Generative ML Pipeline for Screening Process

Authors: Jakub Adamiak, Maciej Bielecki, Marcin Potkański, Paweł Nagórko, Piotr Trzaskowski
Supervisors: Joanna Sułkowska, PhD, DSc, Prof. Tit, Wanda Niemyska, PhD

The focus of this year's project was the development of a novel computational procedure for identifying potential inhibitors of Nsp13, the RNA helicase of SARS-CoV-2. Collaborating with teams from Sorbonne University and Milano University, our goal was to explore generative models as an alternative to traditional virtual screening methods.

Virtual screening (VS) is a common approach in drug discovery, involving the docking of ligands to a receptor structure. However, this process is computationally intensive. In this project, we explored the use of generative models (GM), particularly Pocket2Mol, an autoregressive generative graph neural network model. Despite the promise of GMs, they have limitations, such as generating unrealistic and unsynthesizable molecules. To address this, we proposed a hybrid approach, cross-referencing GM-generated molecules with established compound databases to validate and identify potential candidates.

We utilized Punicalagin (PUG), an Nsp13 inhibitor, as a control and reference for generated ligands. The project workflow began with determining the pocket coordinates within Nsp13's structure based on PUG's inhibitory properties.

We used Pocket2Mol, a pretrained model, to generate ligands targeting Nsp13's binding site, with a generation space of 25 Å.

Generated ligands were filtered by docking to Nsp13 using AutoDock Vina. Ligands with Protein-Ligand Affinity (PLA) and Locality of Binding (LoB) higher than PUG were selected.

The Tanimoto coefficient was used to assess similarity between generated molecules and compounds in Enamine's REAL Diversity Set (50M compounds). Multithreading was employed for efficient computation.

Analogous molecules with a high Tanimoto coefficient underwent binding affinity assessment using AutoDock Vina.

1592 potential ligands were generated, with 33 showing higher PLA and LoB than PUG. The top 10 ligands, ranked by PLA, were identified. Analog search in the database revealed challenges, with Tanimoto coefficients generally below 0.6, indicating dissimilarity. The analogs with Tanimoto coefficient greater than 0.6 had docking scores lower than PUG.

We conclude that the proposed approach is computationally efficient and scalable, running on standard hardware. Utilizing open-source software, including multithreaded and GPU-compatible tools, allows for quick results. The strategy of synthesizing analogs by ENAMINE addresses the cost and risk issues associated with generative models. Challenges include the size-dependent search space, where larger molecules have fewer analogs in the database.