Advanced Computational Methods in Drug Discovery

Name: Mania Habibijouybari

Student number: s3269930

Introduction:

Protein-tyrosine phosphatase (PTP) is an intracellular enzyme superfamily, which involves in regulating the signaling pathways of many biological processes in the cell cycle. PTP enzymes are located in cytoplasmic side of endoplasmic reticulum membrane, and catalyze protein tyrosine dephosphorylation by removing the phosphate groups from phosphorylated tyrosine residues on proteins. In humans, more than a hundred PTPs exist that can function either as negative or positive modulators in various signal transduction pathways.

PTP1B as a key member of this family plays an important role in the signaling pathways of several human diseases, such as diabetes, obesity, and hematopoietic malignancies. Studies have shown, the inhibition of this enzyme can have therapeutic effects against these diseases. Therefore, over the last decades PTP1B has been considered a potential drug target for developing novel therapeutic agents. However, targeting this enzyme for drug discovery is a bit challenging due to the highly conserved and positively charged active-site pocket of this enzyme.

So far, tremendous progress has been made in the development of potent and selective PTP1B inhibitors, resulting in the discovery of few molecules, including: ertiprotafib, trodusquemine, JTT-551, etc. Ertiprotafib is a noncompetitive multiple-action inhibitor and belongs to a novel class of insulin sensitizers, which is developed for the treatment of type 2 diabetes. Trodusquemine selectively targets and inhibits PTP1B, thereby preventing PTP1B-mediated signaling. It has potential hypoglycemic, anti-diabetic, anti-obesity, and antineoplastic activities. JTT-551 is a mixed-type inhibitor that is reported to improve glucose metabolism by enhancement of insulin signaling.

Unfortunately, most of these inhibitors were discontinued due to their insufficient efficiency and lack of specificity and notable side effects.

The PTP1B structure is composed of three domains: an N-terminal catalytic domain (1–300), a regulatory domain (301–400), and a C-terminal domain (401–435) responsible for targeting the enzyme to the endoplasmic reticulum (ER) membrane. All three domains play a critical role in the regulation of PTP1B.

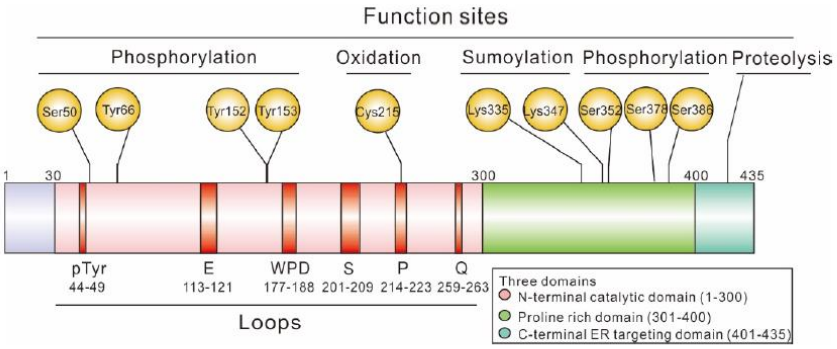

Figure 1. Schematic representation of the domain structures of PTP1B full length. Full-length PTP1B is composed of an N-terminal catalytic domain containing several important loops (1–300) and C-terminal ER targeting domain (401–435), flanking two proline-rich domains (301–400), with multiple functions at different sites of PTP1B.

Practical work:

During this course, we aimed to design a new ligand for our protein target (PTP1B) based on the obtained results of the bio-informatics, chem-informatics, and docking measurements.

The practical work started with retrieving bio-activity data for the target PTP1B from ChEMBL database.

The ChEMBL database is a freely available bioactivity database containing close to 2.5 million compound records on nearly 2 million unique chemical structures. In ChEMBL, bioactivity data of curated marketed drugs and clinical candidates from all sources including scientific articles, deposited datasets and curated drug sources are aggregated according to the chemical structure. It is a useful tool in computational drug discovery that brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

First, ChEMBL was queried using the target ID of our protein target PTP1B, which is P18031 to get the ChEMBL ID used for this target (CHEMBL335).

```
uniprot_id = "P18031" #PTP1B (2QBS)
```

```
# Get target information from ChEMBL but restrict it to specified values only
targets = targets_api.get(target_components__accession=uniprot_id).only(
    "target_chembl_id", "organism", "pref_name", "target_type"
)
print(f'The type of the targets is "{type(targets)}"')
```

```
The type of the targets is ""
```

```
# Select first entry
targets = pd.DataFrame.from_records(targets)
targets
```

	organism	pref_name	target_chembl_id	target_type
0	Homo sapiens	Protein-tyrosine phosphatase 1B	CHEMBL335	SINGLE PROTEIN
1	Homo sapiens	Protein-tyrosine phosphatase 1B	CHEMBL335	SINGLE PROTEIN

Next, the bioactivity data of compounds tested against PTPB1 was queried using this ChEMBL ID which resulted in 3437 rows of data.

```
bioactivities_df.from_records = pd.DataFrame.from_records(bioactivities)
print(f"DataFrame shape: {bioactivities_df.shape}")
bioactivities_df.head()
```

```
DataFrame shape: (3437, 13)
```

	activity_id	assay_chembl_id	assay_description	assay_type	molecule_chembl_id	relation	standard_units	standard_value	target_chembl_id	target_organism	type	units	value
0	33473	CHEMBL772435	In vitro inhibitory activity against recombina...	B	CHEMBL301254	=	nM	1540.0	CHEMBL335	Homo sapiens	IC50	uM	1.54
1	33479	CHEMBL772435	In vitro inhibitory activity against recombina...	B	CHEMBL58435	=	nM	10130.0	CHEMBL335	Homo sapiens	IC50	uM	10.13
2	34712	CHEMBL770122	Inhibition of human Protein-tyrosine phosphata...	B	CHEMBL57157	=	nM	610.0	CHEMBL335	Homo sapiens	IC50	uM	0.61
3	34713	CHEMBL770122	Inhibition of human Protein-tyrosine phosphata...	B	CHEMBL292444	=	nM	1010.0	CHEMBL335	Homo sapiens	IC50	uM	1.01
4	34733	CHEMBL772435	In vitro inhibitory activity against recombina...	B	CHEMBL60707	=	nM	1130.0	CHEMBL335	Homo sapiens	IC50	uM	1.13

The data was filtered to only include data measured in nM which resulted in 3369 rows. However, this data frame consists of multiple entries of the same compound so the data frame was altered to include the mean of the same compounds resulting in 2989 unique compounds. Next up, the corresponding SMILES were added to the table and compounds not displaying a SMILES string were removed. Since the IC50 values are in a large value range and are given in different units (M, nM, ...), the pIC50 values were used to facilitate the comparison of results. Therefore, the IC50-values were converted to pIC50 values. Which finally gave rise to a data frame containing 2983 rows of data.

```
# Apply conversion to each row of the compounds DataFrame
output_df["pIC50"] = output_df.apply(lambda x: convert_ic50_to_pic50(x.IC50), axis=1)
output_df.head()
```

	molecule_chembl_id	IC50	units	smiles	pIC50
0	CHEMBL100267	20000000.0	nM	COc1ccc2cc(C(=O)O)ccc2c1C(=O)O	1.698970
1	CHEMBL101427	19000.0	nM	C[C@H]1CCC[C@@H](C)N1NC(=O)c1ccc(Cl)c(S(=O)(=O...	4.721246
2	CHEMBL102015	24000.0	nM	O=C(NS(=O)(=O)Cc1ccccc1)c1ccc2cc(C(F)(F)P(=O)(...	4.619789
3	CHEMBL103709	25000.0	nM	O=C(NS(=O)(=O)c1cc(C2(O)NC(=O)c3ccccc32)ccc1Cl...	4.602060
4	CHEMBL103942	30000.0	nM	O=C(NS(=O)(=O)c1ccccc1)c1ccc2cc(C(F)(F)P(=O)(O...	4.522879

After this step, the Lipinski-rule was applied to filter only orally bioavailable compounds. The bioavailability of a compound is an important part of pharmacokinetics, which determines the ability of a compound to reach the target and remove from the body in an appropriate period of time. The Lipinski-rule was formulated by Christopher A. Lipinski in 1997, based on the observation that most orally administered drugs are relatively small and moderately lipophilic molecules. The 2983 rows of data were further filtered based on the Lipinski-rule using rdkit dataset. As a result, 1992 of these compounds were in compliance with the Lipinski-rule, thus are likely to be ADME-favourable.

The IC50 value or half maximal inhibitory concentration, is a concentration of a compound inhibiting a specific biological or biochemical function by 50%. Under certain conditions, it can be used to express the affinity of that inhibitor. Here, pIC50 value, the negative log of IC50, is used to facilitate the comparison of different IC50 values. In the next step, the compounds with higher pIC50 that shows higher inhibitory effects are filtered.

An additional column was added to the data which resembled the activity of the compound and was set to True if the compound displayed a pIC50 higher than 7 and False otherwise. The smiles column was converted to MACCS keys fingerprints and a neural network was trained on 70% of this data and validated on 30% of the data. The neural network consisted of two hidden layers with 32 nodes and one output layer with one node and a sigmoidal activation. The model was trained for 100 epochs and reached a validation accuracy of 97% and a training accuracy of 99%. A confusion matrix was generated as well and the results are shown in the table below.

Table 1	Active	In-active
Predicted Active	42	5
Predicted Inactive	29	1916

It seems that this model is more realistic in predicting if a compound is inactive (99.7%) compared to if a compound is active (59.2%).

De novo generation

In order to generate potent and selective ligands, the DrugEx tool developed by Martin Sicho was used. This resulted in 1000 de novo molecules being generated which closely resemble the original dataset, which is illustrated in figure 1. Next, the obtained 1000 compounds were filtered to fulfil all 5 rules of Lipinski, which resulted in 872 compounds. Then, the unwanted substructures which influence activity, known to be toxic or influence absorbance were filtered, which resulted in 391 de novo compounds. The model which was previously been made was used to predict the activity probabilities of these compounds and compounds displaying a probability higher than 0.5 were chosen. Only 4 compounds fulfilled these criteria, shown in Figure 2.

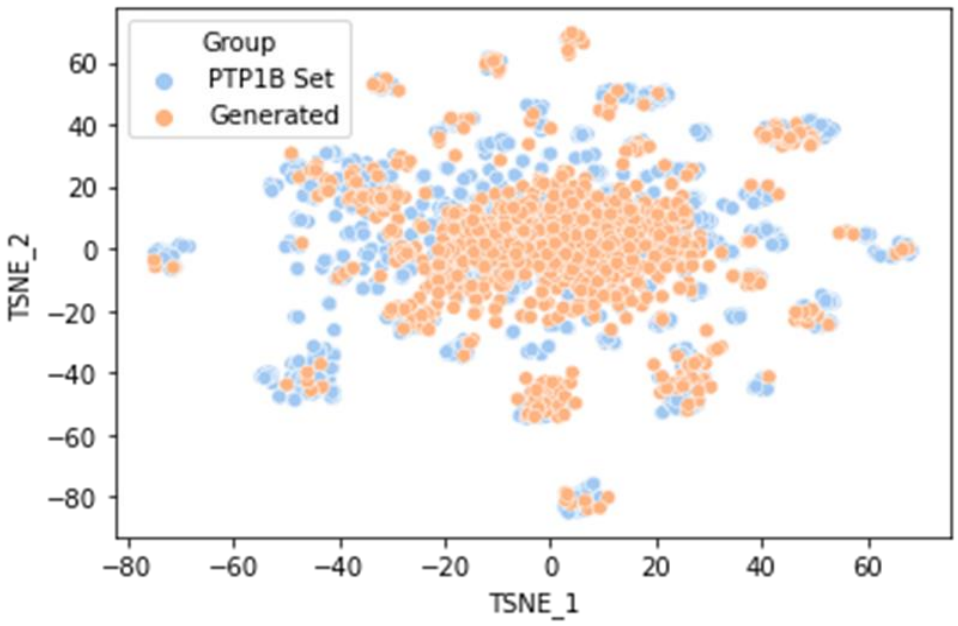

Figure 1. tSNE of the morgan fingerprints of the drugs contained in the original dataset (blue) and in the generated dataset (orange).

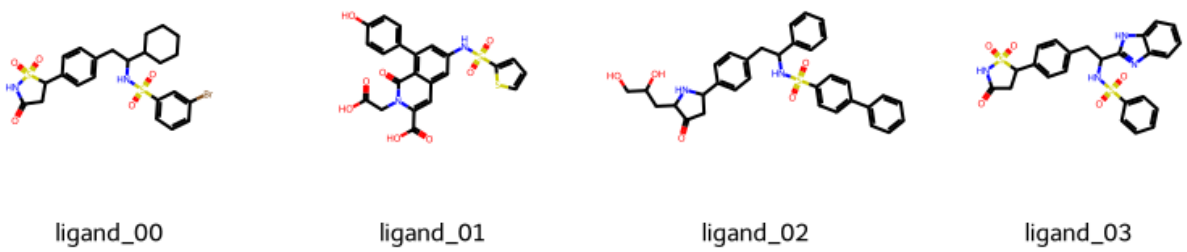

Figure 2. Potential active de novo compounds.

Docking

Molecular docking as a significant part of structure-based drug design (SBDD), is a technology that is being used to predict the fitting (docking) of a virtual molecule into the target protein receptor. Molecular docking software can predict these binding modes by sampling possible conformations of a ligand inside the protein binding pocket. Based on the obtained results, a potential ligand with appropriate binding properties will be recognized and selected for further drug discovery experiments.

The molecular docking studies were performed based on the template of Willem Jespers on GitHub, to predict the binding mode of these four compounds in the protein binding site. The protein code (2QBS) and co-crystalized ligand (024) were used to create the necessary PDB files after which the docking of the 4 de novo ligands were evaluated. The docking poses are illustrated in Figure 3. Ligand 02 seems most interesting since it encapsulates one of the residues of the target site due to its scissor-like appearance. Furthermore, the three highest scoring docking ligands share a common structure (Figure 4). Another benzene ring was added to this common structure to enable the scissor-like bending characteristics of ligand 02. One R group were added to the molecule to give rise to the final lead scaffold compound (Figure 5).

Ligand Name	Kcal/mol	Model Predicted Activity	pIC50
Ligand_00	-7.799	60.1%	5.72
Ligand_01	-6.676	81.9%	4.9
Ligand_02	-7.865	90.6%	5.77
Ligand_03	-7.466	85.8%	5.46

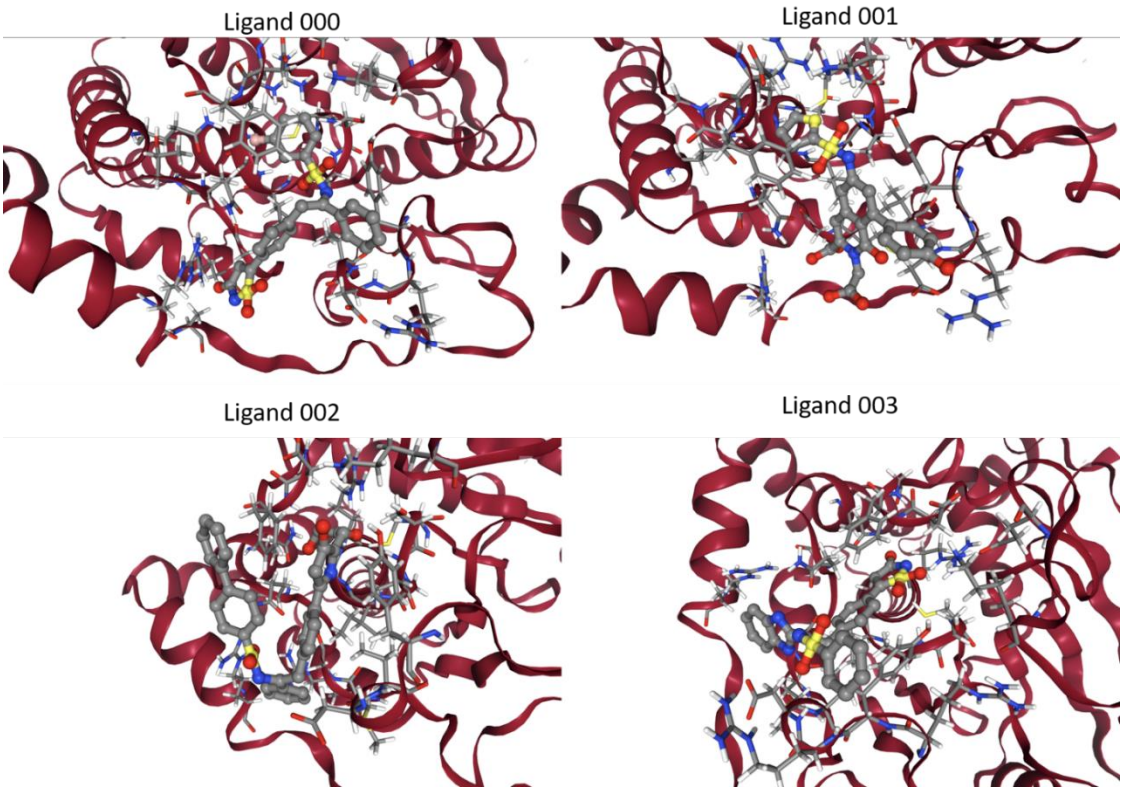

Figure 3. Docking poses of the de novo compounds.

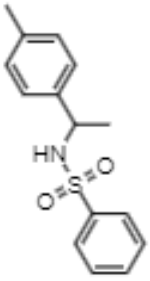

Figure 4. Common structure of 3 highest scoring ligands.

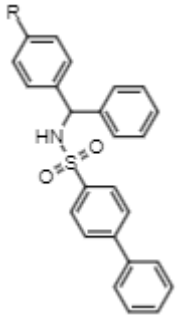

Figure 5. Final lead scaffold compound.

References

1. Zhang, Sheng, and Zhong-Yin Zhang. "PTP1B as a drug target: recent developments in PTP1B inhibitor discovery." Drug discovery today 12, no. 9-10 (2007): 373-381.
2. Liu, Rongxing, Cécile Mathieu, Jérémy Berthelet, Wenchao Zhang, Jean-Marie Dupret, and Fernando Rodrigues Lima. "Human Protein Tyrosine Phosphatase 1B (PTP1B): From Structure to Clinical Inhibitor Perspectives." International Journal of Molecular Sciences 23, no. 13 (2022): 7027.
3. Kumar, Ajay, Divya Rana, Rajat Rana, and Rohit Bhatia. "Protein tyrosine phosphatase (PTP1B): A promising drug target against life-threatening ailments." *Current Molecular Pharmacology* 13, no. 1 (2020): 17-30.
4. https://github.com/martin-sicho/drugex-demo
5. https://github.com/jesperswillem/CBR_teaching
6. ChEMBL bioactivity database: Gaulton et al., Nucleic Acids Res. (2017), 45(Database issue), D945–D954
7. ChEMBL web services: Davies et al., Nucleic Acids Res. (2015), 43, 612-620
8. ChEMBL web-interface
9. ChEMBL bioactivity database: Gaulton et al., Nucleic Acids Res. (2017), 45(Database issue), D945–D954
10. ChEMBL web services: Davies et al., Nucleic Acids Res. (2015), 43, 612-620
11. ChEMBL web-interface
12. Pagadala et al., Biophy Rev (2017), 9, 91-102
13. Meng et al., Curr Comput Aided Drug Des (2011), 7, 2, 146-157
14. Gromski et al., Nat Rev Chem (2019), 3, 119-128
15. Docking and scoring function assessment:
16. Warren et al., J Med Chem (2006), 49, 20, 5912-31
17. Wang et al., Phys Chem Chem Phys (2016), 18, 18, 12964-75
18. Koes et al., J Chem Inf Model (2013), 53, 8, 1893-1904
19. Kimber et al., Int J Mol Sci, (2021), 22, 9, 1-34
20. McNutt et al., J Cheminform (2021), 13, 43, 13-43
21. Visual inspection of docking results: Fischer et al., J Med Chem (2021), 64, 5, 2489–2500