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

1. Introduction

The project's goal was to identify potential Sars-Cov-2 Nsp-13 helicase inhibitors using the chosen approach. At the beginning of the course, we received an article about approaching machine learning-based prediction of drug/component-target interactions [1]. The article presented a proteochemometrics model (PCM) which we decided to pick. PCM modeling is a computational technique that is trained on a dataset composed of a series of target and compound features to predict an output variable of interest [2].

There are three main reasons why we decided to pick the PCM approach. Firstly, many protein structures, especially in the context of emerging pathogens like SARS-CoV-2, still need to be determined. PCM allows for the modeling of protein-ligand interactions even in the absence of detailed 3D structures, using sequence-based information instead. Moreover, PCM is less computationally intensive than methods that require detailed 3D structures, such as molecular docking or molecular dynamics simulations. This is very advantageous when dealing with large datasets or when computational resources are limited. Lastly, PCM can handle a wide range of protein and ligand variations, making it a versatile tool for modeling diverse protein-ligand interactions. This flexibility is crucial in rapidly evolving situations like drug discovery for new viruses, where the target proteins and potential inhibitors can vary significantly.

In the context of proteochemometric modeling, there are two primary approaches: conventional and embedding. The conventional approach is largely based on models that are generated by applying defined rules and/or statistical calculations to sequences, taking into account various molecular properties. This method often involves the use of established algorithms and mathematical formulas to analyze and predict interactions between proteins and chemical compounds, based on the understanding of their molecular structures and properties. The other one, the embedding approach, involves the automatic feature extraction of input samples through the training and application of machine learning models. In this approach, machine learning algorithms, such as deep neural networks, are used to automatically learn and identify relevant features from the data [3]. We decided to use an embedding approach which is particularly useful in handling complex and high-dimensional data, typical in proteochemometrics, where the relationships between proteins, chemicals, and their interactions are intricate.

Our pipeline is presented in Fig. 1. We present more details in the Materials & Methods section.

At mid-course, after a face-to-face meeting, we were paired with the third and the sixth teams from Sorbonne University (S3 and S6).

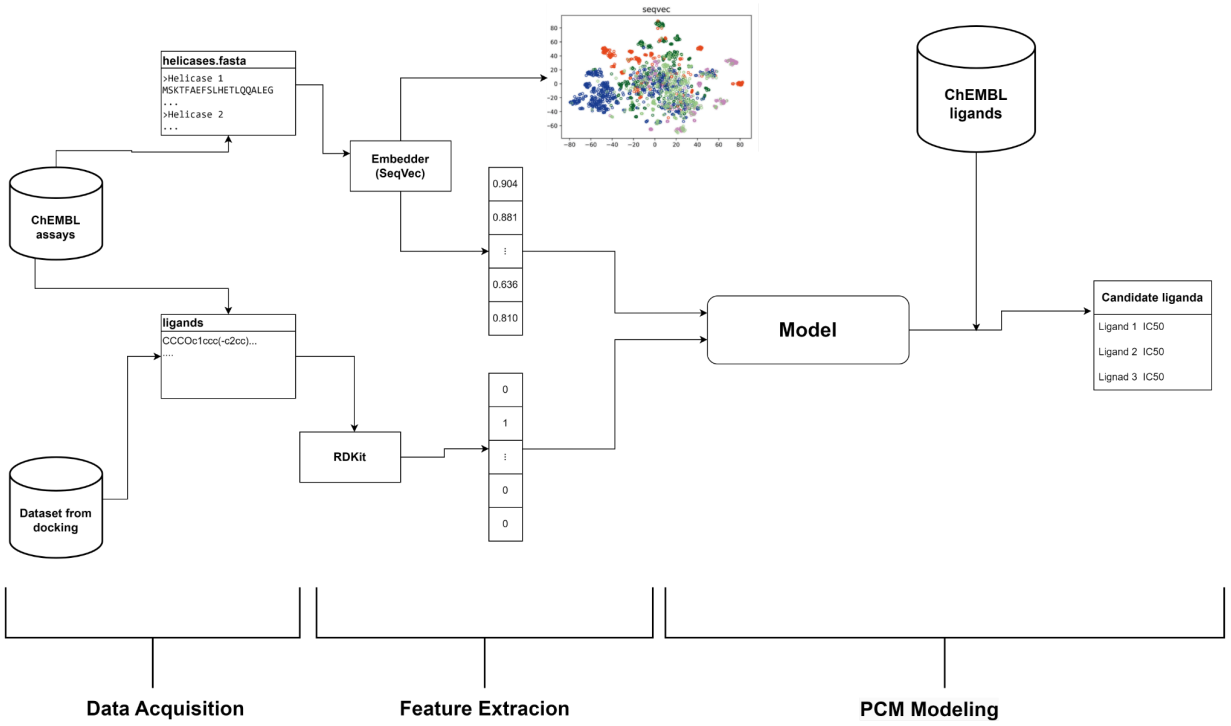

Fig. 1 Project pipeline

2. Materials & Methods

2.1 Data Acquisition

2.1.1 Dataset 1: Data from ChEMBL assays

We started by compiling datasets of helicases with associated ligands. This phase was crucial for us to understand the binding interactions between helicases and ligands. We intended to use binding assay data from the ChEMBL database, focusing on targets containing ligands and corresponding IC50 values. We selected all binding assays for the helicases from viruses, and we found around 150 assays. From them, we got 1038 pairs of ligands and proteins. We decided to use a filter to get the pairs where IC50 values and standard values (numerical outcome of a bioassay in ChEMBL) were specified. However, we encountered a hard challenge: many target proteins from our selected assays are labeled as UNCHECKED in the database and lack sequence data. To address this issue, we wanted to conduct an extensive literature review and database searches. Each binding assay is linked to a corresponding scientific paper, but this method is not only inefficient but also risks introducing inaccuracies into our data. We decided to fill only some data gaps to address the challenges we faced with the ChEMBL database, concentrating mostly on our main helicase. This approach allowed us to fill in missing data accurately, and we had 247 pairs of protein with ligands, ensuring that our dataset was comprehensive.

2.1.2 Dataset 2: Data from docking

At the same time, we went with a new strategy. In addition to conducting an extensive literature review and database searches for helicase sequences and ligand SMILES, we performed a method of docking multiple ligands to Nsp-13, a helicase. We implemented a docking strategy using Maestro, a sophisticated molecular modeling platform. Maestro is renowned for its robustness and accuracy in molecular docking. Before the docking process, we carefully selected the protein structure to which the ligands would be docked. This involved choosing an appropriate conformation of Nsp-13, ensuring it was in a state representative of its active form to maximize the relevance of our docking simulations. We decided to pick the 7RDX structure, the E chain, and the ADP-binding pocket.

Additionally, we used a previously prepared list of ligands for the docking experiments from the ChEMBL database. To realistically model helicase conformations for ligand binding, we set the generation of possible states from 7.0±2.0 pH, and select OPLS3 as the force field for its robustness in molecular interactions and energetics. To exclude large molecules we decided to avoid docking ligands with over 500 atoms and 100 rotatable bonds to reduce computational complexity. For each ligand docked to Nsp-13 using Maestro, we obtained a binding score. This score quantitatively represents the strength of the ligand's interaction with the helicase, providing valuable data for our analyses. By generating our binding data through this method, we effectively navigated around the limitations posed by the ChEMBL database.

2.2 Feature Extraction

We have selected an advanced embedding technique to analyze protein features, specifically focusing on helicase sequences. We employed the SeqVec Embedder from the bio_embeddings.embed library, which is particularly suited for handling the intricacies of protein sequences. The SeqVec Embedder is a specific type of embedder that uses the SeqVec model, a deep learning model trained to understand and encode the contextual information of amino acid sequences into high-dimensional vectors. It's built upon a language model that learns from a large corpus of protein sequences, allowing it to capture the complex relationships and patterns within these sequences.

We parsed the helicase sequences stored in a FASTA format file, converting them into a format that is compatible with our embedder. Once we had the sequences, we utilized the SeqVec Embedder to generate embeddings for each sequence. These embeddings are essentially high-dimensional vectors that encapsulate the essential features of the helicases. By embedding multiple sequences, we obtained a set of feature vectors that represent the diverse characteristics of the helicase proteins. After generating these embeddings, we employed a dimensionality reduction step by transforming the variable-size embeddings into fixed-size vectors.

To evaluate the effectiveness of our embeddings and visualize the high-dimensional data in a lower-dimensional space we use PCA (Principal Component Analysis). Performed clustering is based on the similarity of ligands using the Normalized Hamming (Dice

coefficient) for distance selection and a clustering threshold of 0.7. Fig. 2 shows the dispersion of points which suggests varying degrees of similarity among the ligands.

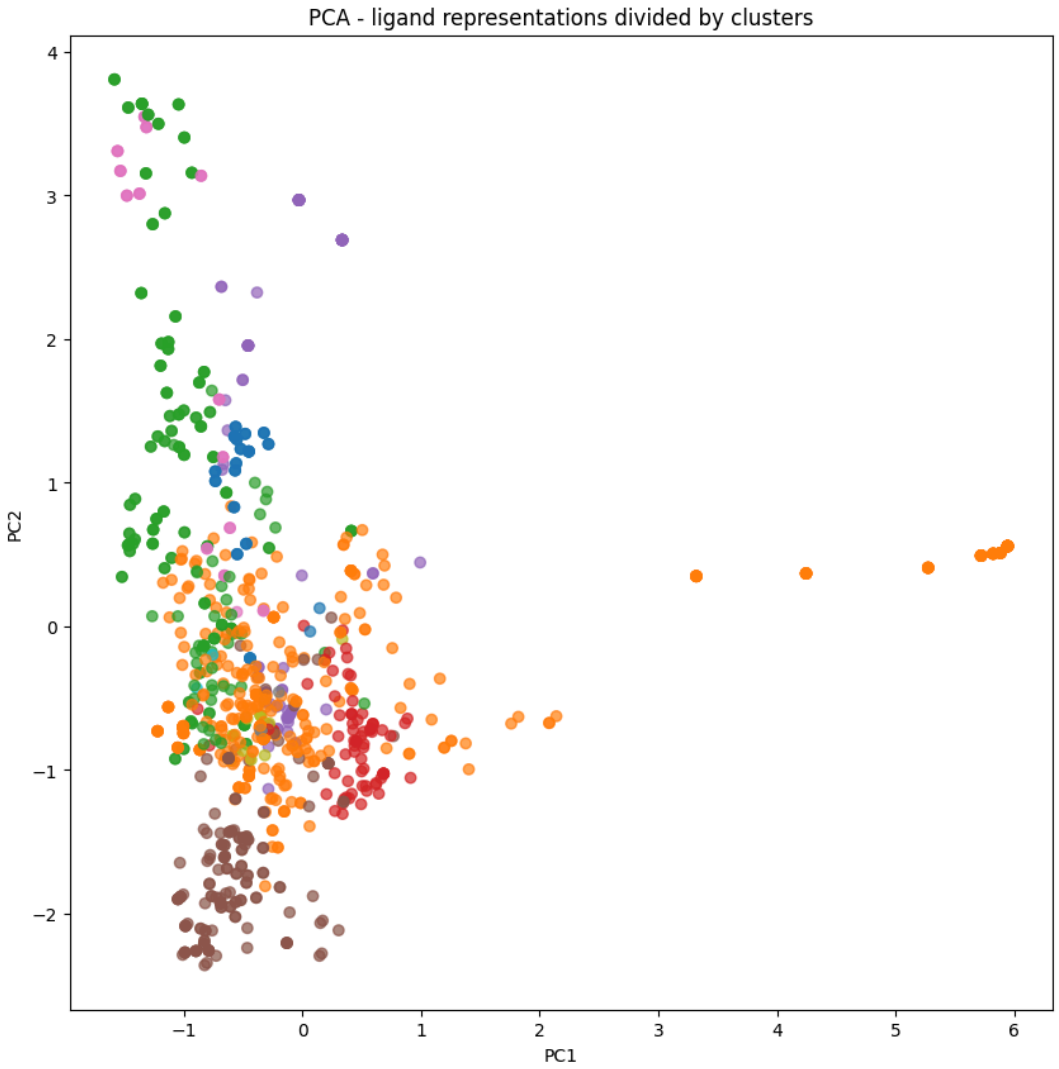

Fig. 2 PCA - ligand representations divided by clusters

In parallel, we obtain feature vectors for ligands using the RDKit library, based on the SMILES notation of ligands from our dataset. By converting the SMILES notation into RDKit molecule objects, we were able to gain a detailed representation of each ligand's chemical structure. We used Extended-Connectivity Fingerprints (ECFPs), specifically ECFP4. ECFP4 generates a fingerprint by circularly examining the molecular structure, centered on each atom, and considering the atom's environment up to two bonds away (since the diameter is four, half of it, which is two, determines the bond distance considered). Each of these environments is then converted into a numerical identifier, which is a part of the fingerprint. This technique of generating ECFP4 fingerprints offers a multidimensional view of the ligands, so it is particularly beneficial for understanding how these ligands might interact with various protein targets. The numerical representation obtained from these fingerprints also allows us to perform and use machine learning algorithms.

2.3 PCM Modeling

In our study, we used Proteochemometric Modeling (PCM) to analyze interactions between proteins and ligands, specifically focusing on the Nsp-13 helicase. PCM is a method that models the joint properties of both ligands and their protein targets to predict their interactions.

For the PCM, we implemented our machine learning strategy on a smaller subset of the data to establish baseline performance metrics. We chose two models: Random Forest Regressor and Support Vector Regressor (SVR). While both models have their strengths, our preliminary analysis revealed that the Random Forest Regressor showed more promising results in this context. The Random Forest is well-regarded for its capability to handle complex datasets with high accuracy, and this was reflected in our initial tests. The SVR, known for its effectiveness in regression tasks, also performed well.

After these initial tests, we proceeded to split the entire dataset into training and testing sets with a 70-30 ratio and trained both models on the larger training set. The performance of our models was then evaluated using metrics like RMSE (Root Mean Square Error), Spearman's rank correlation, and MCC (Matthews Correlation Coefficient). These metrics were instrumental in assessing the accuracy, correlation strength, and overall effectiveness of our models in predicting protein-ligand interactions and we present them in Tab 1.

Tab. 1 Comparision of our models' metrics

	Data from ChEMBL assays [IC50]		Data from docking [docking score]	
	Random Forest Model	SVR Model	Random Forest Model	SVR Model
RMSE	29.56	24.02	0.96	1.06
Spearman	0.54	0.65	0.59	0.52
MCC	0.49	0.53	0.49	0.43

We ended up having four different models, each combining different methods and data sources used for training:

1. Random Forest model trained on data from Dataset 1
2. Support Vector Machine (SVM) model trained on Dataset 1
3. Random Forest model trained on Dataset 2
4. Support Vector Machine (SVM) model trained on Dataset 2

For both the Random Forest and SVM models trained on Dataset 1, we focused on modeling the half-maximal inhibitory concentration (IC50) values. This provided us with quantitative insights into the potency of the ligands. In addition to assay data, we also

analyzed the results from our docking experiments which were included in Dataset 2, using both the Random Forest and SVM models to interpret the docking scores. By applying these diverse approaches, we were able to cross-validate our findings and gain a comprehensive understanding of the potential efficacy of the ligands in interacting with the Nsp-13 helicase. Fig. 3-6 present density plots comparing the distribution of true values and predicted values for each model.

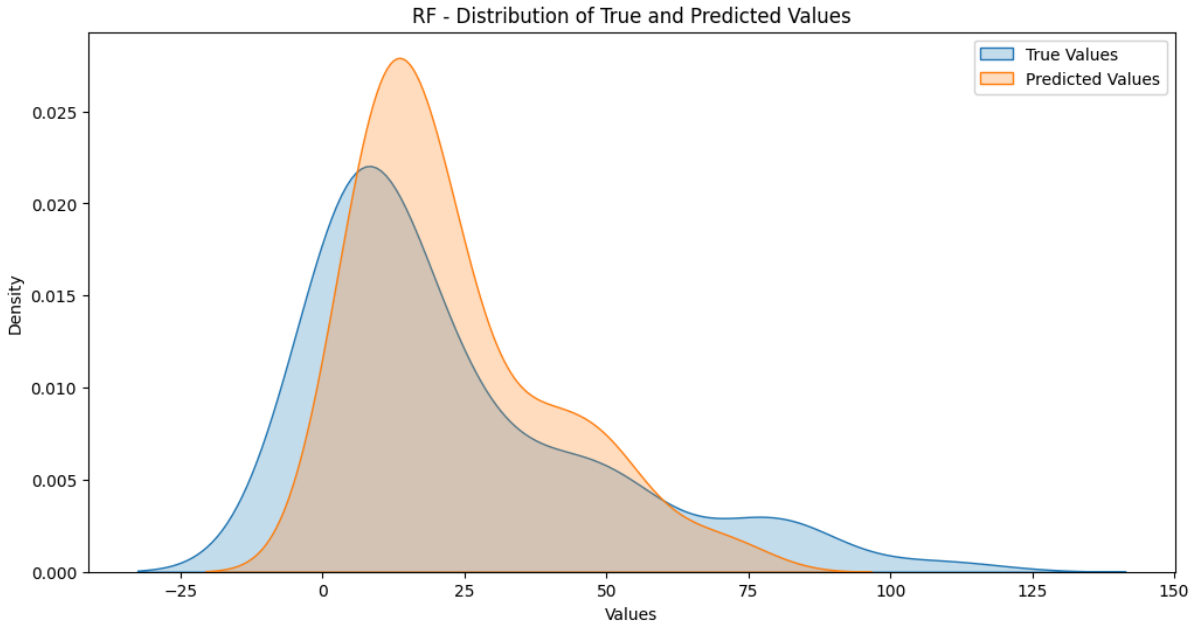

Fig. 3 Distribution of true and predicted values for random forest model on data from ChEMBL assays

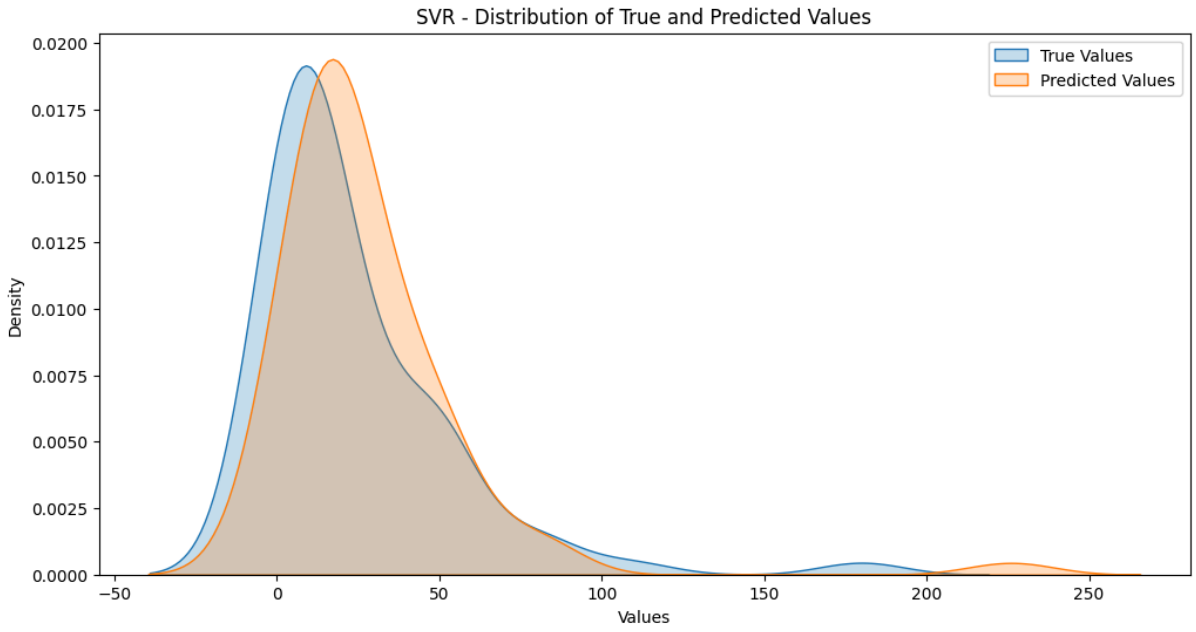

Fig. 4 Distribution of true and predicted values for SVR model on data from ChEMBL assays

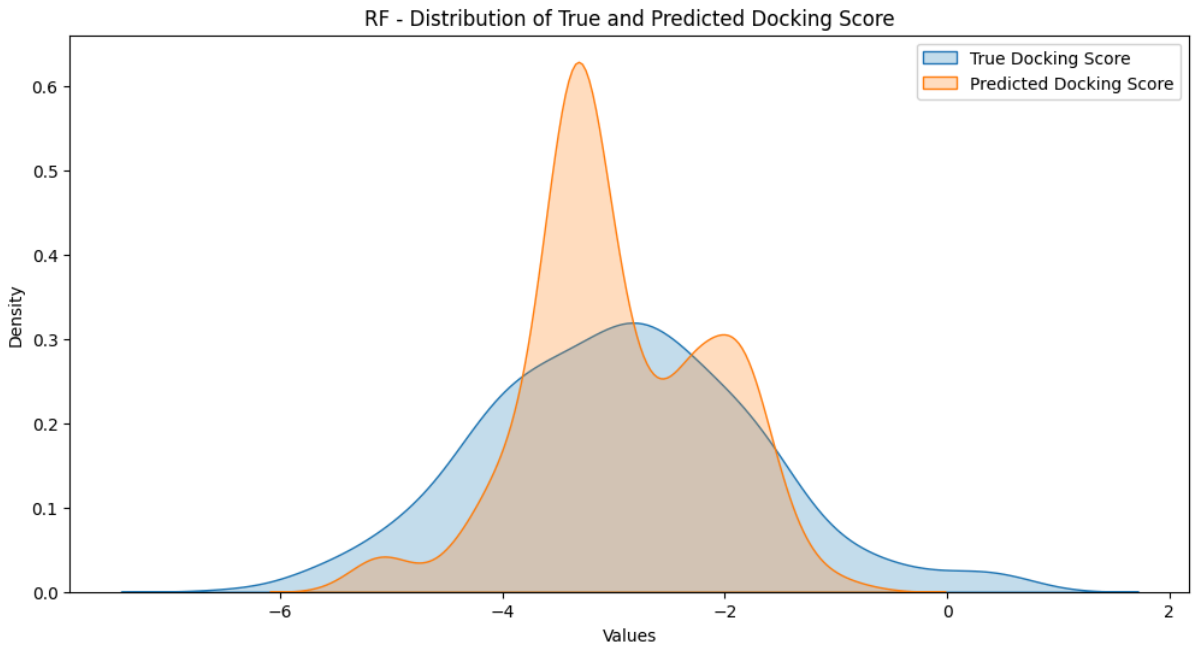

Fig. 5 Distribution of true and predicted values for random forest model on data from docking

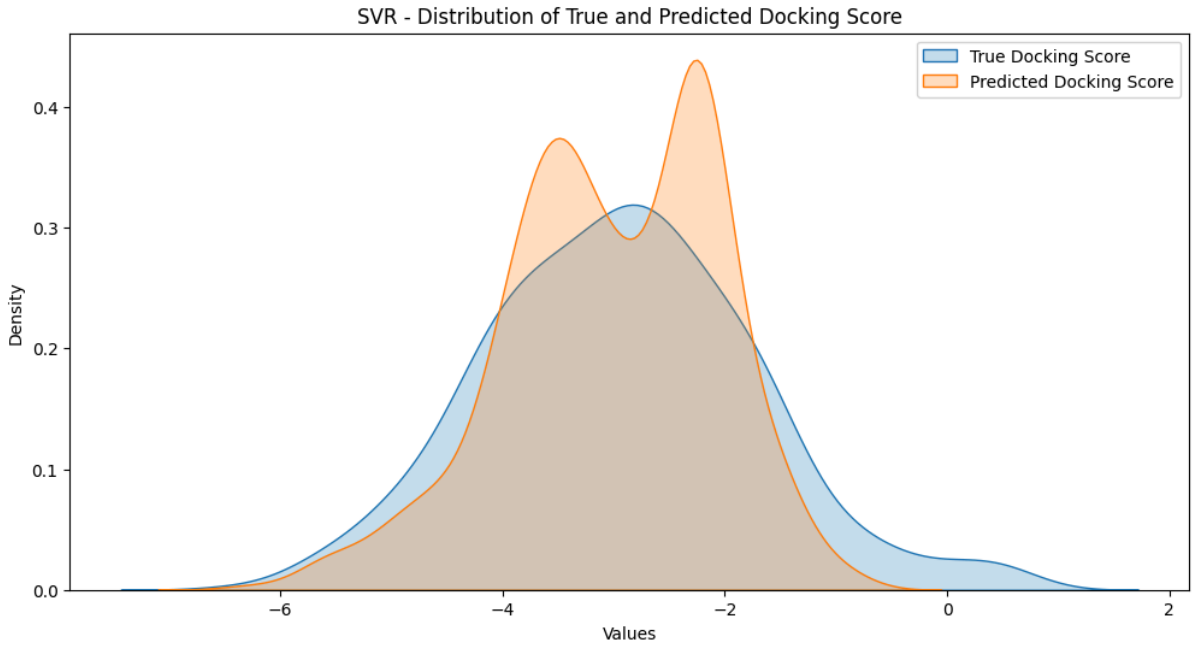

Fig. 6 Distribution of true and predicted values for SVR model on data from docking

After our initial testing and evaluation, we conducted a final, comprehensive analysis on a significantly larger dataset, comprising approximately 2.4 million ligands — all chemical compounds in the ChEMBL database. Based on our results, we decided that Random Forest models appear to be a better choice for further analysis. In this analysis, the helicase Nsp-13 protein was represented using an embedding technique — SeqVec, while the ligands were characterized through feature vectors prepared with RDKit. Our objective was to identify the most promising ligands based on their predicted interactions with the helicase.

We presented a distribution of molecular weights of the compounds from the ChEMBL database in Fig. 7. The histogram shows how the molecular weights of the ligands are spread across different ranges. We also explored the lipophilicity of the compounds, measured by AlogP values. Lipophilicity is a crucial factor in drug design, affecting solubility and permeability. Fig. 8 shows the range of AlogP values among the ChEMBL ligands. We performed this analysis based on our focus on the number of Lipinski's rule violations among the ligands. Lipinski's rule, also known as the "Rule of Five," is a set of criteria to evaluate the drug-likeness of compounds. This rule is a widely accepted set of guidelines in pharmaceutical chemistry, aimed at determining the drug-likeness of compounds based on their pharmacokinetic properties. Specifically, Lipinski's Rule of Five states that an orally active drug is more likely to have: no more than 5 hydrogen bond donors (the sum of OHs and NHs), no more than 10 hydrogen bond acceptors (the sum of Os and Ns), a molecular mass less than 500 daltons and andnoctanol-water partition coefficient (log P) not greater than 5 [4]. The number of Lipinski's rule violations among the ligands is presented in Fig 10.

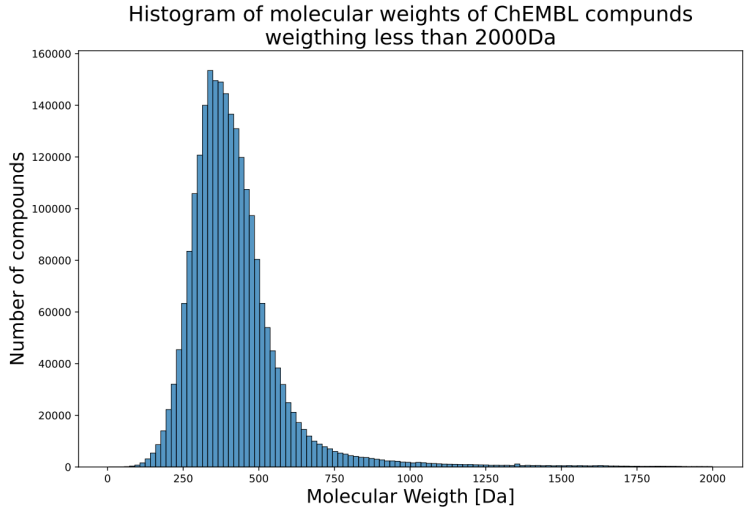

Fig. 7 Histogram of molecular weights of ChEMBL compounds weighting less than 2000Da

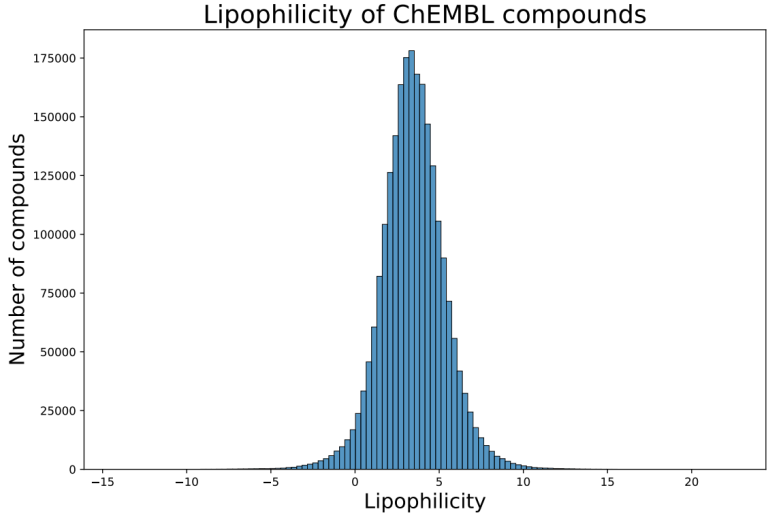

Fig. 8 Lipophilicity of ChEMBL compounds

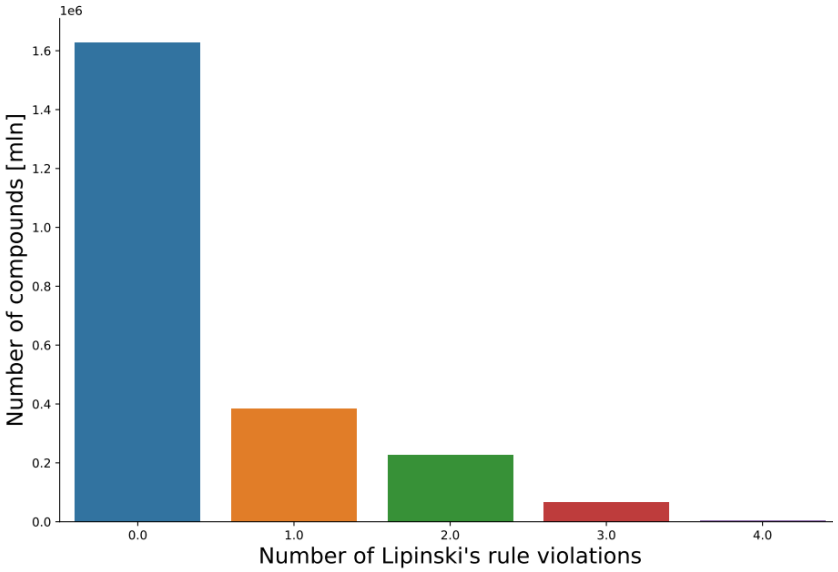

Fig. 10 Number of Lipinski's rule violations among the ligands

We ran trained Random Forest models on a large dataset in which we paired all ChEMBL assays with embeddings of Nsp-13 helicase, enabling us to determine the best ligands based on the lowest obtained IC50 (model trained on Dataset 1), and the best docking score (model trained on Dataset 2). Given that the best outcomes are indicated by both a low docking score and a low IC50 value, we integrated the measures from both models to identify the most promising ligands. The results showcasing the top-performing ligands are presented in Tab. 2. Additionally, the chemical structures of the two best-performing ligands, which emerged as the most promising candidates for further investigation, are illustrated in Fig. 11.

Tab. 2 Comparision of best-performing ligands achieved by Random Forest model

ID of ligand	IC50 [μM]	Docking score
CHEMBL1595621	0.046	-2.904
CHEMBL2228592	1.273	-3.223
CHEMBL5191763	1.615	-3.210
CHEMBL5203212	1.615	-3.210
CHEMBL5175180	1.615	-3.120
CHEMBL5174624	1.615	-3.120

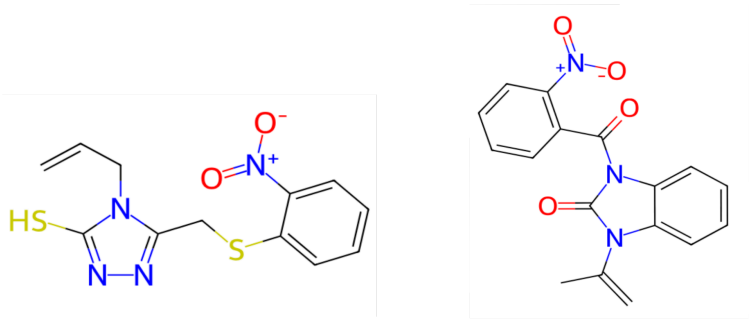

Fig. 11 Structures of the best-performing ligands: CHEMBL1595621 on the left and CHEMBL2228592 on the right [5]

3. Discussion

In our project, we successfully identified potential inhibitors of the SARS-CoV-2 Nsp-13 helicase, which marks a significant achievement. We enhanced the proteochemometric (PCM) modeling by incorporating docking results, overcoming potential data limitations. By modeling both IC50 values and docking scores, we could compare two distinct units, leading to more informed decisions on selecting the best inhibitors. We also conducted a comparison of the Random Forest and SVR models. This allowed us to design a comprehensive pipeline. While we have made significant strides in identifying potential helicase inhibitors using PCM and comparing models like Random Forest and SVR, we understand that further enhancements could be made. This might involve refining our models and for sure incorporating more diverse datasets.

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] Cortés-Ciriano, I., Ain, Q.U., Subramanian, V., et al. (2015). Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. MedChemComm, 6.

[3] Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., Cao, Z., & Zhu, R. (2017). The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope. Briefings in Bioinformatics, 18(1).

[4] Lipinski, C.A., Lombardo, F., Dominy, B.W., & Feeney, P.J. (1997). Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Advanced Drug Delivery Reviews, 23(1-3), 3-25.

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