Identifying inhibitors of SARS-CoV-2 Nsp-13 helicase using a proteochemometrics (PCM) approach

Anastasiia Avdonina, Julia Byrska, Jakub Guzek, Paulina Kucharewicz, Michalina Wysocka
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

The goal of our project was to identify potential inhibitors of the SARS-CoV-2 Nsp-13 helicase. Our approach was initially inspired by an article that discussed machine learning-based prediction of drug/component-target interactions [1]. This article introduced us to the proteochemometrics model (PCM), which we employed in our study. Our pipeline involved the innovative use of proteochemometrics, a computational technique that employs machine learning to forecast the interactions between proteins and chemical compounds.

To gather the necessary data for training our models, we compiled and analyzed two primary datasets. For the first one, we extracted binding assay data for helicases and their ligands from the ChEMBL database, focusing on helicases from viruses. After overcoming challenges with data quality and completeness, we refined our dataset to include 247 well-characterized pairs of proteins and ligands. For the second one, by using Maestro we performed molecular docking of various ligands to the Nsp-13 helicase. We selected the 7RDX structure of Nsp-13 for docking. Ligands were chosen from the ChEMBL database, and docking parameters were set for accurate simulations.

In addition to this, we utilized advanced computational techniques to analyze protein and ligand features. For proteins, we used the SeqVec Embedder to generate high-dimensional embeddings from helicase sequences. For ligands, we employed the RDKit library to obtain feature vectors based on SMILES notation, using Extended-Connectivity Fingerprints (ECFP4) for a multidimensional chemical structure representation.

We decided to implement both Random Forest Regressor and Support Vector Regressor models for predicting the interactions between proteins and ligands. The accuracy and reliability of these models were assessed using several metrics such as RMSE (Root Mean Square Error), Spearman's rank correlation, and MCC (Matthews Correlation Coefficient).

In our study, we successfully identified several potential ligands that showed promise in interacting with the SARS-CoV-2 Nsp-13 helicase. Part of our analysis involved examining the molecular weights and lipophilicity of ligands, as well as applying Lipinski's rule to evaluate their drug suitability. We present some of them in Fig.1.

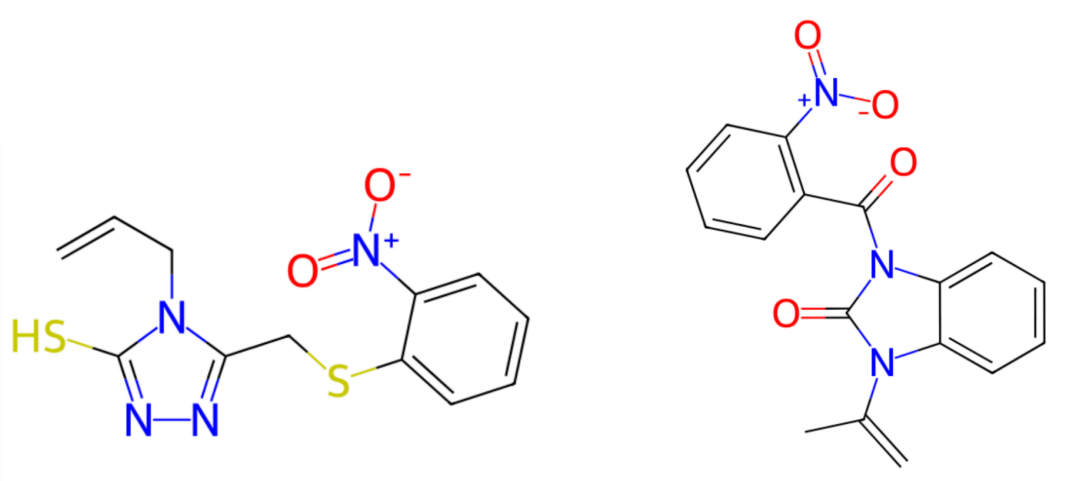

Fig. 1 Structures of the best-performing ligands: CHEMBL1595621 on the left and CHEMBL2228592 on the right [2]

References

[1] Atas Guvenilir, H., & Doğan, T. (2023). How to approach machine learning-based prediction of drug/compound–target interactions. Journal of Cheminformatics, 15, 16.

[2] ChEMBL Database. Version 27. 2023. Available from: https://www.ebi.ac.uk/chembl/