1 Virtual Docking

The goal of this assignment is to evaluate the performance of a docking software in estimating the binding affinity of compounds to Carbonic Anhydrase I (CA1), by comparing the virtual docking results to available bioactivity data.

Task 1

The bioactivity data for CA1 can be found in ChEMBL webpage as CHEMBL261. Currently, 9031 activity types for CHEMBL261 are reported in ChEMBL, see Figure 1.

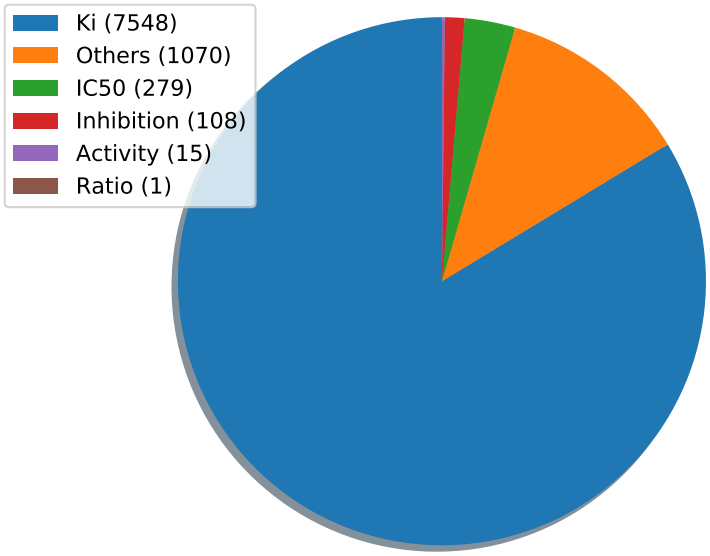

Figure 1: ChEMBL activity types for target CHEMBL261.

CHEMBL261 indicates 24 reported structures which can be cross-referenced with RCSB or PDBe such as: 1AZM, 1BZM, 1CRM, 1CZM, 1HCB, 1HUG, 1HUH, 1J9W, 1JV0, 2CAB, 2FOY, 2FW4, 2IT4, 2NMX, 2NN1, 2NN7, 3LXE, 3W6H, 3W6I, 4WR7, 4WUP, 4WUQ, 5E2M, and 5GMM, with the matching binding affinity data.

Figure 2 indicate the structure of CA1, the details of curing the structure, after downloading from the databank, is given in **Task 4**.

Task 2

Choosing a working set of structural quality criteria depends to what we are planning to do with the PDB once we have it. It is recommended to consider the following:

- Is the structure in native form or in ligand bound form?

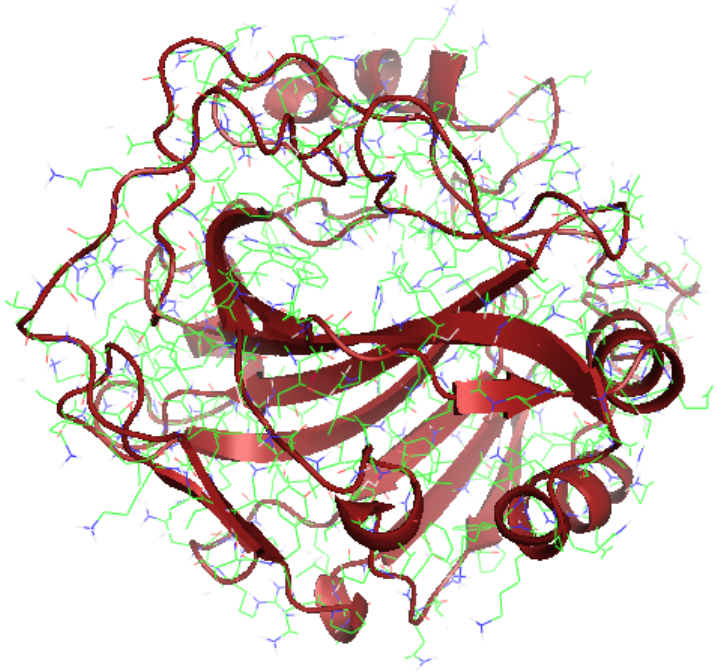

Figure 2: Structure of the receptor CA1. Image has been generated using pymol.

- Is the structure for a normal or mutated protein?
- Is the full structure available?
- How much are the resolution and R-factor?

One also can search and superimpose high resolution structures of the target and select a structure that is a good representative, i.e, not distorted too far away from the others. For more information, refer to [1].

In this exercise, we considered structures with the following criteria:

- resolution < 2Å
- $R_{free} < 0.23$
- Ramachandran outliers $= 0$
- Clashscore < 5
- Sidechain outliers $< 2\%$

With the chosen criteria, two CA1 structures have been downloaded from PDB; 4WR7 and 4WUQ.

Task 3

10 ligands have been chosen for docking, including; ACT, AZM, BCT, FLB, M28, MZM, PEG, TOR, TRS, and V14.

Task 4

Every docking exercise starts with editing the PDB files to make the input satisfy the requirements of the program. The most common adjustments are

- adding Hydrogen atoms
- removal of extraneous waters or other molecules commonly labelled HETATM in the PDB file
- adjustment of pH sensitive protonation
- editing out disordered elements
- selection of the monomer or biological unit important to the calculation

In this exercise, after downloading the selected structures, 4wr7 and 4wuq, from PBD in pdb format, open them using any text editor like *kate* or *vi*. First remove all non-protein atoms such as water, ions, or ligand which labelled as HETAT. Then if the file contains more than one monomer, seperate the monomers to individual files so that each file only contains one monomer structure. In our case, both selected structures have two monomer, by separating them, we will have 4 structures for the receptor.

Another important step is to add hydrogen to the receptor files, since hydrogen atoms are not provided in PDB. To add hydrogen, we use AutoDockTools (MGLTools). Open the program, load the chosen structure by selecting **File** → **Read Molecule**. To add hydrogen select **Edit** → **Hydrogens** → **Add** → **Polar Only** → **OK**.

We will use AutoDock for docking step, this software requires pdbqt format for the receptor and the ligands. To save the structure of the receptor in pdbqt format select **Grid** → **Macromolecule** → **Choose** and select the molecule of interest. A pop-up WARNING window appears, close it and then select the location to save the pdbqt file.

We also need to define the so-called search space for the docking process, which is a grid box where software will use to dock the ligand. The location of the center of the box as well as it's size should be provided for docking process. Select **Grid** → **Grid Box** and change the parameters so that the box cover the target location of docking. There is no need to save anything here, just have a note of the chosen parameters, we will use them later in the docking process. In this work, we used a cubic box of $26 \times 26 \times 26$ Å^3. To have a same location of the box, we alighned the structures using AutoDock. The cneter of box is at $x = 36.8$, $y = 14.7$ and $z = -12.9$.

To prepare the ligands, select **Ligand** → **Input** → **Open**, and choose the ligand file. In the docking process, the program will try to best locate the ligand in the docking site by rotating the molecules around the allowed rotational bonds. To modify the rotational bonds, select **Ligand** → **TorsionTree** → **Choose Torsion**. To save the ligand file in pdbqt format, select **Ligand** → **Output** → **Save as PBDQT** and save the file at the desired location.

Task 5

Having 4 receptor structures and 10 ligands, we should perform 40 docking calculations using *vina*. We write a bash script to carry out the calculations.

Task 6

a) Table 1 indicates the binding affinity of the first mode of 40 structure-ligand complexes, 4 receptors and 10 ligands. For each chosen ligand, we averaged the binding energy over 4 studied structures.

	ACT	AZM	BCT	FLB	M28	MZM	PEG	TOR	TRS	V14
4wr7A	-3.8	-6.0	-3.6	-6.4	-6.6	-6.2	-3.8	-6.5	-3.9	-6.0
4wr7B	-3.7	-5.9	-3.5	-6.4	-6.6	-6.0	-3.8	-6.2	-4.0	-6.1
4wuqA	-3.7	-6.2	-3.5	-6.5	-6.5	-6.1	-3.8	-6.7	-3.7	-5.9
4wuqB	-3.7	-6.0	-3.5	-6.4	-6.5	-6.2	-3.8	-6.4	-3.8	-5.9
ave.	-3.7	-6.0	-3.5	-6.4	-6.6	-6.1	-3.8	-6.4	-3.8	-6.0
s.d.	±0.04	±0.11	±0.04	±0.04	±0.05	±0.08	±0.00	±0.18	±0.11	±0.08

Table 1: Binding affinity in kcal/mol of the first mode for all the studied complexes out of 4 receptors and 10 ligands.

Vina calculates the binding energy in kcal/mol, to compare the calculated values to the experimental results, we convert the reported experimental dissociation constant from Chembl to binding energy using $k_i(k_D) = exp(\Delta G/RT)$, where ΔG is the binding energy, R is the gas constant, and T is the room temperature.

Table 2 shows the binding energy in kcal/mol for both calculated and experimental values. As table indicates, we observe a discrepancy between the experimental values and the calculated ones, we should attention that the calculations are done using a very simple model for the complexes. There is no environmental effect, effect of water molecules and other ions, considered for the calculations. Moreover, all the structures in the model are considered rigid!

	ACT	AZM	BCT	FLB	M28	MZM	PEG	TOR	TRS	V14
calculated	-3.7	-6.0	-3.5	-6.4	-6.6	-6.1	-3.8	-6.4	-3.8	-6.0
experiment	-6.8	-6.1	-2.5	-6.1	-7.1	-5.9	-6.8	-4.9	-8.3	-8.1
error	45.4%	1.0%	40.7%	5.6%	6.5%	4.0%	43.9%	30.2%	54.3%	26.2%

Table 2: Binding affinity in kcal/mol of experimental and calculated values.

The average error in estimating the binding affinity stands about **25.8%**. Please attention that even small error in estimating the binding energies will lead to a large error in estimating the dissociation constants.

b) As depicted in Figure 3 there is a very weak correlation between the calculated and the experimental values. To have a quantitative measure, the correlation of the docking binding affinities versus the experimental binding affinities from Chembl are listed below:

- Coefficient of determination, R^2 = **0.013**. This small value indicates that the model explains almost none of the variability of the response data around its mean.
- Pearson correlation = **0.113**, which shows that there is a very weak association between the two variables.
- Spearman correlation = **-0.003**. There is no relationship exists between the variables.
- Kendall's rank correlation = **-0.047**. Two variables are independent.

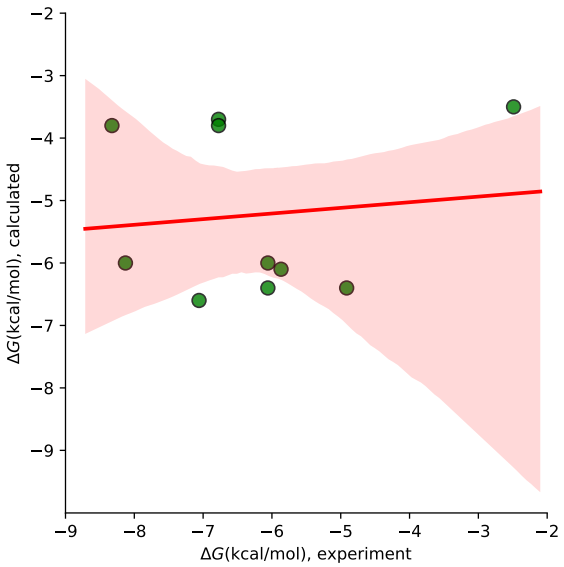

Figure 3: Correlation between calculated binding affinities and the experimental values from Chembl.

Please attention that Pearson correlation coefficients measure only linear relationships while Spearman correlation coefficients measure only monotonic relationships.

2 Library Creation

The goal of this assignment is to construct a library of 10,000 distinct compounds, which are all synthesizable. The library should be optimized for chemical diversity, to span as much of the chemical space as possible.

Task 1

We can download the building blocks from Molport in sdf (Spatial Data File) or smiles (Simplified molecular-input line-entry system) formats. In this work, we use smiles format. After downloading the whole database, we selected the first 200 compounds from the initial file. This set will be used to generate new products using some reactions.

Task 2

We used a set of two reactions;

- formation of amid bond: COOH + -NH- $\rightarrow$ -CONH-, which in smile format can be written as [C:1](=[O:2])O.[N:3]>>[C:1](=[O:2])[N:3]
- alcohol-acid reaction: [CH1:1][OH:2].[OH][C:3]=[O:4]>>[C:1][O:2][C:3]=[O:4]

Task 3

Using rdkit, we performed two type of reactions between the initial set of 200 building blocks. After the reactions, we obtained 26,471 compounds. The python code used for generating the product is attached to this report.

Task 4

Figure 4 and 5 indicate the molecular weight and logP distributions of initial set of building blocks, products, and all the sets together. As the figures show, the set of the products has larger molecular weight and logP values in comparison to the initial set. Since there are more products, and hence statistics, a more symmetric shape is observed.

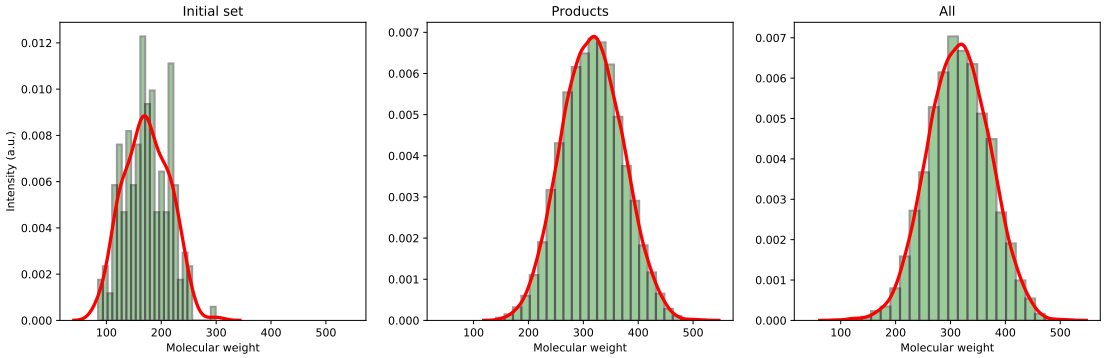

Figure 4: Distribution of molecular weight of initial set of building blocks, products, and all the sets together.

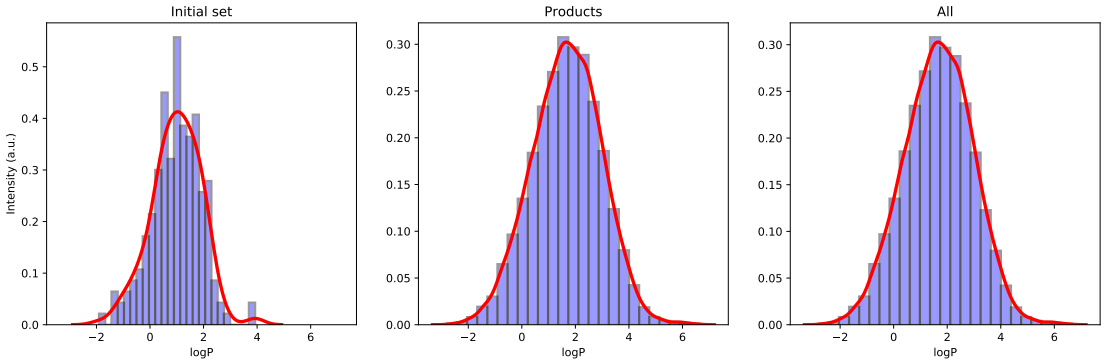

Figure 5: Distribution of logP of initial set of building blocks, products, and all the sets together.

Task 5

To compare the chemical diversity in each set, we define a metric-set including molecular weight (MW), logP, topological polar surface area (TPSA) in $Å^2$, number of hydrogen-bond donors (HBD), number of hydrogen-bond acceptors (HBA), number of aliphatic carbocycles (AliCyc), number of aromatic carbocycles (AroCyc), and number of rotational bonds (RB). Figure 6 indicates the chosen metrics. Comparing these results to the published databanks, one can see that the generated set in this work is in the range of drugbank data at least in the sence of the chosen metrics.

References

[1] G. L. Warren, T. D. Do, B. P. Kelley, A. Nicholls, S. D. Warren, "Essential considerations for using protein–ligand structures in drug discovery", *Drug Discovery Today*, **2012**, *17*, 1270-1281.

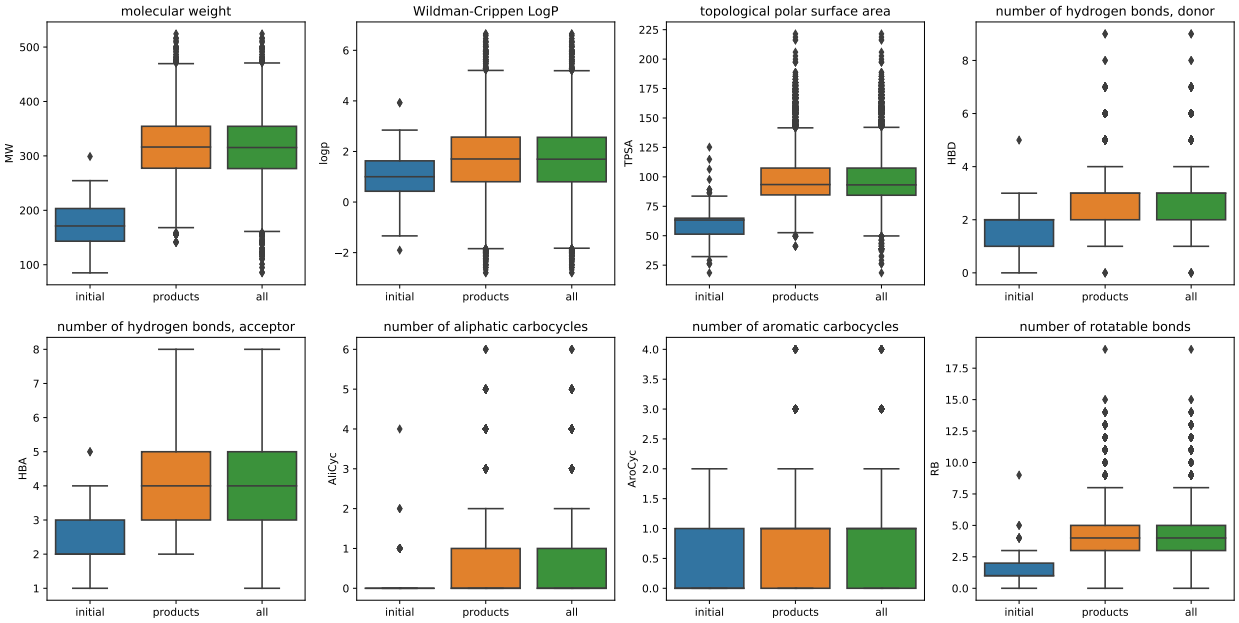

Figure 6: Chemical diversity in each set. The chosen metric includes molecular weight (MW), logP, topological polar surface area (TPSA) in $Å^2$, number of hydrogen-bond donors (HBD), number of hydrogen-bond acceptors (HBA), number of aliphatic carbocycles (AliCyc), number of aromatic carbocycles (AroCyc), and number of rotational bonds (RB).