J Comput Aided Mol Des (2007) 21:223–238
DOI 10.1007/s10822-007-9103-5

ORIGINAL PAPER

TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time

Ingo Schellhammer · Matthias Rarey

Received: 18 August 2006 / Accepted: 5 January 2007 / Published online: 9 February 2007

Abstract Structure-based virtual screening today is basically organized as a sequential process where the molecules of a screening library are evaluated for instance with respect to their fit with a biological target. In this paper, we present a novel structure-based screening paradigm avoiding sequential searching and therefore enabling sublinear runtime behavior. We implemented the novel paradigm in the virtual screening tool TrixX and successfully applied it in screening experiments on four targets from relevant therapeutic areas. With the screening paradigm implemented in TrixX, we propose some important extensions and modifications to traditional virtual screening approaches: Instead of processing all compounds in the screening library sequentially, TrixX first analyzes the geometric and physicochemical binding site characteristics and then draws compounds with matching features from a compound catalog. The catalog organizes the compounds by their physicochemical and geometric features making use of relational database technology with indexed tables in order to support efficient queries for compounds with specific features. A key element of the compound catalog is a highly selective geometric descriptor that carries information on the type of functional groups of the compound, their Euclidian distance, the preferred interaction direction of each functional group, and the location of steric bulk around the triangle. In a re-docking experiment with 200 protein–ligand complexes, we could show that TrixX is able to correctly predict the location of ligand functional groups in co-crystallized complexes. In a retrospective virtual screening experiment for four different targets, the enrichment factors of TrixX are comparable to the enrichment factors of FlexX and FlexX-Scan. With computing times clearly below one second per compound, TrixX counts among the fastest virtual screening tools currently available and is nearly two orders of magnitude faster than standard FlexX.

Keywords Structure-based virtual screening · Molecular descriptor · Relational database · Geometric hashing · Molecular docking · TrixX · FlexX

I. Schellhammer · M. Rarey (✉)
Center for Bioinformatics, Research Group for Computational Molecular Design, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

Introduction

Virtual Screening (VS) has become a central element of modern drug discovery [1, 2]. Often used as a complementary approach to assay-based high-throughput screening (HTS), virtual screening tools support the generation of focused, yet diverse screening libraries and the selection of a small set of hit candidates. Here, virtual screening includes receptor-based screening techniques, ligand-based screening techniques, as well as fast and less accurate selection algorithms bases on molecular properties (e.g., Lipinski's rule of five). In the context of this work, we focus on receptor-based screening techniques, also referred to as molecular docking approaches. With the structural elucidation of ever more screening targets and the growing size of screening libraries, two key challenges arise for modern virtual screening methods.

First, the large number of compounds in today's virtual screening libraries requires screening methods with very high selectivity, or in other words, a high enrichment factor [1, 3]. The selectivity of a virtual screening method is primarily driven by the quality of the underlying scoring scheme for estimating the relative binding affinity of all screened compounds [4].

The second challenge consists in increased runtime requirements for virtual screening methods. Today, screening methods have to process hundreds of thousands of compounds. In order to address protein flexibility and drug selectivity issues, these compounds are often docked against a number of protein structures (receptor-based screening) or conformers of known actives (ligand-based screening). In the context of the functional elucidation of proteins, small molecules are docked into the binding sites of proteins and receptors in order to investigate their function [5]. In the past years, increased hardware performance and massive parallelization allowed for keeping pace with the rising throughput requirements. By contrast, little progress has been made in improving the fundamental runtime behavior of molecular docking beyond the sequential approach of most screening techniques.

Researchers have proposed various concepts for speeding up the time-consuming tasks of sampling the binding mode and estimating binding affinity. The majority of approaches aim at the simplification of the docking problem itself. Most docking methods today make use of discretized models for conformational space and for possible functional group positions (e.g. even-spaced grids or spheres of different sizes populating the binding site [6]). Furthermore, pharmacophoric constraints defining necessary interactions to certain amino acids of the protein help to reduce the search space [7–9]. Alternatively, simplified compound representations or scoring functions can be used to bridge the gap to similarity and pharmacophore searching [10–15]. Models of increasing complexity can then be combined to a single tool applying the models in a hierarchical fashion. Prominent examples for hierarchical screening protocols include Glide [16] and HierVLS [17].

Although these developments represent important improvements with respect to the efficiency of modern molecular docking tools, none of these concepts renovates the sequential approach of molecular docking. There are only few approaches that deviate from the strict sequential docking protocol and consequently achieve substantial runtime improvements. Special-purpose docking tools for combinatorial libraries exploit the redundant structure of the compounds [18, 19]. This approach is capable of speeding up the docking process by one to two orders of magnitude, depending on the degree of redundancy in the combinatorial library. Cluster-based screening approaches exploit redundancies in non-combinatorial compound libraries by grouping compounds, fragments, or pharmacophores into distinct clusters. During screening, only one representative of each cluster needs to be docked into the protein-binding site. Thus, the overall computational cost of library screening, especially for very large libraries, is significantly decreased. A method following this approach is NWDOCK developed by Lorber et al. [20] and Su et al. [21]. The former approach by Lorber puts all compounds into ensembles by identifying a common rigid fragment. The rigid fragment of such an ensemble is placed into the receptor-binding site only once. Next, all compounds of this ensemble are overlaid with the placed rigid fragment and conformationally expanded. Su et al. extended this method by grouping similar conformers of different expanded molecules and docking representative instances of these conformer ensembles. Joseph-McCarthy et al. [22] have published another cluster-based docking approach called PHDOCK. The algorithm describes compound conformers by means of a three-dimensional pharmacophore of functional ligand groups and their relative spatial arrangement. In a preprocessing step, PHDOCK conformationally expands all compounds of the database, determines the pharmacophore descriptor of each conformer, and clusters similar pharmacophores that share the same types of functional groups in a similar spatial arrangement. PHDOCK docks all pharmacophores into the receptor-binding site using an adapted version of the DOCK 4.0 docking procedure. If a pharmacophore is successfully placed into the binding site, all associated conformers of that pharmacophore are overlaid with the pharmacophore and scored. The authors report a speed-up factor between 5 and 8 compared to a standard DOCK 4.0 screening run solely due to the overlay of similar conformers in one pharmacophore. Schnecke et al. developed the molecular docking program SLIDE [23, 24] that uses a four-level hashing scheme for placing ligands into a binding site with flexible side chains. In their first version SLIDE 1.0, a grid-based search method identifies favorable hydrophilic and hydrophobic site points. A geometric hashing algorithm superposes these site points with triplets of complementary functional groups of ligand conformers. The hashing scheme includes the types of matched site points, the perimeter of the triangle, the longest and the shortest triangle side. The anchor fragment containing the superposed triangle is considered rigid, whereas all outgoing single bonds to adjacent fragments of the ligand conformer are

considered rotatable. A mean force-based fitting algorithm optimizes the rotation of protein side chains and adjacent ligand fragments and displaces water particles as necessary. The authors report of docking 175,000 compounds within minutes or a day, depending on the complexity of the binding site template. The most time consuming part of the docking algorithm is the conformational sampling of ligand substituents and protein side chains.

With TrixX, we introduce a prototype for a structure-based virtual screening tool featuring a target-driven screening paradigm instead of a sequential compound-by-compound screening paradigm. The novel approach is to overcome the direct linear dependence of the runtime requirements of sequential molecular docking tools from the number of screened compounds. Essentially, TrixX describes the target by a set of triplets of favorable spots for ligand functional groups and then uses an indexed molecular descriptor for drawing compounds from a pre-calculated relational database that satisfy the local preferences and constraints of the spot triplets.

In the remainder, we describe this novel paradigm and the key elements of TrixX in detail. We then provide a first validation of the different novel elements of TrixX and discuss the performance of our prototype in re-docking experiments and in enrichment experiments. A comparison with FlexX allows for delineating the specific strengths and weaknesses of TrixX. In a series of enrichment experiments with increasing library sizes, we investigate the runtime behavior of TrixX and discuss its potential future use in large-scale structure-based virtual screening.

Materials and methods

General overview of the TrixX approach

The overall process flow of a virtual screening experiment with TrixX divides into five phases (see Fig. 1).

The first phase, *compound cataloging*, is only performed once for registering all compounds of the virtual screening library in the compound database. Subsequent screening experiments then directly extract all compound information from the compound database. TrixX reads the structure of a compound from a flat file and identifies functional groups of the compound (compound interaction centers, CIACs). The algorithm then decomposes the compound into partly overlapping fragments such that each fragment covers a number of the functional groups of the compound (now called fragment interaction centers, FIACs). The unique SMILES notation [25, 26] identifies a fragment unambiguously by applying a canonical ordering scheme to the fragment atoms and bonds. This allows TrixX to check whether the fragment is already in the database. If that is not the case, a fragment is conformationally sampled and TrixX calculates the resulting fragment interaction triangle descriptor of each FIAC triplet for each fragment conformation. The triangle descriptor contains the types of the covered FIACs, the triangle side lengths, the interaction directions, and the location of steric bulk of the fragment exceeding the triangle boundaries. Compound, CIACs, fragments, FIACs, fragment conformations, and fragment interaction triangles are written to the compound database.

Phases two to five are target-specific and are therefore repeated in each new virtual screening experiment. In phase two, *site analysis*, TrixX reads the structure of the receptor and its binding site from a flat file and identifies favorable positions (site interaction centers, SIACs) where functional groups of ligands can interact with functional groups of the receptor. Triplets of site interaction centers form so-called site interaction triangles. Here, the triangle descriptor of TrixX encodes the interaction center types, side lengths, interaction directions, and the presence of steric bulk of the receptor extending into the triangle boundaries. All site interaction triangle descriptors are written to another table of the compound database.

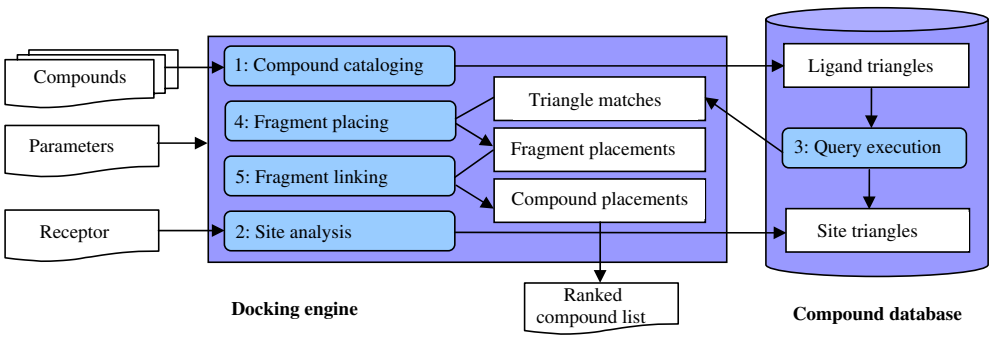

Fig. 1 Five phases of the overall TrixX process flow shown by light blue boxes

After registering both fragment and site triangles in the compound database, TrixX looks up pairs of matching fragment triangles for each site triangle (*query execution*, phase three). For this purpose, a script in the relational database system loops over all registered site triangles. The geometric properties of a site triangle plus some tolerance values define value ranges for the side lengths and interaction directions of matching fragment triangles. Together with the FIAC types that are complementary to the SIAC types of the site triangle these values and value ranges are used as constraints for a query on the fragment triangle table. Steric bulk constraints are not part of the query. For each site triangle, this query returns a number of hits, i.e. fragment triangles that more or less match the physicochemical and geometric constraints of the site triangle. The database script then checks for each hit whether the steric bulk descriptors of the fragment triangle and the site triangle indicate a clash. Non-clashing hits with sufficient agreement to the geometric constraints of the site triangle are returned to the docking engine as triangle matches.

In phase four, *fragment placing*, TrixX transforms triangle matches into fragment placements. In order to do so, the algorithm reads the fragment conformer associated to the fragment triangle from the compound database. The superposition of the site triangle with the fragment triangle defines the translation and rotation for placing the fragment conformer in the receptor-binding site. Clashing fragment placements with a too large receptor–ligand overlap are discarded. TrixX scores the interaction contributions from the superposed FIACs and SIACs and stores each fragment placement in fragment-specific priority queues based on their score.

After all triangle matches have been transformed into fragment placements, TrixX links the placements of fragments of the same *compound to compound* placements (*fragment linking*, phase five). For this purpose, the algorithm queries for each placed fragment in which compounds it occurs. The placements of two different fragments of the same compound are merged into *one compound placement* if their overlapping parts lie approximately upon each other. Also, placements of non-overlapping fragments of the same compound can be merged if a compound conformation exists which would make the relative spatial arrangement of the FIACs of each fragment possible. For this purpose, TrixX holds a matrix of minimal and maximal pairwise distances of the interaction centers of a compound over all conformations. The resulting compound placements are tested for intra- and inter-molecule clashes. Moreover, the algorithm searches for additional interactions between unsaturated interaction centers of the compound and the receptor. Finally, compound placements are scored and the best placements are stored in compound-specific priority queues. The scores of the best compound placements allow for compiling a ranked list of all compounds with respect to their estimated binding affinity to the target. Note that this list will only contain those compounds for which TrixX found at least one triangle match and subsequently a valid placement.

Fragment clustering approach

Fragment clustering addresses the redundancy of large compound collections where many compounds share common molecular substructures. Similar to other cluster-based screening approaches, TrixX decomposes compounds into smaller, overlapping molecular fragments. Here, a fragment typically covers only a fraction of the compound atoms but contains at least three fragment interaction centers. Dummy atoms mark bonds to adjacent parts of a compound. Hence, the representation of the fragment is independent of the compound it has been derived from. TrixX uses the unique SMILES notation extended by an outgoing bond symbol for describing a fragment and for identifying common fragments of different compounds in a library (see Fig. 2).

The idea is that the more compounds are in the library, the more the fragments repeat and the less new fragments have to be added to the library. The description of a fragment is independent of the screening target, so compound cataloging is a one-time effort and does not affect docking runtime. After placing all fragments into the receptor-binding site (phase four), TrixX tries to link the placements of fragments belonging to the same compound in order to generate solutions containing two, three, or all compound fragments, respectively (phase five). This fragment placing and linking strategy is a major difference to FlexX and other prominent molecular docking tools that mostly apply incremental build-up, random search, or multi-conformer docking strategies. It implies that TrixX can only generate a complete docking solution for a compound if all its fragments could both be placed and linked successfully. The fragment linking step is only necessary for compounds where at least one of its fragments could be successfully docked, all other compounds will not be processed during a virtual screening run.

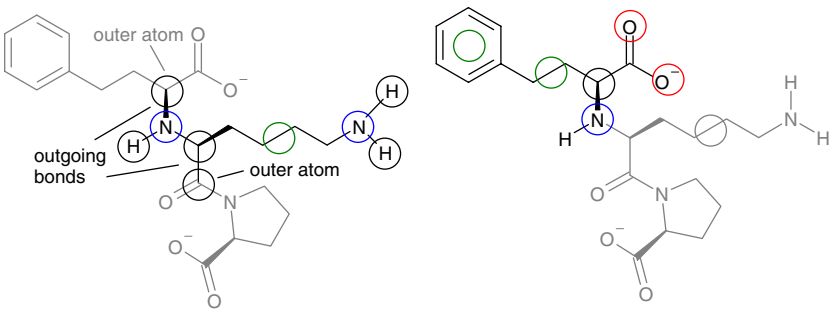

Fig. 2 Two sample fragments of a compound, outgoing bonds (transition from black to gray lines) are marked in the unique SMILES notation by additional '°' symbols

Comprehensive yet compact molecular descriptor

TrixX features a novel molecular descriptor for representing the physicochemical and geometric properties of a compound, i.e. of its fragments. The typical atom-and-bond representation of a compound is reduced to a graph of functional groups, here called compound interaction centers (CIACs). Figure 3 provides an example for the graph reduction of a compound.

TrixX distinguishes hydrogen bond donors, hydrogen bond acceptors, metals, hydrophobic groups, and dummy interaction centers. The latter represent outgoing bonds of fragments that have been cut from larger compounds. The same interaction center types are used for the representation of favorable interaction spots in the target-binding site (except for dummy interaction centers). The algorithm for detecting those spots has been described in detail in our previous publication on FlexX-Scan [15]. Table 1 shows which compound functional groups carry what type of interaction center and where the center of an interaction center is located.

Each interaction center of a compound has an interaction geometry associated to it. The surface of the geometry defines the location of a putative counter-group in order to form a hydrogen bond, realize a hydrophobic contact, or coordinate a metal ion, respectively. Figure 4 shows the basic CIAC interaction geometries in TrixX.

Table 1 Types of compound interaction center (CIAC) types, their functional groups, and the location of their center

ID	CIAC type	Functional groups	Location of the CIAC center
0	Hydrogen bond donor	NH^+, NH, OH, COOH	Hydrogen atom
1	Hydrogen bond acceptor	COO^-, CO, COOH, N_{aryl}, N_{amino}, OH, COC	Nitrogen/oxygen atom
2	Metal ion	N/A	N/A
3	Hydrophobic group	Aromatic rings	Ring center
		Methyls (CH_3)	Carbon
		Ethyls (C_2H_2, C_2H_3, C_2H_4)	Bond center
4	Dummy center	Outgoing bonds of a fragment	Outer bond atom

These interaction geometries are a simplification of the interaction geometries used in FlexX. Here, all geometries are rotationally symmetric and can be described by the location of the interaction center, a main interaction direction, an opening angle, and the radius of the conical interaction surface (see Table 2).

TrixX selects a number of triplets of fragment interaction centers (FIACs) for describing the physicochemical and geometric properties of a conformationally flexible fragment. Note that a *FIAC triplet*

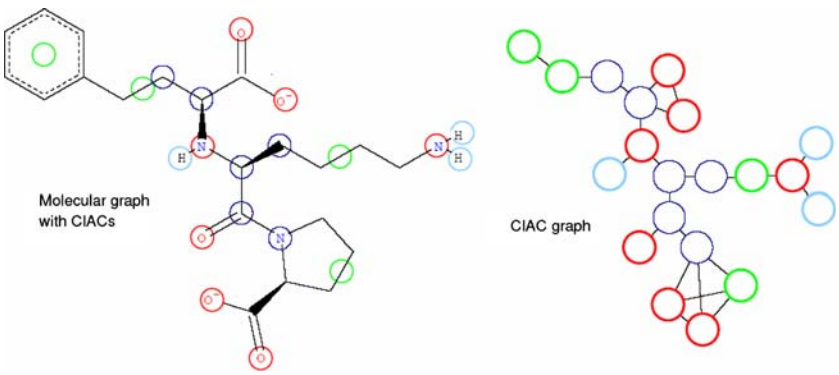

Fig. 3 Representation of compounds as CIAC graphs

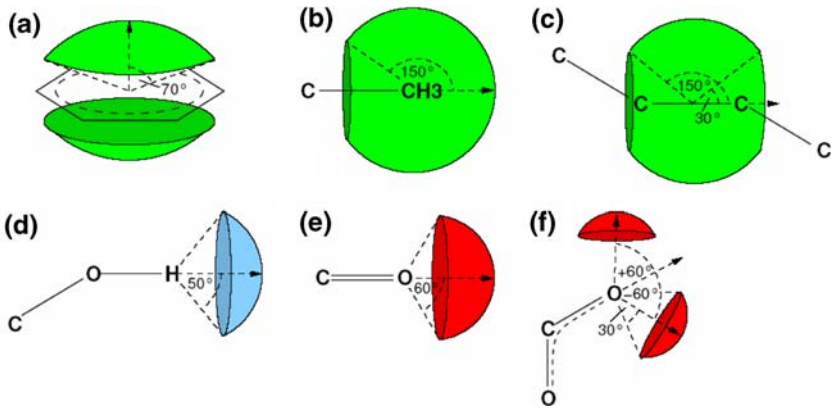

Fig. 4 Representation of compounds interaction geometries in TrixX: (**a**) Aromatic rings, (**b**) methyl groups, and (**c**) ethyl groups on aliphatic carbon chains represent hydrophobic groups of a compound. (**d**) Hydrogen donor geometries are centered in the hydrogen atom, whereas the interaction geometry of hydrogen bond acceptor groups is centered in the electronegative hetero atom (**e–f**). TrixX represents the characteristic geometry of carboxylate groups by two cones for each oxygen atom (**f**)

Table 2 Opening angles and radii of different types of compound interaction centers

CIAC type	Subtype	Opening angle (degrees)	Radius (Å)
Hydrogen bond donors	Charged groups	50	1.8
	Uncharged groups	50	1.9
Hydrogen bond acceptor	Carboxylate groups	30 (two cones)	1.8
	Other	60	1.9
Metal ions	All	360	1.9
Hydrophobic groups	Phenyl rings	70	4.5
	Ethyl groups	30–150	4.0
	Methyl groups	150	4.0
	Halogens	150	4.8

does not depend on the fragment conformation. TrixX uses the methods of FlexX for a thorough conformational sampling of the fragment. For each fragment conformer, TrixX calculates the molecular descriptor of each *FIAC triangle*. In contrast to a FIAC triplet, a FIAC triangle belongs to a specific fragment conformation and thus has a defined triangle geometry. This includes the pairwise distances of the FIACs, the directional orientation of each FIAC, and the location of steric bulk of the fragment conformer (see Fig. 5).

For the purpose of quick comparability, the attributes of the molecule descriptor must refer to a local coordinate system where no rotation or translation is

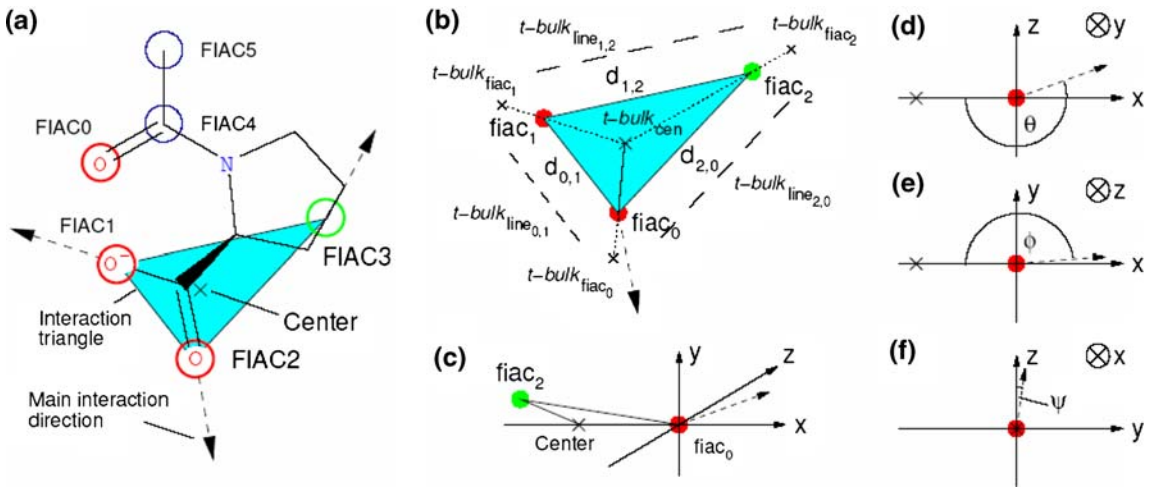

Fig. 5 A triplet of FIACs (**a**) forms a fragment interaction triangle (**b**). For representing the directional interaction preferences of a FIAC, a local coordinate system is placed onto the FIAC (**c**). Euler angles describe the interaction direction relative to the axis of the local coordinate system (**d–f**)

necessary for comparing the physicochemical and geometric properties of two FIAC triangles. TrixX uses a canonical ordering scheme for sorting the FIACs in different triangles in a consistent way. The vertices of a fragment triangle, that is FIACs i_0, i_1, and i_2, are in canonical order if the types of the FIACs t_{i0}, t_{i1}, and t_{i2} and the distances to the subsequent FIAC $d_{i0,i1}$, $d_{i1,i2}$, $d_{i2,i0}$ adhere to a lexicographical order

$$(t_{i0}, d_{i0,i1}) \leq_{\mathrm{L}} (t_{i1}, d_{i1,i2}) \leq_{\mathrm{L}} (t_{i2}, d_{i2,i0}).$$

In the example of Fig. 5b), the canonical ordering of the FIACs would be fiac_0, fiac_1, fiac_2, or $i_0 = 0$, $i_1 = 1$, $i_2 = 2$.

All elements of the molecular descriptor are arranged according to this canonical order. Using the centers of the ordered FIACs, one can define a local coordinate system for encoding the directional orientation of each FIAC: Its origin coincides with the FIAC center (e.g., fiac_0 in Fig. 5c)). The x-axis points from the triangle center through the FIAC center. The z-axis runs perpendicular to the x-axis through the triangle plane such that the subsequent FIAC (e.g., fiac_2 in Fig. 5c)) has a positive z-value. The direction of the y-axis corresponds to the cross-product of x- and z-axis (right-hand rule). We use three Euler angles θ, ϕ, and ψ referring to this local coordinate system for encoding the directional orientation of an interaction center (cf. Fig. 5d–f). The use of Euler angles allows for directly comparing the directional orientation of two interaction centers. This is a key property for an efficient look-up of fragment triangles with matching directional preferences from the database.

Finally, the descriptor encodes where steric bulk of the fragment conformer exceeds the triangle sides. For this purpose, each triangle side is shifted horizontally (in the triangle plane) by 2.0 Å from the triangle center, and then shifted vertically (perpendicular to the triangle plane) by 1.5 Å both atop the triangle plane (labeled *t-bulk*) and below the triangle plane (labeled *b-bulk*). As a result, there are three shifted triangles sides above the triangle plane and three shifted triangle sides below it. All of these are equally divided into nine line segments. Figure 5b illustrates the shifted and segmented sides atop of an example triangle. The bit vectors $tbulk_{\mathrm{line}_{0,1}}$, $tbulk_{\mathrm{line}_{1,2}}$, and $tbulk_{\mathrm{line}_{2,0}}$ represent each of the nine line segments by one bit. A bit is set if the van der Waals volume of any fragment atom intersects or exceeds the respective line segment. In addition to the three triangle sides, the descriptor also uses the similarly shifted triangle vertices ($tbulk_{\mathrm{fiac}_0}$, $tbulk_{\mathrm{fiac}_1}$,$tbulk_{\mathrm{fiac}_2}$) and the vertically shifted triangle center ($tbulk_{\mathrm{cen}}$) as probe points for identifying fragment bulk that significantly exceeds the shifted triangle boundaries. In total, the descriptor requires 31 bits for encoding steric bulk above the triangle plane (3*9 line segments, three triangle vertices, one triangle center) and another 31 bits for steric bulk below the triangle plane. Assuming that Euler angles and triangle side lengths will be stored with 32 bit precision and that the combination of FIAC types can also be encoded in 32 bits, the complete triangle descriptor takes about 480 bits or 60 bytes. The low space requirement makes the descriptor applicable to cataloging large virtual compound libraries.

Indexed relational database technology

The third central method of TrixX is the use of fast and scalable relational database technology with table indices. Upon cataloging, all necessary information about compounds, fragments, and fragment conformers is written to tables of a relational database, the so-called compound database. Figure 6 shows the entity relationship diagram of the compound database.

The advantage of storing compound information in indexed relational database tables is the ability to directly access entries satisfying certain search attributes from the tables. TrixX exploits this feature when looking up fragment triangles that match the physicochemical and geometric properties of a site triangle of the target. The compound database organizes the attributes FIAC types, side lengths and Euler angles of all fragment triangles in a B*-tree table index [27]. B*-trees are balanced search trees that allow accessing tree entries in logarithmic time. In addition, this index structure supports fast range queries that take tolerance values for acceptable triangle side lengths and Euler angle deviations into account. Due to the canonical ordering scheme, interaction center types and triangle side lengths of the search query can directly be compared with the entries in the fragment triangles table. Likewise, Euler angle intervals of the search query are directly comparable with the entries in the fragment triangles table since all Euler angles are defined with respect to the same triangle-dependent local coordinate system. Note that the steric bulk descriptor cannot be used for the B*-tree table index since assessing the compatibility of two steric bulk descriptors requires a one-to-one comparison of all positions of the bit vector. Hence, TrixX checks the steric bulk compatibility in a postprocessing step for each query hit.

Scoring function

Due to the non-availability of individual atom coordinates, TrixX uses a slightly modified version of

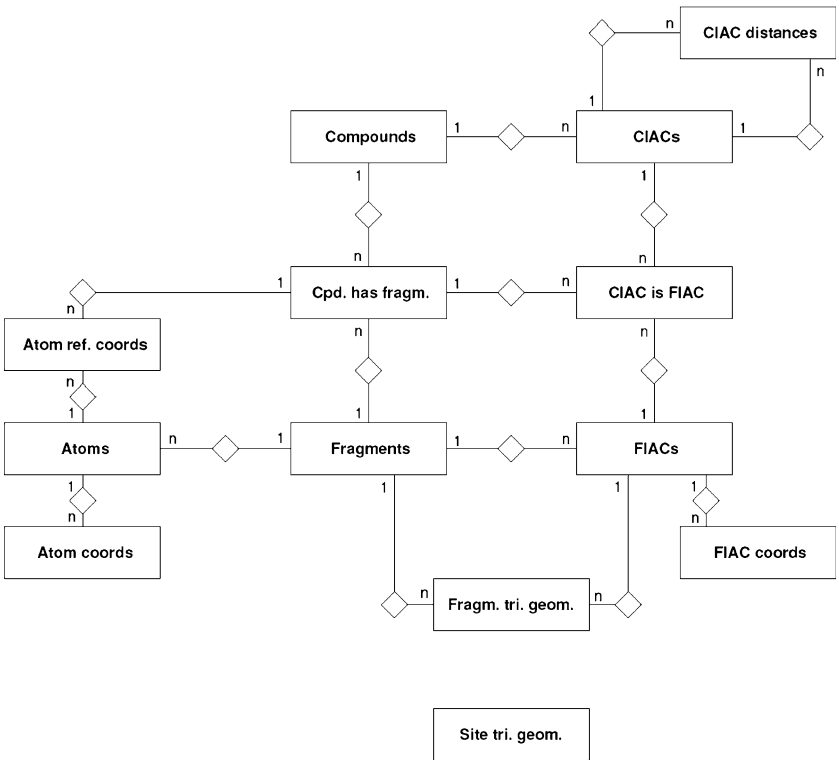

Fig. 6 Tables and relationships in the compound catalog database

the empiric scoring function of FlexX based on the CIAC–SIAC matching of a docking solution. The score includes the sum of the pairwise interaction energies of all matched site and compound interaction centers, scaled by the deviation from the ideal interaction geometry. The score contribution of charged groups (0.0–8.3) is higher than for uncharged hydrogen bonds (0.0–4.7). The contribution of hydrophobic groups scales with a discrete approximation of the contact surface between matched compound and site interaction center (aromatic compound rings: 0.0–7.0; other hydrophobic compound interaction centers: 0.0–4.0).

The sum of interaction energies is reduced by a penalty for each hydrophilic, unsaturated site interaction center that is partly or fully buried by a hydrophobic compound interaction center (hydrogen bond acceptors: 0.0–2.35; hydrogen bond donors or metal ions: 0.0–4.7). Furthermore, TrixX approximates the steric receptor-compound and intra-compound overlap and subtracts the corresponding van der Waals energy contribution from the score (0.5 per A^3). For this purpose, TrixX calculates the steric volume of each CIAC based on its location, type and directional orientation. Finally, there is a penalty for each rotatable bond rigified by the receptor–ligand complex (0.7 per bond). For scoring fragment placements, TrixX neglects the term for rotatable bonds and does not compute the intra-compound clash volume.

Results

We validated the performance of TrixX in several aspects: First, we analyzed how TrixX can successfully reproduce the binding mode of known co-crystallized receptor–ligand complexes. Here, we used a set of 200 complexes from the PDB [28]. Second, we tested the enrichment behavior of TrixX for four different target proteins (Cyclin Dependent Kinase 2 (1jvp), Thrombin (1dwd), Dihydrofolate Reductase (4dfr), and Angiotensin Converting Enzyme (1o86)). Using a subset of the World Drug Index, we composed a screening library with about 8,200 drug-like compounds and known active compounds of the four targets. Third, we investigated the runtime and space behavior of TrixX with larger screening libraries. For this purpose, a collection of 130,000 compounds collected from various public sources was augmented by the active compounds of the four targets.

TrixX re-docking performance

First, we validated the ability of TrixX to correctly reproduce the binding mode of co-crystallized receptor–ligand complexes. For TrixX, it is essential that the fragment interaction centers lie close enough to the site interaction centers that they interact with in the receptor–ligand complex. The coordinates of

Table 3 Percent of correctly placed fragment instances per compound for different RMSD classes

RMSD (Å)	Fragments docked within RMSD class				
	No. of solutions (total: 200)				
	1	≥25%	≥50%	≥75%	All
≤1.0	60	49	28	18	18
≤1.5	97	84	61	37	36
≤2.0	120	107	77	49	44
≤2.5	140	126	96	63	57

These results are based on the placements of each fragment with lowest RMSD to the crystal structure (at any rank). Twelve out of the 200 receptor–ligand complexes did have no placement for any of their fragment instances

non-interacting atoms are not available to TrixX during screening. Therefore, we only used the fragment interaction centers instead of all fragment atoms for the calculation of the RMSD from the crystal structure. We counted how many fragments of a compound TrixX is able to correctly place with different RMSD accuracy. The test dataset comprises 200 co-crystallized complexes [28].

As one can see from Table 3, TrixX successfully places all fragments of a compound in 44 cases (RMSD ≤ 2.0 Å). In the majority of all cases, TrixX does not place all fragments of a compound correctly. This can be explained by the fact that in most complexes one or more substituents of the compound lie in unspecific areas of the binding site where they possess a general steric fit but do not build strong pairwise interactions. Table 3 also shows, that in 120 cases, TrixX places at least one fragment of a compound successfully. In the remaining cases, the triangle matching algorithm of TrixX fails to reproduce the correct binding mode or the specific fragment lacks the necessary number of at least three fragment interaction centers.

Overall, these results prove the general ability of TrixX to produce valid fragment placements and at the same time show the intrinsic limitation of the triangle matching algorithm.

Protein targets for virtual screening

We provided a detailed analysis of the characteristics of the binding site descriptor in our previous paper on FlexX-Scan [15]. We could show that the new binding site descriptor represents the physicochemical and spatial constraints of a binding site by a set of only 90 SIACs per site on average. The four targets considered in this validation study were taken from the PDB structures listed in Table 4. In all cases, default protonation rules were applied. Protonation of histidine residues and the torsion angle at hydroxyl groups (either 0 or 180 degrees) were adapted manually. In all cases, the active site was set to all atoms being in distance of up to 6.0–10.0 Å to any atom of the co-crystallized ligand. The four targets used for our screening experiments with TrixX feature 95 SIACs on average, two thirds of which are hydrophilic and one third hydrophobic. In addition, the target Angiotensin Converting Enzyme has a metal ion in its binding site, represented by two metal SIACs. Triplets of these SIACs build site interaction triangles. As Table 4 shows, the defined constraints for site triangles reduce the number of theoretically possible combinations of SIAC triplets to a much smaller number of site triangles that are actually retained for querying the compound database. On average, the four screening targets feature between 3,848 and 12,841 retained site triangles.

Ligand dataset for virtual screening

We composed a screening library of about 8,200 compounds for the validation experiments. The library contains about 7,900 drug-like compounds drawn from the World Drug Index (WDI) [29]. In addition, the library contains 342 known active compounds for four drug targets from public sources (Table 5). Figure 7 shows the molecular weight and calculated log *P* distribution [30] of the four sets of active compounds and the set of compounds from the WDI (referred to as random set). Only the *c*log*P* curve for 4dfr ligands separates clearly from the random set. As will be

Table 4 Number of different site interaction center (SIAC) types per target and number of possible and retained site interaction triangles per target

Target	Number of site interaction centers (SIACs)				Number of site interaction triangles	
	Donor	Acceptor	Metal	Hydrophobic	Possible	Retained
1dwd	16	57	0	33	197,160	12,633
4dfr	14	35	0	32	82,751	8,009
1jvp	22	76	0	32	344,372	12,841
1o86	7	23	2	13	12,958	3,848

Table 5 Four targets and active compound sets

Target			Active compounds	
PDB code	Source	Name	Number	Source
1dwd	[31]	Thrombin	144	[32]
4dfr	[33]	Dihydrofolate reductase (DHFR)	68	[34]
1jvp	[35]	Cyclin dependent kinase 2 (CDK2)	72	[9]
1o86	[36]	Angiotensin converting enzyme (ACE)	58	[9]

shown in the screening results, 4dfr ligands can in fact be easily detected, although the major reason for this is the common binding motif of the anchor region.

We used TrixX for preprocessing the compound library and setting up a relational database with the compound catalog. On average, a compound carries 16.2 CIACs and is decomposed into 3.2 fragments. About 85.8% of those fragments mutually overlap having on average 4.4 CIACs in common. The typical properties of a fragment are as follows: a fragment of average size has 28.2 hetero atoms and is represented by 9.0 FIACs. There are 1.8 outgoing bonds and 83.1 conformations per fragment. As expected, a number of fragments occur in multiple compounds: Each fragment on average occurs in 1.53 compounds. While 14,081 of in total 25,710 fragments occur in just one compound, there are 16 highly repetitive fragments with more than 50 occurrences in all compounds. Among the latter, there are for example phenyl rings with one or two oxygen substituents. An average number of 14 triplets densely cover the nine interaction centers of the typical fragment such that each FIAC is part of about 4.6 triplets. In contrast to 83 conformations per fragment, a typical triplet has only 43 different triangle geometries. This can be explained by the fact that a triplet covers only a fraction of the whole fragment. If the conformational changes occur in uncovered parts of the fragment, the triangle geometry will remain the same. In total, a fragment is described by about 600 triangle descriptors.

Selectivity of the triangle descriptor

All triangle geometries are indexed using a B*-tree index structure on the interaction center types, the triangle side lengths, and the Euler angles of the main interaction directions. The efficiency of the B*-tree index depends on a balanced distribution of the descriptor values. Therefore, we analyzed the selectivity of the descriptor for different query value ranges. For this purpose, we counted the number of triangle hits for different types of queries on the fragment triangle table.

The first type of queries defined constraints on the FIAC types and on the side lengths of the triangle. Note that the FIAC types of a triangle hit must exactly match the query constraints (e.g. two hydrogen donor groups and one hydrophobic group), while the side lengths of the triangle hit can deviate up to 1.0 Å from the query template triangle. For this query type, we found that queries for triangles with relatively short triangle sides return between 0.1% and 1.0% of all triangles in the catalog, whereas queries for triangles with longer triangle side lengths return only between 0.01% and 0.1% of all triangles.

The second type of query enhances the selectivity by specifying constraints for the interaction direction of hydrophilic FIACs. For the purpose of this analysis, we set the query intervals for the Euler angles of both hydrogen acceptor and hydrogen donor groups to the typical ranges of $90° \le \theta \le 190°$, $0° \le \Phi \le 360°$, and $90° \le \Psi \le 190°$. The queries including these additional constraints result in much lower hit rates between 0.001% and 0.01% of all triangles in the catalog. Unlike hydrophilic FIACs, hydrophobic FIACs do not have a directional preference. Therefore, queries including hydrophobic FIAC types have a lower selectivity than queries on triangles with exclusively hydrophilic FIACs.

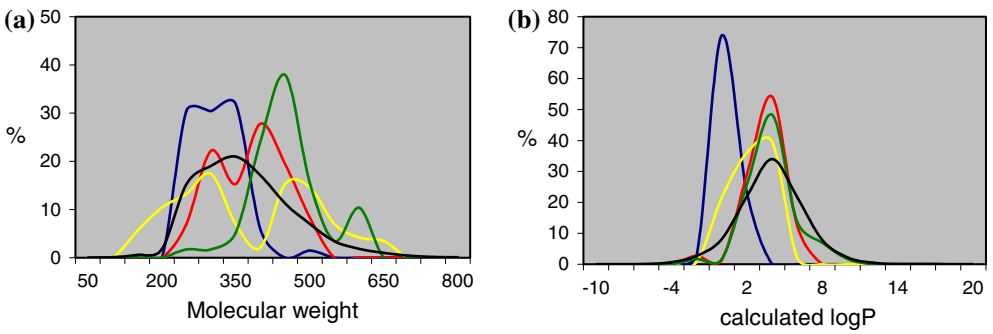

Fig. 7 Molecular weight (**a**) and calculated log *P* distribution (**b**) of the four active sets (1dwd: yellow, 4dfr: blue, 1jvp: red, 1o86: green), versus the random set (black)

Enrichment performance of TrixX

Figure 8 shows the enrichment performance of TrixX compared to standard FlexX and FlexX-Scan for the four targets.

As expected, the enrichment curves of TrixX deviate stronger from the enrichment curves of standard FlexX than the enrichment behavior of FlexX-Scan from FlexX.

At 10% of the top ranked compounds, between 50% (51) and 99% (100) of all active compounds can be found with FlexX (with FlexX-Scan). FlexX performs slightly better with on average 77% of all active compounds among the top 10% of all ranked compounds (FlexX-Scan: 74%).

TrixX shows an overall similar enrichment performance with placing between 31% and 100% of all active compounds among the top 10% of all ranked compounds. On overage, TrixX places 65% of all active compounds among the top 10% of all ranked compounds (enrichment factor of 6.5). TrixX performs particularly well compared to FlexX and FlexX-Scan in case of Dihydrofolate Reductase (4dfr). Here, TrixX successfully identifies all active compounds among the top 1% of all ranked compounds as opposed to 10% with FlexX. A negative case example for the performance of TrixX is the Angiotensin Converting Enzyme (1o86) where FlexX and FlexX-Scan achieve enrichment factors of 7.1 and 5.9, respectively, and TrixX has an enrichment factor of 3.1.

We found that essentially four types of compound properties account for the observed score and rank differences between FlexX-Scan and TrixX:

- Number of fragments per compound
- Number of hydrophobic CIACs per compound
- Number of rotatable bonds within the compound
- Number of compound hetero atoms

We used these dimensions for disaggregating the compound library into classes of compounds with common properties along one dimension. For instance, 92% of all compounds have between one and six fragments. For the compounds of one class, e.g. all compounds having only one fragment, we calculated the average rank difference of a compound if screened with TrixX or with FlexX-Scan. As Fig. 9 shows, compounds of that class have an average rank advantage in TrixX over FlexX-Scan of about 259.

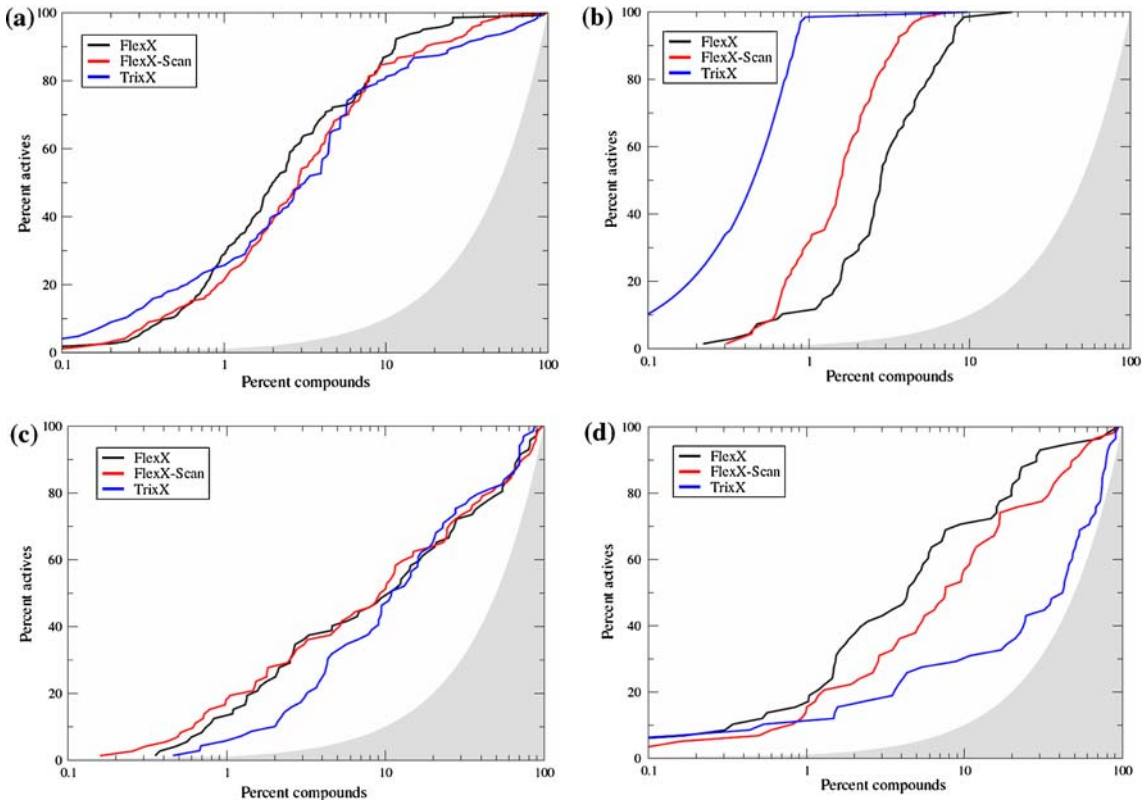

Fig. 8 Enrichment behavior of FlexX, FlexX-Scan, and TrixX for targets with PDB codes (**a**) 1dwd, (**b**) 4dfr, (**c**) 1jvp, and (**d**) 1o86 (*x*-axis in logarithmic scale). The area shaded in light gray indicates the enrichment that would be obtained by a random ranking of all compounds in the screening library

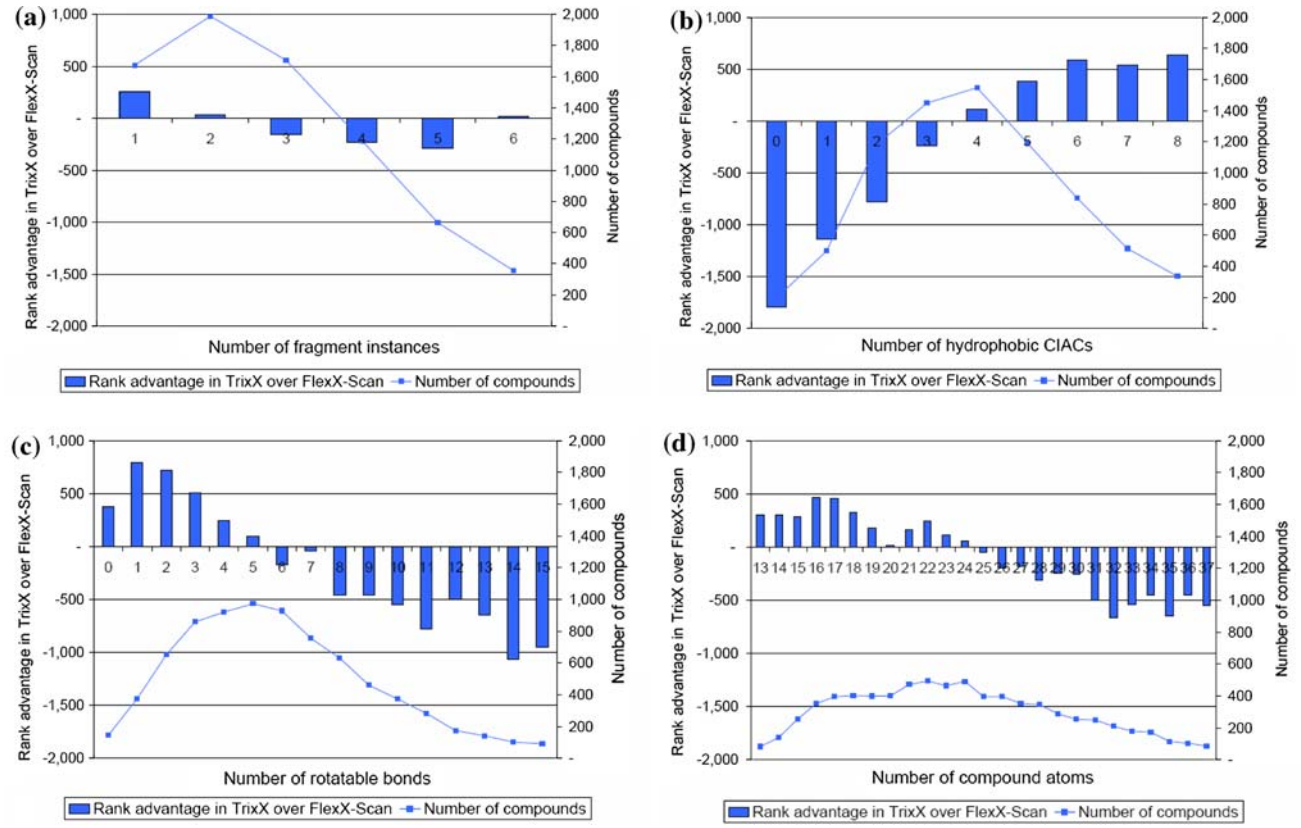

Fig. 9 Rank advantages of compounds in TrixX over FlexX-Scan. Rank differences between TrixX and FlexX-Scan are shown depending on (**a**) the number of fragment instances, (**b**) the number of hydrophobic CIACs, (**c**) the number of rotatable bonds, and (**d**) the number of hetero atoms of a compound

In other words, TrixX ranks compounds of that class 259 ranks better (closer to rank 1) than FlexX-Scan would do. As a consequence, those compounds are more likely to be retained as hits with TrixX than with FlexX-Scan. The results shown in the above figure show that TrixX has a relative tendency towards placing small (few fragments, few hetero atoms), rigid (few rotatable bonds) and rather hydrophobic (many hydrophobic CIACs) compounds on the top ranks of a screening experiment.

Depending on the properties of the target, this can be an advantage or a disadvantage for the purpose of high enrichment factors. Table 6 below shows the average properties of the active compounds of the four screening targets.

These results are in line with the observed biases of the four active sets with respect to the calculated log *P* and the molecular weight of the active compounds: The 1dwd and 4dfr ligands have lower calculated log *P* values than the random set (see Fig. 7a; in Table 6 expressed by low number of hydrophobic CIACs) and thus incur higher enrichment factors with TrixX due to the relative tendency of TrixX to rank hydrophobic compounds lower than hydrophilic compounds. Likewise, the 1o86 ligands have a higher molecular weight than the random set (see Fig. 7b; in Table 6 expressed

Table 6 Key properties of known actives used in virtual screening experiments

Target	Molecular properties of known actives							
	Number of fragments		Number of hydrophobic CIACs		Number of rotatable bonds		Number of hetero atoms	
1dwd	2.9		3.4	–	6.5		25	
4dfr	2.1		3.6	–	4.9	+	20	
1jvp	3.3		4.3		5.0	+	25	
1o86	4.1	–	4.6		10.5	–	30	–

A trailing plus or minus indicates that a property implies a relative rank advantage or disadvantage, respectively, for TrixX over FlexX-Scan of 200 ranks or more (double plus or minus: 400 ranks or more)

by higher number of hetero atoms) and thus incur lower enrichment factors with TrixX due to the relative tendency of TrixX to rank large compounds with many fragments lower than small compounds consisting of only 1 or few fragments.

Database size

We used a different compound library for investigating the growth behavior of the compound database. This collection has been kindly provided by BioSolveIT GmbH and contains 1.3 million compounds from publicly available vendor catalogs (including Aldrich Rare Chemicals and Chemstar). We used TrixX for preprocessing this compound library and creating a relational database with the compound catalog. When populating the database, we monitored the number of compounds, the total number and the number of non-redundant fragments, the number of triangles and the actual file size of the relational database (see Fig. 10a).

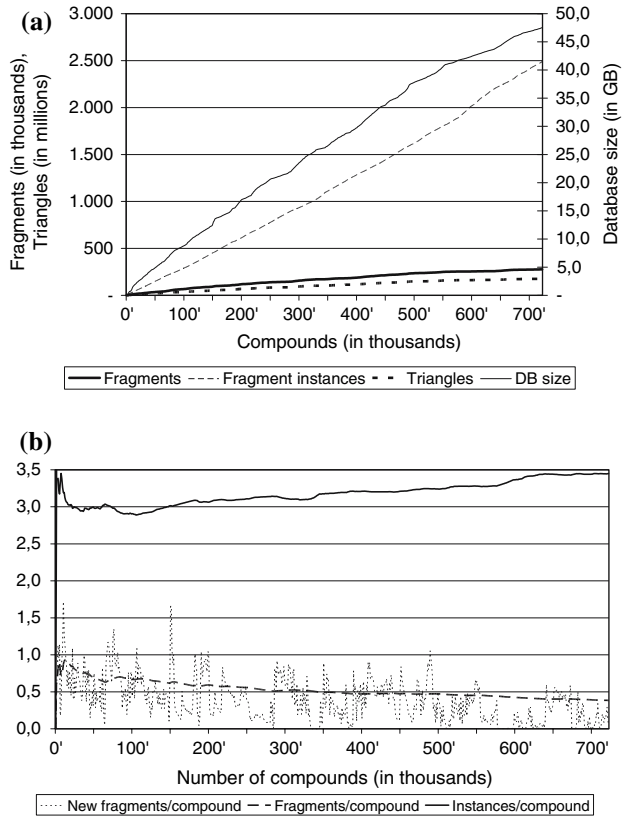

Fig. 10 (**a**) Growth of database size and number of triangles, unique fragments, and total fragment instances for increasing numbers of registered compounds. (**b**) Average number of fragments (chemically different molecular structures) and fragment instances (all substructures including multiple occurrences of a fragment per compound) per compound; average number of additional unique fragments identified among the last 1,000 registered compounds

We stopped the preprocessing after cataloging 700,000 compounds, yielding a total database size of 275,000 fragments (excl. redundant fragment instances) and 200 million triangles. The more compounds have been registered, the less novel fragments are detected among additional compounds (see Fig. 7b). For instance, after registering 700,000 compounds, each additional compound results in a mere 0.38 fragments that are not yet in the catalog.

Space and runtime requirements

The runtime and space analyses have been performed with a 64-bit version of TrixX on a SUN Fire server with four CPUs and 32 GB main memory. While using the full memory of the machine, TrixX runs sequentially on a single CPU. A varying number of compounds from the large collection with 1.3 million compounds have been registered to five compound databases. In the smallest database, every 200th compound of the large collection has been registered. For the other databases, every 100th, 50th, 25th and tenth compound was used, resulting in five databases, each with an approximately doubling number of compounds. Here, the average runtime per compound will be used for a comparison of TrixX with the runtime requirements of the sequential docking protocols FlexX and FlexX-Scan.

Figure 11 shows the development of the average runtime per compound for the four targets with increasing sizes of the screening library.

The runtime of TrixX ranges between 1.2 s and 0.2 s per compound. The analysis reveals two factors that influence the average runtime. First, the runtime is strongly influenced by the complexity of the target. In fact, the number of queried site triangles correlates

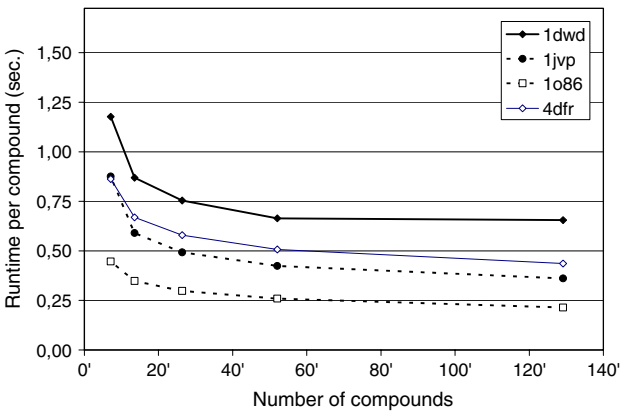

Fig. 11 Average runtime per compound of TrixX in seconds for an increasing number of compounds in the screening experiments

Table 7 Runtime per compound of FlexX, FlexX-Scan (mode 2: fast complex build-up), and TrixX (for largest screening experiment with 130,000 compounds)

Tool	Runtime per compound (s)					Average speed-up factor over FlexX
	1dwd	1jvp	1o86	4dfr	Average	
FlexX	37.30	31.00	10.30	22.20	25.20	N/A
FlexX-Scan	9.60	10.50	5.00	6.90	8.00	3.2
TrixX	0.66	0.36	0.21	0.44	0.42	60.4

The runtimes for FlexX and FlexX-Scan have been achieved on a Dual Xeon 2.4 GHz processor Linux workstation

with the observed average runtime by a factor of 0.934 (average correlation coefficient for all five database sizes). Second, the number of compounds in the screening library drives the decrease of the average runtime per compound. When doubling the size of the screening library, the average runtime per compound decreases by approximately 9–27%. One can assume a correlation with the logarithm of the number of compounds. Indeed, the average runtime per compound is negatively correlated with the logarithm of the number of compounds (correlation coefficient: –0.933). In other words, the runtime per compound in the presented experiments approximately decreases linearly with the logarithm of the number of screened compounds.[1]

Finally, we compared the average runtime of TrixX with the average runtime of FlexX-Scan and FlexX measured in a previous publication [15]. In the latter, we assessed the average runtime for FlexX and FlexX-Scan in screening experiments on four targets (1dwd, 1jvp, 1o86, 4dfr) with a 34,000-compound-library. As FlexX and FlexX-Scan apply a sequential docking strategy, larger screening libraries will not affect the average runtime per compound such that we can use these runtime values for the performance comparison. For TrixX, we measured the average runtime per compound for the same targets with the largest library size of about 130,000 compounds. We used a 64-bit version of TrixX on a SUN Fire server with four CPUs (running on a single CPU) and 32 GB of main memory for these experiments. This hardware environment shows similar performance like that the hardware used for our FlexX and FlexX-Scan performance tests.[2] Table 7 provides a summary of the observed runtimes:

Overall, for a 130,000-compound-library, TrixX achieves a speed-up factor of 60 compared to FlexX, which places TrixX among the fastest docking codes currently available.

In terms of space requirements, TrixX requires between 6 GB and 13 GB of main memory in screening experiments with 130,000 compounds. We observed a linear increase in space requirements with respect to the number of compounds in the library.

The largest database created included 700,000 compounds and resulted in a total database size of 46 GB on hard disk. This large size mainly results from the comprehensive information stored in the triangle descriptors of all fragment conformers (60 bytes per triangle descriptor). The resulting disk space requirements restrict the use of TrixX to modern high-capacity storage technology. The application of data compression techniques and advancements in storage technology will further alleviate this constraint.

Discussion

Our screening experiments with the TrixX prototype show that the novel docking paradigm of TrixX yields on average similar albeit differently biased enrichment performances for the four targets with respect to FlexX. Two characteristic features of TrixX are of particular interest for this observation: the fragment-based docking approach and the simplified compound representation.

First, the fragment-based approach fosters the enrichment of screening results with rather small compounds. TrixX decomposes large and flexible compounds into fragments. Based on an analysis of the binding site, TrixX searches for fragments that match the physicochemical and geometric constraints of different parts of the site best. The results show that usually not all fragments of a compound can be placed successfully. Since TrixX does not have an incremental build-up phase, the score has to be estimated on the placed part only which might result in large individual scoring errors. So far, this results in a good enrichment

[1] This runtime behavior is valid within the tested library sizes. Libraries with more than 130,000 compounds demand more main memory and may require even more selective molecule descriptors. Further, experiments need to prove or disprove whether the observed runtime behavior can be further extrapolated.

[2] Hardware environment A: 32-bit version of TrixX on a 2.4 GHz Dual Xeon workstation with 4 GB of main memory. Hardware environment B: 64-bit version of TrixX on a SUN Fire server using a one of four CPUs and 32 GB of main memory: We observed similar average runtimes of TrixX on both hardware environments.

behavior for test cases in which the actives are relatively small or have a highly characteristic fragment in most actives. Test cases with large active compounds, however, have a high chance of scoring errors due to unplaced fragments.

The second observation deals with the effects of the simplified compound representation in TrixX. The compound descriptor of TrixX reduces the atom-bond representation of compounds to a graph of functional groups (FIACs or CIACs), standardizes the interaction geometries of compounds functional groups to symmetric cones (i.e. no spherical rectangles), and uses a discrete binning scheme for describing the distances between functional groups. At the same time, TrixX applies more restrictive constraints on these simplified compound features than FlexX-Scan and FlexX, e.g. smaller opening angles for interaction geometries and small distance tolerances for superposing pairs of functional receptor and ligand groups. As a result, TrixX is less successful in finding the optimal pose of an active compound in re-docking experiments than FlexX or FlexX-Scan.[3] The overall similar performance of TrixX and FlexX/FlexX-Scan in enrichment experiments suggests that those restrictive constraints of TrixX affect active and inactive compounds in the same way. This conclusion is particularly important since it questions the role of re-docking experiments as a stand-alone performance indicator for virtual screening tools. Good or poor re-docking performance does not necessarily imply high or low enrichment factors in screening experiments because re-docking experiments can only evaluate the handling of active compounds and do not take into account the consequences for inactive compounds.

The runtime analyses underline the main strength of the novel docking paradigm introduced with the TrixX prototype: TrixX is about 60 times faster than FlexX. The speed advantage of TrixX over standard FlexX decomposes into a factor of about 3 related to the advances in receptor representation with FlexX-Scan and into a factor of about 20 related to the new docking paradigm of TrixX. The compact binding site representation of FlexX-Scan based on hot spots is a mandatory prerequisite for a manageable number of site triangles in TrixX. Still, the resulting speed advantage is also amenable to sequential docking strategies like FlexX-Scan and should therefore not be assigned to the novel docking paradigm of TrixX. The paradigm-related sources of speed advantage of TrixX are the non-redundant compound catalog, the use of relational database technology with efficient indices and the target-driven fragment matching approach.

[3] The gap of re-docking performance between TrixX and FlexX/FlexX-Scan (number of correct predictions) is rather large for very accurate solutions (≤1.0 Å) and tends to close for less accurate solutions (≤2.5 Å).

Conclusions

With TrixX, we developed a prototype for a novel virtual screening paradigm. In contrast to classical sequential screening, TrixX is based on a geometric index allowing direct access to matching compounds. Our initial validation experiments underline that TrixX is able to process large virtual screening libraries substantially faster than traditional sequential docking tools like FlexX while yielding similar enrichment rates. TrixX offers effective approaches to speeding up virtual screening beyond better hardware performance and massive parallelization. The runtime performance of TrixX will be of particular value in research areas like chemical biology where very large compound libraries need to be screened against multiple proteins and receptors. Furthermore, if protein flexibility and selectivity issues are addressed with multiple target structures, TrixX database-driven approach is advantageous. While our TrixX prototype proves the general feasibility of the novel virtual screening paradigm, some limitations will have to be addressed in future research for the practical applicability of TrixX.

The main application of an approach like TrixX is its usage as an efficient pre-filter to more accurate molecular docking tools in large virtual screening experiments. Yet, our initial experiments revealed an overly good enrichment of small and hydrophobic compounds compared to FlexX. TrixX is likely to discard large hydrophilic compounds that would be ranked as high affinity compounds by FlexX. For the practical use of TrixX as a pre-filter to FlexX, future research will have to focus on mitigating this relative bias of TrixX towards small, hydrophobic compounds. Potential levers are a different parameterization of the scoring scheme and a revision of the fragment placing and linking approach. An alternative to the latter approach might be a rough multi-conformer-sampling of each entire compound. After placing a conformer initially into the binding site, ''tweaking'' techniques could be used for optimizing conformation and pose of the placed conformer within the binding site.

A second challenge of TrixX with high practical relevance is parallelizing the software for large compute clusters. Most prominent molecular docking tools support parallel environments with a hundred or more nodes, which outweighs the computing time advantage

of TrixX. Deploying TrixX in a parallel environment requires splitting the compound catalog on multiple, distributed database instances on the one hand and in a parallel execution of the docking engine on several nodes on the other. The discrete features of the TrixX triangle descriptor (interaction center types, side length bins) are an effective mean for splitting up the compound catalog in multiple, non-overlapping parts. Likewise, different, non-overlapping regions of the binding site could be processed on different nodes, each node searching for matches to the site triangles within its binding site region.

In summary, TrixX can be seen as a successful proof of concept for database-driven structure-based virtual screening. Further research is necessary to address important issues like an improved enrichment behavior and parallelization. Nevertheless, TrixX opens a promising route to a new class of efficient tools for structure-based drug design.

Acknowledgment The authors thank BioSolveIT GmbH (St. Augustin, Germany) and AstraZeneca (Mölndal, Sweden) for funding our work. We are grateful for constructive discussions on the molecule descriptor and on method validation with colleagues from AstraZeneca and BioSolveIT, especially Jens Sadowski and Christian Lemmen.

References

1. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nat Rev Drug Discov 3(11):935
2. Bajorath J (2002) Nat Rev Drug Discov 1(11):882
3. Wang R, Lu Y, Wang S (2003) J Med Chem 46(12):2287
4. Wang R, Lu Y, Fang X, Wang S (2004) J Chem Inf Comput Sci 44(6):2114
5. Stockwell BR (2004) Nature 432(7019):846
6. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) J Mol Biol 161(2):269
7. Hindle SA, Rarey M, Buning C, Lengauer T (2002) J Comput Aided Mol Des 16(2):129
8. Good AC, Krystek SR, Mason JS (2000) Drug Discov Today 5(12):S61
9. Claussen H, Gastreich M, Apelt V, Greene J, Hindle SA, Lemmen C (2004) Curr Drug Discov Technol 1:49
10. Zuccotto F (2003) J Chem Inf Comput Sci 43(5):1542
11. Renner S, Schneider G (2004) J Med Chem 47(19):4653
12. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW (2000) J Comput Aided Mol Des 14:731
13. Kraemer A, Horn HW, Rice JE (2003) J Comput Aided Mol Des 17:13
14. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF (1999) J Med Chem 42:3251
15. Schellhammer I, Rarey M (2004) Proteins 57(3):504
16. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) J Med Chem 47(7):1739
17. Floriano WB, Vaidehi N, Zamanakos G, Goddard III WA (2004) J Med Chem 47(1):56
18. Rarey M, Lengauer T (2000) Persp Drug Discov Des 20:63
19. Sun Y, Ewing TJ, Skillman AG, Kuntz ID (1998) J Comput Aided Mol Des 12(6):597
20. Lorber DM, Shoichet BK (1998) Protein Sci 7:938
21. Su AI, Lorber DM, Weston GS, Baase WA, Matthews BW, Shoichet BK (2001) Proteins 42(2):279
22. Joseph-McCarthy D, Thomas IV BE, Belmarsh M, Moustakas D, Alvarez JC (2003) Proteins 51:172
23. Schnecke V, Swanson CA, Getzoff ED, Tainer JA, Kuhn LA (1998) Proteins 33(1):74
24. Schnecke V, Kuhn LA (2000) Persp Drug Discov Des 20:171
25. Weininger D, Weininger A, Weininger JL (1989) J Chem Inf Comput Sci 29:97
26. Weininger D (1988) J Chem Inf Comput Sci 28(1):31
27. Bayer R, McCreight E (1972) Acta Informatica 1(3):173
28. Kramer B, Rarey M, Lengauer T (1999) Proteins 37:228
29. Stahl M, Rarey M (2001) J Med Chem 44:1035
30. Wildman SA, Crippen GM (1999) J Chem Inf Comput Sci 39:868
31. Banner DW, Hadvary P (1991) J Biol Chem 266(30):20085
32. Böhm HJ Thrombin-Inhibitors, collected experimental data, personal communication
33. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J (1982) J Biol Chem 257(22):13650
34. Selassie CD, Fang ZX, Li RL, Hansch C, Debnath G, Klein TE, Langridge R, Kaufman BT (1989) J Med Chem 32(8):1895
35. Furet P, Meyer T, Strauss A, Raccuglia S, Rondeau JM (2002) Bioorg Med Chem Lett 12(2):221
36. Natesh R, Schwager SL, Sturrock ED, Acharya KR (2003) Nature 421(6922):551