Active learning to select the most suitable reactants for synthesizing target-specific ligands

Vladimir KOZYREV[1], François SINDT[1], Didier ROGNAN [1]

[1]*Laboratoire d'Innovation Thérapeutique (LIT), UMR7200 CNRS-Université de Strasbourg, Illkirch, France 67400*

Structure-based virtual screening is an important tool in early-stage drug discovery that scores the interactions between a target protein and candidate ligands. Ultra-large 'on-demand' combinatorial libraries [1] are revolutionizing virtual screening strategies aimed at identifying innovative hit compounds or guiding fast hit-to-lead optimization. Due to their size (several billion molecules), these libraries are encoded as fragment spaces defined by starting building blocks and organic chemistry yielding the fully enumerated compounds [2].

Spacedock [3] is a structure-based approach to ultra large chemical space screening in which commercial chemical reagents are first docked to the target of interest and then directly connected according to organic chemistry and topological rules to enumerate drug-like compounds under the three-dimensional constraints of the target.

In this study, we investigate how active learning can help prioritize chemical reagents for recombination by Spacedock in the target binding site. We utilize Morgan fingerprints of the reagents and employ various classification models (Logistic regression, Random Forest, Multilayer perceptron) to dynamically select, by active learning, a subset of compounds from the library for screening. We iteratively refine the model's predictive capabilities by actively querying the most informative samples. Our dataset comprises 33,726 different amines and 19,887 carboxylic acid reagents, resulting in a chemical space of 670,708,962 unique carboxamides. By using a multilayer perceptron model, we were able to retrieve 90.41% of the top 0.1% scoring compounds (scored by Tanimoto IFP similarity to the PDB reference ligand) after just exploring only 5% of the full chemical space (see Figure 1).

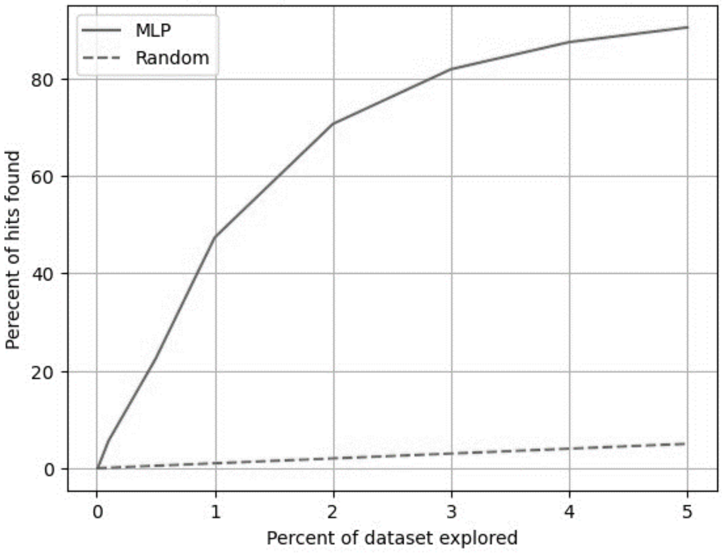

Figure 1. Percent of top 0.1% scoring compounds selected as a function of the percent of the chemical space explored for multilayer perceptron model.

Bibliography:

[1] HoffmanT; Gastreich M. Drug Discov. Today. (2019), 24(5), 1148–1156.
[2] Warr W; Nicklaus M; NicolaouC; Rarey M. J. Chem. Inf. Model. (2022), 62, 9, 2021–2034.
[3] Sindt F; Seyller A; Eguida M; Rognan D. ACS Cent. Sci. (2024), 10, 3, 615–627.