ref
CDK2	h_don	N LEU83	*e*	1, 2	40
	h_acc	O LEU83	*o*		
	h_acc	O GLU 81	*o*		
DHFR	h_acc	_OD2 ASP26	*e*		37
	h_acc	O LEU4	*e*		
	phen_center	_CG PHE30	*e*		
ER agonist	h_don	_NH2 ARG394	*e*		40
	h_acc	_OE1 GLU353	*e*		
	spatial	1.7, −1.4, −3.4 (2.5 Å)	*e*		
ER antagonist	h_don	_NH2 ARG394	*e*		40
	h_acc	_OE2 GLU353	*e*		
	spatial	34.1, 0.5, 27.9 (2.5 Å)	*e*		

[a] For interaction constraints, the name of the receptor atom (PDB nomenclature: atom name, amino acid code, amino acid number) is given. For a spatial constraint, the coordinates and the sphere radius are given. [b] Denotes an essential constraint, *o* an optional constraint. [c] P_{min} is the minimum number of optional constraints allowed. P_{max} is the maximum number of optional constraints allowed.

compounds from the ZINC database[38] and the runtimes of the enrichment experiments are presented.

5. The last part demonstrates the *scalability* of TrixX BMI on a medium-sized compute cluster. We use about 1.7 million random leadlike compounds from the ZINC database with the corresponding compound indices being distributed on 48 cluster nodes. Those indices are used to perform virtual high-throughput screening against the four targets used in our enrichment studies.

Data Preparation. All experiments on the DUD targets were performed using the protein targets from the PDB as they are. Default protonation rules were applied, and the active site was defined using all atoms within a radius of 6.5 Å around any atom of the cocrystallized ligand. In case of the Astex Diverse Set we included essential metals, altered protonation states, and torsional angles in order to capture the binding mode of the cocrystallized ligand. Concerning the ligands two scenarios have to be differentiated. During enrichment studies using targets from the DUD database and the redocking experiments using the Astex Diverse Set, we refrained from using any molecular property filters in order to keep our results comparable to the corresponding publications. In contrast to that, all runtime and scalability experiments were run on catalogs using only leadlike ligands. Leadlikeness is here defined according to "The Oprea criteria"[39] which means the following: molecular weight no more than 450, logP between −3.5 and 4.5, no more than 5 rings, no more than 10 nonterminal single bonds, no more than 5 hydrogen bond donors, and no more than 8 hydrogen bond acceptors. These criteria were derived from known drug molecules respectively their corresponding lead structures and therefore represent reasonable ranges which help to identify promising leads with potential for later optimization. Conformational ensembles used during compound indexing were calculated for all ligands using an in-house conformer sampling generator which takes torsion angles at acyclic bonds as well as flexible rings with up to 8 bonds into account.

1. Redocking Performance. In the redocking experiments TrixX BMI, FlexX, and GOLD are compared using the 85 protein−ligand complexes from the Astex Diverse Set. All tools use Corina structures as a starting point: This represents a real world virtual screening scenario where the cocrystallized conformation of the ligand is unknown. The results (see Table 1) show that for accurate placements below 1.0 Å FlexX outperforms TrixX BMI. Concerning predictions within 2.0 Å rmsd TrixX BMI is on a par with FlexX and misses only two poses correctly predicted by GOLD. Pose predictions below 2.5 Å, which we believe to be a useful starting point for subsequent postoptimizations, are found for 67 ligands compared to 62 found by FlexX.

These results show that our new index-driven approach is able to reproduce binding poses within an rmsd of 2.0 Å in about as many cases as FlexX and GOLD. In order to further improve pose prediction we have identified two important points. First, the conformational ensembles need to hold a conformation close to the cocrystallized ligand. This allows the descriptor to identify a valid pose prediction. Second, a postoptimization routine able to tweak our predictions toward the native ligand pose needs to be incorporated into the workflow. Since 74 of 85 ligands, corresponding to seven additional useful predictions, are placed within 2.5 Å of the native structure, if all ranks are considered, the adaptation toward a multilevel postoptimization scheme shows a promising route for further improvements.

2) Enrichment Studies. A common problem when setting up enrichment studies is how to choose a library of inactive compounds to avoid an artificial enrichment of known actives. We therefore use the DUD database which supplies tailored sets for different target proteins. Each target has its own set of actives and decoys which have similar molecular properties but differ in chemical structure thus posing a challenging test set. We chose four targets (Cyclin Dependent Kinase 2 (PDB entry 1ckp), Dihydrofolate Reductase (3dfr), Estrogen Receptor (ER) Agonist (1l2i), and ER Antagonist (3ert)) representing different classes of proteins being of interest in pharmaceutical research. Furthermore, pharmacophores for those four targets are available from the literature (see Table 2), and the used actives and decoys represent a diverse set of compounds.

Figure 5 shows the enrichment performance of TrixX BMI compared to that of FlexX and Dock. All tests were run with and without consideration of pharmacophore constraints. As further validation we provide TrixX BMI pose predictions for each of the four targets in Figure 5. All targets are shown with the corresponding pharmacophore, the crystal structure in orange, and the TrixX BMI pose prediction. The constraints are adapted from the literature (see Table 2) and generated using the FlexX-Pharm interface.

In the case of CDK2 interactions to the hinge need to be established. This is a well-suited scenario for TrixX BMI. It consists of rather rigid actives and a hydrophilic pharmacophore which is present in numerous queries describing the active site. The enrichment plot for CDK2 illustrates that TrixX BMI outperforms both FlexX and Dock. The enrichment factor (EF) at 1% is considerably higher for TrixX BMI ($EF_{1\%}$ 17) than for the other tools ($EF_{1\%}$ ~9). DHFR is known to perform rather well in combination with FlexX, and also TrixX BMI shows good enrichment behavior: All tools achieve an EF of about 20 at 1% of the database. Dock basically is in the same range with a slightly diminished performance after the first percent. In case of the ER agonist, the pharmacophore consists of a spatial constraint which is introduced due to a nonspecific flexible

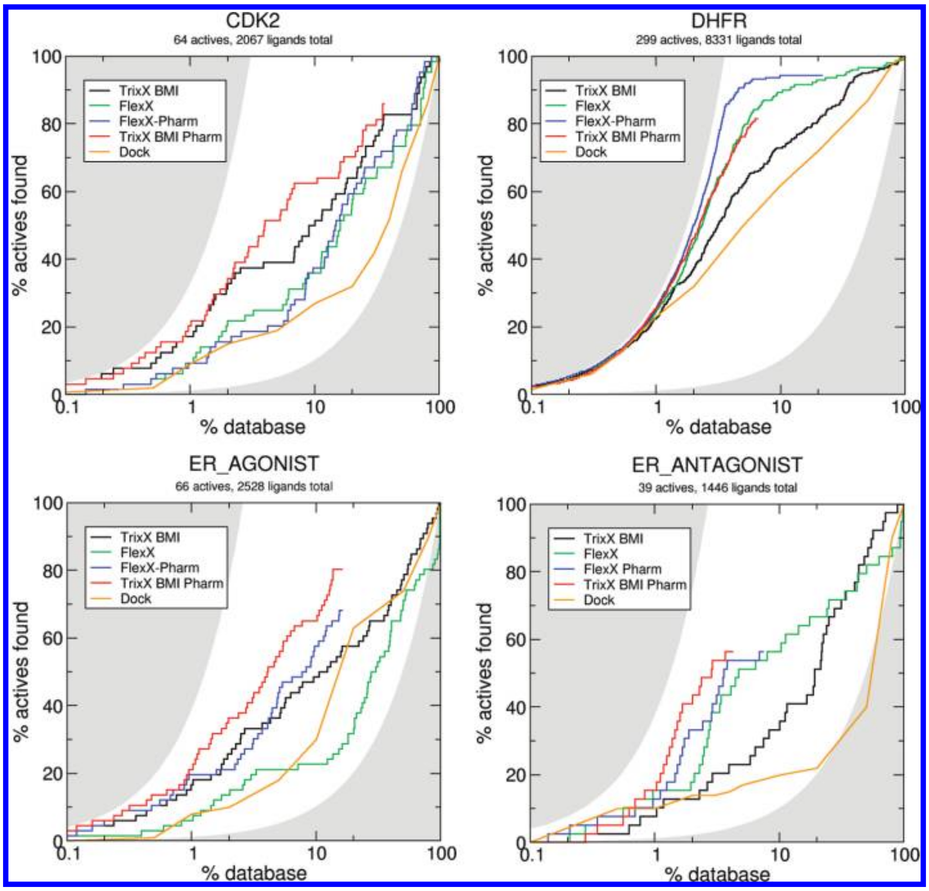


Figure 5. Enrichment plots for four targets using TRIXX BMI, FLEXX, and DOCK. Enrichment curves for DOCK are estimated using the original DUD publication. The gray shaded areas represent maximal (top) and random enrichment (bottom). The *x*-axis is scaled logarithmically to focus on the important range of percentages.

part of the pocket and two hydrophilic interactions. The spatial constraint forces the ligands to be in a rather strained conformation thus filling the entire active site. TRIXX BMI outperforms FLEXX and DOCK with and without usage of this pharmacophore. TRIXX BMI enriches with an $EF_{1\%}$ 18 compared to an $EF_{1\%}$ of less than 10 for all other tools in the unconstraint search. In the constraint search both TRIXX BMI and FLEXX improve their performance significantly ($EF_{1\%}$ 23 compared to $EF_{1\%}$ 20). We chose the ER antagonist as the last target to show that TRIXX BMI is also able to dock larger compounds which are not necessarily leadlike. Enrichment at 1% shows the best results for FLEXX ($EF_{1\%}$ 15) followed by DOCK ($EF_{1\%}$ 10) and TRIXX BMI ($EF_{1\%}$ 10). The introduction of pharmacophore constraints again significantly improves the results for TRIXX BMI ($EF_{1\%}$ 18) outperforming FLEXX ($EF_{1\%}$ 13) in the constraint case.

All four protein targets have active compounds which cannot be placed using the TRIXX BMI constrained search. This is also true for FLEXX-PHARM. In some cases this might be due to a different binding mode of the active ligands which means that the given pharmacophore cannot be obeyed by all of them. Another aspect connected to TRIXX BMI is the quality of the conformational ensembles. If none of the conformations within an ensemble represents a conformation close to the bioactive one, this can also prevent a correct pose prediction due to clashes or a wrong geometry arrangement of the required pharmacophore interactions.

3) Runtime and Space Requirement. All experiments were run in a single thread on an Intel Core2 Duo with 3 GHz and 4 GByte of main memory. The runtime measurements were taken using the target specific catalogs from the DUD and the Z_2 data set which represents a standard leadlike screening library of a productive screening setup. Figure 7 shows distributions of molecular weight, number of rotatable bonds, and calculated logP for all five ligand data sets. Table 3 shows statistics of the DUD targets and ligands. In addition, average values of the Z_2 compound library and theAFX$_{285}$ set are presented. This set consists of all protein targets from the FLEXX200[41] augmented by Astex Diverse Set. It can be seen that different target proteins and compound libraries yield significant differences in the number of site descriptors and compound descriptors. This is why we averaged runtime measurements over 285 targets from the AFX$_{285}$ and used a leadlike compound index based on the Z_2 compound library.

The introduction of pharmacophore constraints reduces the number of query triangles significantly. Concerning the reported runtimes using the DUD ligands it is necessary to point out that those ligands are not necessarily leadlike (see Figure 7) and also highly compatible with the corresponding active site. This means that the descriptor selectivity in this artificial scenario is not as high as in a real world VHTS scenario. Therefore we also include runtimes using the Z_2 as compound catalog thus providing more realistic runtimes for each target.

Table 4 shows the observed runtimes of all experiments. It is obvious that TRIXX BMI offers a substantial speed-up over FLEXX. The average runtime on all 285 targets of AFX$_{285}$ drops by a factor of about 20 to 0.81 s per ligand. This demonstrates that the descriptor based lookup technol-

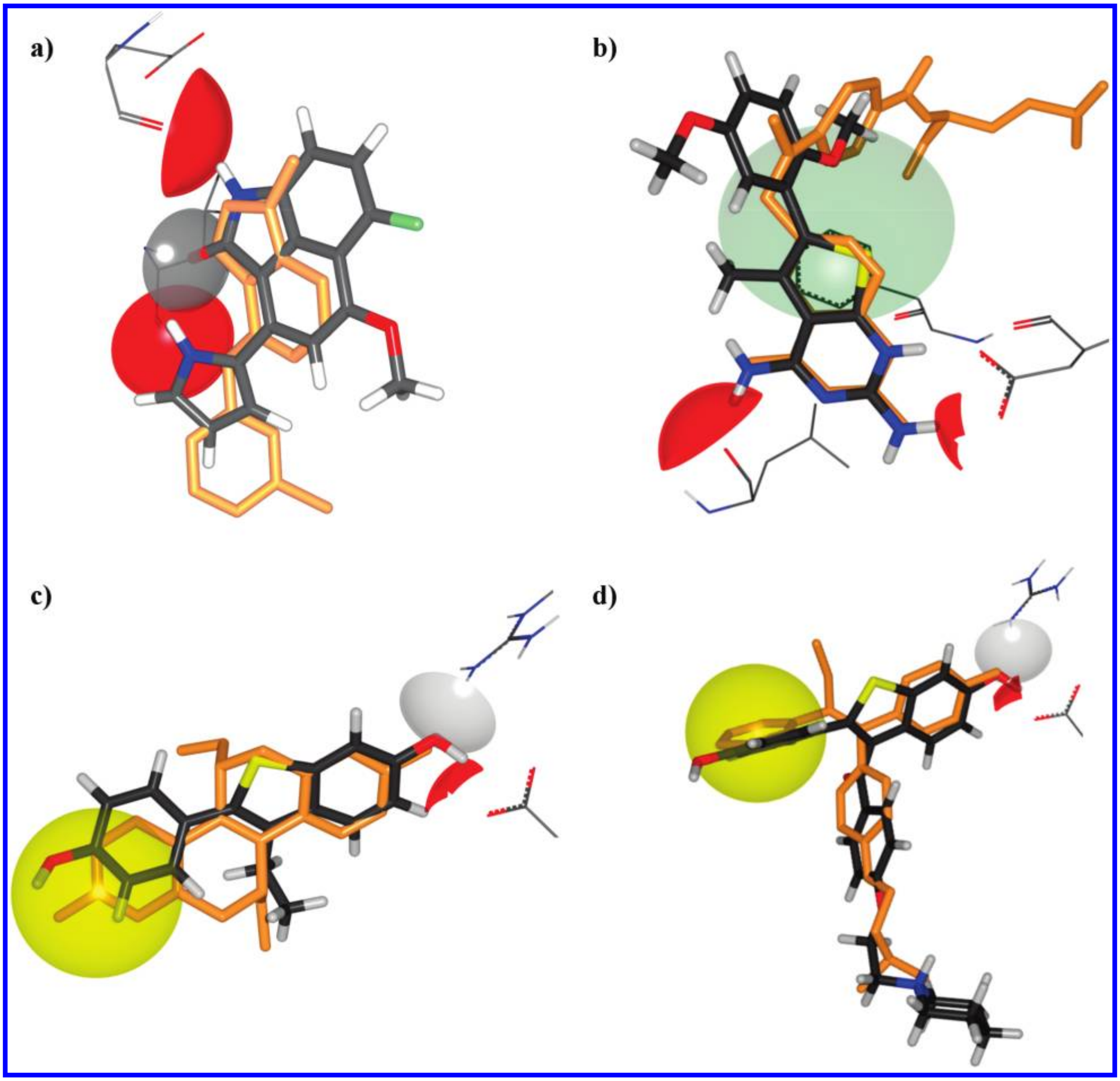


Figure 6. TRIXX BMI poses of the top ranked active after screening the DUD actives and decoys (rank without | with constraints): a) CDK2 (2 | 1), b) DHFR (1 | 1), c) ER agonist (3 | 1), and d) ER antagonist (10 | 5).

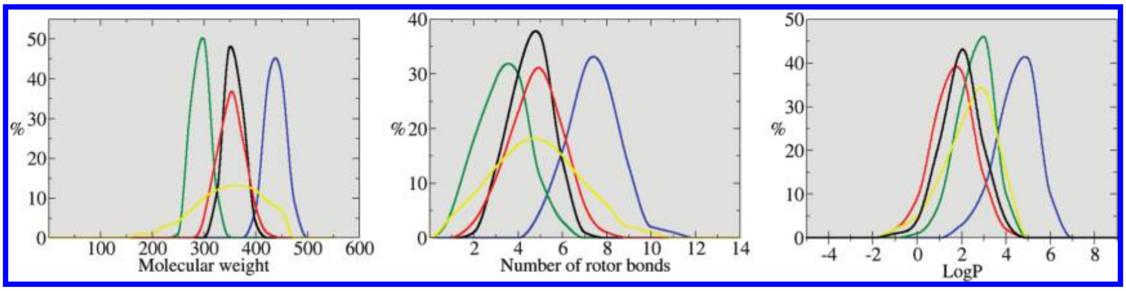


Figure 7. Property distributions of the used ligand sets (actives+decoys). Left: molecular weight, center: number of rotatable bonds, right: logP. (CDK2: black, DHFR: red, ER agonist: green, ER antagonist: blue, Z_2: yellow).

ogy yields an improvement in runtime of 1 order of magnitude to iterative screening approaches.

The targets from the DUD show that TRIXX BMI has good runtime behavior on the given data sets even though they are not strictly leadlike. If pharmacophore constraints are supplied, the already shown reduction of site triangles and the increase of the descriptor selectivity further reduce the runtime. TRIXX BMI in combination with pharmacophore constraints needs only 80 ms for each ligand in the database averaged over all targets. This is true for the target specific catalogs from the DUD database as well as the leadlike Z_2 catalog.

Due to nonavailable pharmacophores only unconstrained runtimes against the Z_2 are supplied for the AFX_{285} set. These values represent a typical screening run against a leadlike compound catalog.

Table 3. Protein and Target/Decoy Data for the Chosen Targets[a]

				no. of site descriptors	
target	actives + decoys	av no. of conformations	av no. of descriptors	without	with constraints
CDK2	64 + 2003	8.32	147	6519	242
DHFR	299 + 8132	8.95	181	24105	59
ER (agonist)	66 + 2452	6.35	98	2544	194
ER (antagonist)	39 + 1407	20.39	155	12477	146
AFX_{285} (averaged)	Z_2	10.46	108	13778	N.A.

[a] The first column shows the number of active and decoy compounds for the given target, the second one the average number of conformations generated. The third column gives the number of descriptors averaged over all ligands in the specific set, and the last two columns give the number of descriptor queries for the given targets, first without and then with pharmacophore constraints.

Table 4. Average Runtimes of FLEXX and TRIXX BMI (in s) for each Compound in the Index, without (and with) Pharmacophore Constraints

	runtime [s] on							
	target specific DUD index				Z_2 index			
target	TRIXX BMI		FLEXX		TRIXX BMI		FLEXX	
CDK2	0.36	(0.05)	6.41	(5.52)	0.22	(0.03)	5.21	(4.07)
DHFR	1.83	(0.02)	9.15	(14.39)	0.97	(0.03)	8.26	(2.66)
ER(agonist)	0.13	(0.06)	6.23	(6.18)	0.09	(0.06)	5.71	(4.65)
ER(antagonist)	0.93	(0.10)	16.61	(21.72)	0.59	(0.10)	7.04	(6.41)
AFX_{285}	N.A.	(N.A.)	N.A.	(N.A.)	0.81	(N.A.)	19.40	(N.A.)

Table 5. Total VHTS Runtimes Using the DUD Targets and 1.7 Million Compounds on 48 Cluster Nodes

	TRIXX BMI (max. runtime [h:min])		FLEXX (runtime [h:min] estimated based on representative results)	
target	without	with constraints	without	with constraints
CDK2	2:04	0:09	37:54	24:20
DHFR	9:43	0:05	58:20	17:06
ER (agonist)	0:53	0:27	43:45	28:36
ER (antagonist)	6:01	0:25	48:37	33:03

The speed-up from standard to pharmacophore mode depends on the reduction of query triangles. If the active site already has a few triangles and if those triangles reside in a closed receptor pocket (e.g., ER agonist) the speed-up is smaller because the unconstrained setup of TRIXX BMI already is rather fast.

4) Scalability. In case of an index-driven screening approach it is important to show the applicability in a parallel computing environment without generating significant overhead thus retaining space and runtime boundaries. We therefore setup TRIXX BMI on a 48 node compute cluster of 2.4 GHz Dual Xenon CPUs with 4 GByte of main memory on each node. In total we use a compound catalog of about 1.7 million compounds. In order to reduce network load and to reduce the system's response time under heavy cluster load, we split the catalog into 96 packages and distributed those to local hard drives, such that each node manages six different packages. This leads to 30 GByte on each node's hard drive which adds up to about 1.5 TByte on the whole compute cluster. The division of the library into packages enables the nodes to run almost independently. Only the final step of merging the cluster results needs to be coordinated.

Table 5 shows the maximal runtime for each target on 48 cluster nodes. We are able to perform virtual screening of 1.7 million ligands using four different target proteins in between 5 and 27 min with pharmacophore constraints. Without constraints it takes between 53 min and just below 10 h. In comparison, a sequential docking tool like FLEXX in the same parallel setup needs at least 17 h in the constraint search up to more than 2.5 days in the unbiased search.

Another important aspect is the time needed to build the necessary indices, which are the basis for all subsequent screening experiments. It is obviously correlated to the average number of conformations and the size of the ligands within the library. For the Z_2 data set the compound indexing phase takes about one second per ligand, which means that a single conformation of a leadlike ligand is processed in about 0.1 s. Thus the setup of a large compound collection consisting of millions of individual compounds can be accomplished overnight on a reasonable sized compute cluster.

CONCLUSIONS AND OUTLOOK

TRIXX BMI has been developed on the basis of FLEXX using the concept of index driven virtual screening. Geometric and chemical properties are used to perform initial descriptor matches, thus selectively identifying hits strictly based on active site properties. With respect to a very early version of TRIXX certain weaknesses are removed. The problem of incompatible fragment placements is solved by preprocessed conformational ensembles and leadlike fragment sizes which can be incrementally constructed to rebuild larger molecules. A new shape descriptor introducing descriptor-based search to structure-based drug design and the use of a bitmap indexing system significantly improved the overall performance with respect to prediction quality and speed.

TRIXX BMI performs on a comparable level to FLEXX and GOLD concerning redocking accuracy; concerning enriching known actives the same holds in comparison to FLEXX and DOCK. Its novel methodology allows the incorporation of more sophisticated calculations, for instance the introduction of receptor flexibility.

The above results were produced within fractions of seconds per ligand in the library. In comparison with FLEXX, already the standard TRIXX BMI approach offers a speed-up of 1 order of magnitude. In combination with additional pharmacophore constraints the runtime decreases by 2 orders of magnitude. Further experiments on a large compute cluster show that TRIXX BMI scales well, thus enabling virtual high-throughput screening of millions of compounds within minutes rather than days.

The experiments performed suggest that TRIXX BMI in its current state is suited as a fast screening tool suggesting reasonable poses. To the best of our knowledge, there is currently no other virtual screening tool supporting a strictly structure-based approach and a protein−ligand clash detection based on descriptors, without direct access to individual compounds. These features allow fast screening of large compound libraries. If the application calls for more accurate poses the decrease in runtime leaves enough time for a more detailed analysis. First postoptimization runs of all TRIXX BMI poses using YASARA[42] show promising results. Each complex was energy-minimized with the AMBER03 force field,[43] using a 7 Å force cutoff and subsequently minimized using steepest descent and simulated annealing. This procedure results in a 10% increase of poses redocked within 2 Å rmsd and even a 30% increase for real accurate pose predictions below 1.0 Å rmsd. Therefore, we believe that TRIXX BMI is a valuable tool, especially if pipelined with a more accurate postoptimization docking tool.

A challenging goal for the future will be to use multiple protein target structures to address protein flexibility. The identification of rigid protein parts and therefore identical descriptors across multiple protein conformations as well as the independent preprocessing of the protein structure suggest that TRIXX BMI screening technology is also suited for flexible protein docking.

REFERENCES AND NOTES

(1) Rester, U. From Virtuality to Reality - Virtual Screening in Lead Discovery and Lead Optimization: A Medicinal Chemistry Perspective. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (4), 559–568.

(2) Gribbon, P.; Lyons, R.; Laflin, P.; Bradley, J.; Chambers, C.; Williams, B.; Keighley, W.; Sewing, A. Evaluating Real-Life High-Throughput Screening Data. *J. Biomol. Screening* **2005**, *10*, 99–107.

(3) Leach, A.; Hann, M. The in Silico World of Virtual Libraries. *Drug Discovery Today* **2000**, *5* (8), 326–336.

(4) Reynolds, C.; Tounge, B.; Bembenek, S. Ligand Binding Efficiency: Trends, Physical Basis, and Implications. *J. Med. Chem.* **2008**, *51* (8), 2432–2438.

(5) Hann, M.; Leach, A.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Model.* **2001**, *41* (3), 856–864.

(6) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6* (3), 211–219.

(7) 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.

(8) 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.

(9) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM -- A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15* (5), 488–506.

(10) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(11) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489.

(12) Warren, G. L.; Andrews, C. W.; Cpelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2005**, xx. in press.

(13) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.

(14) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49* (20), 5851–5.

(15) Rarey, M.; Degen, J.; Reulecke, I. Docking and Scoring for Structure-Based Drug Design. In *Bioinformatics - From Genomes to Therapies*; Lengauer, T., Ed.; Wiley-VCH: Weinheim, 2005; Vol. 2, pp 541−600.

(16) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J. Med. Chem.* **2008**, *51* (20), 6237–6255.

(17) Teodoro, M. L.; Kavraki, L. E. Conformational flexibility models for the receptor in structure based drug design. *Curr. Pharm. Des.* **2003**, *9*, 1635–1648.

(18) Osman, G. *Pharmacophore Perception, Development, and Use in Drug Design*; International University Line: La Jolla, CA, 2000.

(19) Langer, T.; Hoffmann, R.; Mannhold, R. *Pharmacophores and Pharmacophore Searches*; Wiley-VCH: Weinheim, Germany, 2006.

(20) Joseph-McCarthy, D.; Thomas(IV), B. E.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 172–188.

(21) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13* (4), 505–524.

(22) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411–428.

(23) *Discovery Studio, version 2.1*; Accelrys: San Diego, CA, 92121, U.S.A, 2008.

(24) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–9.

(25) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(26) *Unity, version 8*; Tripos Inc.: St. Louis, MO, U.S.A., 2008.

(27) Ebalunode, J. O.; Ouyang, Z.; Liang, J.; Zheng, W. Novel Approach to Structure-Based Pharmacophore Search Using Computational Geometry and Shape Matching Techniques. *J. Chem. Inf. Model.* **2008**, *48* (4), 889–901.

(28) *OEShape Toolkit, version 1.6*; OpenEye Scientific Software: Santa Fe, NM, 87507, U.S.A., 2006.

(29) Schellhammer, I.; Rarey, M. TrixX: Structure-Based Molecule Indexing for Large-Scale Virtual Screening in Sublinear Time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223–238.

(30) Schellhammer, I.; Rarey, M. FlexX-Scan: Fast Structure-Based Virtual Screening. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 504–517.

(31) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. *Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998)*; IUPAC: 1998; Vol. 70, pp 1129−1143.

(32) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.

(33) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63* (4), 892–906.

(34) Wu, K. FastBit: an efficient indexing technology for accelerating data-intensive science. *J. Phys.: Conf. Ser.* **2005**, *16*, 556–560.

(35) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; 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.

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

(37) Hindle, S. A.; Rarey, M.; Buning, C.; Lengauer, T. Flexible Docking under Pharmacophore Type Constraints. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129–149.

(38) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–82.

(39) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(40) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H.-J.; Dean, P. M. A validation study on the practical use of automated de novo design. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 459–478.

(41) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228–241.

(42) *YASARA, version 8*; YASARA Biosciences: 8042 Graz, Austria 2008.

(43) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R. T. L. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

CI9000212