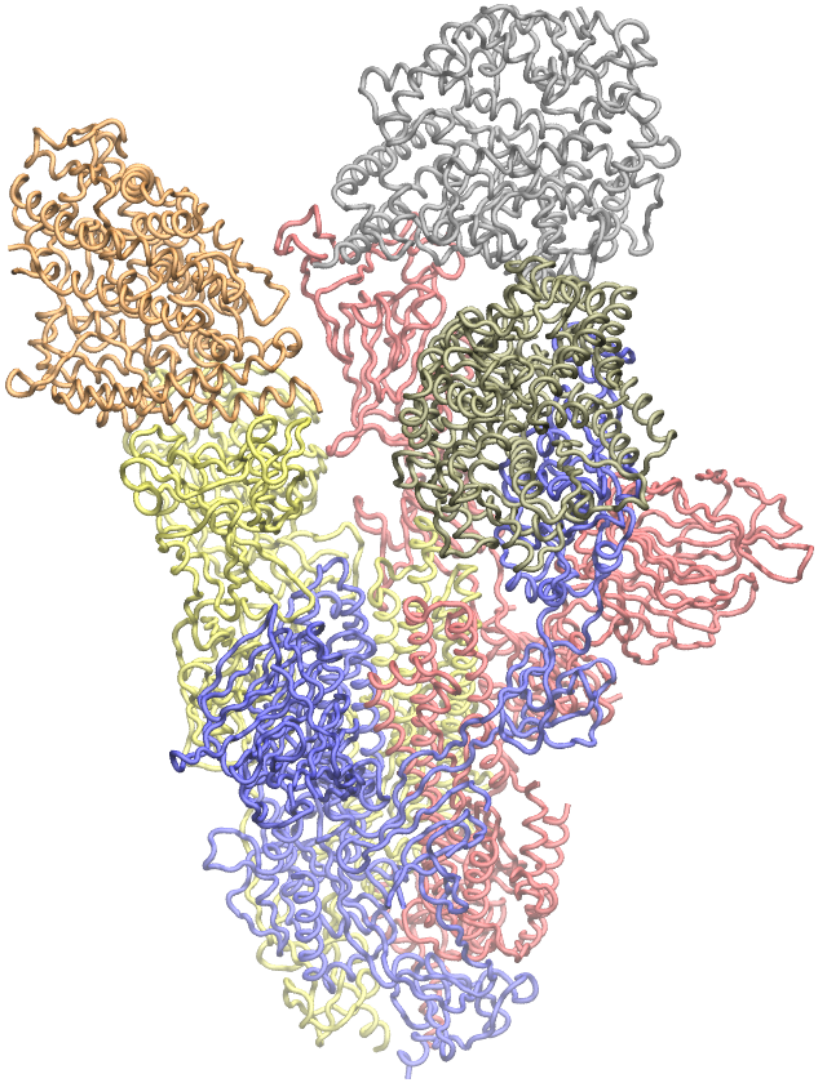
EPIMOL-2310-1: Fundamentals of Molecular Modelling

15 August 2021

Presented by Julia Liang and Eleni Pitsillou



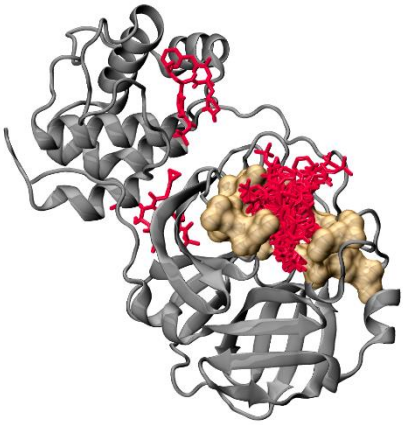
PDB ID: 7A98
Trimeric SARS-CoV-2 Spike protein with 3 ACE2 bound

EPIGENOMIC MEDICINE

EPIMOL-2310-1: Fundamentals of Molecular Modelling (August 15, 2021)

Instructors

Julia Liang



Julia is a senior student in the Epigenomic Medicine Laboratory with broad experience in molecular modelling. She completed her Honours in 2015, studying anti-pneumococcal effects of L-sulforaphane. Julia completed a Master of Science by Research in 2018, using molecular dynamics simulations to characterise anti-inflammatory action of dietary compounds. Currently she is undertaking a PhD, investigating molecular pathways and compounds associated with longevity.

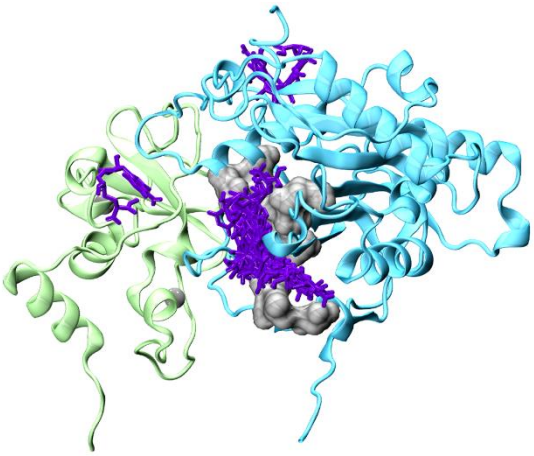
Publication highlights:

Liang, J. et al. Site mapping and small molecule blind docking reveal a possible target site on the SARS-CoV-2 main protease dimer interface. Computational Biology and Chemistry vol. 89 (2020): 107372.

Liang, J. et al. In silico characterisation of olive phenolic compounds as potential cyclooxygenase modulators. Part 1. Journal of Molecular Graphics & Modelling vol. 101 (2020): 107719.

Liang, J. et al. Investigation of potential anti-pneumococcal effects of l-sulforaphane and metabolites: Insights from synchrotron-FTIR microspectroscopy and molecular docking studies. Journal of Molecular Graphics & Modelling vol. 97 (2020): 107568.

Eleni Pitsillou



Eleni completed her Honours in 2019 and her project focused on identifying potential lead dietary compounds for proteins implicated in the pathophysiology of Major Depressive Disorder (MDD). She is now a Master of Science (Applied Chemistry) Candidate and is investigating the relationship between the circadian rhythm and its associated comorbidities.

Publication highlights:

Pitsillou, E., et al., Molecular docking utilising the OliveNetTM Library reveals novel phenolic compounds which may potentially target key proteins associated with major depressive disorder. Computational Biology and Chemistry, 2020. **86**: p. 107234.

Pitsillou, E., et al., The circadian machinery links metabolic disorders and depression: A review of pathways, proteins and potential pharmacological interventions. Life Sciences, 2021. **265**: p. 118809.

Pitsillou, E., et al., Identification of small molecule inhibitors of the deubiquitinating activity of the SARS-CoV-2 papain-like protease: in silico molecular docking studies and in vitro enzymatic activity assay. Frontiers in Chemistry, 2020. **8**.

EPIGENOMIC MEDICINE

EPIMOL-2310-1: Fundamentals of Molecular Modelling

Contents

Week 1: Introduction to Molecular Modelling

1.0 Introduction to Molecular Docking and Molecular Dynamics Simulations

Drugs elicit their effects on biological systems by binding to molecules, such as proteins, and this can directly affect their function. These interactions can also result in changes to signalling pathways and the expression of genes. Establishing the mechanisms of action of compounds is an important part of the drug discovery and drug development process, and this can be a challenging task. In addition to investigating novel drugs, drug repurposing also has its advantages. This involves taking existing drugs and identifying new therapeutic uses.

The drug development process is divided into different stages and it can take several years for a compound to be tested and approved. It has become common practice to initially examine large compound libraries against biological targets and determine whether there are any potential leads that can be evaluated further. *In silico* methods, which are also known as computational methods, can assist with this and chemical libraries that consist of a large number of compounds can be screened in a timely and cost-effective manner.

Molecular docking is an *in silico* tool that can be used to predict how strongly a compound binds to a particular site within a biological molecule of interest (ie. protein, DNA, RNA etc.). This process requires the three-dimensional (3D) structure of the drug target and molecular docking also allows for the interactions that are formed between the compound and the drug target to be predicted at the molecular level. In other words, we are able to gain further insight into the intermolecular bonds that are formed between the ligand and the protein residues, and properties of the amino acids in the binding site. The top scoring molecules can be evaluated further using more advanced *in silico* techniques, such as molecular dynamics (MD) simulations.

In this tutorial, we are focusing on rigid molecular docking and this means that biological molecules are treated as rigid objects, while the ligands are flexible – eg. the protein does not change shape during the docking process. With MD simulations, we are able to examine the conformational changes that occur in a protein-ligand complex over a period of time and can change the conditions of the system to mimic a physiological environment. We will discuss MD simulations in more detail at a later stage.

1.1 Protein Data Bank

The RCSB Protein Data Bank (PDB) is the database where the structures of proteins can be obtained (https://www.rcsb.org/) (1). If the structure of a protein has been made available, it will be assigned a PDB identification code (PDB ID). The PDB ID can be typed into the search bar and a page displaying information about the structure will appear. You can also type in the name of the protein of interest in the search bar to determine if the structure is available.

For this tutorial, we will be focusing on the structure of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) chimeric receptor-binding domain in complex with the human angiotensin-converting enzyme 2 (ACE2) receptor (Figure 1.1.1) (2). The PDB ID for this structure is 6VW1. To download the structure, go to "Download Files" > Select "PDB Format" > Press "Save".

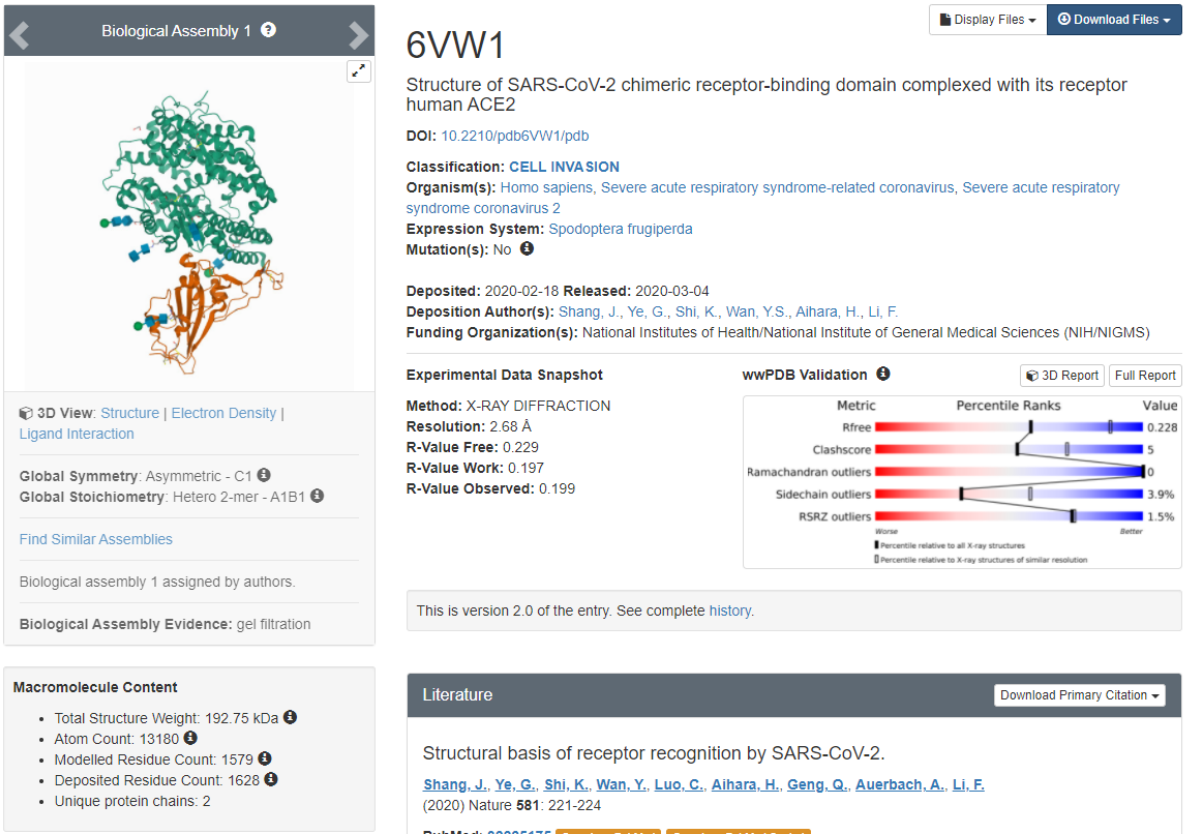


Figure 1.1.1: Structure Summary page that is displayed for the SARS-CoV-2 RBD in complex with the ectodomain of the human ACE2 receptor (PDB ID: 6VW1). At the top of the page, an overview is provided about the classification of the macromolecule, organisms the macromolecules are from, the expression system, whether there are mutations present in the structure, the dates that the structure was deposited and released, the authors, and the funding organisation. The experimental data snapshot section provides information about the quality of the structure. If available, the paper that is linked with the structure will be listed underneath.

Structure Summary

The SARS-CoV-2 spike protein attaches to the human ACE2 receptor and this interaction mediates viral entry into the cell. The 6VW1 crystal structure is comprised of the receptor binding domain (RBD) of the spike protein and the ectodomain of the human ACE2 receptor. We can see that the protein complex is comprised of components from Homo sapiens and severe acute respiratory syndrome coronavirus/severe acute respiratory syndrome coronavirus 2. This can be seen in the "Organism(s)" line. The "Expression System" is Spodoptera frugiperda and these insect cells are used for the production of proteins and complexes. We can also see that there are no mutations in the protein complex.

Biological Assembly and Asymmetric Unit

In the "Biological Assembly" section, the 3D structure of the macromolecule(s) will be displayed (1). An asymmetric unit is the smallest part of a crystal structure and is used to build the complete structure (1). An asymmetric unit may consist of one biological assembly, a portion of a biology assembly, or multiple biological assemblies (1). The biological assembly is believed to be the functional form of the molecule. It can be built from one copy of the asymmetric unit, multiple copies of the asymmetric unit, or a portion of the asymmetric unit (1).

Experimental Data Snapshot

Method

A number of techniques can be used to determine the structure of biological molecules:

- X-ray crystallography
    - This technique provides the detailed atomic information of proteins or nucleic acids, as well as any molecules (ie. ligands, ions) that may be present in the crystal (1). The resulting protein is purified and crystallised, and exposed to X-ray beams. The crystalline atoms cause the beam of X-rays to diffract and the pattern of spots that is produced can be analysed to determine the distribution of electrons (1, 3). The electron density can be used to generate a 3D picture of the structure and the accuracy depends on the quality of the crystals (1, 3).
- NMR spectroscopy

    - NMR spectroscopy can also be used to determine the structure of proteins. Once a protein has been purified, it is placed in a strong magnetic field and probed with radio waves (1, 4). The signal is produced by the excitation of the atomic nuclei and the observed resonances can be analysed to characterise the conformation of atoms (1, 4). This information can be used to build a model of the protein.
- Electron microscopy
    - This process utilises a beam of electrons and system of electron lenses to image the biomolecule (1).
    - Cryo-electron microscopy is a commonly used technique, which involves flash-freezing solutions of biomolecules, and the frozen hydrated samples can be visualised by transmission electron microscopy (TEM) (1, 5). The images generated can then be assembled to reconstruct the 3D structure of the molecule.

Resolution and R-value

In order to determine the accuracy of a structure, we can look at the resolution and R-value. The resolution is a measure of details that can be seen in the diffraction pattern and electron-density map (1). High resolution structures (resolution values that are small: 1Å or lower) are highly ordered and it is easier to see the location of atoms in these structures (1). In lower resolution structures (resolution values that are large: 3Å or higher), the basic contours of the protein shape are shown (1). The R-value measures how well the simulated/predicted diffraction pattern (calculated from the atomic model that is initially built by the researcher) matches the experimentally-observed diffraction pattern (evaluates the quality of the model) (1). The atomic model can be refined to make it fit the experimental data in a better way and improve the R-value (1). To minimise bias, 10% of the experimental observations are removed from the data set before refinement begins (1). The remaining 90% of experimental observations are then used for the refinement process (1). The R-free value is calculated by seeing how well the model predicts the 10% not used in refinement and for an ideal model, the R-free will be similar to the R-value (1).

Macromolecules and Ligands

Macromolecules

This section will provide details about the macromolecules in the structure. The name of each molecule will be provided and the representative protein chains will be listed. The sequence

length will show how many amino acids the protein consists of and the name of the organism that the protein is from will be provided. There may also be information about the number of mutations, gene names, and function of the protein. You may also be able to access the protein sequence on the Universal Protein Resource Knowledgebase (UniProtKB) by pressing the code that is next to "Go to UniProtKB" (6). By clicking on the accession code that is next to "Find proteins for", a list of the structures that have been made available on the RCSB PDB with the same amino acid sequence will be displayed.

For the 6VW1 crystal structure, entity 1 is the human ACE2 receptor (membrane protein). Chains A and B of the structure are the ACE2 receptor, there are 597 amino acids in the ACE2 ectodomain, and no mutations present (Figure 1.1.2). Entity 2 is the SARS-CoV-2 chimeric RBD (membrane protein). Chains E and F of the structure are the RBD, there are 217 amino acids in the RBD, and no mutations. In this example, there is also an *Oligosaccharides* section. Entities 3, 4, and 5 are the oligosaccharides and they are important for the interaction between the spike protein and ACE2 receptor.

Note: On the RCSB PDB page for the 6VW1 structure, the SARS-CoV-2 RBD chains are labelled as C[auth E], D[auth F]. This occurs when the two ID chains assigned by the PDB and the authors do not coincide. In this case, chains E and F have been selected by the author (in the crystal structure, chains E and F are the SARS-CoV-2 RBD). Chains C and D have been assigned by the PDB.

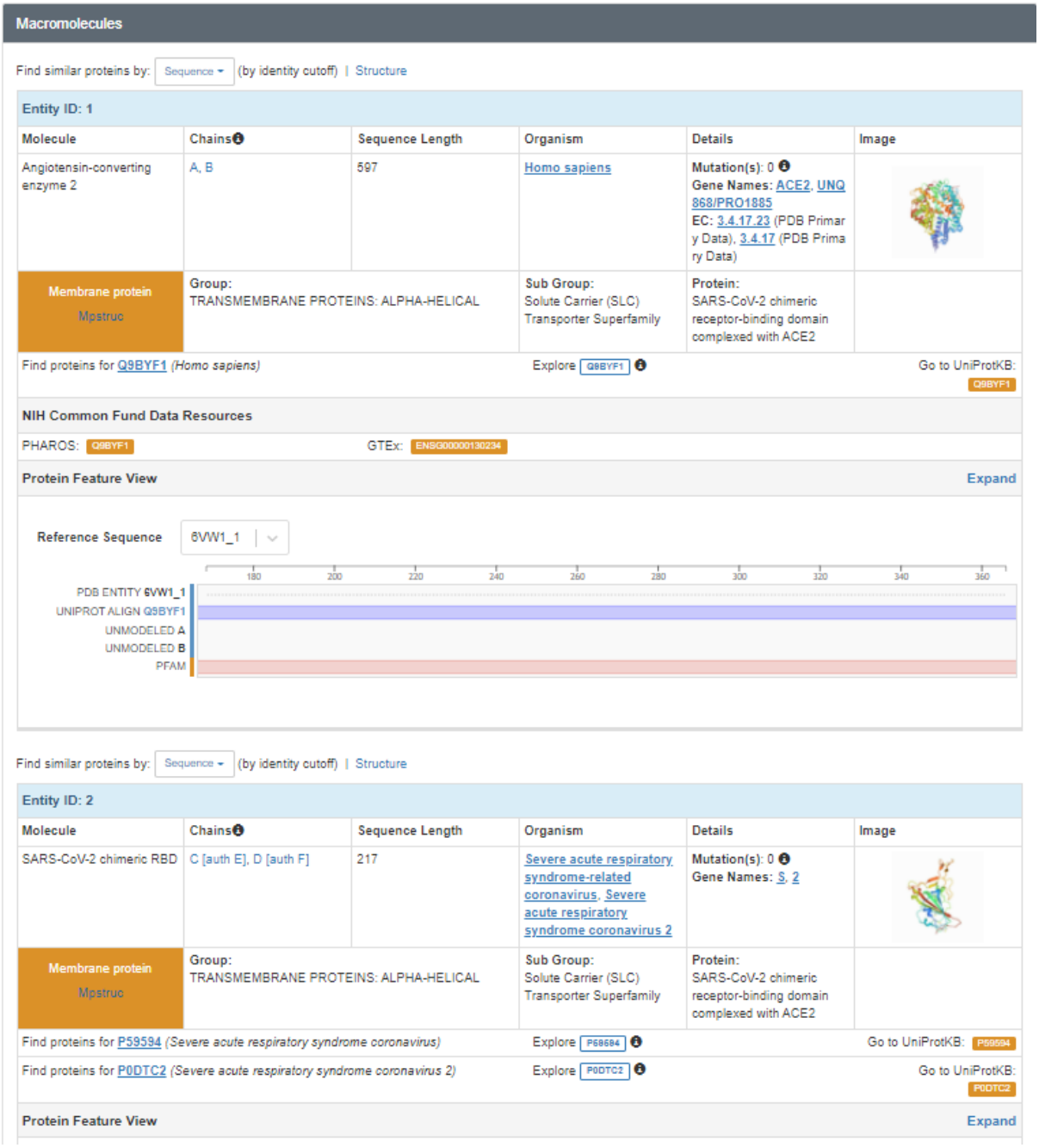


Figure 1.1.2: The Macromolecules section of the Structure Summary page is shown. The molecules that are in the biological assembly are listed and a description is provided for each entity. For the 6VW1 crystal structure, the molecules are the human ACE2 receptor and SARS-CoV-2 chimeric RBD.

Small Molecules

This section will list the ligands (ie. inhibitors, activators, cofactors) and other small molecules that are present within the relevant chains of the structure. The small molecules that are present for chains A and B of the 6VW1 crystal structure are 2-acetoamido-2-deoxy-beta-D-glucopyranose, zinc, chloride, and 1,2-Ethanediol (Figure 1.1.3).

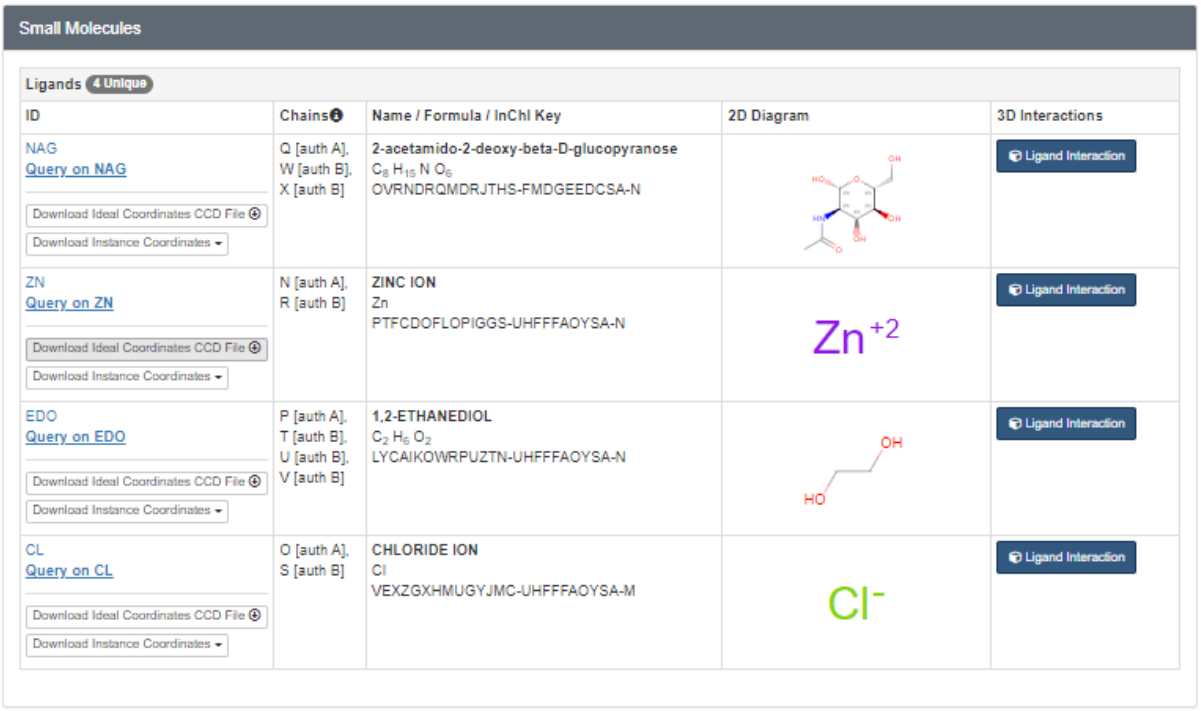


Figure 1.1.3: The Small Molecules section of the Structure Summary page is displayed. The ligands that may be present in the biological assembly are provided. In the complex that the SARS-CoV-2 chimeric RBD forms with the human ACE2 receptor (PDB ID: 6VW1), the co-crystallised ligands are 2-acetoamido-2-deoxy-beta-D-glucopyranose, zinc, chloride, and 1,2-ethanediol.

1.2 Ligand Databases

There are a number of ligand databases available and the chemical structures of compounds can be obtained from these libraries. This includes PubChem, ChEMBL and ZINC to name a few. In 2018, our lab developed the OliveNet^{TM} database and this is a curated library of 676 compounds from *Olea Europaea* (7-10). The compounds from these databases can be downloaded and can be used in the molecular docking process.

- https://pubchem.ncbi.nlm.nih.gov/
- https://www.ebi.ac.uk/chembl/

- https://zinc.docking.org/
- https://mccordresearch.com.au/

PDB, SDF and MOL Files

Ligand files can be downloaded from databases in different formats and this includes the PDB (Program Database), SDF (Standard Database Format) and MOL (Molfile) files. The chemical structure file formats contain structural information about the compound. For this tutorial, we will be downloading the structure of the anti-malarial drug chloroquine from the National Centre for Biotechnology Information (NCBI) PubChem Database (Figure 1.2.1).

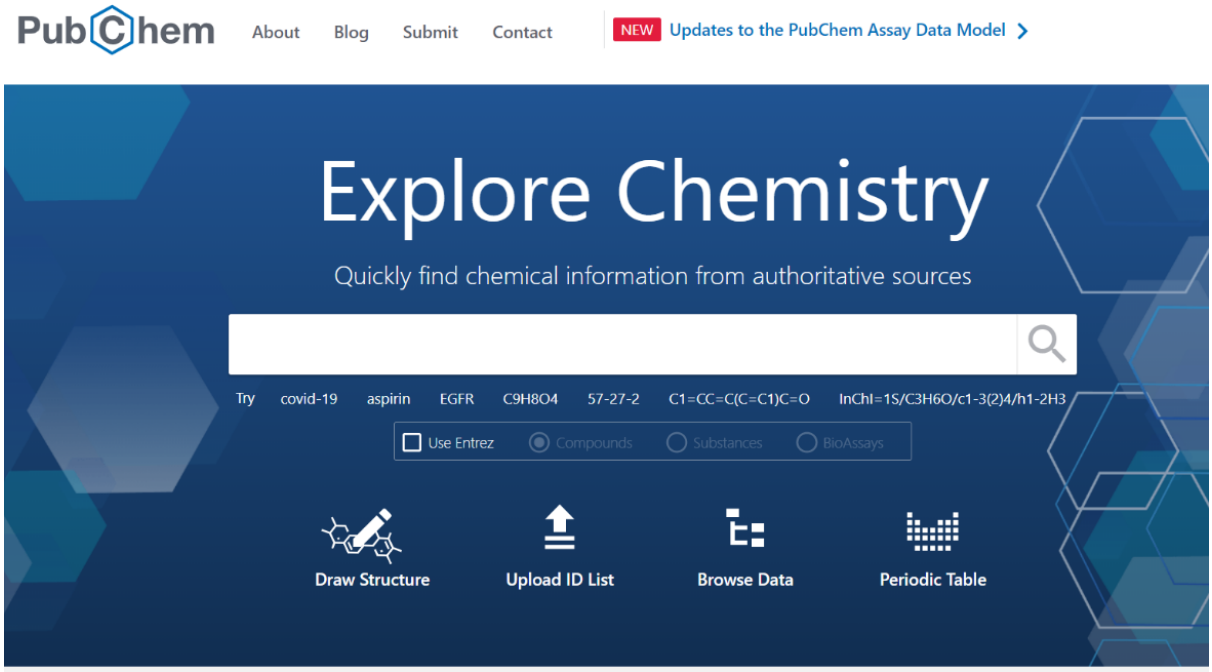


Figure 1.2.1: The home page of the NCBI PubChem Database is displayed. The search bar can be used to find the chemical structures of compounds of interest. Alternatively, the PubChem Database can be searched by drawing the structure of a compound.

In the search bar that appears on the home page of PubChem, the compound name can be entered. For chloroquine, the results demonstrate that there are 65 compounds that contain "chloroquine" within their name. The structure of chloroquine comes under the "Compound Best Match" section and when selected, the "Compound Summary" page will appear. Under the "Structures" section, the two-dimensional (2D) and 3D structures of the compound are provided. In order to download the files, click on the "Download" button and press "Save" for

the sdf option. As shown in Figure 1.2.2, additional information about the compound is provided and these details can accessed from the “Contents” list.

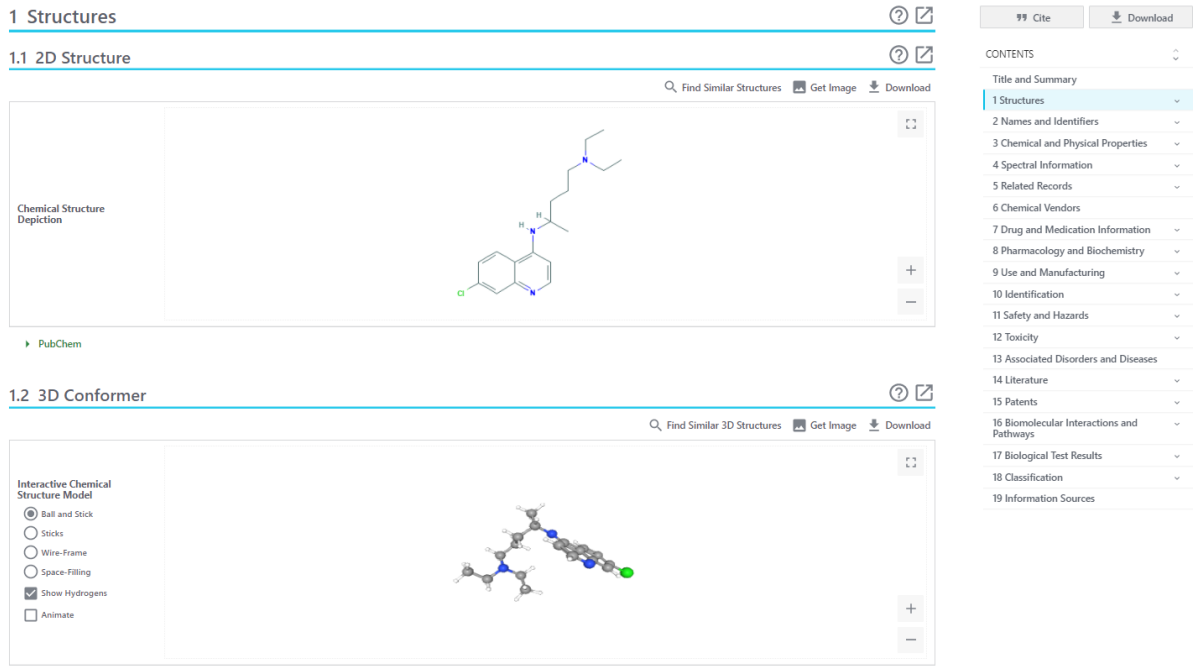


Figure 1.2.2: In the “Structures” section of the “Compound Summary” page, the 2D and 3D structures of the ligand can be obtained.

1.3 Freely Available Software

If the chemical structure of the compound is unavailable on PubChem, then other databases can be utilised. The chemical structure of the ligand can also be drawn in programs such as Chem3D (Perkin Elmer, Massachusetts, USA) or ChemSketch, which is a freely available program (11). In order to visualise the protein and ligand structures, freely available programs including PyMOL Academic, Maestro Academic and Visual Molecular Dynamics (VMD) can be downloaded and installed (12-14). Similarly, there are freely available molecular docking programs such as PyRx and AutoDock Vina (15, 16).

Week 2: Manipulating 3D Structures

2.0 Overview of Protein Structures

Proteins are complex macromolecules that play a crucial role in a number of biological processes. Protein synthesis occurs in cells in two main stages: transcription and translation. During transcription, deoxyribonucleic acid (DNA) is used as a template to make messenger RNA (mRNA) and this occurs in the nucleus of cells. This single-stranded mRNA molecule (pre-mRNA) is complementary to one of the DNA strands and it must undergo processing before leaving the nucleus (ie. 5' capping, 3' cleavage and polyadenylation, and RNA splicing).

The next stage is translation and this occurs in the cytoplasm. The mature mRNA leaves the nucleus and moves to the ribosome, which is the site of protein synthesis. With the help of transfer RNAs, the correct sequence of amino acids are brought to the ribosome and corresponding mRNA codons. Ribosomal RNAs (rRNAs) catalyse the formation of peptide bonds between amino acids and a polypeptide chain is produced.

The sequence of amino acids within a protein is called the primary structure. As seen in the table below, there are 20 types of amino acids and they exhibit different properties (Table 1). The secondary structure refers to the local folding patterns that form from the interactions that occur between atoms of the protein backbone. This includes alpha-helices (□-helix) and beta-pleated sheets (□-sheets). The overall 3D structure of a protein is called its tertiary structure, while the quaternary structure refers to a protein that is made up of many subunits (many polypeptide chains).

Table 1: The properties of the amino acids are provided, along with the residue codes.

Type		**Name**	**3 letter and 1 letter code**
Charged amino acids	Positive	Arginine Histidine Lysine	ARG (R) HIS (H) LYS (K)
	Negative	Aspartic acid Glutamic acid	ASP (D) GLU (E)
Polar		Serine Threonine Asparagine	SER (S) THR (T) ASN (N)

		Glutamine	GLN (Q)
Hydrophobic	Aliphatic	Alanine Valine Isoleucine Leucine Methionine	ALA (A) VAL (V) ILE (I) LEU (L) MET (M)
	Aromatic	Phenylalanine Tryptophan Tyrosine	PHE (F) TRP (W) TYR (Y)
Others		Cysteine Glycine Proline	CYS (C) GLY (G) PRO (P)

Modulation of Proteins by Small Molecules

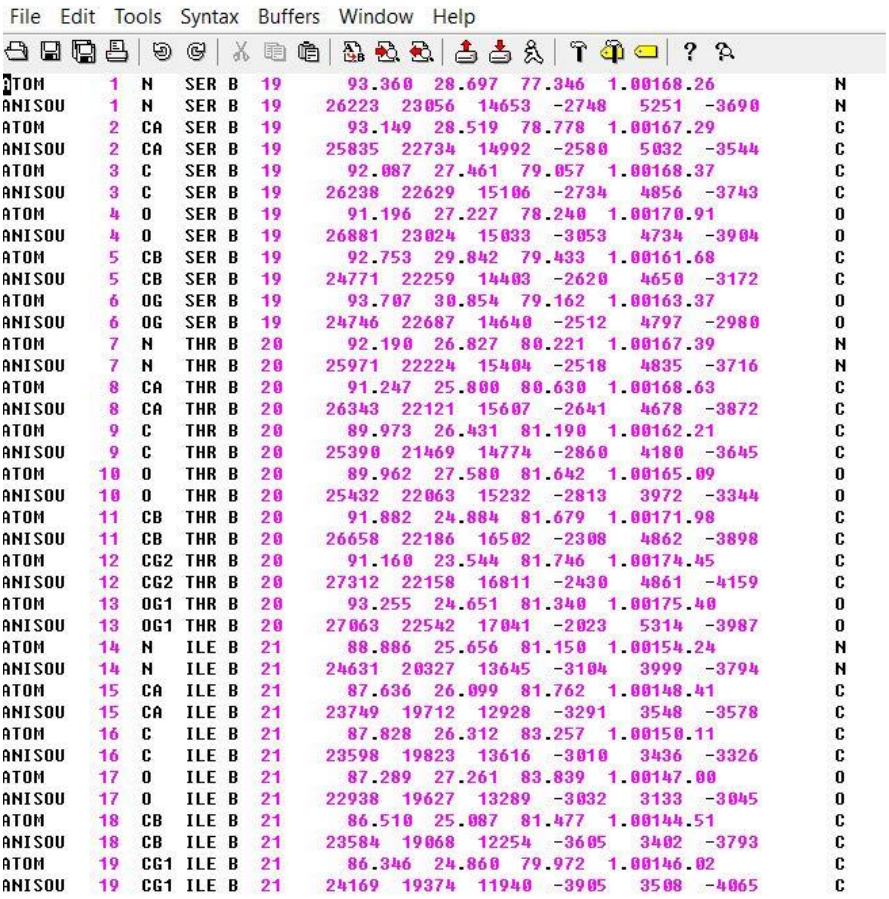
The function and activity of proteins can be regulated by small molecules and the ligand may interact with the protein in several ways:

- Active site
    - If a drug binds to the active site of an enzyme, the shape of this region may change and the natural substrate may no longer be able to bind. The inhibitor could also interact with the cofactors within the active site, such as a metallic ion, and disrupt the catalytic activity of the enzyme.
    - It is important to note that not all molecules that bind to proteins are inhibitors; there are activators as well.
- Allosteric
    - If a drug binds to a protein at a site other than the active site, then this interaction is known as allosteric regulation.
- Protein-protein
    - There are also inhibitors of protein-protein interactions. For example, peptide inhibitors are being developed to target the interaction between the SARS-CoV-2 spike protein and human ACE2 receptor.

2.1 PDB Files

The structural information of a protein that has been downloaded from the RCSB Protein Data Bank can be examined using a text editor such as Notepad, Wordpad, or Vim (Figure 2.1.1). In the PDB file, the following details are provided:

- Record – ATOM, HETATM
    - ATOM refers to the standard residues of a protein whereas HETATM applies to non-standard residues (ie. ligands, cofactors, ions, and solvent)
- Atom number and atom type
    - Each atom is designated a number and the properties of each atom are provided (1, 2, 3, 4, 5, 6… and N, CA, C, O, CB…)
- Residue
    - The residue that each atom corresponds to is listed (SER, GLY, PHE)
- Chain
    - The chain name is provided next to the corresponding residues – if a protein consists of more than one chain then they may be labelled A, B etc. All the amino acids belonging to chain A will be labelled “Residue name A” and all of the amino acids belonging to chain B will be labelled “Residue name B”
- Residue number
    - Each residue is assigned a number
- XYZ coordinates
    - The atomic coordinates are provided
- Occupancy
    - The fraction of atoms that appear at that location
- Beta factor
    - Average displacement of the atoms
    - < 10 – very sharp model, atoms not moving
    - > 50 – atoms moving, potentially flexible region



File Edit Tools Syntax Buffers Window Help

```
ATOM      1  N   SER B  19      93.360  28.697  77.346  1.00168.26           N
ANISOU    1  N   SER B  19   26223  23056  14653  -2748   5251  -3690       N
ATOM      2  CA  SER B  19      93.149  28.519  78.778  1.00167.29           C
ANISOU    2  CA  SER B  19   25835  22734  14992  -2580   5032  -3544       C
ATOM      3  C   SER B  19      92.087  27.461  79.057  1.00168.37           C
ANISOU    3  C   SER B  19   26238  22629  15106  -2734   4856  -3743       C
ATOM      4  O   SER B  19      91.196  27.227  78.240  1.00170.91           O
ANISOU    4  O   SER B  19   26881  23024  15033  -3053   4734  -3904       O
ATOM      5  CB  SER B  19      92.753  29.842  79.433  1.00161.68           C
ANISOU    5  CB  SER B  19   24771  22259  14403  -2620   4650  -3172       C
ATOM      6  OG  SER B  19      93.707  30.854  79.162  1.00163.37           O
ANISOU    6  OG  SER B  19   24746  22687  14640  -2512   4797  -2980       O
ATOM      7  N   THR B  20      92.190  26.827  80.221  1.00167.39           N
ANISOU    7  N   THR B  20   25971  22224  15404  -2518   4835  -3716       N
ATOM      8  CA  THR B  20      91.247  25.800  80.630  1.00168.63           C
ANISOU    8  CA  THR B  20   26343  22121  15607  -2641   4678  -3872       C
ATOM      9  C   THR B  20      89.973  26.431  81.190  1.00162.21           C
ANISOU    9  C   THR B  20   25390  21469  14774  -2860   4180  -3645       C
ATOM     10  O   THR B  20      89.962  27.580  81.642  1.00165.09           O
ANISOU   10  O   THR B  20   25432  22063  15232  -2813   3972  -3344       O
ATOM     11  CB  THR B  20      91.882  24.884  81.679  1.00171.98           C
ANISOU   11  CB  THR B  20   26658  22186  16502  -2308   4862  -3898       C
ATOM     12  CG2 THR B  20      91.160  23.544  81.746  1.00174.45           C
ANISOU   12  CG2 THR B  20   27312  22158  16811  -2430   4861  -4159       C
ATOM     13  OG1 THR B  20      93.255  24.651  81.340  1.00175.40           O
ANISOU   13  OG1 THR B  20   27063  22542  17041  -2023   5314  -3987       O
ATOM     14  N   ILE B  21      88.886  25.656  81.150  1.00154.24           N
ANISOU   14  N   ILE B  21   24631  20327  13645  -3104   3999  -3794       N
ATOM     15  CA  ILE B  21      87.636  26.099  81.762  1.00148.41           C
ANISOU   15  CA  ILE B  21   23749  19712  12928  -3291   3548  -3578       C
ATOM     16  C   ILE B  21      87.828  26.312  83.257  1.00150.11           C
ANISOU   16  C   ILE B  21   23598  19823  13616  -3010   3436  -3326       C
ATOM     17  O   ILE B  21      87.289  27.261  83.839  1.00147.00           O
ANISOU   17  O   ILE B  21   22938  19627  13289  -3032   3133  -3045       O
ATOM     18  CB  ILE B  21      86.510  25.087  81.477  1.00144.51           C
ANISOU   18  CB  ILE B  21   23584  19068  12254  -3605   3402  -3793       C
ATOM     19  CG1 ILE B  21      86.346  24.860  79.972  1.00146.02           C
ANISOU   19  CG1 ILE B  21   24169  19374  11940  -3905   3508  -4065       C
```

Figure 2.1.1: The PDB file of the SARS-CoV-2 RBD-ACE2 ectodomain complex (PDB ID: 6VW1) was opened using the Vim text editor and the structural information is displayed.

2.2 Visualising and Manipulating Molecules

Once the structure of a protein has been downloaded from the RCSB PDB, it can be modified and visualised using different programs.

PyMOL

Open up the PyMOL program and two windows should appear. The smaller window contains the menu bar, shortcut buttons, and the command line. The second window is the PyMOL Viewer and this is where the 3D models are displayed. In the smaller window, go to “File” > “Open” > Select the pdb file in the relevant folder where it was saved > Press “Open”. We will be using the crystal structure of the spike protein in complex with the ACE2 receptor (PDB ID: 6VW1) and the default settings display the protein in line representation.

The 6VW1 crystal structure is comprised of multiple chains. Based on the “Structure Summary” page on the RCSB PDB, chains A and B are the ACE2 receptor. Chains E and F

are the SARS-CoV-2 spike protein RBD. For this example, we are interested in the ACE2 receptor and we will select chain B (Figure 2.2.1). To select the chains of interest, go to "Display" > Select "Sequence" (ensure that this option is ticked) > "Sequence Mode" > "Chain Identifiers". To save a single chain or multiple chains, make sure the chains of interest are selected and go to "File" > "Save Molecule" > Choose "sele" > Press "Ok" > Access the relevant folder where you want to save the structure > Press "Save".

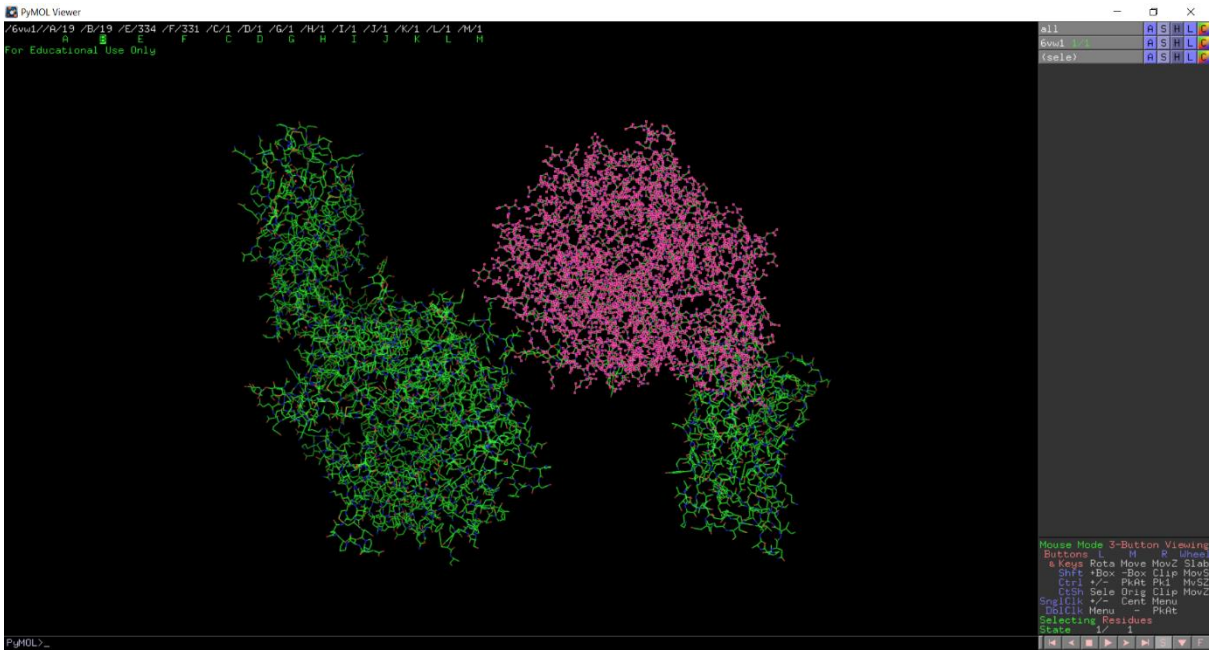


Figure 2.2.1: The SARS-CoV-2 spike protein RBD in complex with the human ACE2 receptor is shown in the PyMOL viewer. Chain B, which is ACE2, is selected and is coloured red in the complex (PDB ID: 6VW1).

Go to "File" > "Reinitialize". A blank window will appear. Go to "File" > "Open" > Access the relevant folder where chain B of the crystal structure was saved > Select "Open". To check if the structure has any water molecules or co-crystallised ligands, go to "Display" > "Sequence" > "Sequence Mode" > "Residue Codes". If you move to the end of the sequence, you will see that chain B consists of ZN, CL, EDO, NAG and water molecules (O).

To remove the water molecules, go to the "Action" button (A) in the Object Control Panel on the right-hand side of the PyMOL Viewer and select "remove waters". We can also remove the NAG and EDO molecules from the crystal structure. Select these ligands in the sequence and a new line that is named "(sele)" should appear in the Object Control Panel. Go to the "Action" button in the 'sele' line, and press "remove atoms". To save the modified structure, go to "File" > "Save Molecule" > Rename the file (eg. 6VW1chainBmodified) > Press "Save".

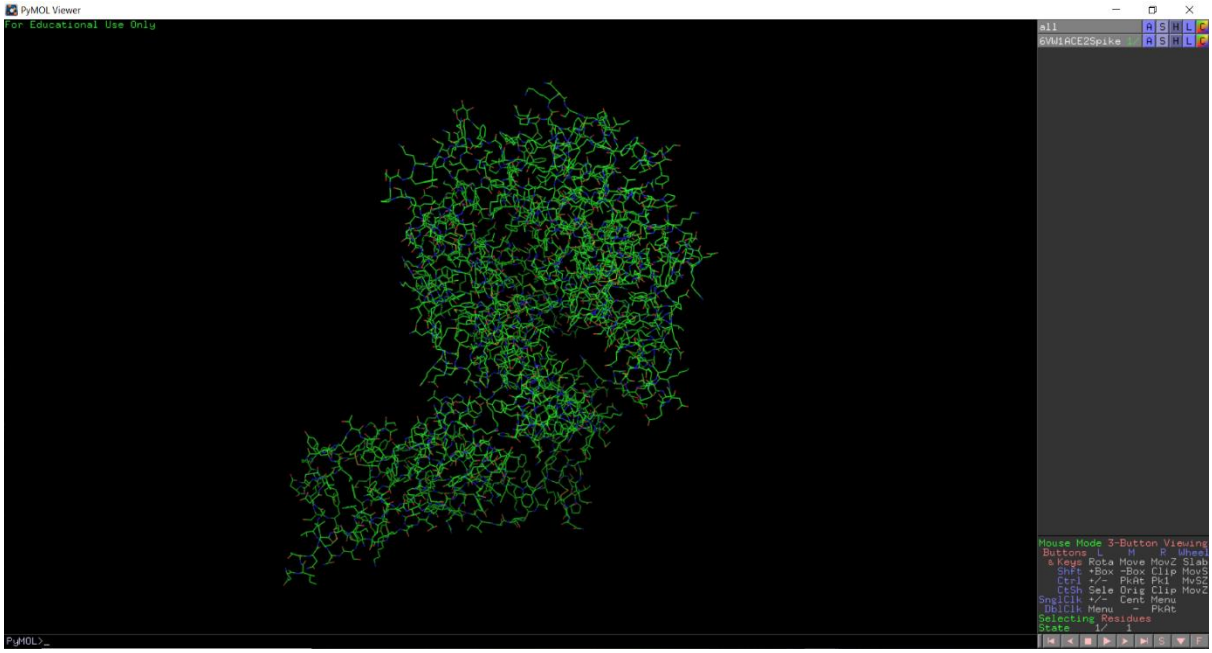


Figure 2.2.2: Chains B and F have been selected from the crystal structure of the SARS-CoV-2 spike protein RBD in complex with the human ACE2 receptor (PDB ID: 6VW1). The default settings have been used and the complex appears in "line" representation.

In this example, the RBD of the spike protein and the ACE2 ectodomain can be seen and the structure is depicted in line representation (PDB ID: 6VW1) (Figure 2.2.2). Chains B and F have been isolated from the original crystal structure and were selected for use in this section. Chain B is the ACE2 ectodomain, while chain F is the SARS-CoV-2 spike RBD. To display the sequence of amino acids, go to "Display" > "Sequence". If the "Sequence Mode" option is selected, you can see that there are different options to display the components of the protein (Residue Codes, Residue Names, Chain Identifiers, Atom Names and States). In the "Chain Identifiers" mode, you can see that the protein is comprised of two chains.

If chain B is selected, then all of the residues that are part of chain B are highlighted. To change the display of the selected molecule, go to the Object Control Panel and press the "Show" button (S). We are using the "cartoon" mode in this tutorial. To hide the lines, which is the default representation, select "lines" from drop-down menu that appears when pressing the "Hide" (H) button > To colour the chain, select the "Color" button. The same settings can be applied to chain F (Figure 2.2.3).

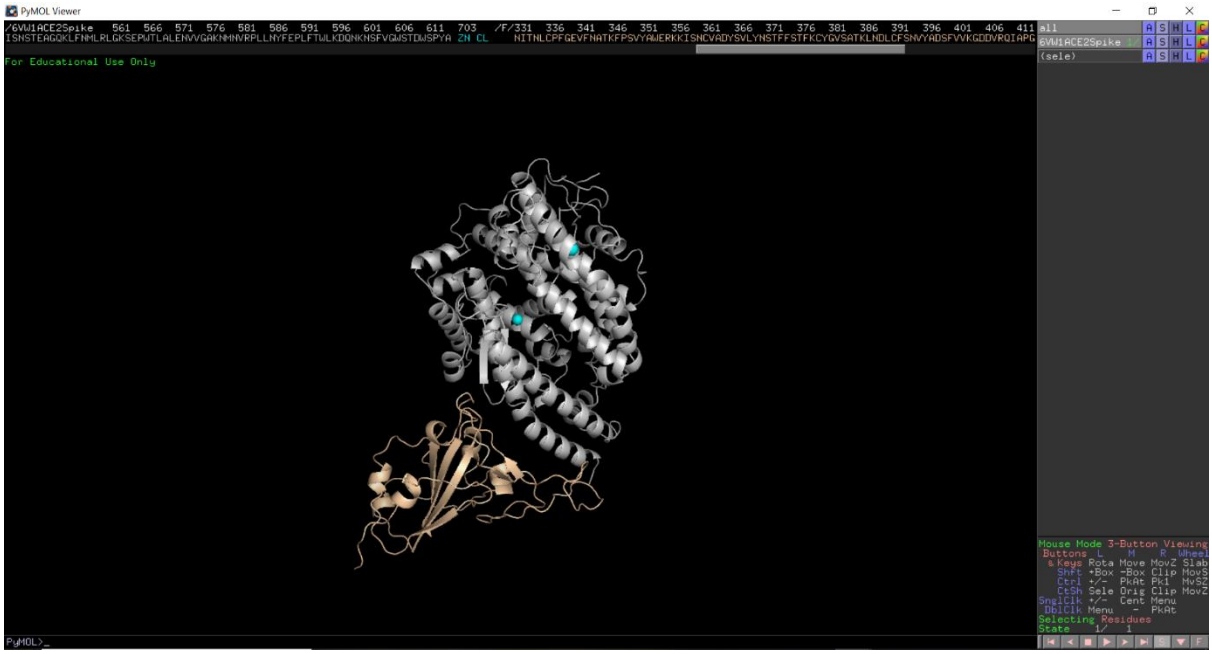


Figure 2.2.3: The SARS-CoV-2 spike protein in complex with the human ACE2 receptor is shown. The spike protein is coloured light brown and the ACE2 component is coloured silver. Both molecules are depicted in cartoon representation. The zinc and chloride ions can be seen as spheres and are coloured light blue.

If there are small molecules (ie. cofactors, ions, ligands) in the protein structure, they can also be coloured differently to the main protein chains and shown in a different style. The ACE2 receptor has ions and they can be displayed as spheres. Go to “Display” > “Sequence Mode” > “Residue Codes”. Select ZN and CL in the sequence. Go to the “Show” (S) button in the Object Control Panel and select “spheres”. To colour the spheres differently, select the “Color” button (Figure 2.2.3).

Manipulating and Visualising Crystal Structures With Ligands Present

The crystal structure of ACE2 in complex with the inhibitor (S,S)-2-{1-Carboxy-2-[3-(3,5-dichloro-benzyl)-3H-imidazol-4-yl]-ethylamino}-4-methyl-pentanoic acid will be used (PDB ID: 1R4L) for this section (Figure 2.2.4). Open the structure in PyMOL and go to “Display” > “Sequence”. To isolate chain A, which is the ACE2 receptor, go to “Display” > “Sequence Mode” > “Chain Identifiers”. Select and save chain A. Go to “File” > “Reinitialize” > Open the saved structure of chain A. Remove the water and NAG molecules from the structure.

To prepare the crystal structure for molecular docking, we need to firstly save the co-crystallised inhibitor as a ligand. Select the XX5 ligand in the sequence and go to “File” > “Save Molecule” > Choose the ‘sele’ option > Press “Ok” > Access the relevant folder where you want to save the structure > Press “Save”. For molecular docking, we need the protein

structure without the co-crystallised inhibitor present. To remove XX5, the ligand needs to be highlighted in the sequence. Go to the 'sele' line in the Object Control Panel > Press the "Action" (A) button > Choose "remove atoms" from the drop-down menu.

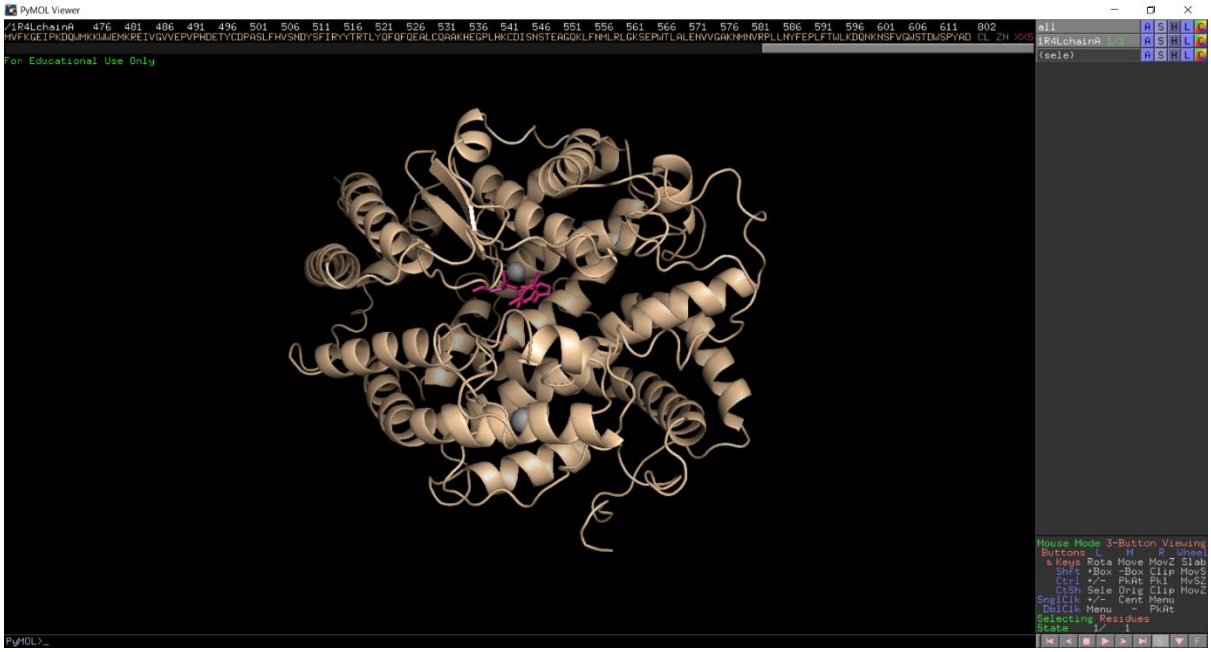


Figure 2.2.4: The ACE2 receptor in complex with the inhibitor (S,S)-2-{1-Carboxy-2-[3-(3,5-dichloro-benzyl)-3H-imidazol-4-yl]-ethylamino}-4-methyl-pentanoic acid can be seen. ACE2 is coloured light brown and the inhibitor is coloured pink.

Note: To create figures with the co-crystallised inhibitor present, you can represent the ACE2 receptor chain in "cartoon" format, the ions as "spheres", and the XX5 ligand (inhibitor) as "sticks". The colours can also be changed.

Importing and Visualising Ligands

To visualise the structure of a ligand in PyMOL, in this case the co-crystallised inhibitor that was isolated from the ACE2 receptor, go to "File" > "Open" > Access the relevant folder where you have saved the structure > Press "Open". To change the representation of the ligand, press the "Show" (S) button in the Object Control Panel and you can select a format from the drop-down menu (Figure 2.2.5).

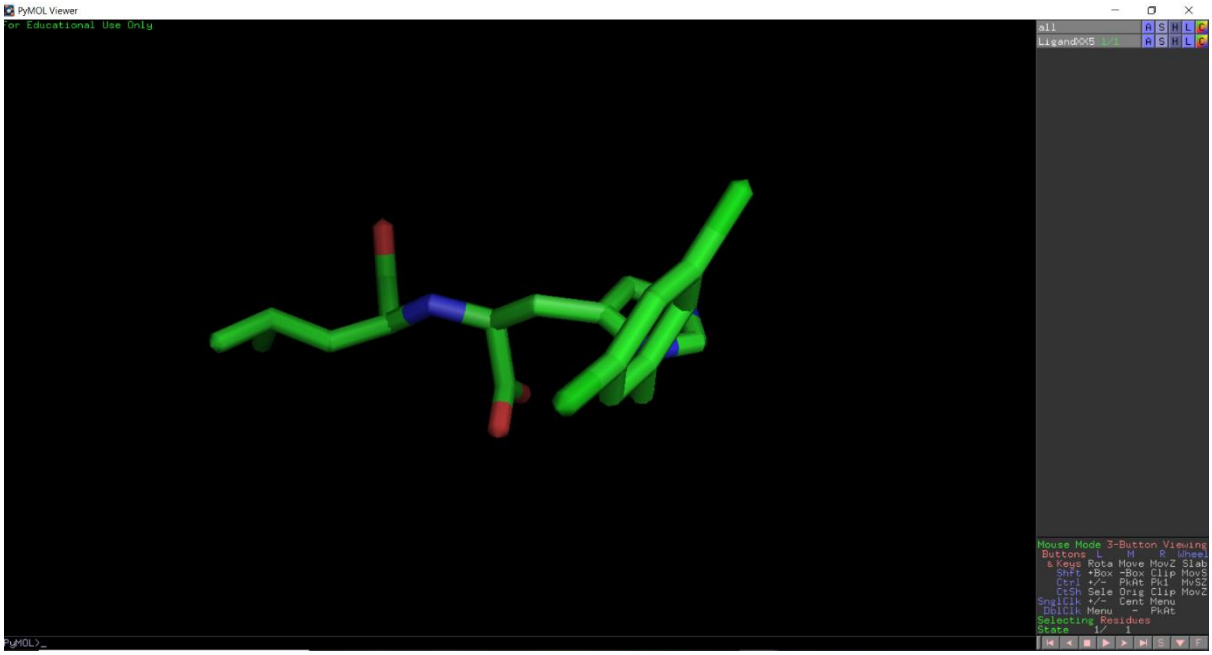


Figure 2.2.5: Chemical structure of (S,S)-2-{1-Carboxy-2-[3-(3,5-dichloro-benzyl)-3H-imidazol-4-yl]-ethylamino}-4-methyl-pentanoic acid. This compound is an inhibitor of ACE2 (PDB ID: 1R4L).

Selecting Residues of Interest

In PyMOL, you are also able to highlight and show residues that may be of interest in the protein (Figure 2.2.6). For this example, the interface residues of ACE2 will be selected. Go to "Display" > "Sequence" > Select the residues His34, Glu35, Glu37, Asp38, Leu39, and Tyr41. To represent these residues as "sticks", press the "Show" (S) button in the 'sele' line of the Object Control Panel. The colour can be changed by selecting the different options from the drop-down menu that appears when pressing the "Color" (C) button.

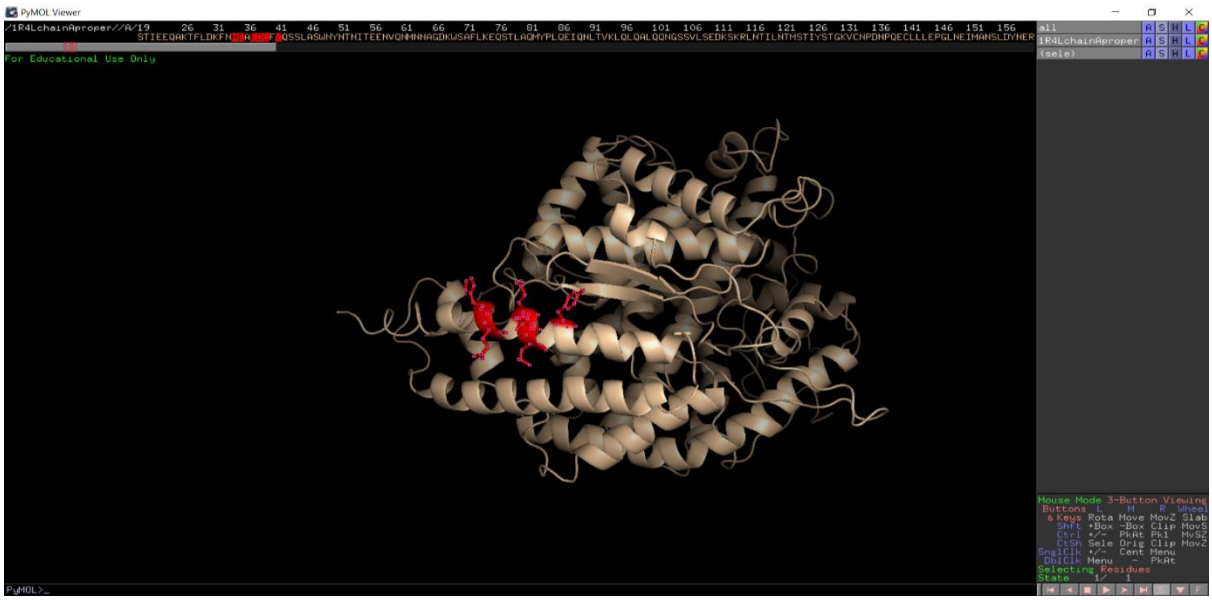


Figure 2.2.6: The ACE2 receptor in complex with the co-crystallised inhibitor is shown (PDB ID: 1R4L). Several residues that are in the target binding site have been selected in the sequence and are depicted in red.

Visual Molecular Dynamics (VMD)

Structures With Multiple Chains (eg. PDB ID: 6VW1)

Using PyMOL, chains B and F of the 6VW1 crystal structure have been selected, the water molecules and ligands have been removed, and the complex has been saved as a .pdb file (Figure 2.2.7). To visualise the structure in VMD, open the VMD program and go to "File" > "New Molecule" > Select "Browse" in the Molecule File Browser > Access the relevant folder where you have saved the modified structure > Press "Load".

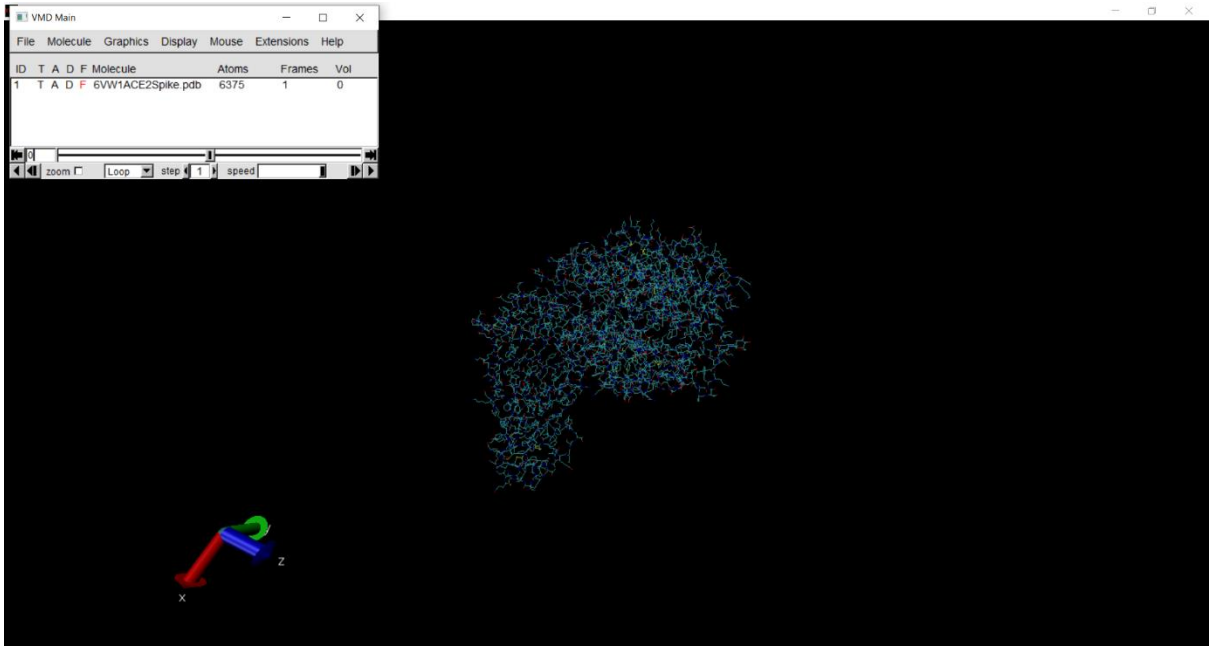


Figure 2.2.7: The SARS-CoV-2 spike protein RBD (chain B) and human ACE2 receptor (chain F) that was isolated from the 6VW1 crystal structure using PyMOL was imported into VMD. The default settings display the complex as "lines", the background is black, and the axes is present.

To change the colour of the display background, go to "Graphics" > "Colors" > "Display" > "Background". For this example, we will be using a white background and you can select "White" from the menu. To remove the axes from the screen, go to "Display" > "Axes" > "Off" (Figure 2.2.8).

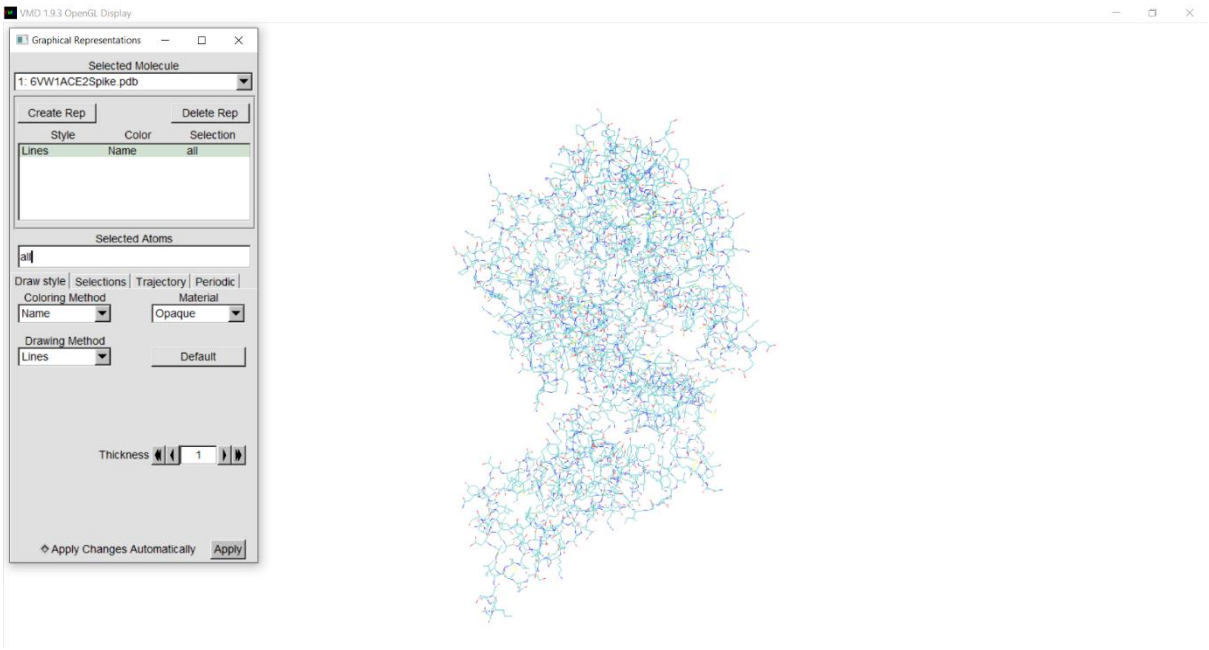


Figure 2.2.8: The background of the VMD Display has been changed to white and the Graphical Representations window has been opened. Using the drop-down menus in this window, the model can be displayed in various ways.

To change the representation of the protein complex, go to “Graphics” > “Representations”. When a structure is imported into VMD, the default drawing method is “Lines”. Under the “Drawing Method” section, there is a drop-down menu and the protein can be represented in various ways depending on the option selected. If the “NewCartoon” option is selected from the “Drawing Method” drop-down menu, then the □-helices and □-pleated sheets become apparent. In the “Coloring Method” section, select “Color ID” and a menu containing a list of colours will appear. You can also change the material (ie. how shiny or transparent) that is used for the molecular model via the “Material” button (Figure 2.2.9).

To show the Zn^{2+} and Cl^{-} ions in the crystal structure, press the “Create Rep” button > In the “Selected Atoms” section, type in “resname ZN”. Go to the “Drawing Method” drop-down menu and select “VDW”. The zinc ion should appear as a sphere. To display the chloride ion, press the “Create Rep” button > In the “Selected Atoms” section, type in “resname CL”. You can change the style and colour of the ions (Figure 2.2.9).

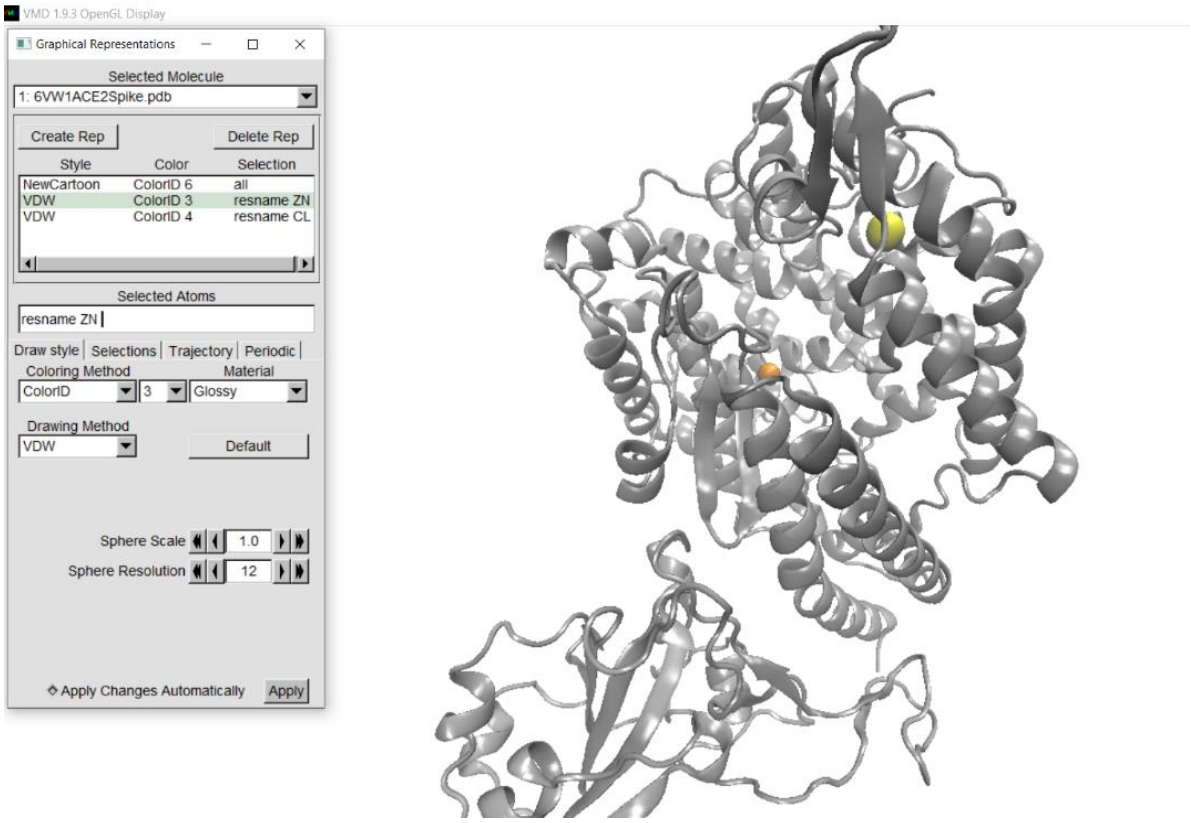


Figure 2.2.9: The "Drawing Method" for the SARS-CoV-2 spike protein RBD and ACE2 receptor have been changed to "NewCartoon" and are coloured silver. The ions are shown as spheres by choosing the "VDW" option in the "Drawing Method" drop-down menu and are also coloured differently to the main protein chain.

To differentiate between the ACE2 receptor and the RBD of the spike protein, we can specify the relevant chains (Figure 2.2.10). In the "Selected Atoms" section of the Graphical Representation box, type in "Chain B". Press the "Create Rep" button and type in "Chain F" in the "Selected Atoms" section. You can also display the ions as previously described.

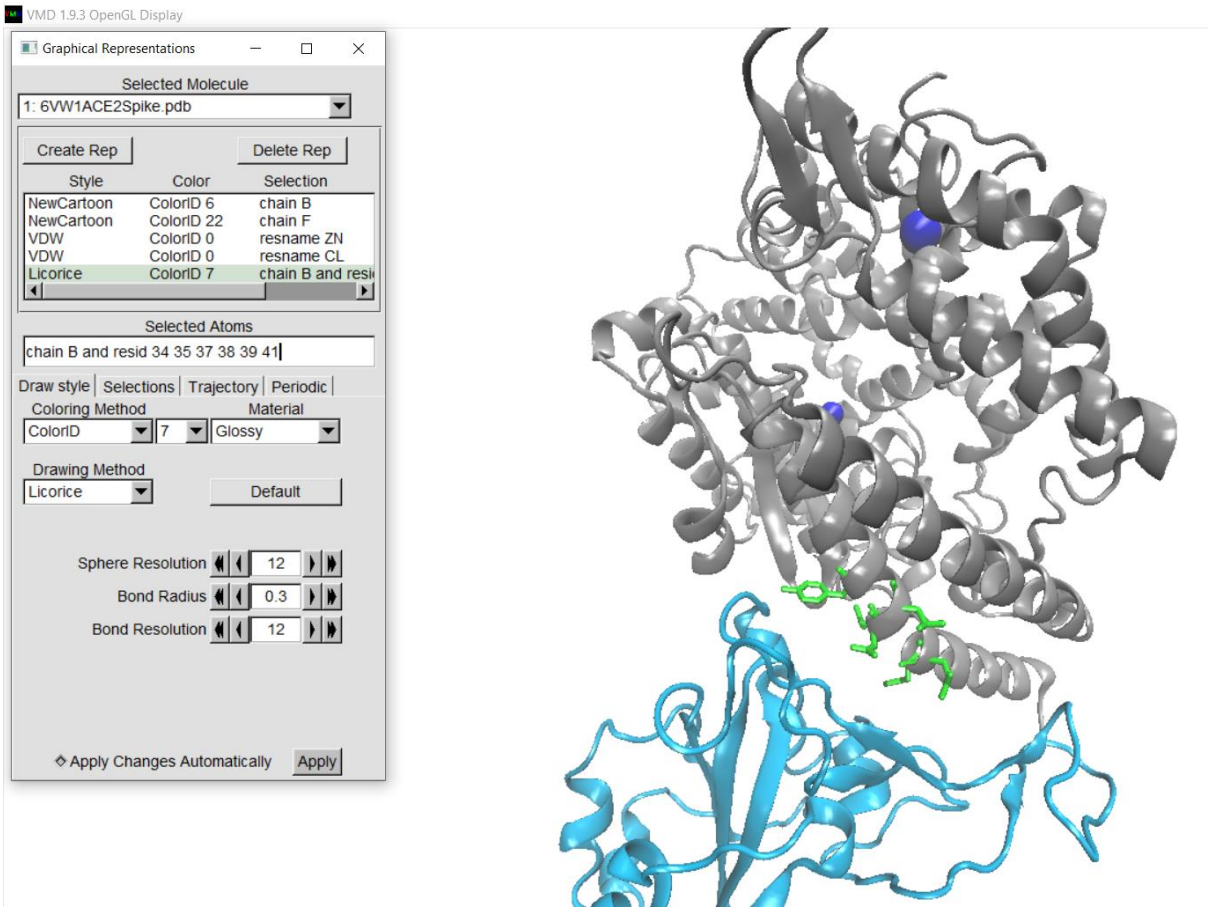


Figure 2.2.10: Key residues of the ACE2 receptor (chain B) have been specified in the "Selected Atoms" section of the "Graphical Representations" window and are coloured green. The "Drawing Method" that was chosen for the residues was "licorice".

To highlight certain residues of interest, press the "Create Rep" button and type in the residue numbers in the "Selected Atoms" section ie. "chain B and resid 34 35 37 38 39 41". The residues have been coloured green and are displayed in "licorice" mode (Figure 2.2.10).

Protein-Ligand Complexes

If a protein were to have a co-crystallised ligand in the structure, the small molecule can be displayed by typing in "resname" followed by the name of the ligand ie. "resname XX5". XX5 is the co-crystallised inhibitor in the 1R4L structure of ACE2. XX5 is the ligand ID and this can also be found on the RCSB PDB page. The chemical structures of ligands can also be visualised in VMD and can be imported as new molecules. Go to "File" > "New Molecule" > Select "Browse" in the "Molecule File Browser" > Access the relevant folder where you have saved the ligand > Press "Load". The ligand structure will appear in the display window and its representation can be modified by using the "Graphical Representations" tool (Figure 2.2.11).

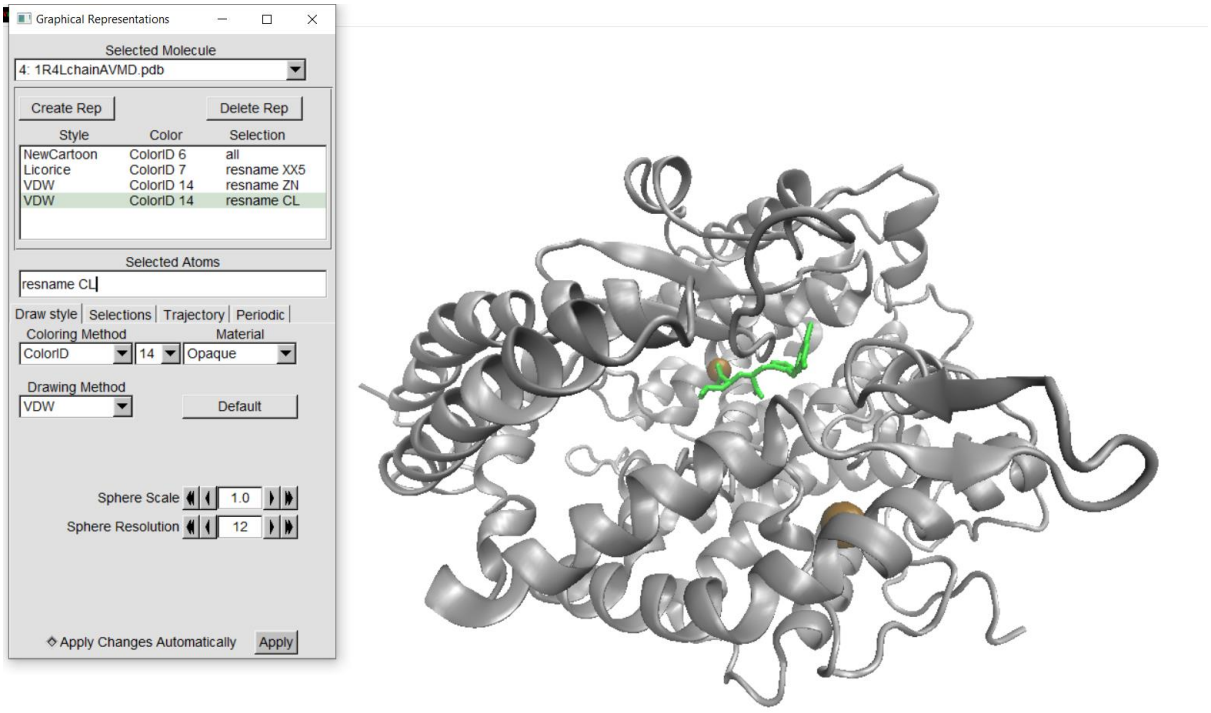


Figure 2.2.11: The human ACE2 receptor with the co-crystallised ligand (S,S)-2-{1-Carboxy-2-[3-(3,5-dichloro-benzyl)-3H-imidazol-4-yl]-ethylamino}-4-methyl-pentanoic acid is shown. The inhibitor is coloured green, while the protein chain is coloured silver.

To save the structure in the display window as an image, go to “File” > “Render” > Select the “Tachyon (internal in-memory rendering)” option from the “Render the current scene using:” drop-down menu. Press “Browse” next to the “Filename” option and select the folder that you want the image to be saved in. Name the image and add .bmp (ie. structure.bmp) > Press “Save”.

You can also save the VMD window as a visualisation state that can be accessed at a later time. Go to “File” > “Save Visualization State” > Select the folder that you want the workflow to be saved in > Name the file and for the “Save as type” option, select “.vmd” > Press “Save”. To load the visualisation state at a later time, go to “File” > “Load Visualization State” > Select the .vmd file in the folder that it was saved in.

Maestro – Schrödinger Suite

Open Maestro and the 6VW1 crystal structure that was modified in PyMOL can be imported (consists of chain B and chain F for this example) (Figure 2.2.12). Go to “File” > “Import” > Access the relevant folder where you have saved the modified structure > Press “Open”.

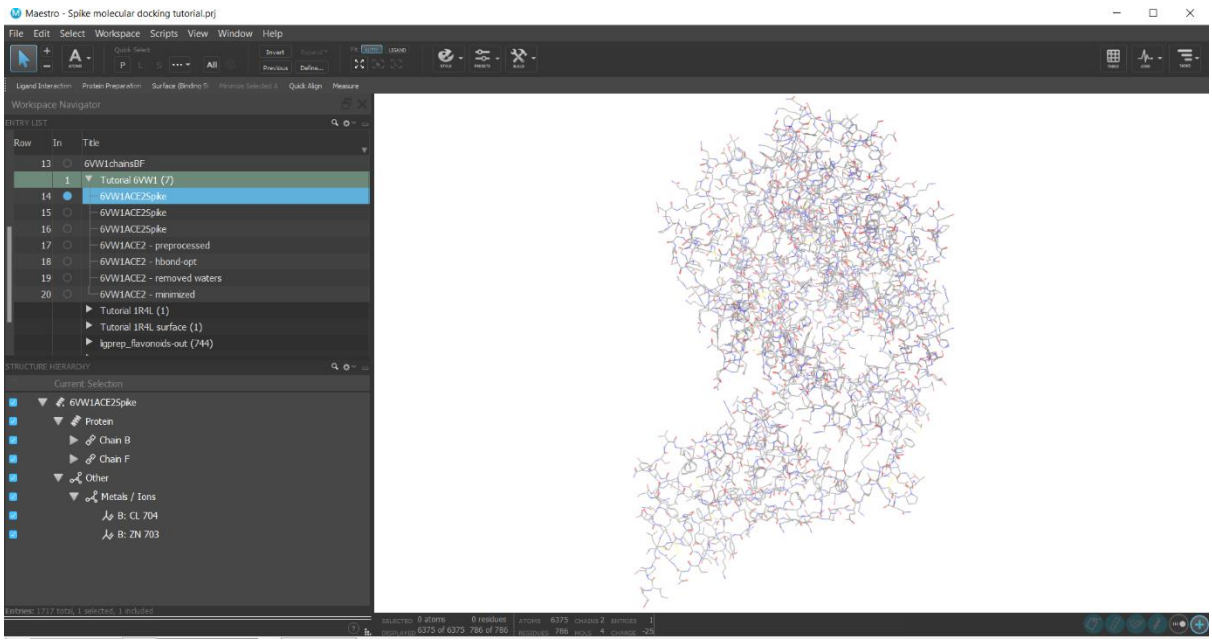


Figure 2.2.12: The 6VW1 crystal structure was modified in PyMOL (chains B and F were isolated) and imported into Maestro. The default settings can be seen.

To change the style of the molecular model, make sure that it is included and selected in the entry list. The circle must be blue for the structure to be included and the file name should also be blue for it to be selected. In the "Structure Hierarchy" section, you can press the arrow next to the name of the structure and a drop-down menu will appear. In Figure 2.2.12, you can see that the structure consists of two protein chains (chain B and chain F), as well as metals/ions (Zn^{2+} and Cl^-). Press the "Style" button at the top of the window and select the "Ribbons" option. In the "Edit Ribbons" section underneath, select "Single Color", and you can choose the colour of the ribbon. Hide the lines by pressing the "eye" symbol in the "Style" drop-down menu (Figure 2.2.13).

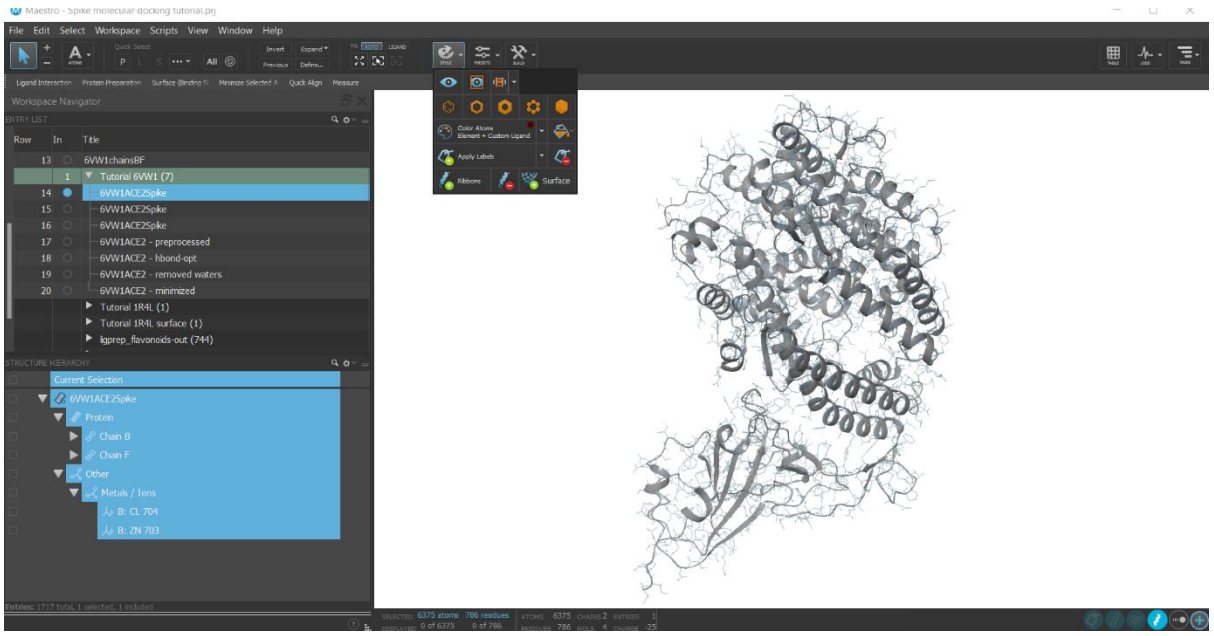


Figure 2.2.13: The menu that appears when pressing the "Style" button is shown and the complex is displayed as ribbons.

To colour the spike protein (chain F) differently from the ACE2 receptor (chain B), chain F needs to be highlighted/selected in the "Structure Hierarchy" section. Press the "Style" button at the top of the window and select the "Ribbons" option. In the "Edit Ribbons" section underneath, you can choose the colour of the ribbon (Figure 2.2.14).

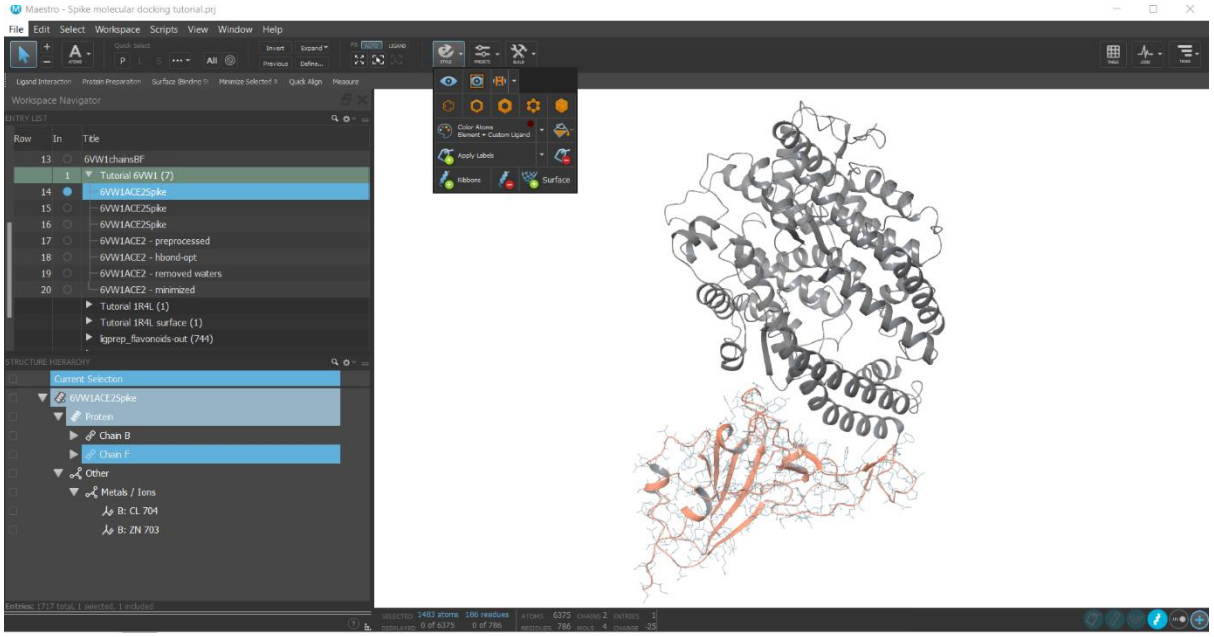


Figure 2.2.14: The SARS-CoV-2 spike protein RBD (chain F) is selected in the "Current Selection" section and has been coloured orange.

To display the metals/ions, highlight/select them in the “Structure Hierarchy” section and the corresponding boxes also need to be ticked. Press the “Style” button at the top of the window and apply the “CPK” representation. Press the “Paint” button to choose a colour for the ions (Figure 2.2.15).

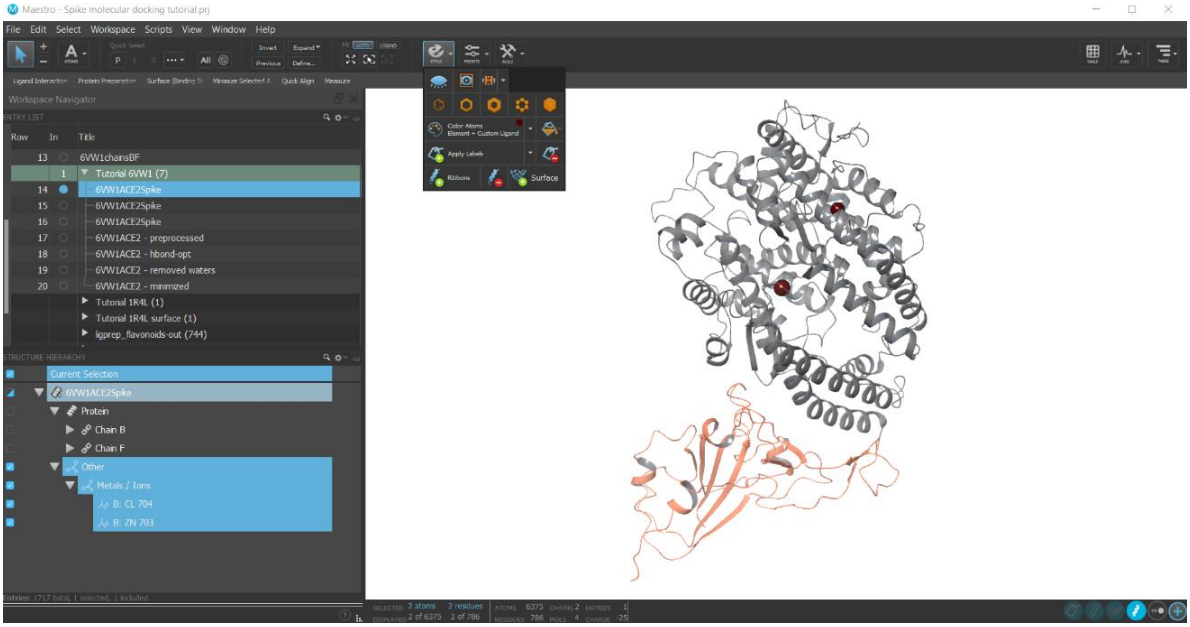


Figure 2.2.15: The zinc and chloride ions that are in the complex have been selected in the “Current Selection” section and are displayed as spheres. They have been coloured maroon.

To select residues of interest in chain B (ACE2 receptor), press the arrow next to chain B in the “Structure Hierarchy” section so that the residues are listed. Highlight/select the residues that you want to display and ensure that the boxes are ticked next to the residue names. Press the “Style” button at the top of the window and apply the “Thick Tube” representation. Press the “Paint” button to choose a colour for the residues (Figure 2.2.16).

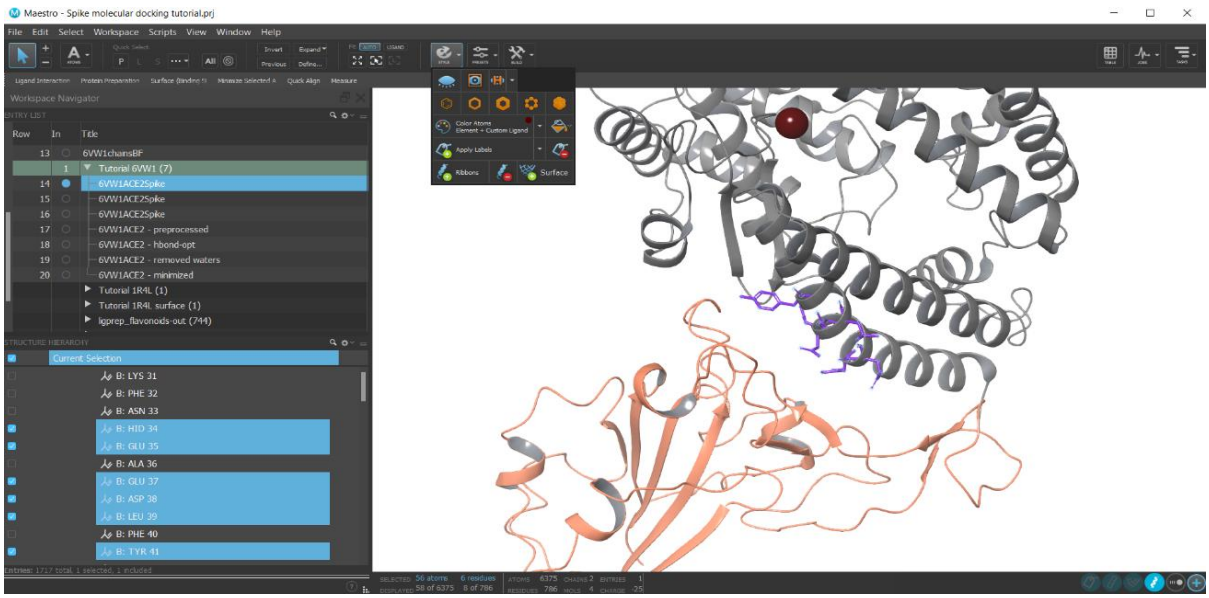


Figure 2.2.16: Several residues of the human ACE2 receptor have been selected and are shown in purple.

For the 1R4L crystal structure of the ACE2 receptor, the co-crystallised inhibitor can be represented in the "Thick Tube" format from the "Style" tab. The atoms of the ligands can also be depicted and there are several different options under the "Color Atoms" drop-down menu (Figure 2.2.17).

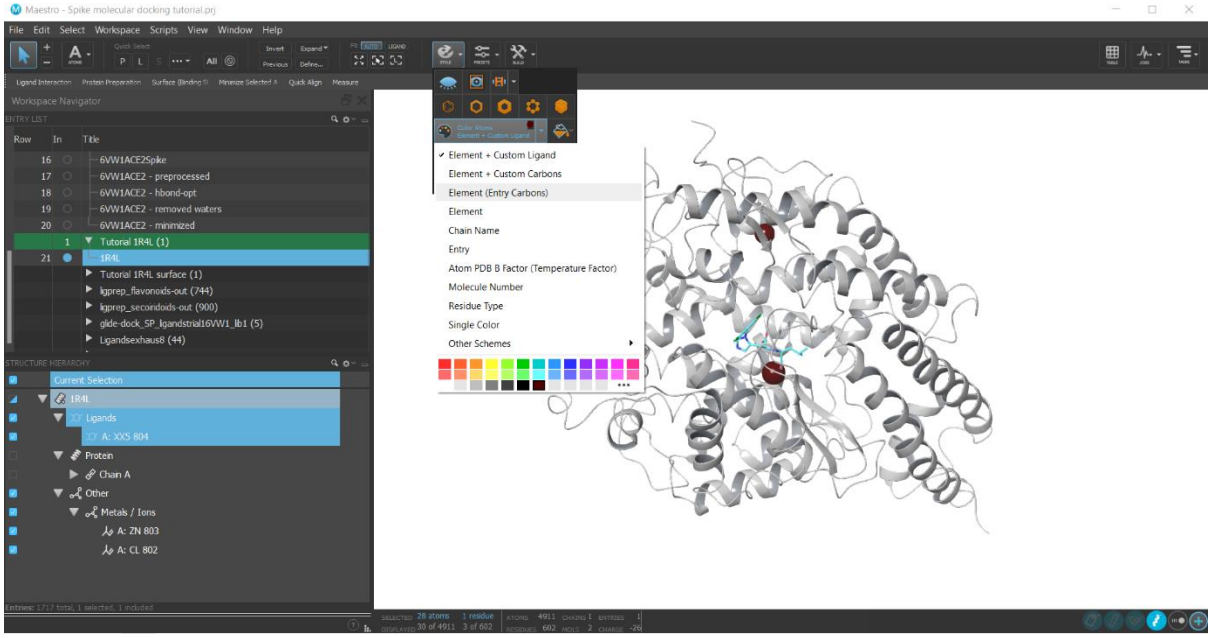


Figure 2.2.17: The crystal structure of the human ACE2 receptor with a co-crystallised inhibitor can be seen (PDB ID: 1R4L). The main protein chain is coloured silver, the inhibitor is coloured light blue, and the ions are coloured maroon.

Ligand Interaction Diagrams

The "Ligand Interaction Diagram" tool can be used to visualise the protein-ligand interactions that occur in a complex (Figure 2.2.18). Press the "Ligand Interaction Diagram" button that is positioned above the "Workspace Navigator". Go to "View" > "LID Legend". As described in the legend that appears, hydrophobic residues will be shown in green, polar residues will be shown in blue, positively charged residues will be shown in purple, and negatively charged residues will be shown in red. Various intermolecular bonds will also be displayed. The cutoff can also be changed to show the residues within a certain distance of the ligand.

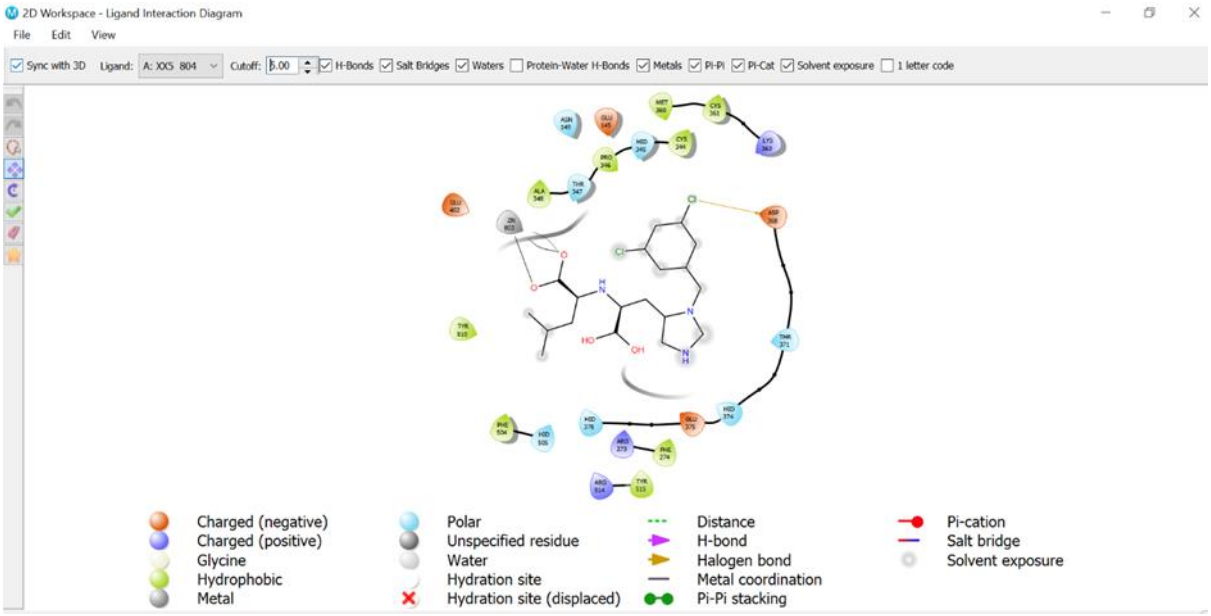


Figure 2.2.18: Protein-ligand interaction diagram for the co-crystallised inhibitor that is in the active site of the human ACE2 receptor (PDB ID: 1R4L). The residues that are within 5 Angstroms (Å) of the ligand are depicted.

2.3 Drawing Chemical Structures

If the 3D structure of a compound is unavailable from ligand databases, then the chemical structure can be drawn using freely available software such as ChemSketch (Figure 2.2.19). It's important to note that the chemical notation of each compound can be described by a Simplified Molecular Input Line Entry System (SMILES). The SMILES for chloroquine, for example, would be CCN(CC)CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl.

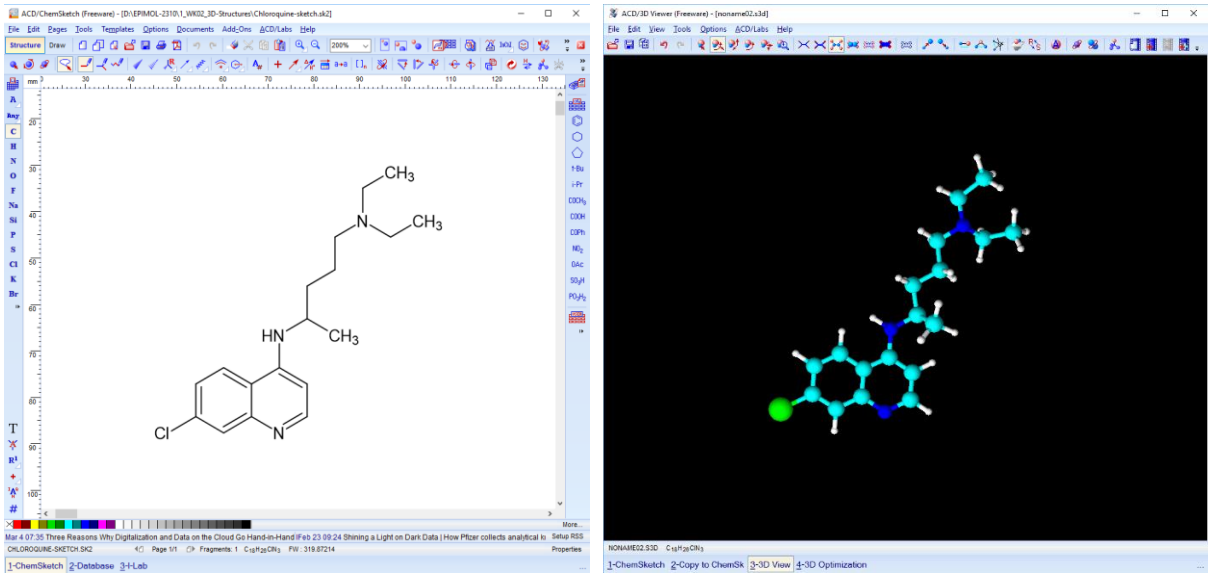


Figure 2.2.19: The 2D structure of the anti-malarial agent chloroquine has been drawn in ChemSketch and the 3D structure has been obtained.

Open up ChemSketch > Sketch the structure in 2D > Open the 3D viewer > Copy the 2D structure to 3D > Save the 3D structure as a .mol file. The 3D structure can also be optimised by adding hydrogens and correcting the bond angles. Using PyMOL, the 3D structure can be saved as a .pdb file.

Week 3: Building Protein Structures

3.0 Protein Sequences

The Universal Protein Resource (UniProt) can be used to obtain the sequences and annotation data of proteins from various organisms (https://www.uniprot.org/). The name of the protein can be typed into the search bar and any relevant results will appear. Each entry that is assigned the "Reviewed" symbol will have been manually annotated (based on the literature and curator-elevated computational analysis) and those that are "Unreviewed" await full manual annotation. This can be seen in the example below and the ACE2 receptor is the protein of interest (Figure 3.0.1).

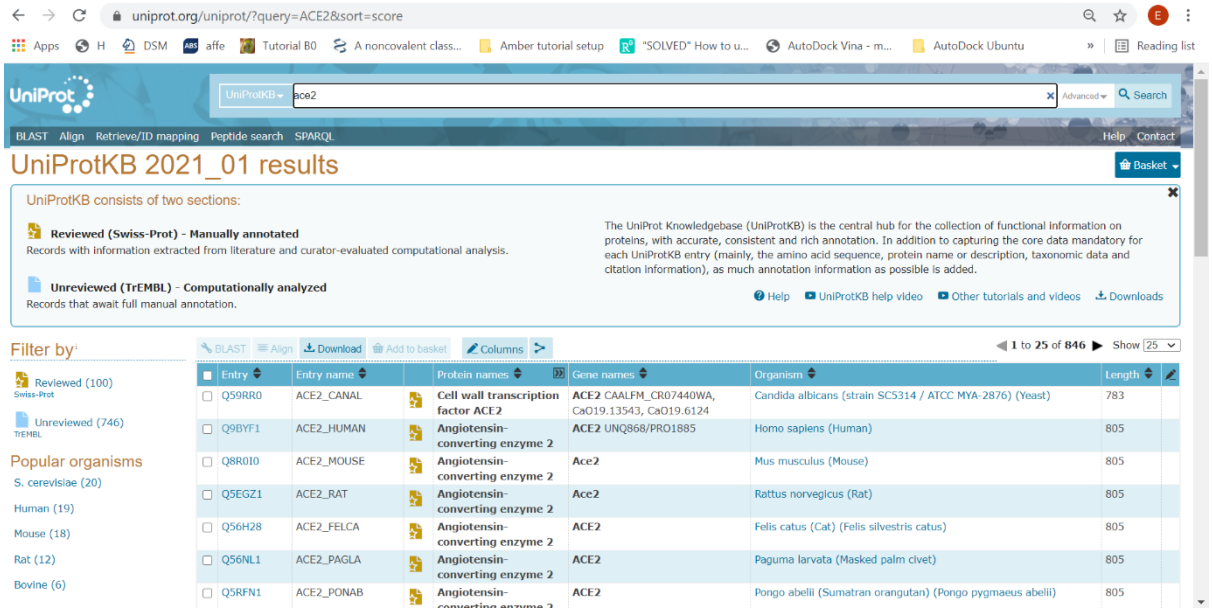


Figure 3.0.1: The UniProt database was used to obtain the sequence of the ACE2 receptor and the results page is displayed. The ACE2 protein can be found in a number of organisms and we are interested in the human ACE2 receptor. The ID is Q9BYF1 and the entry has been "Reviewed".

The human ACE2 receptor is of interest for this stage of the tutorial. The UniProt entry ID is Q9BYF1 and the status is "Reviewed". Information is provided about its function, names and taxonomy, subcellular location, pathology and biotechnological use, post-translational modifications/processing, expression, interaction, structure, family and domains, sequences, and similar proteins. There are also the cross-references, entry information and miscellaneous sections.

Under the “Sequences” section, there are two isoforms of ACE2 that are shown (isoform 1 and isoform 2). Isoform 1 has been characterised as the canonical sequence and the FASTA file can be downloaded. To download the sequence, press the “FASTA” button with the downwards arrow symbol. A page that displays a single line description followed by the amino acid residues or base pairs should appear (single letter codes) (Figure 3.0.2). This information can be selected, copied and pasted into a text file, and saved in .fasta format (Text Editor).

```
>sp|Q9BYF1|ACE2_HUMAN Angiotensin-converting enzyme 2 OS=Homo sapiens OX=9606 GN=ACE2 PE=1 SV=2
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQ
NMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTIL
NTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLY
EEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHL
HAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQ
AWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWDLGKGDFRILM
CTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKS
IGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEM
KREIVGVVEPVPHDETYCDPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLH
KCDISNSTEAGQKLFNMLRLGKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNK
NSFVGWSTDWSPYADQSIKVRISLKSALGDKAYEWNDNEMYLFRSSVAYAMRQYFLKVKN
QMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEVEKAIRMSRSRINDAFRLNDN
SLEFLGIQPTLGPPNQPPVSIWLIVFGVVMGVIVVGIVILIFTGIRDRKKKNKARSGENP
YASIDISKGENNPGFQNTDDVQTSF
```

Figure 3.0.2: The sequence of the human ACE2 receptor is shown in FASTA format.

In this example, we will be utilising the protein sequences of ACE2 from Felis catus, Gorilla gorilla gorilla, Rhinolophus ferrumequinum, Homo sapiens, Mus musculus, Macaca mulatta, and Pongo abelii.

3.1 Sequence Alignment and Protein Structure Alignment

Pairwise and multiple sequence alignment can be used to compare nucleotide or amino acid sequences. In Maestro, the protein structures of the ACE2 receptors can be imported into the workspace. Select and include each of the proteins. Select the “Multiple Sequence Viewer (Deprecated)” tool from the “Tasks” tab (Figure 3.1.1). Go to “Alignment” > Select “Multiple Alignment” since there are more than three biological sequences. In order to visualise the residues that are the same in each sequence, you can select the “Color Matching Residues Only” option. To compare the sequences, go to “Tools” > “Compare Sequences” > and you can obtain the values of the % identity, % similarity, and % homology. Instead of importing the structures and performing sequence alignment, there is also an option to import the FASTA sequences into the “Multiple Sequence Viewer (Deprecated)” window. Go to “File” > “Import Sequences” > Select the relevant .fasta files > Press “Open”. You may need to rename the sequences to make it more intuitive.

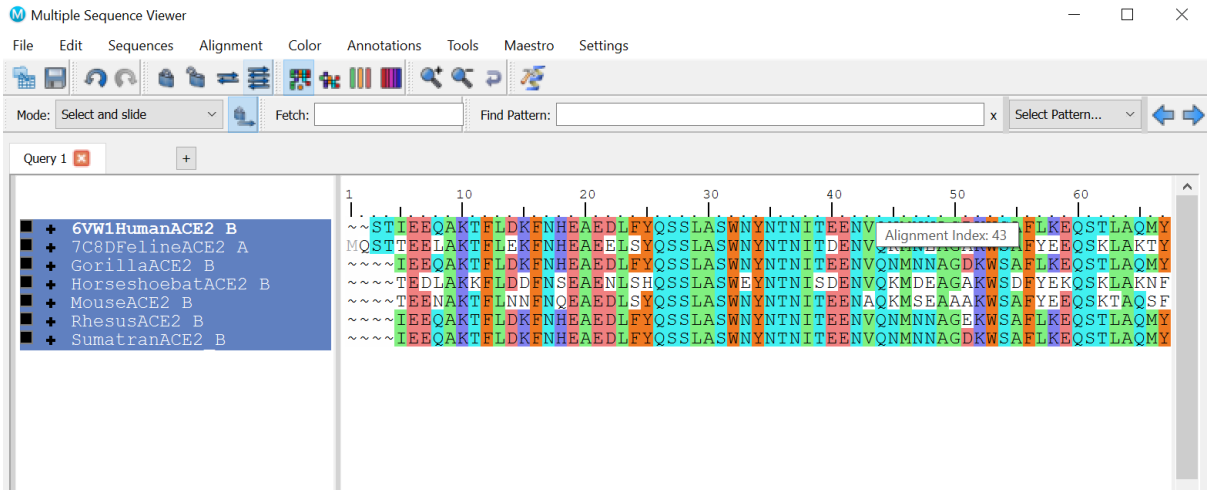


Figure 3.1.1: The Multiple Sequence Viewer window is displayed and the sequences of interest are shown. Multiple sequence alignment was performed in this example and the matching residues are coloured.

Additionally, the protein structures can be aligned and this can be performed using the "Protein Structure Alignment" tool that can be accessed through the Tasks tab (Figure 3.1.2). In PyMOL, the sequence alignment can also be performed using the crystal structures that are available. Import the proteins that you want to align and in this example, we want to align the protein sequences to the human ACE2 receptor. If we want to align the Gorilla gorilla gorilla sequence to the human ACE2 receptor, go to the "Action" (A) button of the Gorilla gorilla protein. From the drop-down menu, select "Align" > Select "To molecule" > Select the human ACE2 receptor.

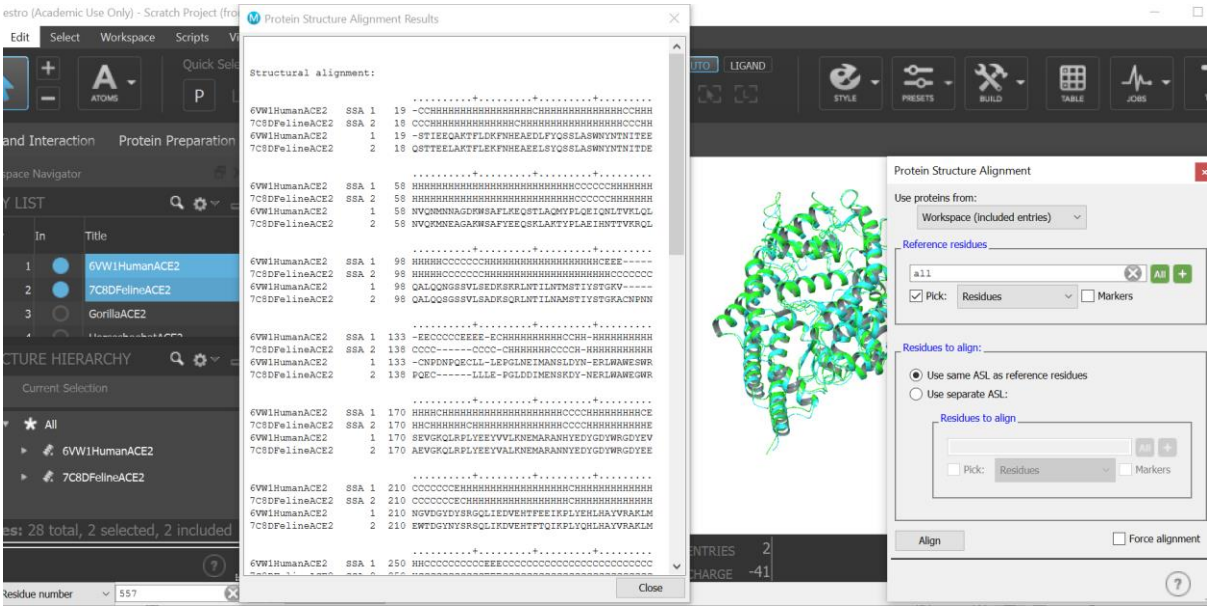


Figure 3.1.2: The structure of the feline ACE2 receptor was aligned to the human ACE2 receptor using the "Protein Structure Alignment" tool.

Basic Local Alignment Search Tools (BLAST) also allow for regions of similarity between biological sequences to be identified. The National Center for Biotechnology Information has a freely available BLAST server for nucleotides and proteins (17). Go to “Protein BLAST” > Enter the FASTA sequence of the protein into the text box or you can upload the file > In the “Database” component of the “Choose Search Set” section, select “Protein Data Bank proteins (pdb)”. In the “Program Selection” section, select “blastp (protein-protein BLAST)”, and then press the “BLAST” button. This can also be performed in the “Multiple Sequence Viewer (Deprecated)” window in Maestro. Go to “Tools” > “Find Homologs (BLAST)” > In the “Blast Search Settings” window > Select “Remote (NCBI)” > Select “Start Job”.

3.2 Homology Modelling

In order to gain further insight into the protein-ligand complex, a reliable 3D structure of the macromolecule is required. If a macromolecule is difficult to crystallise (ie. the protein is embedded in a membrane or has flexible regions) or is novel (ie. the structure has not been determined), then homology modelling can be used. Homology modelling is a method that allows for proteins to be constructed from amino acid sequences and requires a suitable template. The template is a 3D structure of a related homologous protein and in general, evolutionary related proteins share a similar structure. The structural conformation is more highly conserved than the amino acid sequence in proteins, as small or medium changes in the sequence usually have little impact on the overall 3D structure. Homology modelling can be used for a number of different functions including drug design, substrate specificity, function annotation and generating starting models for solving structures (X-ray crystallography, NMR and electron microscopy).

Homology modelling consists of four main steps and they are template identification, target-template sequence alignment, model building and model evaluation. As aforementioned, one or more existing experimental structures similar to the protein of interest can be used as a template to construct a model of the target sequence. The target sequence is used as a query to search for a suitable template that has an experimentally determined structure. This involves searching structural databases, such as the Protein Data Bank, or performing a protein-BLAST. The next step involves examining the identity between the target and template sequence, and this is a determinant in model quality. The sequence identity should be greater than 30% and to be used in molecular docking, the identity should be greater than 70%.

The model is then built using atomic and residual information from aligned sequences. There are four main aspects of model building and this includes backbone construction, side chain modelling, loop modelling and model optimisation. Several protein structure prediction resources are available for use:

- SWISS-MODEL (http://swissmodel.expasy.org/)
- I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)
- MODELLER (https://salilab.org/modeller/)

Once the model has been obtained, it is important to assess its quality. Ramachandran plots can be used to assess the stereochemistry of a model (bonds, bond angles, dihedral angles, non-bonded atom distances) and can allow for the distribution of the □□and ψ torsional angles to be visualised (Figure 3.2.1). In order for a model to be considered good quality, more than 90% of the residues should be in the favoured regions. Model evaluation resources, such as PROCHECK, can be used to generate the Ramachandran plots (https://saves.mbi.ucla.edu/).

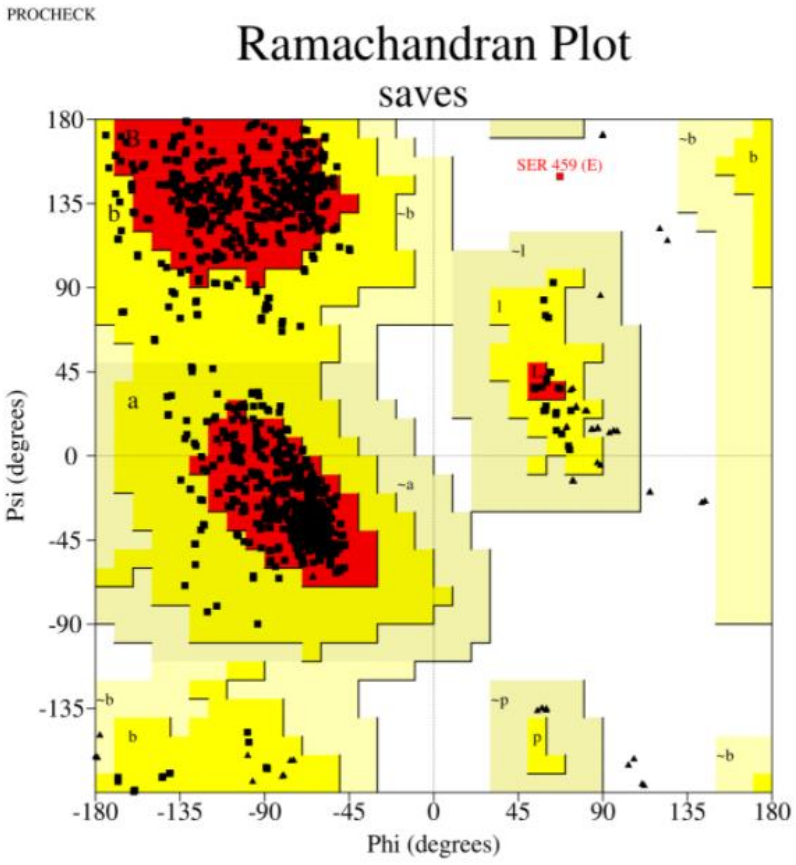


Figure 3.2.1: The crystal structure of the SARS-CoV-2 spike protein RBD in complex with the human ACE2 receptor was uploaded to PROCHECK and the Ramachandran plot was obtained (PDB ID: 6VW1).

Once again, we will be focusing on the protein sequences of ACE2 from Gorilla gorilla gorilla, Rhinolophus ferrumequinum, Mus musculus, Macaca mulatta, and Pongo abelii. The crystal structure of the ACE2 receptor for each of these organisms is unavailable on the Protein Data Bank and as a result, a model can be built using homology modelling. The crystal structures of the ACE2 receptors for Homo sapiens and Felis catus are available (their structures have been solved experimentally). We will be using the SWISS-MODEL server and the sequence of the ACE2 receptor from Macaca mulatta (18). Once on the website, press "Start Modelling" and you can either paste the FASTA sequence obtained from UniProt into the text box or upload the target sequence file. Select "Search for Templates".

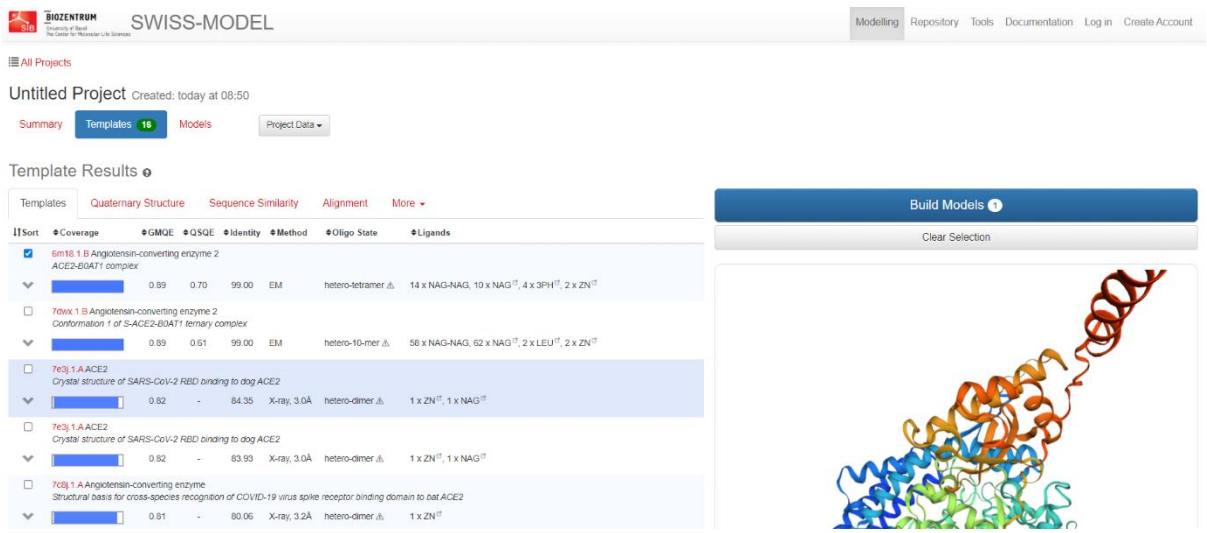


Figure 3.2.2: The template results for the Gorilla gorilla gorilla ACE2 receptor are shown. The human ACE2-B0AT1 complex (PDB ID: 6M18) was the top-ranking template and it has a sequence identity of 99.0%.

A page displaying the template results will appear and details about the quaternary structure, sequence similarity and alignment can also be found (Figure 3.2.2). For each template, information is provided about the target sequence coverage, GMQE, QSQE, the sequence identity to the target, the experimental method used to obtain the structure, the oligomeric state, the ligands (if any), the sequence similarity to the target, and the template search method used. In homology modelling, it is preferable to use structures that are determined by X-ray crystallography with a resolution higher than 2.2 Å as templates (it may be necessary to trade-off between high sequence similarity and experimental resolution). The Global Model Quality Estimation (GMQE) is expressed as a number between 0 and 1 and the value reflects the expected accuracy of a built model (18). Higher numbers indicate higher reliability. The Quaternary Structure Quality Estimate (QSQE) is a number between 0 and 1 that reflects the

expected accuracy of interchain contacts for a built model (18). This number is only calculated if it is possible to generate an oligomer and only for the top ranked templates (18). A high score (above 0.7) is generally better and the model can be considered reliable.

Once the appropriate templates are selected, press “Build Models”. On the “Model Results” page, information is provided about the predicted oligomerisation state, ligands, the GMQE and the QMEAN (Figure 3.2.3). The built model is compared to experimentally determined structures and the QMEAN Z-score is provided (18, 19). If the QMEAN Z-score is around 0, it indicates good agreement between the model structure and experimental structures of similar size (18). Scores that are below -4.0 indicate that the model is low quality (can be seen by the thumbs up and thumbs down symbol) (18). The “Structure Assessment” option can also be selected and the Ramachandran plot, quality estimate and residue quality results will appear.

Figure 3.2.3: The “Model Results” page is displayed and the homology model of the Gorilla gorilla gorilla ACE2 receptor that was built from the 6M18 cryo-EM template can be seen.

The results from homology modelling can be downloaded and the quality of the model can be assessed further using PROCHECK, which is a structure validation server (20). The model (.pdb file) that was generated from SWISS-MODEL can be uploaded to the server and the PROCHECK option can be selected.

3.3 Mutations

Protein residues can be mutated in programs such as Maestro and PyRx (Figure 3.3.1). Once the crystal structures of interest have been imported into Maestro, go to “Tasks” > select “Mutate Residues”. In the “Mutate Residues” window, you can specify the protein chain that consists of the residues of interest, the residue number, and the new amino acid for the mutated

residue. In PyMOL, import the protein of interest and display the sequence of amino acids. Go to “Wizard” > “Mutagenesis”. In the menu on the right-hand side, select “No Mutation” > Select an amino acid from the drop-down menu > Select “Apply” and “Done”.

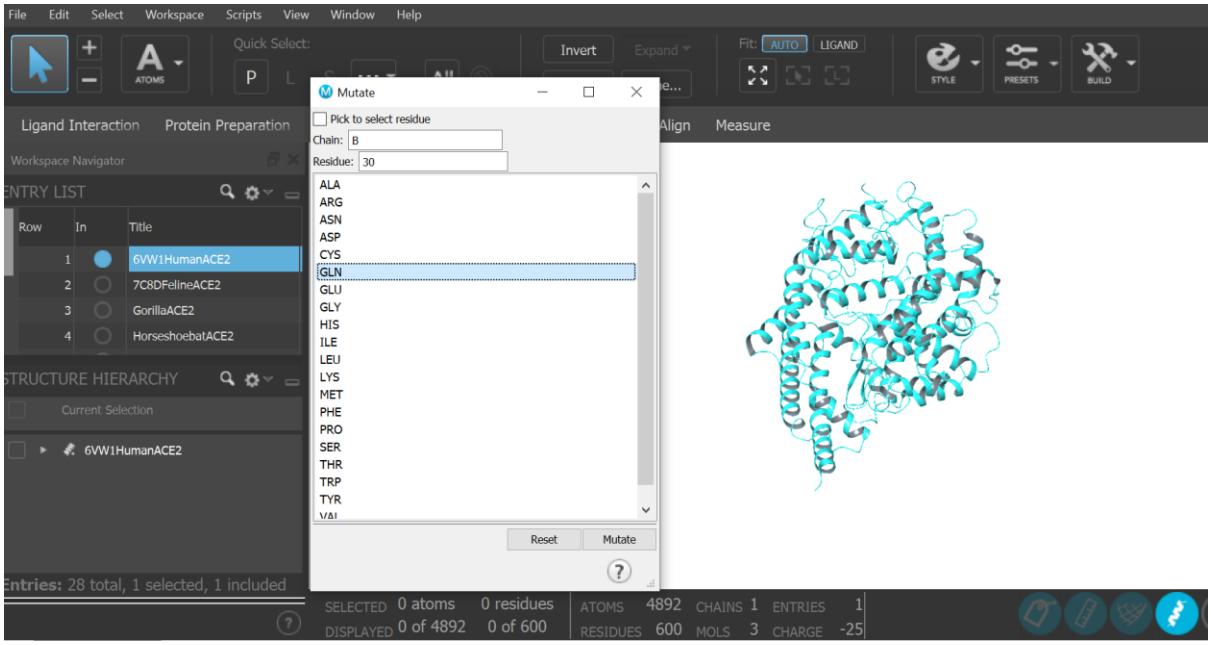


Figure 3.3.1: The amino acids of a protein can be mutated in Maestro using the “Mutate Residues” tool.

Week 4: Molecular Docking Using PyRx

4.0 PyRx

PyRx is a virtual screening software tool that can be used for drug discovery. It uses a large body of open source software such as AutoDock Vina, Python and Open Babel. The structure of the protein and the chemical structures of the compounds must first be converted to PDBQT format for docking with AutoDock Vina (Figure 4.0.1). The 3D structure of the protein and the compounds can be imported into the "Molecules" window of the Navigator section. In the "Molecules" window, right click, and select "Load Molecule" > Access the folder with the structures of interest > Choose "Open". The .sdf file of the ligands that have been downloaded from PubChem can be converted to .pdb format in programs such as PyMOL or Maestro. The structure of the protein must be prepared as a macromolecule. Right click on the protein in the "Molecules" workspace > Choose the "AutoDock" option > In the menu that appears, select "Make Macromolecule". To prepare compounds as ligands, right click on its structure in the "Molecules" workspace > Choose the "AutoDock" option > In the menu that appears, select "Make Ligand". In the "Preferences" tab, expand the AutoDock section and go to "Ligand Preparation" > To generate flexible ligands, ensure that all torsions are activated.

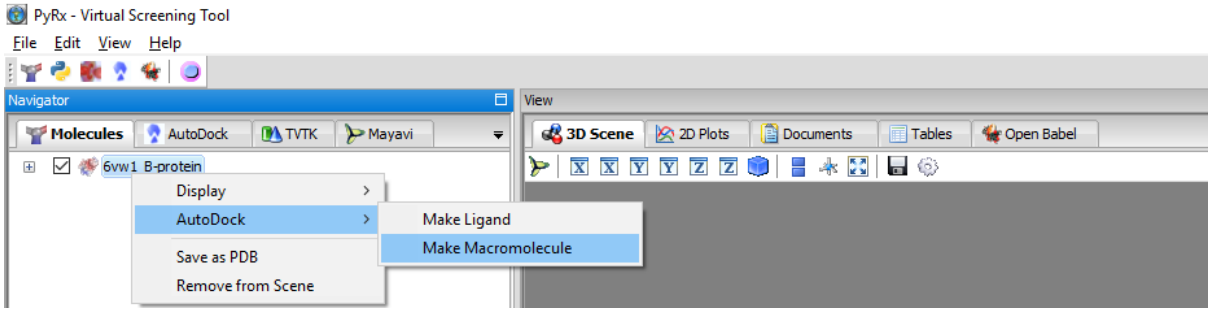


Figure 4.0.1: The PyRx Virtual Screening Tool window is displayed. Protein structures and compounds can be imported into the "Molecules" section and can be prepared as macromolecules and ligands, respectively.

Prior to molecular docking, the ligands can also be energy minimised (Figure 4.0.2). In the "Open Babel" window of the "Controls" section, the structures of the ligands can be imported by selecting the symbol with the addition sign ("insert new item"). Right click on the compound > Choose "Minimize Selected" > Once the ligand has been minimised, you can choose the "Convert selected to AutoDock Ligand (pdbqt) option". If there are multiple ligands, they can

all be selected > Right click and choose “Minimize All” > Once minimised, choose the “Convert all to AutoDock Ligand (pdbqt) option”.

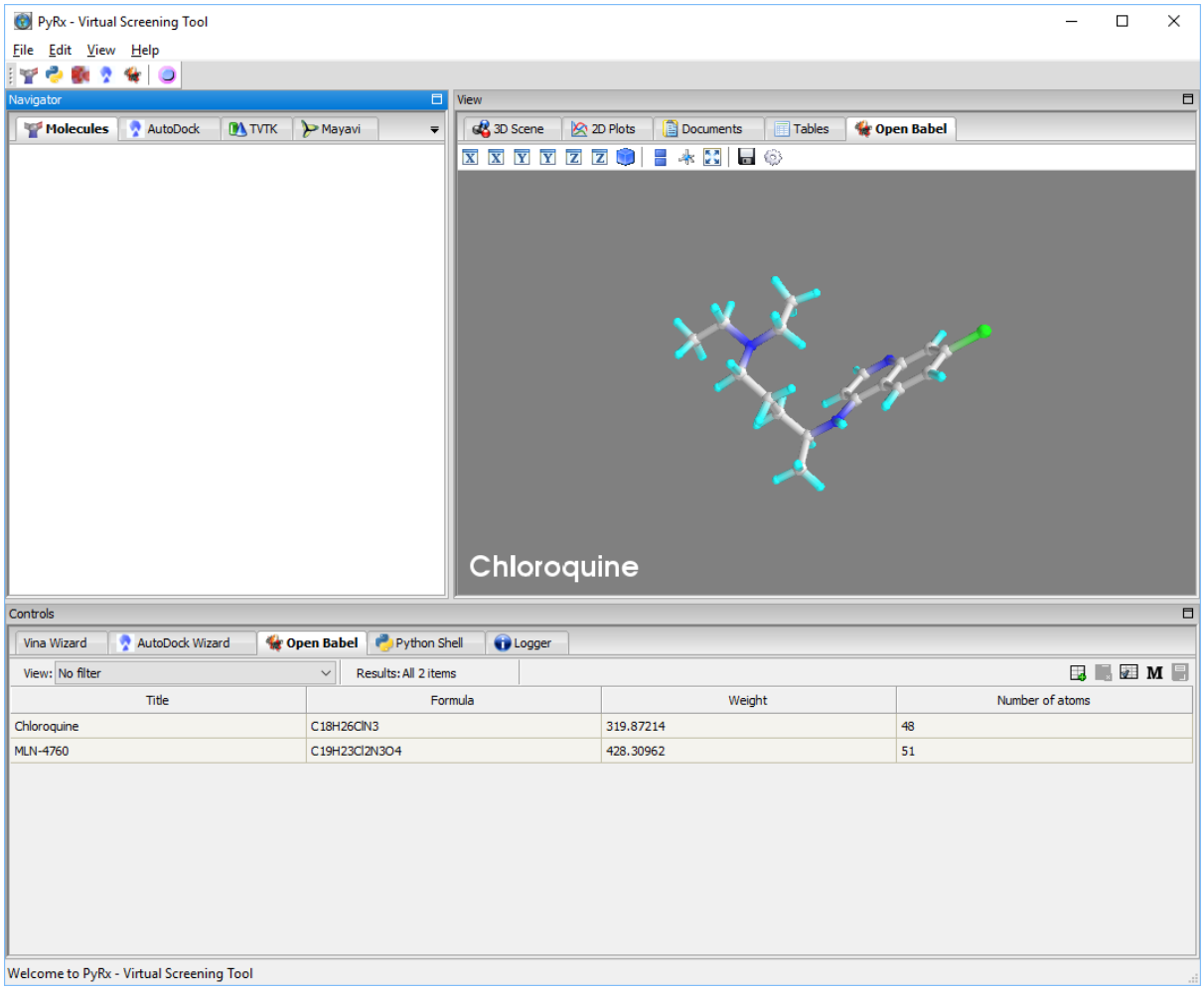


Figure 4.0.2: The ligands can be imported into the “Open Babel” window in the “Controls” section and can be energy minimised. Once minimised, they can be saved as .pdb files and can be converted to an AutoDock ligand (.pdbqt file).

Select the “Vina Wizard” tab under the “Controls” panel. Select “Local” for the “Vina Execution Mode” > Press “Start”. For the “Select Molecules” tab, the prepared protein and compounds can be selected from the “Ligands” and “Macromolecules” folders in the “AutoDock” window > Press “Forward” (Figure 4.0.3).

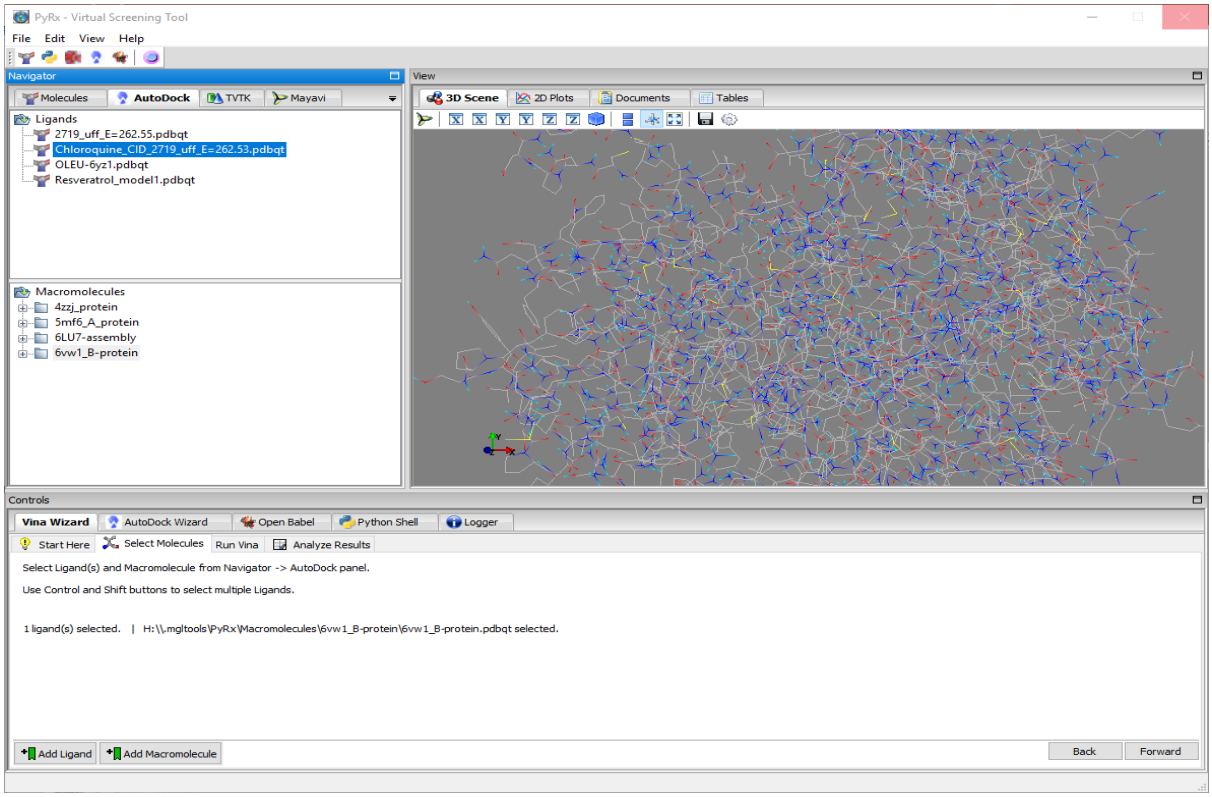


Figure 4.0.3: The energy minimised ligand (chloroquine) and ACE2 receptor that are found in the "Ligands and "Macromolecules" folders, respectively, can be seen. In the "Selected Molecules" tab of the "Vina Wizard", it is evident that 1 ligand and 1 macromolecule are selected.

In the "Run Vina" tab of the "Controls" panel, the receptor grid can be generated and the size and position of the grid can be altered (Figure 4.0.4). If the binding site of interest is known, the receptor grid can be positioned around the residues that form that region. To highlight protein residues, press the addition symbol next to the structure in the "Molecules" window and the sequence will appear. Select the residues of interest and press the "Toggle Selection Spheres" button. If the binding site is unknown, blind docking can be performed. The grid can be generated around the entire protein by maximising its size. The exhaustiveness can also be changed from the default number of 8.

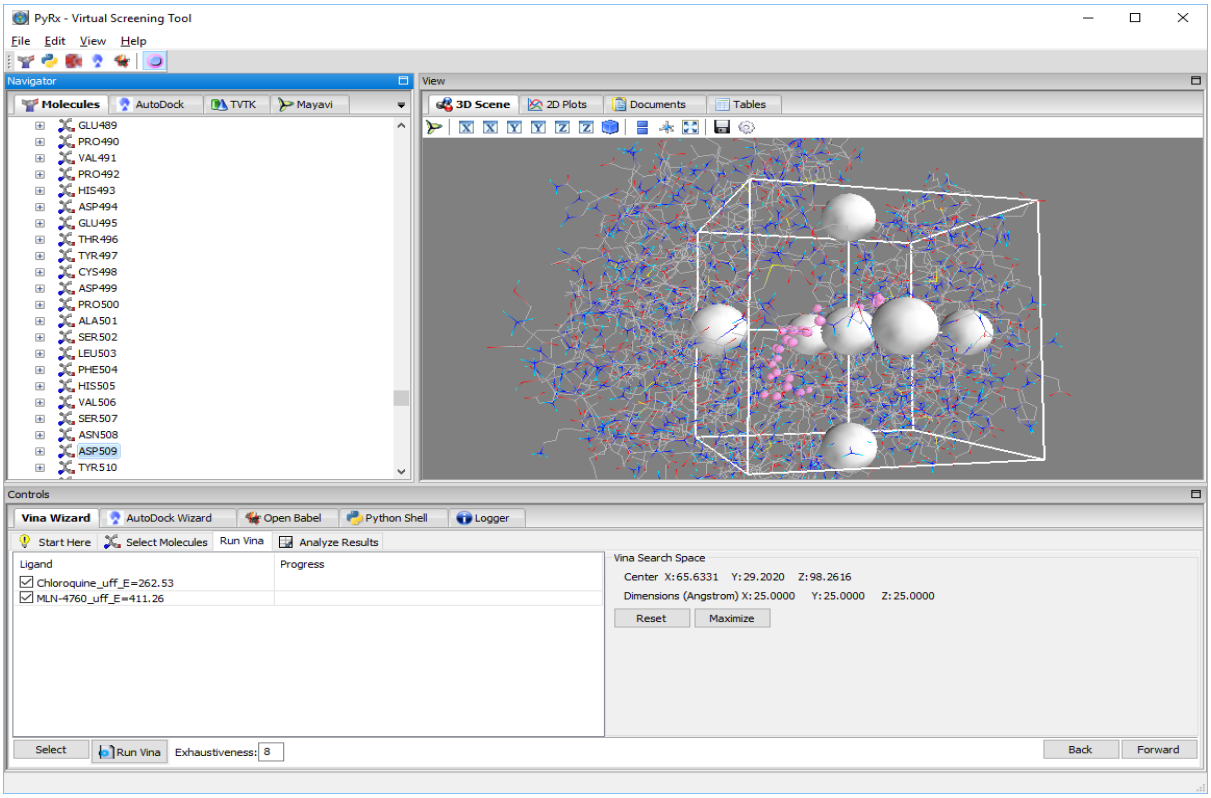


Figure 4.0.4: The receptor grid has been generated around several residues in the target binding site in the human ACE2 receptor and are coloured pink.

Click “Forward” to start the docking calculations and after virtual screening is completed, PyRx will go to the “Analyse Results” page. The nine best binding poses for each ligand are shown and the more negative the binding affinities are (kcal/mol), the stronger the predicted affinity. The binding affinities can be exported in a spreadsheet as a .csv file and the output structures can be saved as a .sdf file. To save all poses, change the preferences. Go to “Edit” > “Preferences” > “Open Babel” > “AutoDock Ligand” > “Number of Poses to Retain: 0”. The docking outputs will appear under the “Macromolecules” folder. The structures can be opened in PyMOL and converted to .pdb format. They can be imported into Maestro for analysis.

4.1 Binding Site Prediction

Ligand binding site prediction tools such as PrankWeb can be used to predict potential ligand binding sites in protein structures (Figure 4.0.5) (21). The PDB code or the file of the protein of interest can be uploaded to the PrankWeb server and the results can be downloaded as a compressed (zip) folder. The predictions spreadsheet contains the pocket scores, residues, and surface atoms. There is also a residues spreadsheet and a visualizations folder that contains the structure.pdb.pml file that can be opened in PyMOL.

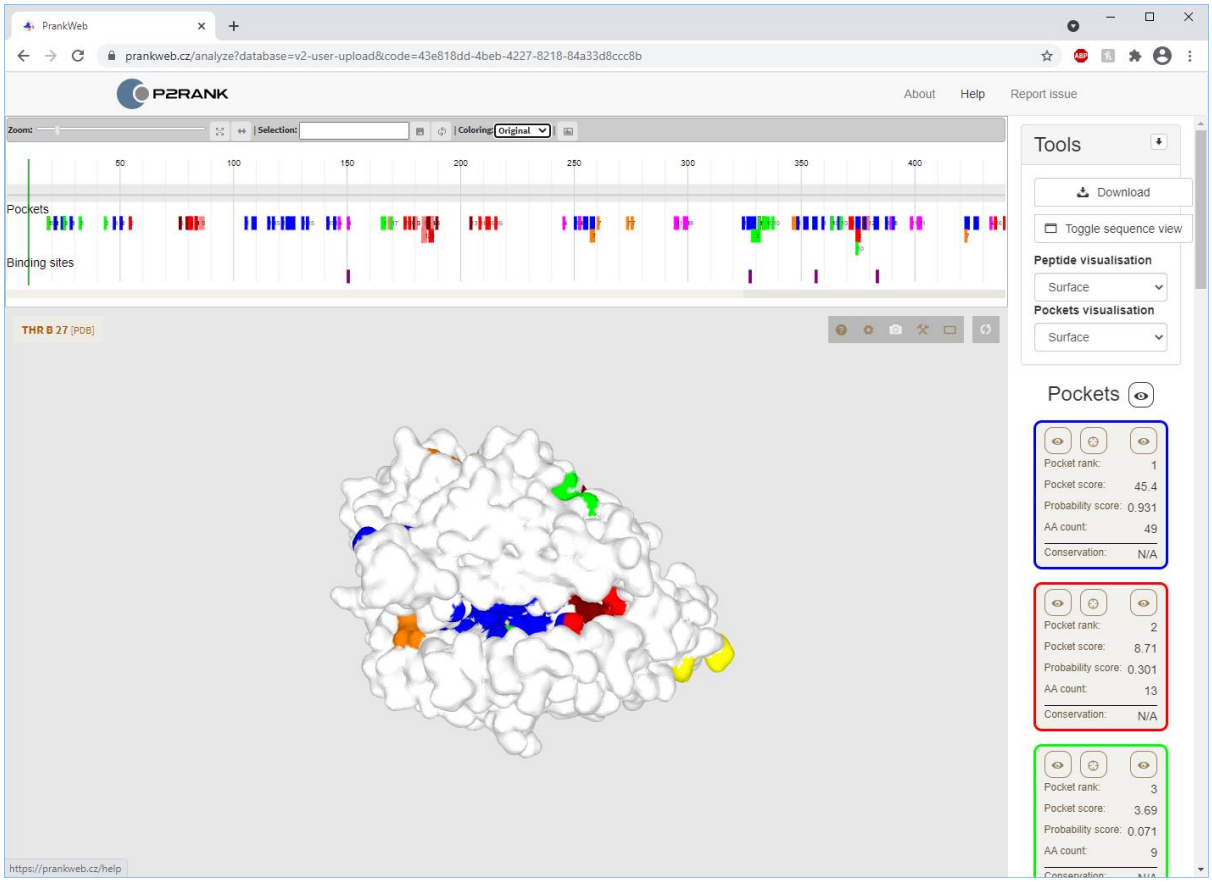


Figure 4.0.5: The human ACE2 receptor was uploaded to the PrankWeb server and the results are shown. The pocket that is coloured blue is predicted to be the top-ranking ligand binding site.

Week 5: Introduction to the Command Line

5.0 Overview of the Command Line

To generate the receptor grid in PyRx, we need to change the size of the box and move it around to the target region manually. To be consistent and more accurate, the coordinates of the grid should be the same each time and this can be a tedious process. Alternatively, the command line can be used.

The graphical user interface (GUI) allows users to interact with system graphical elements such as windows, icons, and menus. The command line is an interface (CLI) that allows users to view and manage computer files, and run applications by executing text commands (Table 2). There are different command line interpreter applications (command line shells) and examples include Windows Powershell, Command Prompt, Ubuntu, and Cygwin. There are also different operating systems. Linux, which is derived from UNIX (UNIX is not free to use and not open source) is one of the most commonly used operating systems. Windows is a commercial operating system and users do not have access to the source code. On a MacOS operating system, commands can be run through the Terminal.

File systems are arranged in a hierarchy, with the working directory being the current folder where commands will take place (Figure 5.0.1). The root directory, which is denoted by the slash (/) sign, is the highest directory in the hierarchy and a user has control of all files and folders in the system. The home directory is a subdirectory of the root directory and it is denoted by "**/home**". The files for a given user can be found in the home directory.

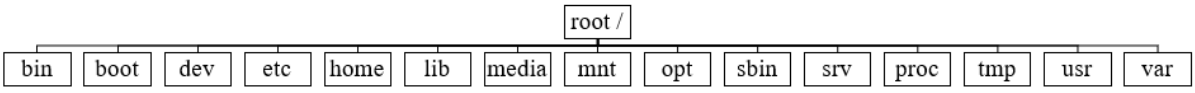


Figure 5.0.1: Files and folders are organised in a hierarchical system.

By typing in the command "**pwd**" (print working directory), it will tell you the current working directory (Figure 5.0.2). To change the working directory, type in the command "**cd**" (change directory). To go back to the parent directory (one level above), type in the command '**cd ..**' (change directory with two dots). You can use this command multiple times to move through multiple levels and a single dot (.) means the current directory. Previous commands use relative paths and this means that it depends on the current working directory. The absolute paths will

allow you to navigate from anywhere. To switch to the root directory, which is the top-level directory of a file system, type in the command "**cd** /". To switch to the home directory, type in the command "**cd** ~". To create a folder, type in the command "**mkdir**" (make directory). To look at files and folders in a directory, type in the command "**ls**" (list). The "**cat**" (concatenate) command can be used to look at the contents of a file.

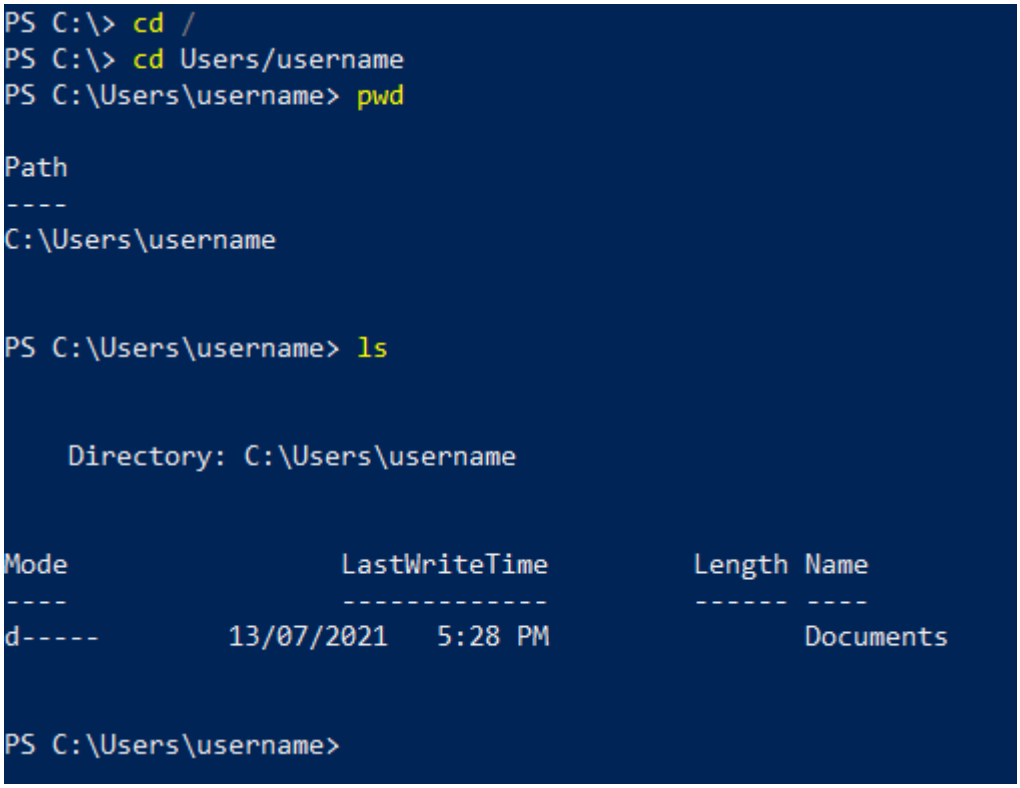


Figure 5.0.2: Commands executed in Windows Powershell. The C: drive is the root directory and the home directory is "C:\Users\username." The "Users" directory will contain a list of the users that have access to the computer and "username" will be used in this example. Once in the "username" directory, the command "**ls**" can be used to display the contents and it can be seen that the "Documents" folder is within the "username" directory.

To copy files, type in the command "**cp**". A copy of file 1 can be made in the current working directory and can be called file 2 by typing in the following command "**cp file1 file2**". To move or rename files, the command "**mv**" can be used. To move or rename file 1 to file 2, type in the command "**mv file1 file2**". In order to carefully delete files and delete folders, the commands "**rm**" and "**rmdir**" can be used respectively. Use these commands with caution as the files will be deleted forever. It is important to name your folders and files carefully, as commands and file names are case sensitive (Figure 5.0.3). Punctuation should be avoided and this is also the case with spaces. Spaces can be used if necessary but they need to be put in quotation marks.

Slashes can also be used if necessary: forward slashes are used in Linux (/), back slashes (\) are specific for Windows.

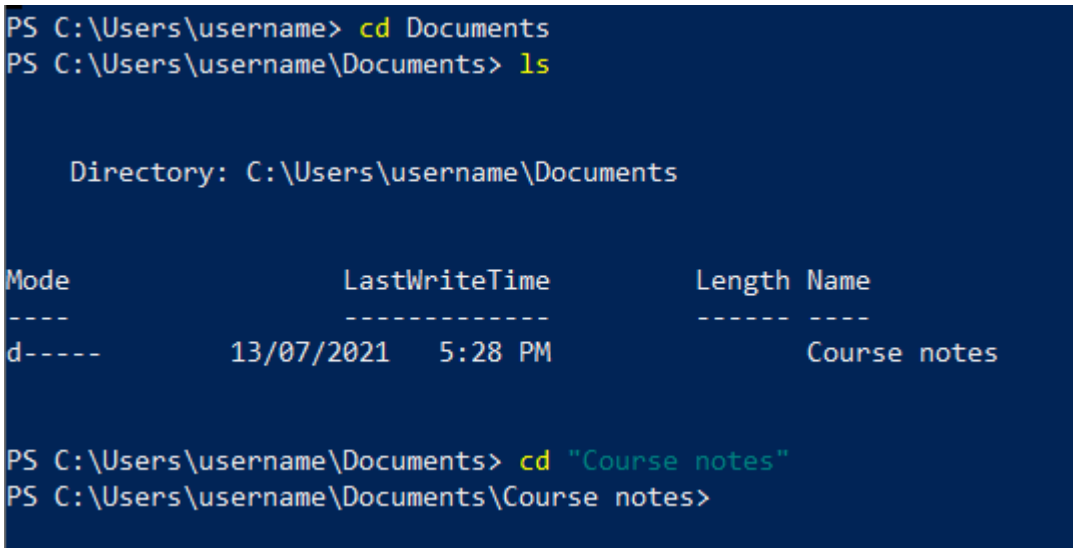


Figure 5.0.3: The "Course notes" folder has been accessed from the "Documents directory" and the notation that is used is depicted.

Table 2. Summary of the commands that are commonly used.

pwd	Print working directory	Display path of current directory
cd *dir*	Change directory	Change working directory
cd **.**	Single dot	(change to) Current directory
cd **..**	Two dots	(change to) Parent directory
ls	List	Lists files and folders
mkdir *dir*	Make directory	Creates new folder
cat *file*	Concatenate	Display contents of file on screen
cp *file1 file2*	Copy	Copy a file
mv *file1 file2*	Move	Moves or renames a file
rm *file*	Remove	Deletes files
rmdir *dir*	Remove directory	Deletes folders

5.1 Molecular Docking Using Vina Through the Command Line

The prepared protein and ligand PDBQT files that were generated through PyRx can be obtained from the "Macromolecules" and "Ligands" folders in the "mgltools" directory (Figure 5.1.1). The path can be checked in PyRx by going to: "Edit" → "Preferences" → "Workspace" (eg. H: \\.mgltools\PyRx).

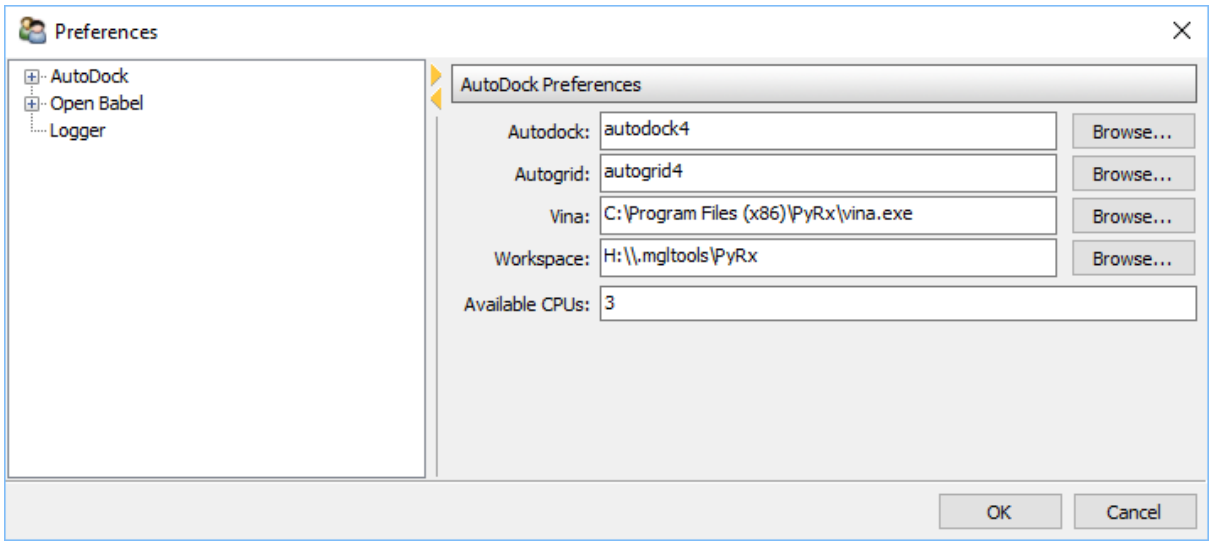


Figure 5.1.1: The "Preferences" window of PyRx is displayed and these settings can be changed.

The input files that are needed for docking include the ligand pdbqt file, protein pdbqt file, configuration text file, and vina.exe application file. The configuration file is a text file that specifies the docking input (conf.txt). This includes the receptor and ligand PDBQT files, the receptor grid coordinates and dimensions, the docking options and the output files. The same configuration file can be used to dock ligands to the same protein and the ligand input/output file name will need to be changed as required. The coordinates and size of the receptor grid can be obtained from Vina Search Space section in PyRx (Figure 5.1.2).

receptor = receptorname.pdbqt
ligand = ligandname.pdbqt

center_x = x coordinate of the center
center_y = y coordinate of the center
center_z = z coordinate of the center
size_x = x direction
size_y = y direction

size_z = z direction

out = vina_nameofligand.pdbqt

log = vina_nameofligand.txt

exhaustiveness = represents the computational effort

num_modes = represents the number of ligand conformations to be produced – the default is 9

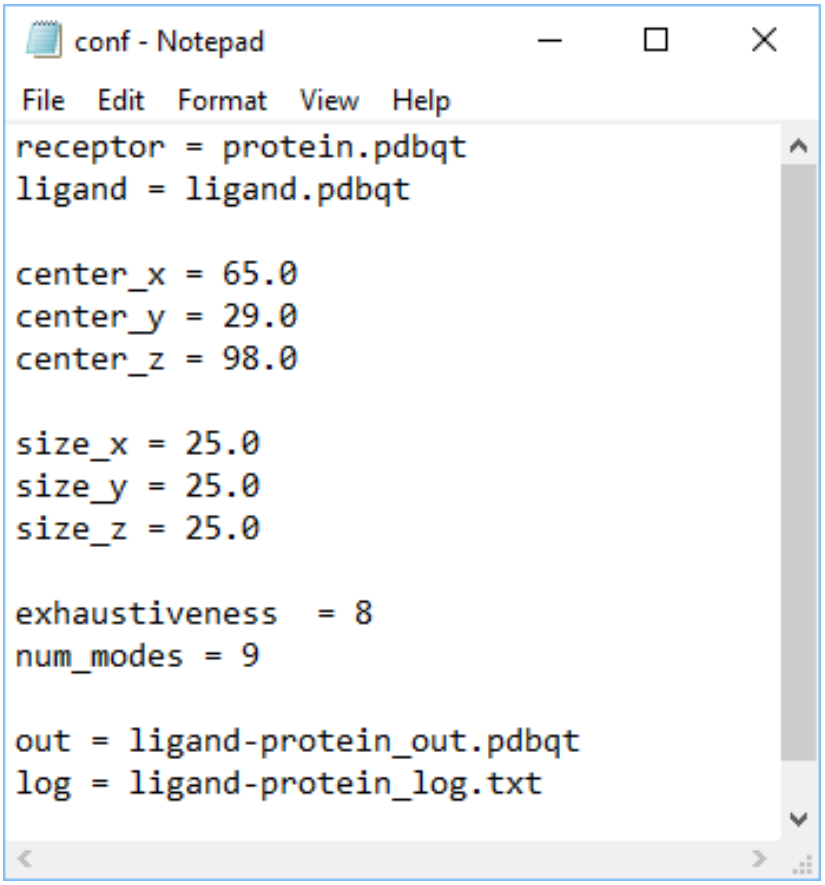


Figure 5.1.2: The configuration file (conf.txt) that can be prepared using a text editor is shown. The names of the receptor and ligand must be specified, as well as the coordinates and dimensions of the grid. The exhaustiveness and the number of conformations produced for each ligand can be changed.

To run the docking calculations, the command "**./vina –config conf.txt**" should be executed (Figure 5.1.3). The docking progress will be displayed on the screen. The docked ligand structure will appear as an output file (ligand-protein_out.pdbqt) and this contains all binding modes in a single file. This can be opened in PyMOL. The ligand-protein_log.txt file lists the binding affinities of the binding modes in kcal/mol.

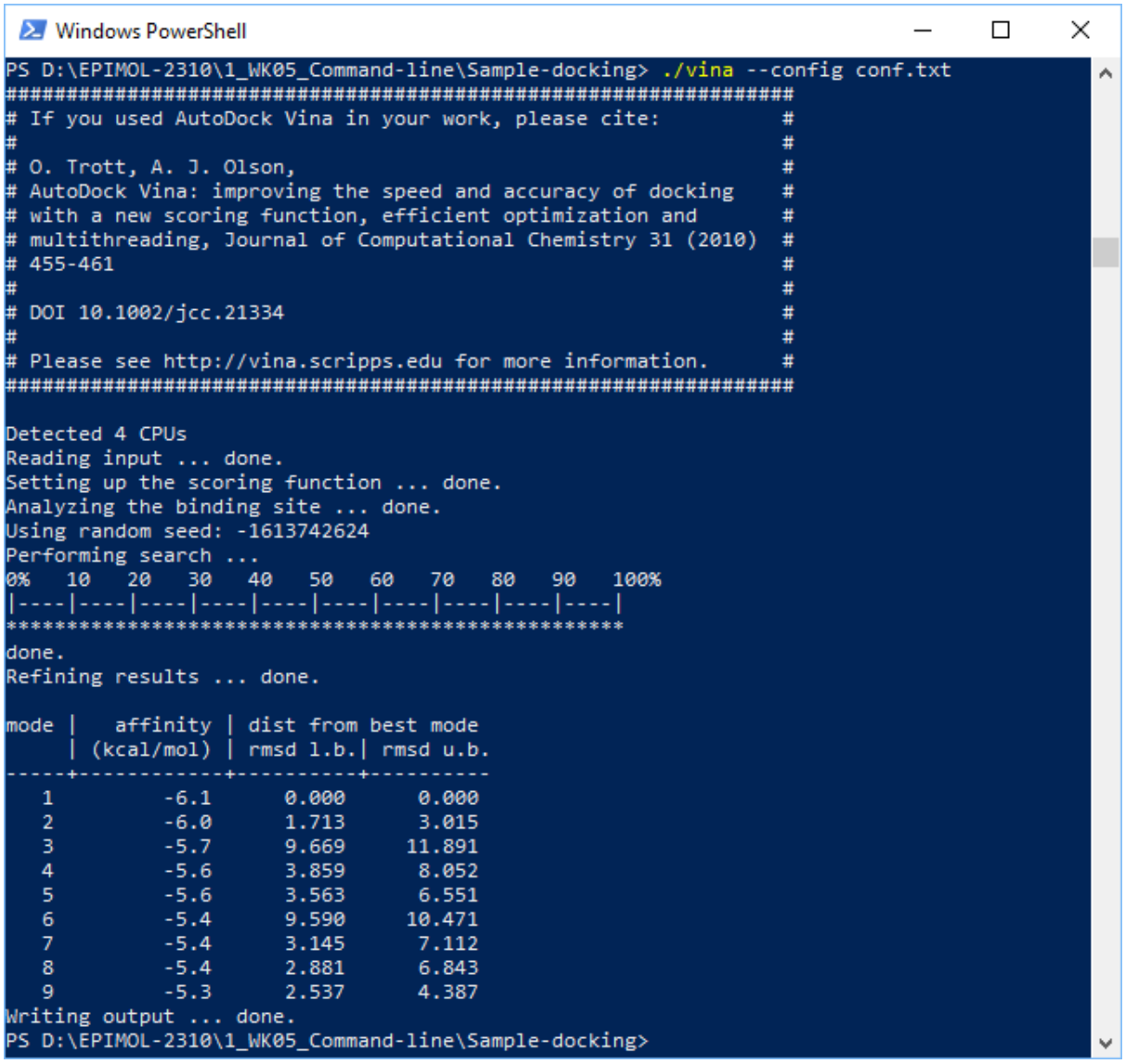


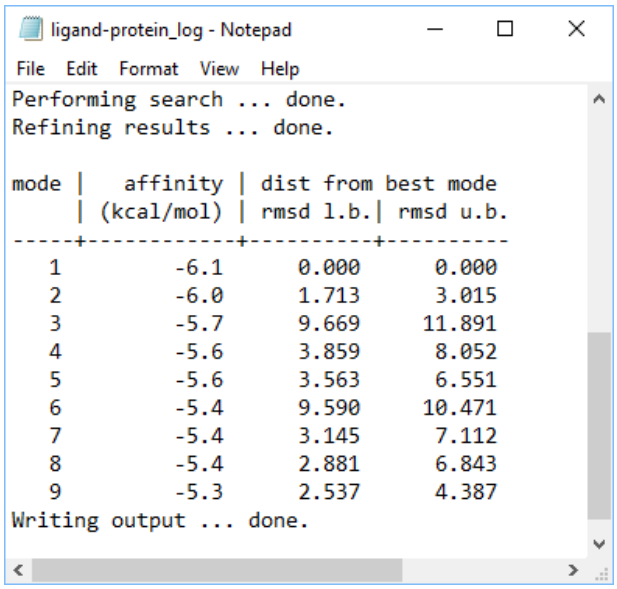
Figure 5.1.3: Chloroquine was docked to the target binding site in the human ACE2 receptor using AutoDock Vina via the command line (Windows Powershell). The binding affinities (kcal/mol) are displayed.

These values can be recorded for further analysis. Once again, the poses can be saved as separate files in PyMOL. Go to “File” > “Save Molecule” > “Save to “Multiple Files” > Save state “All” > Save as a .pdb file. The docking poses and interactions formed with the protein can be visualised in Maestro (Figure 5.1.4).

Example:

- C:Users\username\> cd Documents\Coronavirus\Course\Docking\Chloroquine
- C:Users\username\Documents\Coronavirus\Course\Docking\Chloroquine> ls

- C:Users\username\Documents\Coronavirus\Course\Docking\Chloroquine> ./vina.exe –config conf.txt



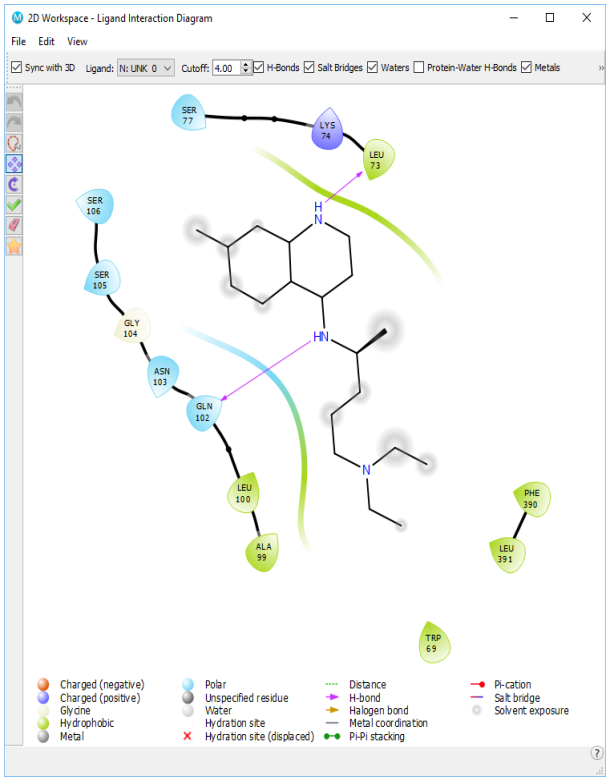


Figure 5.1.4: The ligand-protein_log.txt file that was produced from molecular docking is shown and the binding affinities of chloroquine are listed. The structures of the poses produced from blind docking were imported into Maestro and the "Ligand Interaction Diagram" tool was used to examine the protein-ligand interactions.

Week 6: Docking Multiple Ligands Via the Command Line

6.0 Multiple Ligand Docking

Multiple ligands can also be docked using a script that can be run using Vina through the command line. We will be using an SH script named "vina_screen_local.sh" that has been developed for the Bash language. Bash is a command line interpreter and a UNIX scripting language. In order to perform virtual screening on Windows using the Bash script, the "Windows Subsystem for Linux" setting must be on. This allows for a Linux environment to be run directly on Windows. To install the "Windows Subsystem for Linux", go to "Settings" > "Turn Windows features on or off" > Check "Windows Subsystem for Linux" > Press "Ok". Open the Microsoft Store and download Ubuntu, which is another Linux open source operating source system. The Ubuntu console window will open and it will take a few minutes to install. A new username/password can be made however, it is not necessary to use the program.

The ligands of interest need to be prepared as pdbqt files in PyRx (energy minimised using OpenBabel and saved as pdbqt files) and the structures can be located under the "Ligands" folder in the PyRx working directory. The files need to be renamed as ligand_1.pdbqt, ligand_2.pdbqt…etc. It's very important to keep track of the ligand names and a spreadsheet can be created for this purpose (Figure 6.0.1).

	A	B
1	**Drug name**	**Ligand number**
2	amentoflavone	ligand_1
3	brompheniramine	ligand_2
4	chloroquine	ligand_3
5	cimicoxib	ligand_4
6	delavirdine	ligand_5

Figure 6.0.1: An Excel spreadsheet that contains the names of the ligands can be created and this is important when screening a library of compounds.

The files that are required include the protein pdbqt file, named ligand pdbqt files, vina.exe application file, the configuration file, and the virtual screening script. The vina_screen_local.sh file is a Bash script that can be used to dock multiple ligands using Vina (Figure 6.0.2).

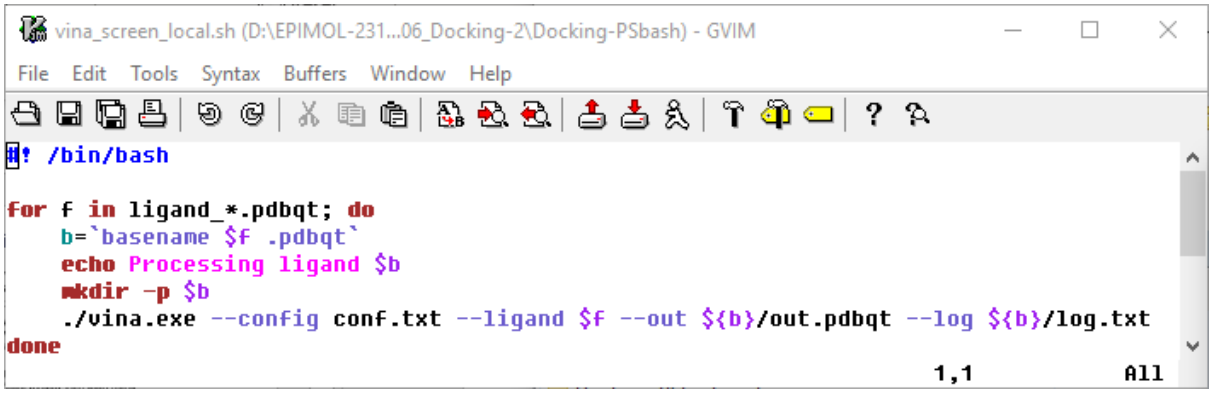


Figure 6.0.2: The contents of the vina_screen_local.sh script are shown.

In order to execute the script in Windows Powershell, type "**bash**" to enter the Linux environment and to leave the Linux environment, type in "**exit**". In the directory containing the relevant files, type in the command "**./vina_screen_local.sh**" to execute the script. The docking calculations will be performed on ligands one after another. Individual folders will be created for each ligand and they will consist of the out.pdbqt file (all binding modes for the ligand in a single file) and a log.txt file (lists the binding affinities in kcal/mol).

The Ubuntu console window can also be opened and the relevant folder can be accessed by navigating through the files. In Ubuntu, the C: drive can be accessed executing the "**/home**" command followed by "**/mnt/c**". The directory containing the relevant files can be specified and the "vina_screen_local.sh" script can be executed.

```
:~# cd /home
:/home# cd /mnt
:/mnt# cd c
:/mnt/c# cd Users2
:/mnt/c/Users2# cd Username
:/mnt/c/Users2/Username# cd Documents
:/mnt/c/Users2/Username/Documents# cd Examplecommand
:/mnt/c/Users2/Username/Documents/Examplecommand# cd Ligandname
:/mnt/c/Users2/Username/Documents/Examplecommand/Ligandname# ./vina_screen_local.sh
```

Figure 6.0.3: The commands that have been used to navigate to the directory of interest and run the script for multiple ligand docking in Ubuntu are shown.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28(1):235-42.
2. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581(7807):221-4.
3. Smyth MS, Martin JH. x ray crystallography. Mol Pathol. 2000;53(1):8-14.
4. Marion D. An introduction to biological NMR spectroscopy. Mol Cell Proteomics. 2013;12(11):3006-25.
5. Cabra V, Samsó M. Do's and don'ts of cryo-electron microscopy: a primer on sample preparation and high quality data collection for macromolecular 3D reconstruction. J Vis Exp. 2015(95):52311-.
6. Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2020;49(D1):D480-D9.
7. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research. 2021;49(D1):D1388-D95.
8. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Research. 2017;45(D1):D945-D54.
9. Irwin JJ, Shoichet BK. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. Journal of Chemical Information and Modeling. 2005;45(1):177-82.
10. Bonvino NP, Liang J, McCord ED, Zafiris E, Benetti N, Ray NB, et al. OliveNet™: a comprehensive library of compounds from Olea europaea. Database (Oxford). 2018;2018:bay016.
11. Advanced Chemistry Development I. ACD/ChemSketch 2020.2.1. Toronto, ON, Canada. 2021.
12. Schrödinger LLC. The PyMOL Molecular Graphics System, Version 1.2r3pre.
13. Schrödinger LLC. Schrödinger Release 2021-2: Maestro. New York, NY, 2021.
14. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics. 1996;14(1):33-8.
15. Dallakyan S, Olson AJ. Small-Molecule Library Screening by Docking with PyRx. In: Hempel JE, Williams CH, Hong CC, editors. Chemical Biology: Methods and Protocols. New York, NY: Springer New York; 2015. p. 243-50.
16. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.
18. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research. 2018;46(W1):W296-W303.
19. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27(3):343-50.
20. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 1993;26(2):283-91.
21. Jendele L, Krivak R, Skoda P, Novotny M, Hoksza D. PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic acids research. 2019;47(W1):W345-W9.