

## Index-driven structure-based virtual screening

Jochen Schlosser<sup>1</sup>

Matthias Rarey<sup>1</sup>

<sup>1</sup> Center for Bioinformatics (ZBH), Bundesstr. 43, 20146 Hamburg, Germany

The standard approach to structure based high-throughput virtual screening nowadays is a sequential procedure. Each molecule of a given library is individually docked into the target protein in order to produce a ranked hit list. With the TrixX approach, we introduce a new paradigm avoiding the iterative process of virtual screening. The non-sequential character of our workflow allows a substantial speedup while yielding comparable re-docking results and enrichment rates.

In order to avoid the iterative processing of molecules, the TrixX method is based on a novel descriptor capable to cover pharmacophoric as well as shape information. Applying standard database technology, TrixX is able to retrieve active compounds in sets of up to several ten thousand ligands. Here, we introduce the next generation of our index-driven virtual screening technology named TrixX BMI with multiple new developments. We replaced the placing and linking procedure of small molecular fragments by rigid body docking of pre-processed conformational ensembles of small molecules or molecular fragments with up to ten rotatable bonds. Furthermore we extended the descriptor significantly by introducing an 80 dimensional steric bulk vector in addition to interaction types, directions and triangle side lengths (Fig. 1). We kept the promising idea of splitting virtual screening into disjoint phases. In the *Data Pre-Processing* phase, descriptors are computed based on conformational ensembles and stored in a database. This is a one-time effort. In the *Virtual High-Throughput Screening* phase, a given protein active site is used to generate complementary descriptors as query templates which are used to identify potential hit candidates within the database. Query matches are then translated into initial fragment placements which are then extended, optimized and scored. Because of the enormous amount of descriptors and their high-dimensional content there is the need for an efficient and overhead-free decision support system. This functionality is realized using compressed bitmap indices supplied by Fastbit.

Re-docking experiments on 115 protein-ligand complexes show that TrixX BMI correctly predicts the pose of the bioactive conformation within a RMSD of less than 2.5 Å of the co-crystallized ligand in 100 cases, thus achieving typical values for current docking tools (Table 1) and improving the runtime by about one order of magnitude. In addition to that several enrichment experiments demonstrate that TrixX BMI is competitive to current methods. TrixX BMI is especially suited for structure-based virtual screening under pharmacophoric constraints (Table 2). To show the systems scalability, a large test set consisting of 1.2 million random lead-like compounds, distributed over a 94-node computing cluster, is used. Four different targets (CDK2, DHFR, ER(agonists), ER(antagonists)) from the DUD, together with pharmacophores from the literature are used as benchmark set. TrixX BMI is able to finish the VHTS runs on all four targets in less than 20 minutes, whereas the average time is below 12 minutes with comparable enrichment rates (see Table 2). Due to its speed, index-based docking opens a new route for modelling protein flexibility in structure-based virtual screening.

1. Schellhammer I, Rarey M. TrixX. Structure-Based Molecule Indexing for Large-Scale Virtual Screening in Sublinear Time. *J. Comp. Aided Mol. Design*, **2007**, 1573-4951
2. Wu K, Otoo E, Shoshani A. An Efficient Compression Scheme for Bitmap Indices. *ACM Transactions on Database Systems*, **2006**, 31, 1-38.
3. Huang, Shoichet, Irwin. Benchmarking Sets for Molecular Docking. *J. Med. Chem.*, **2006**, 49(23), 6789 - 6801.

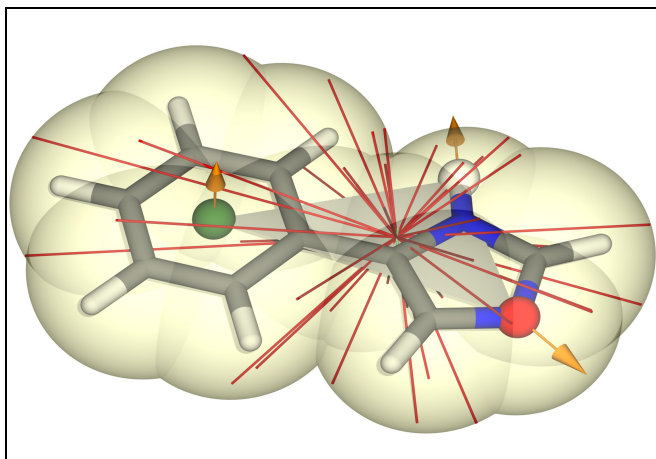


Figure 1: Example of a ligand descriptor

Table 1: Root-mean-square-deviation histogram of the re-docking experiments and the average runtime using all targets against a lead-like dataset of 12600 compounds.

RMSD [ $\text{\AA}$ ] $\leq$	1.0	1.5	2.0	2.5	Avg. runtime
FlexX 2.2.0	76	88	94	96	9.55 [sec/lig]
TriX BMI	48	77	92	100	0.29 [sec/lig]

Table 2: Enrichment factors at 2% [with | without] pharmacophore constraints. In addition [TriX BMI | FlexX] runtimes using pharmacophore constraints on the target specific dataset from the DUD and on a random lead-like dataset of 12600 compounds.

Target	$E_{[2\%]}$ TriX BMI	$E_{[2\%]}$ FlexX 2.2.0	$E_{[2\%]}$ DOCK 3.5.54 (estimated)	Runtime [DUD]	Runtime [rand.]
ER(agonists)	39.39   7.57	21.21   15.15	n.a.   8	0.05   6.18	0.04   4.65
ER(antagonists)	20.51   2.56	33.33   17.94	n.a.   12	0.13   21.72	0.04   6.41
DHFR	44.48   40.13	50.16   44.81	n.a.   35	0.03   14.39	0.01   2.66
CDK2	28.13   26.56	15.62   21.87	n.a.   15	0.04   5.52	0.01   4.07