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

University of Warsaw

Faculty of Mathematics, Informatics and Mechanics

Paulina Kucharewicz

Anastazja Avdonina

Julia Byrska

Jakub Guzek

Michalina Wysocka

Proteochemometric (PCM) approach

Our 3 main reasons

Lack of Protein Structures

Many protein structures, especially in the context of emerging pathogens like SARS-CoV-2, are not yet determined. PCM allows for the modeling of protein-ligand interactions even in the absence of detailed 3D structures, using sequence-based information instead.

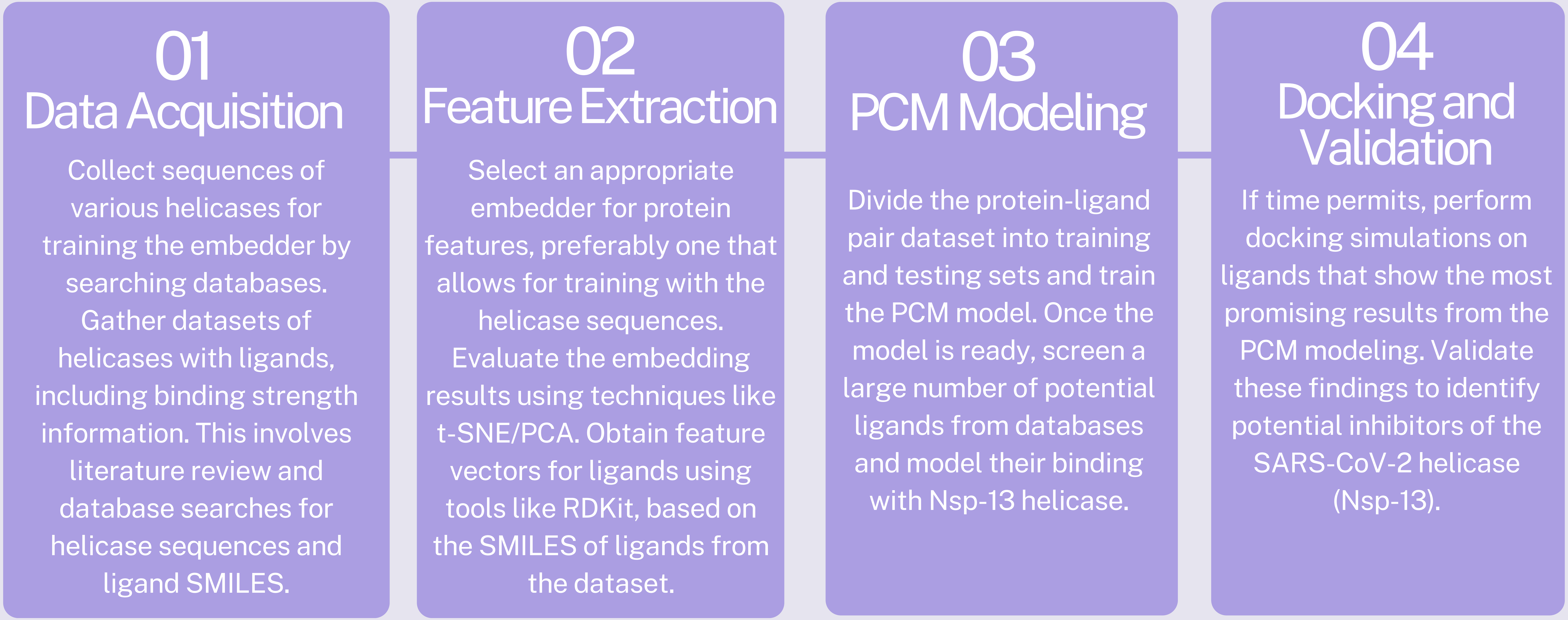
Reduced Computational Burden

Compared to methods that require detailed 3D structures, such as molecular docking or molecular dynamics simulations, PCM is less computationally intensive. This is particularly advantageous when dealing with large datasets or when computational resources are limited.

Flexibility and Generalizability

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.

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



01 Data Acquisition

Collect sequences of various helicases for training the embedder by searching databases. Gather datasets of helicases with ligands, including binding strength information. This involves literature review and database searches for helicase sequences and ligand SMILES.

02 Feature Extraction

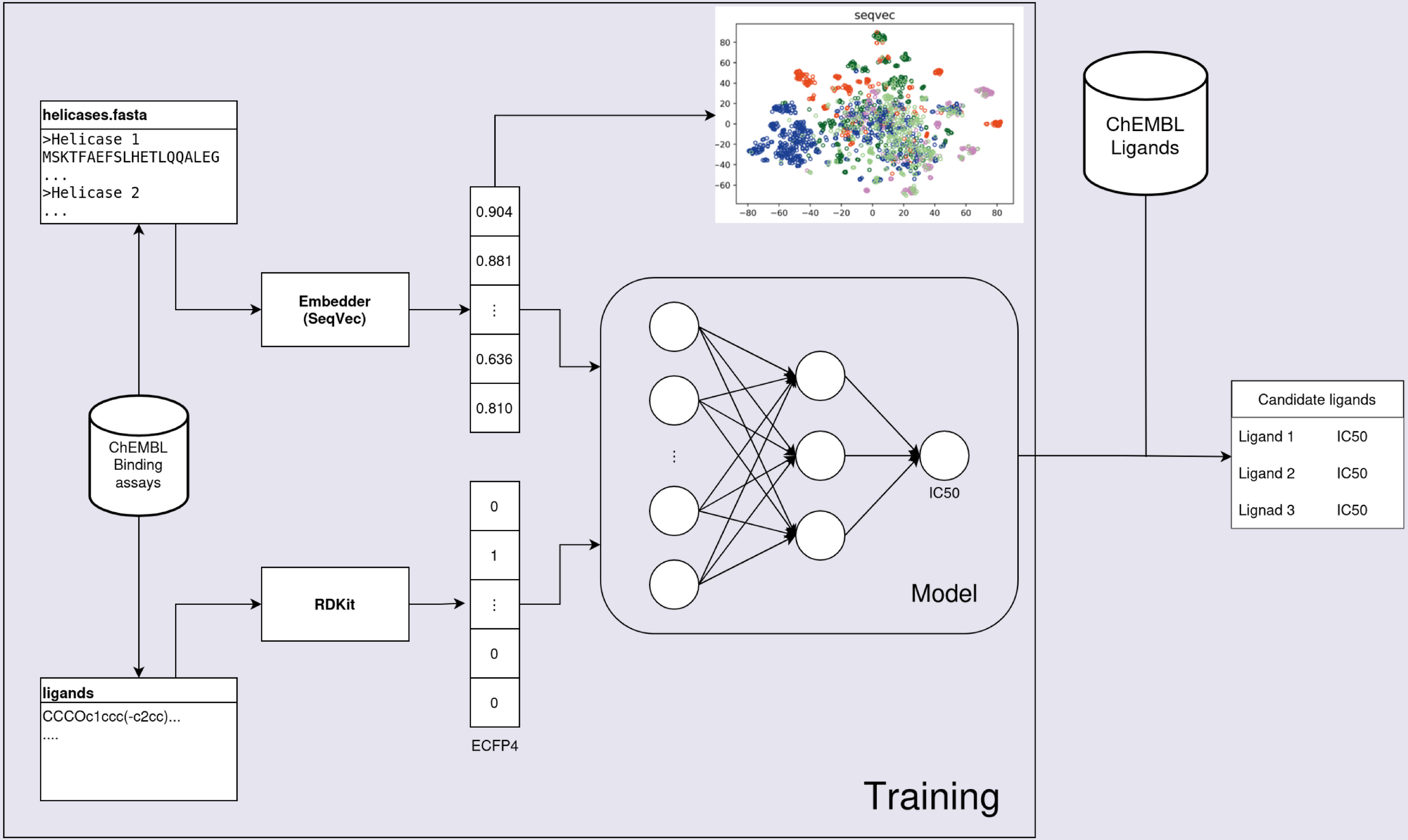
Select an appropriate embedder for protein features, preferably one that allows for training with the helicase sequences. Evaluate the embedding results using techniques like t-SNE/PCA. Obtain feature vectors for ligands using tools like RDKit, based on the SMILES of ligands from the dataset.

03 PCM Modeling

Divide the protein-ligand pair dataset into training and testing sets and train the PCM model. Once the model is ready, screen a large number of potential ligands from databases and model their binding with Nsp-13 helicase.

04 Docking and Validation

If time permits, perform docking simulations on ligands that show the most promising results from the PCM modeling. Validate these findings to identify potential inhibitors of the SARS-CoV-2 helicase (Nsp-13).



helicases.fasta

```
>Helicase 1
MSKTFAEFSLHETLQQALEG
...
>Helicase 2
...
```

ligands

CCCOc1ccc(-c2cc)...
....

Candidate ligands	
Ligand 1	IC50
Ligand 2	IC50
Lignad 3	IC50

Results

01 Data Acquisition

Sequences of various helicases have been collected from the Unirep database for potential embedder training. Data from binding assays have been downloaded and extracted from the ChEMBL database.

02 Feature Extraction

Multiple candidates for embedding protein features have been evaluated, with several being tried out. The plan is to use SeqVec, hence there's no immediate need for embedder training. Ligand representation has been generated using RDKit.

03 PCM Modeling

One variation of analysis was conducted using data and a script made available by the authors of the paper: 'How to approach machine learning-based prediction of drug/compound–target interactions' by Atas Guvenilir, H., and Doğan, T.

04 Docking and Validation

Not done yet