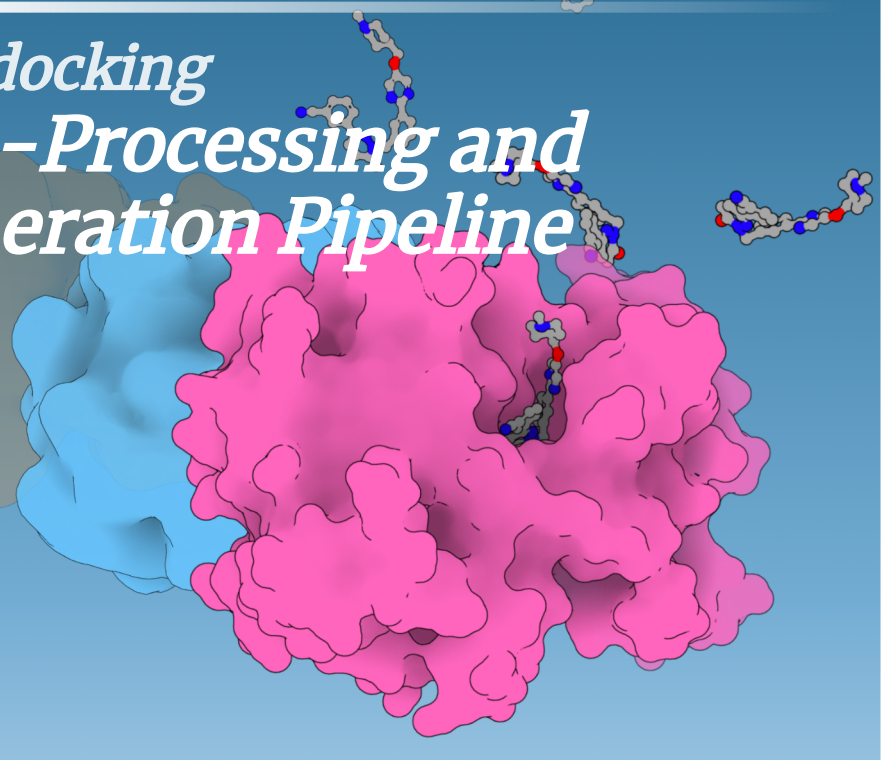
EVARISTE TECHNOLOGIES

Evariste Technologies aims to increase productivity in drug discovery using automated molecular design, integrating bespoke modelling solutions into its proprietary software platform to perform molecular evolution. Using multiobjective optimization scoring functions, Evariste is able to arrive at candidate molecules more quickly than the traditional design cycle.

Improving protein-ligand docking with an Automated Pre-Processing and Constraint Generation Pipeline

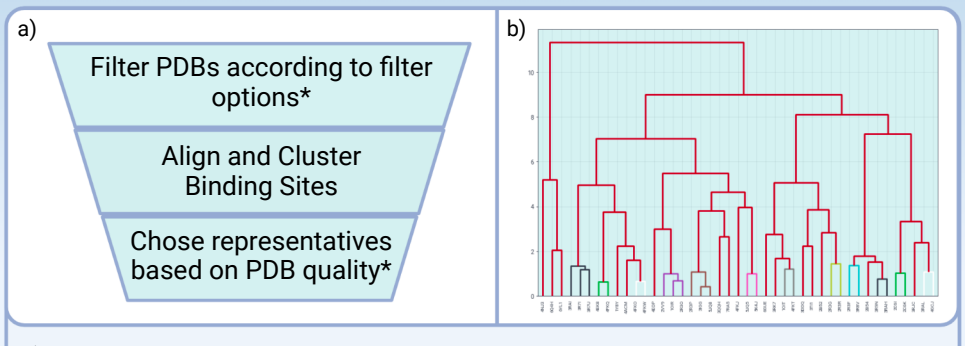


Abstract

We have developed an automated preprocessing pipeline that attempts to optimise the efficiency, reliability, and effectiveness of our virtual screens.

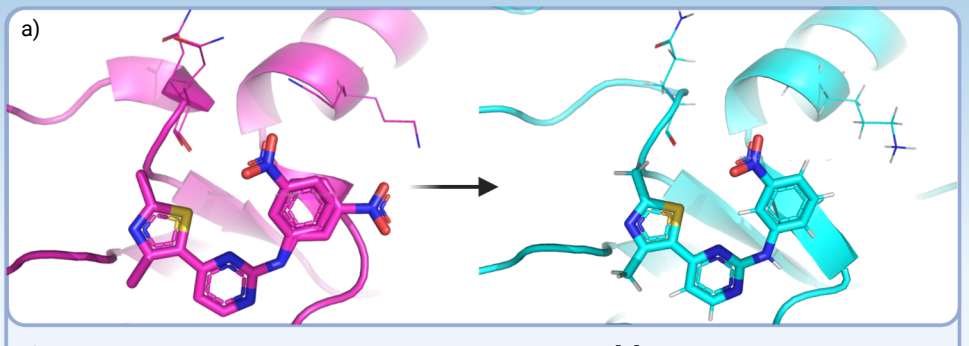
1) The modular pipeline is able to specifically search and filter the protein data bank and carry out structure-based analysis to collect high quality, diverse representatives of the target under investigation.
2) A processing module utilises various tools to protonate, repair and re-annotate the protein files which are then provided to a constraint generator.
3) Key protein-ligand interactions and water molecules extracted through this generator are added to a large parameter set that contains various other iterable features, such as choice of PDB and scoring functions.
4) Hyper-optimisation over this parameter space, with evaluation through various enrichment metrics, allows selection of an optimised set of constraints.
5) Combining our synthon-based docking algorithms with these optimised knowledge-based constraints we are able to efficiently search through ultra-large chemical libraries, implicitly scoring billions of compounds and rapidly enriching high-scoring molecules.
6) Post-processing provides a final refinement stage in which only the most promising of compounds rise to the top.

1) Filter and Cluster PDBs



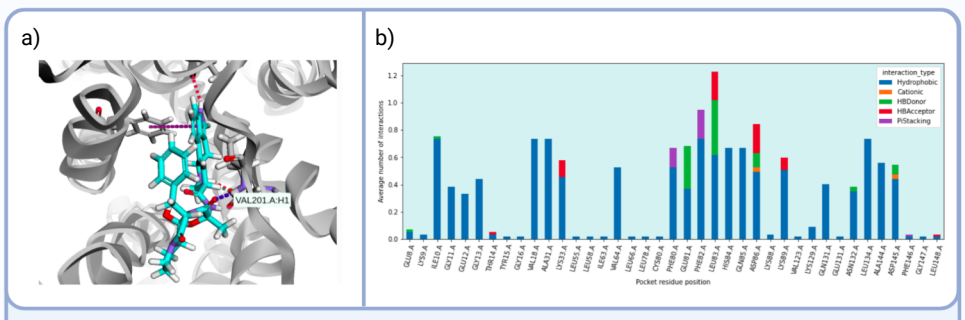
a) PDB selection funnel. *Filter options include: Resolution, orthosteric inhibitor bound, wildtype structure, binding site residues resolved. *Representatives are chosen based on ranking by PDB quality metrics, including Rfree, clashscore, Ramachandran outliers, Sidechain outliers. b) Dendrogram representing the agglomerative clustering of PDBs based on inter-protein binding site RMSD

2) Process PDBs



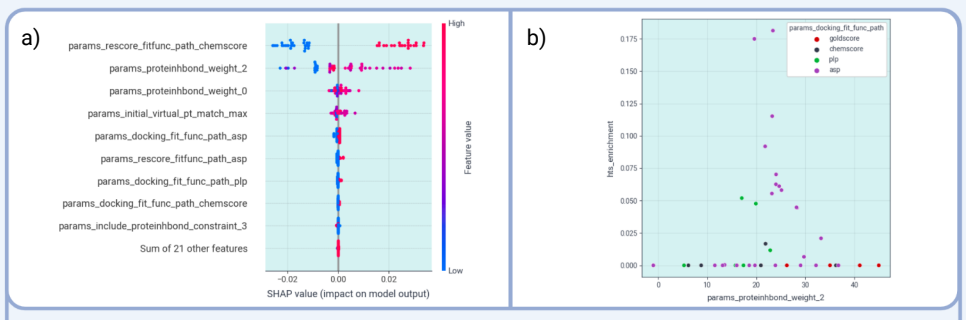
a) The PDB processing pipeline consists of a PDBFixer [1] that builds in missing atoms, residues and loops and selects single alternative locations (pictured). A protonation tool [2] protonates the protein and the ligand, as well as re-orientates flexible sidechains to maximise H-bond geometries. A renumbering tool [3] sets all PDBs to consensus residue numbering, based on SIFTS data.

3) Extract Constraints



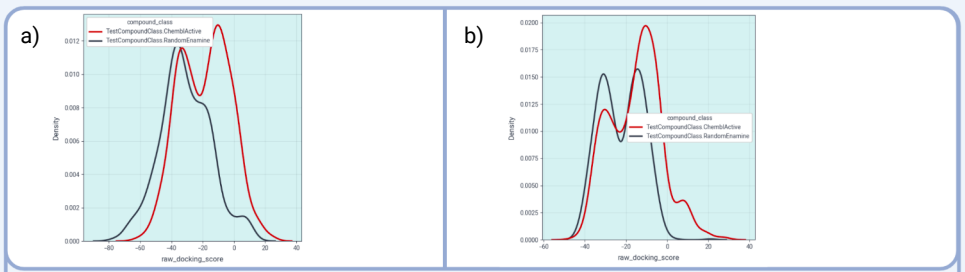
The python package ProLIF [4] is used to evaluate the protein-ligand interactions. These can be: a) visually inspected; b) The relative propensity of the interaction quantified across the set of processed PDBs. Evaluation of water molecules is also carried out at this stage. First, the reliability of water placement is confirmed by electron density analysis [5]. Water molecules are then identified as conserved by clustering by distance and picking a minimum cluster size. A third party tool [6] makes minor translational and rotational adjustments to optimise H-bonding geometries.

4) Optimise Constraints



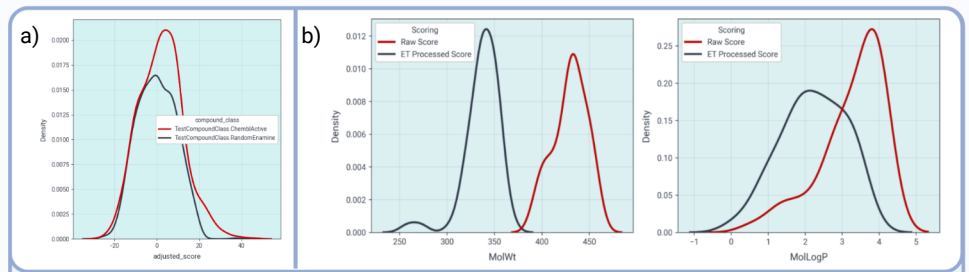
The constraints generated in step 3) are then added to a search space for an optimisation function. Other parameters in this search space include scoring function, re-scoring function, search efficiency. An evaluation metric estimates the hit-rate at points across the search space. a) The contribution of individual parameters to the overall docking performance is assessed. b) The variation in estimated hit-rate can be assessed across all possible values of a search space parameter to find the optimal.

5) Dock



The docking score distribution of active compounds in Chembl compared with random compounds selected from Enamine. a) The score distribution from a low-scoring parameter set. b) The score distribution from a high-scoring, and thus more highly-optimised parameter set. Note the greater enrichment of active compounds at the high scoring tail of the distribution.

6) Post-process



Post processing of docking scores rewards molecules with hit-like properties. This leaves scope for more development. a) The post-processed docking score ('adjusted_score') of active and random Enamine compounds. b) The drug-like properties of top scoring compounds after processing docking scores

References:
[1] 10.1371/journal.pcbi.1005659, [2] 10.1186/1758-2946-6-12, [3] github.com/Faesov/PDBRenum, [4] 10.1186/s13321-021-00548-6, [5] 10.1186/s13321-021-00548-6, [6] 10.1021/acs.jcim.8b00271