

Ensemble Docking with GOLD

2020.0 CSD Release

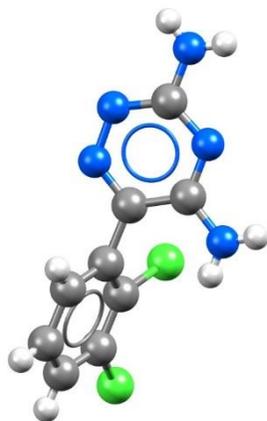


Table of Contents	
Ensemble Docking	2
Case Study	3
Introduction.....	3
Provided Input Files.....	5
Superimposing Protein Structures	6
Protein Setup.....	9
Exploring the Ligand for Ensemble Docking.....	10
Defining the Binding Site	10
Selecting Ligand(s) for Docking	11
Setting Water in the Binding Site (Active Waters)	12
Setting Ligand Flexibility.....	14
Setting Docking Parameters	15
Selecting a Fitness Function	15
Specifying GA Settings.....	15
Run the Ensemble Docking.....	16
Analysing Docking Results	18
Visualising Docking Results	18
Conclusions.....	20

Ensemble Docking

In the last decade, the importance of being able to model protein flexibility has been widely recognised¹. The incorporation of receptor flexibility in automated docking algorithms enables more accurate binding pose prediction and better virtual screening enrichments, in addition to providing a more realistic description of the physics of the protein-ligand binding interaction.

Diverse methodological approaches have been proposed; some of these treat flexibility explicitly, allowing extra degrees of freedom in the search space to perform direct changes of the binding site conformation. In contrast to explicitly modelling flexibility, so-called “ensemble docking” methodologies aim to address the issue of protein flexibility by adding multiple conformations of the target protein rather than just the single rigid receptor structure used in standard docking (Figure 1). This ensemble of protein conformations mimics the conformational equilibrium which characterises the native state of the target protein and provides a structural degree of freedom by which the conformation of the protein model may be matched to fit any particular ligand. There are several cases in which the ensemble docking approach can be useful:

- Using crystal structures of the same target that are isolated and/or co-crystallised with various ligands in order to account for induced fit and explore the potential flexible range of the receptor site.
- Using snapshots of the protein across molecular dynamic runs to explore possible conformations across time.
- Using crystal structures from various groups of the same protein-ligand complex to account for technique differences.

In all cases, GOLD will look for the best single protein-ligand docking result, it does not treat the set in any kind of ‘averaged’ capacity.

Please download the tutorial input files [here](#).

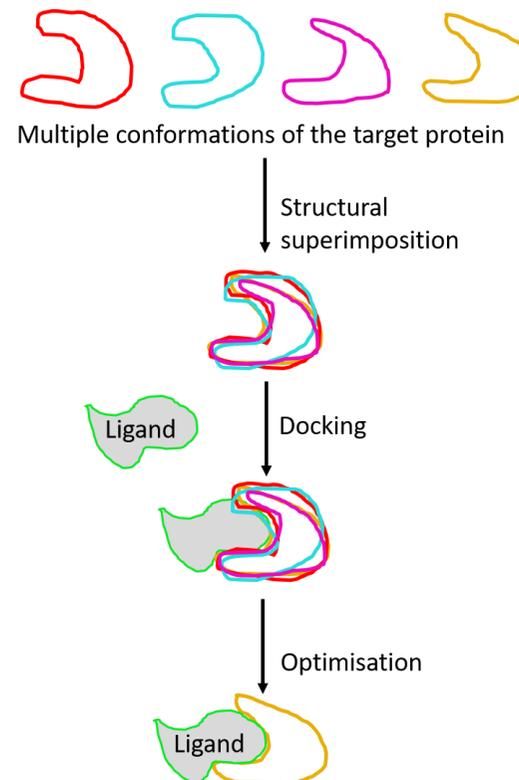


Figure 1. Cartoon illustration of the ensemble docking procedure. The ensemble consists of four conformations of the target protein coloured in red, cyan, magenta and orange.

¹Korb, O. *et al.* "Potential and Limitations of Ensemble Docking." *J. Chem. Inf. Model.* (2012). **52**, 1261-1274.

Case Study

Introduction

Thymidine kinase (TK) is the key enzyme in the pyrimidine salvage pathway catalysing the phosphorylation of thymidine to thymidine monophosphate (dTMP). In the cell, dTMP is then triphosphorylated and used as a DNA-building block. In contrast to the cellular enzyme, the viral thymidine kinase from *Herpes simplex virus type 1* (TK_{HSV1}) exhibits a broad range of acceptance for nucleosides that makes it of interest for the enzyme-prodrug gene therapy of cancer. Thus, dividing cells that express TK_{HSV1} convert nontoxic nucleoside prodrug into their active form which inhibits cellular polymerases leading to cell death and consequently tumour ablation. The broad clinical use of guanosine prodrug analogues has led to the emergence of drug-resistance and to the urgent need of novel series of potent and conformationally different drugs.

Your protein:

TK_{HSV1} has been determined in both unligated form (apo) and in complex with different nucleoside prodrug ligands. The unit cell is composed of a homodimer TK_{HSV1}, where the two asymmetric subunits are named A and B (Figure 2).

Your ligand:

2'-*exo*-methanocarpa-thymidine (MCT) is a potent antiviral compound characterised by high activity against HSV1 and HSV2. The co-crystallised complex TK_{HCV1}-MCT is available at 1.7 Å resolution (PDB code 1e2k). The thymidine ring of MCT is stacked between Met128 and Tyr172, and fixed by a complex hydrogen bonding network. Direct hydrogen bonds between one carbonyl and ammonia group of the nucleobase and the side chain of Gln125 and two water-mediated hydrogen bonds from one carbonyl of the nucleobase to the side chain of Arg176 are tightly fixing the nucleobase within the active site (Figure 3).

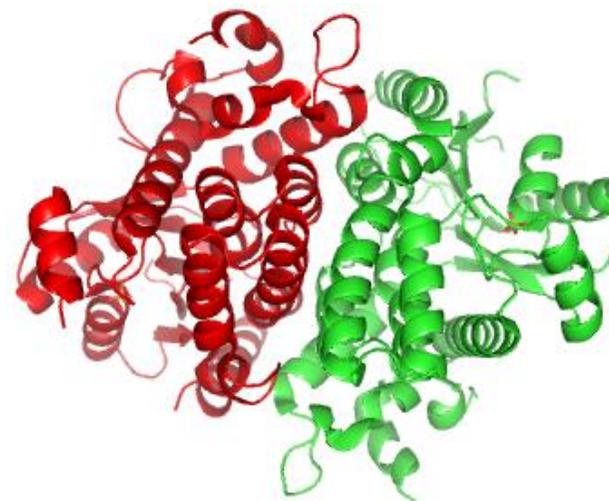


Figure 2. Crystallographic structure of TK_{HSV1} (apo). PDB code 1e2h. The two asymmetric subunits are coloured red and green.

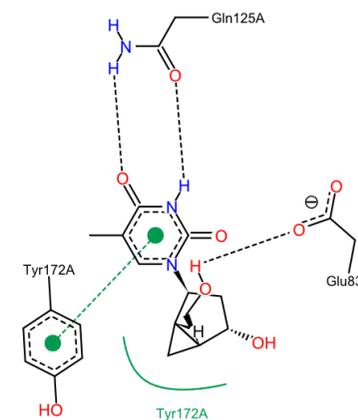


Figure 3. 2D interactions diagram of MCT in complex with TK_{HCV1}, PDB code: 1e2k.

Your task:

In this tutorial, we will use GOLD to perform a non-native docking of MCT (conformation extracted from the PDB entry 1e2k) into an ensemble of four different protein conformers of TK_{HCV1} (PDB entry codes 1e2k, 1e2i, 1of1 and 4ivq). This way we can investigate how this inhibitor would fit into the protein target by accounting for its flexibility.

This example assumes you are already familiar with how to setup protein(s) and ligands for docking calculations. If not, please refer to the following sections of the [GOLD User Guide](#):

- Setting Up the Protein
- Essential Steps

Challenges:

The experimentally determined crystal structures of TK_{HSV1} in complex with several ligands reveal a common binding site for different classes of nucleoside. TK_{HSV1} shows an extremely plastic binding site able to adapt a wide variety of purine and pyrimidine analogues. The binding site includes some water molecules that mediate hydrogen bonds between the nucleobase and the protein contributing to stabilisation of the ligand binding.

Crystal structures of TK_{HSV1} show that Gln125 can adopt two different conformations (Figure 4):

- A conformation where the amido group of Gln125 forms a dimer of hydrogen bonds with the ligand (see 1e2k and 1of1);
- A conformation where only one of these hydrogen bonds is formed and a water molecule mediates the other hydrogen bond between the protein and the ligand (see 1e2i). Interestingly, the apo structure of TK_{HCV1}, 1e2h (not shown in the figure) also contains a water molecule placed in a similar orientation as above.

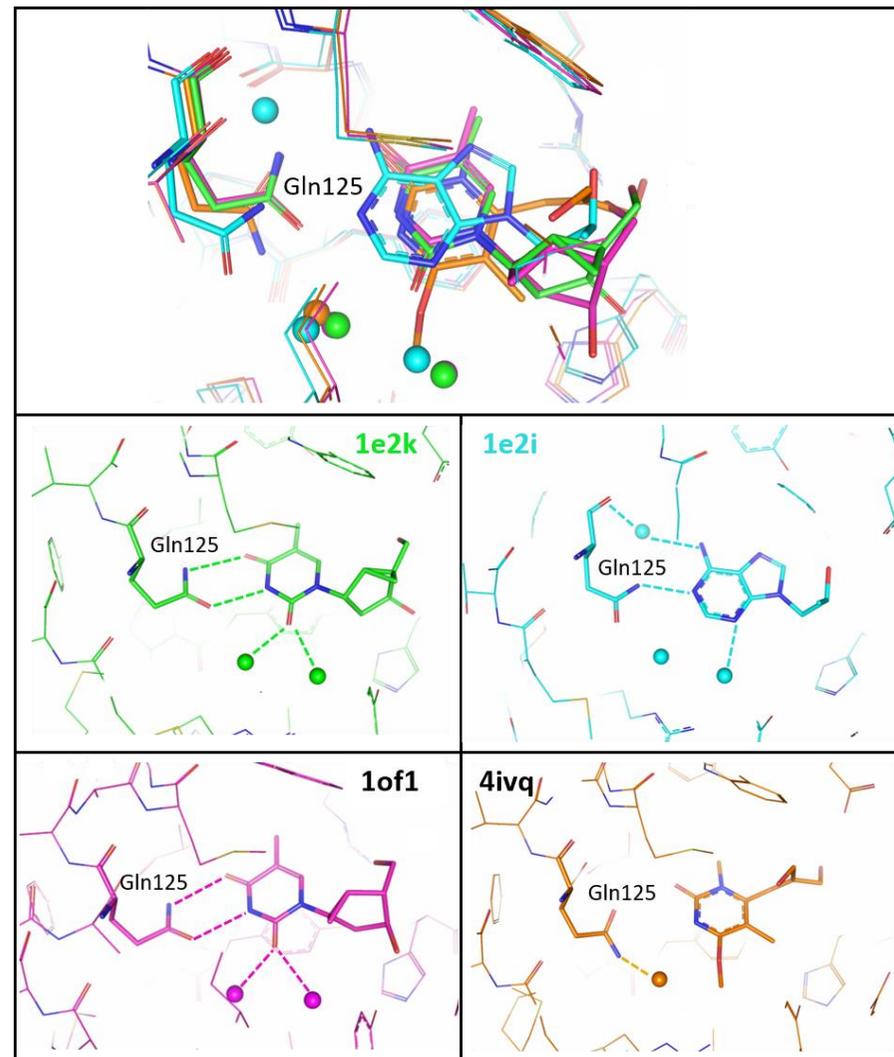


Figure 4: Superimposition of co-crystallised TK_{HCV1}-ligand complexes with the following PDB codes: 1e2k (green), 1e2i (cyan), 1of1 (magenta) and 4ivq (orange). Individual TK_{HCV1}-ligand complexes are also shown to highlight the hydrogen bond network involving Gln125, the nucleobase and water molecules. Water molecules in the binding site are represented as spheres and are coloured accordingly to the colour code used for the respective protein and ligand.

In this tutorial, we will perform a non-native ensemble docking of MCT into an ensemble of four TK_{HCV1} conformers.

In order to represent the two possible binding modes, three water molecules will be used during docking: two molecules conserved across all TK_{HCV1} models plus one molecule observed in 1e2i and in the apo structure (1e2h). The positions of these active waters will be explored to evaluate:

- How MCT binds to TK_{HCV1} and what is the preferred protein conformers of the four included in the ensemble.
- The displacement of the water molecules induced upon binding of MCT.

Provided Input Files

- One protein (*4IVQ.mol2*) will be used to guide you through the key steps required to prepare a protein for ensemble docking.
- The remaining three protein files provided (*1E2H.mol2*, *1E2I.mol2*, *1OF1.mol2*) have been prepared following the same steps.
- The ligand file (*1E2K_ligand.mol2*) has been set up in accordance with the guidelines for the preparation of input files (Setting Up the Protein(s) and Setting Up Ligands) and will be used to define the shape of the binding site as well as for predicting its binding conformation.
- The active waters (i.e. those that you would like GOLD to consider during docking) are provided as separate files (*water_1.mol2*, *water_2.mol2* and *water_3.mol2*).

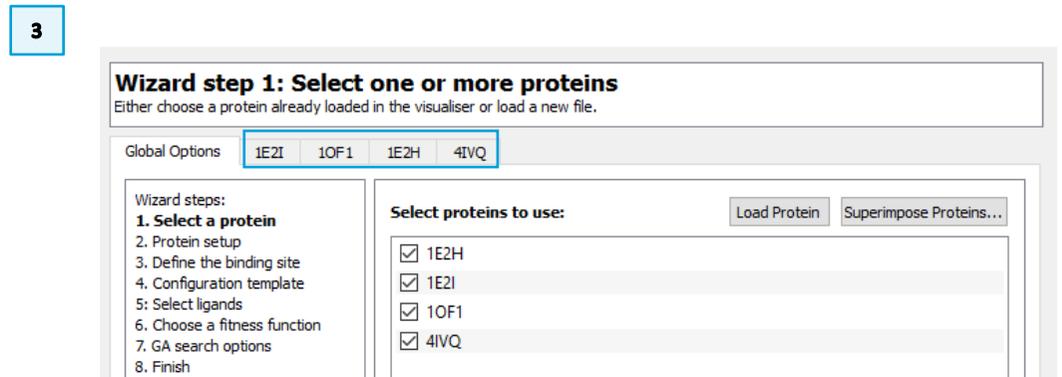
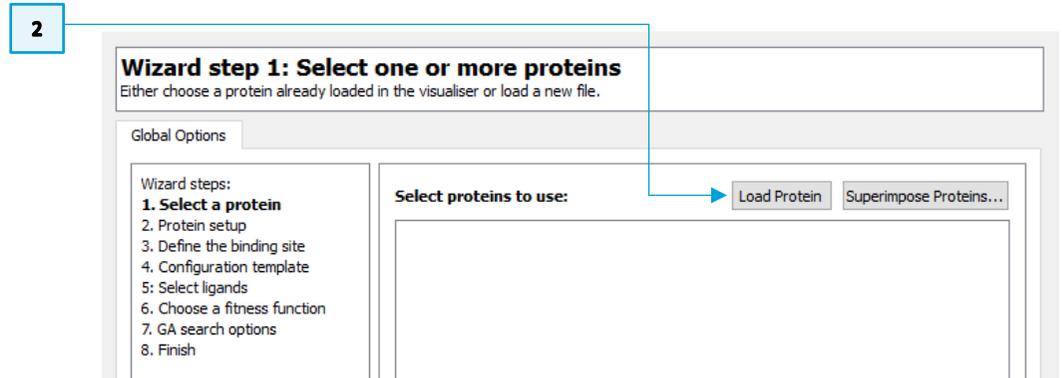
The original PDB complexes (1e2h.pdb, 1e2i.pdb, 1e2k.pdb, 1of1.pdb and 4ivq.pdb) are also available, should you wish to prepare the proteins and ligand from scratch.

Superimposing Protein Structures

An essential step of protein set-up for ensemble docking is that the proteins are superimposed. This is because there can only be a single binding site definition applicable across the whole ensemble. It is necessary to specify the approximate centre and extent of the protein binding site.

Brief details follow; complete details are provided in the [Hermes User Guide](#). A wizard is provided to facilitate protein superimposition. Proteins can be overlaid by matching residues based on label, matching residues based on sequence number or by matching residues based on sequence alignment. Optionally, a component of FASTA (called *ggsearch2*) can be used for the sequence alignment of the proteins to be superimposed. The package can be downloaded from http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml. In both cases above the wizard guides you through the superimposition process.

1. To access the wizard, launch GOLD  and click **Wizard** or launch Hermes , click on **GOLD** from the top-menu bar and select **Wizard...**
2. Click on **Load Protein** in *GOLD Setup* window and load the four protein files provided in this tutorial (*1E2H.mol2*, *1E2I.mol2*, *1OF1.mol2* and *4IVQ.mol2*), one at a time. Please note that whilst there is a maximum limit of 20 proteins when using ensemble docking, we do not recommend using more than 10 proteins.
3. Note that as each protein is separately added, a tab corresponding to that protein appears to the right of the *Global Options* tab and it is labelled with the name taken from the protein file e.g. *1E2H*.



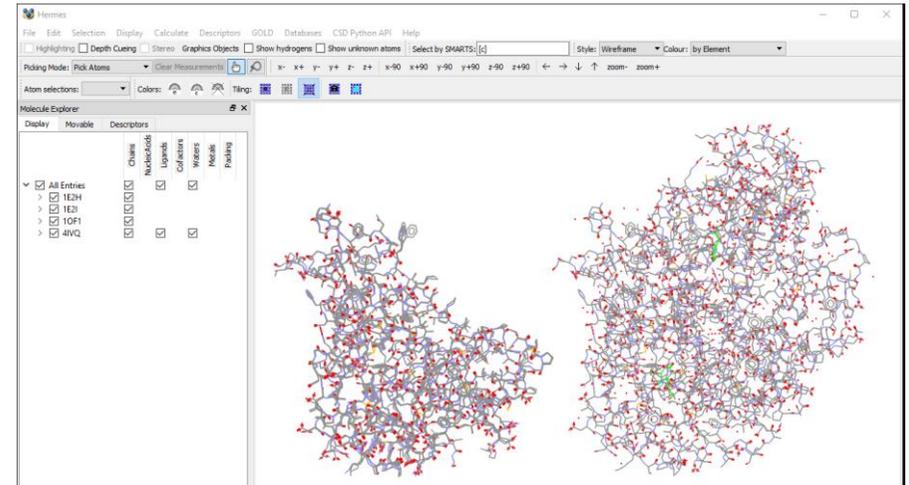
4. In Hermes 3D view, you will notice that three of the proteins (1E2H, 1E2I and 1OF1) are superimposed and have had hydrogen atoms added. The fourth protein, 4IVQ, has not been prepared: it is in a different frame of reference with respect to the other proteins, it has no hydrogen atoms, and still contains chain B and ligands.

5. To superimpose 4IVQ on top of one of the other three proteins, click **Superimpose Proteins** and then follow the onscreen instructions.

A wizard window will ask you if you want to use a component of the FASTA package or, if binaries can't be found, the default is to use Needleman-Wunsch algorithm. If you do not have FASTA installed, click **No** to use the default. Both FASTA and the Needleman-Wunsch algorithm do the same thing, i.e. they generate global sequence alignments which give a pair-wise matching of one residue to another and this can then be used for overlay.

6. Click to select 1OF1: A to use as the reference chain, and click on **Next** to proceed to the *Superimpose Proteins* dialogue.

4



5

Wizard step 1: Select one or more proteins
Either choose a protein already loaded in the visualiser or load a new file.

Global Options: 1E2I 1OF1 1E2H 4IVQ

Wizard steps:
1. Select a protein
2. Protein setup
3. Define the binding site
4. Configuration template
5. Select ligands
6. Choose a fitness function
7. GA search options
8. Finish

Select proteins to use:

- 1E2H
- 1E2I
- 1OF1
- 4IVQ

Buttons: Load Protein, Superimpose Proteins...

6

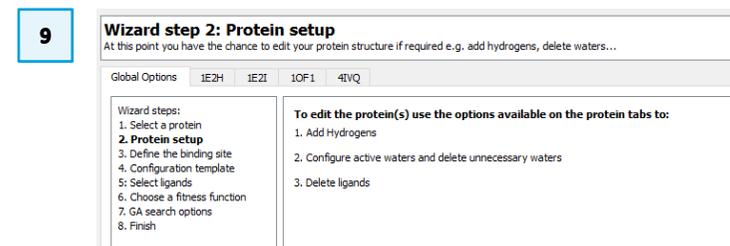
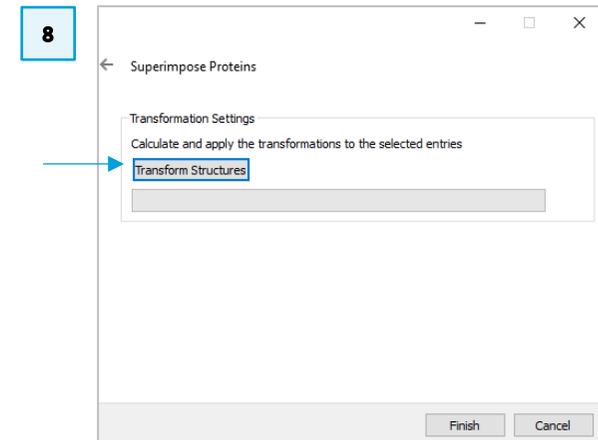
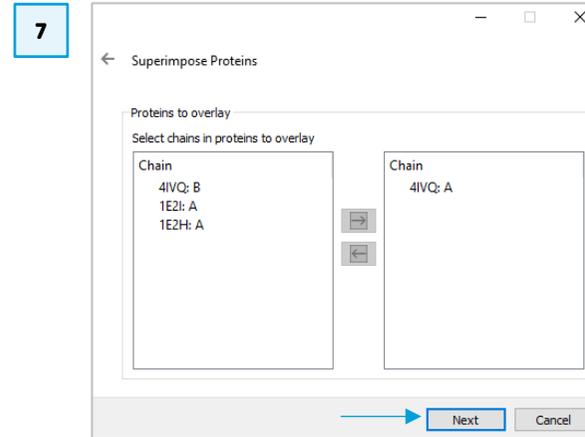
← Superimpose Proteins

Select target chain
Select a chain that will be used as the reference to superimpose one or more proteins

- 1
- 1E2H: A
- 1E2I: A
- 1OF1: A**
- 4IVQ: A
- 4IVQ: B

Buttons: Next, Cancel

7. Click to select 4IVQ: A, then click the right-arrow to choose to overlay this chain only. Then click **Next** to proceed. In the next pop-up window leave the **Use whole protein** activated and the default **Superimposition Weighting** value and click **Next**.
8. Click **Transform Structures** in *Superimpose Proteins* window to proceed to the overlay. When completed click on **Finish**.
9. Return to the *GOLD Setup* window. Click **Next** to proceed to *Protein setup* step to additionally edit the protein structure(s) if required.

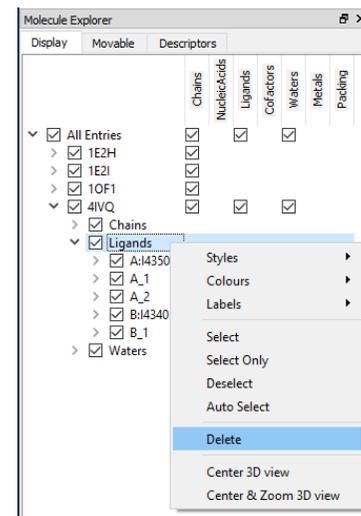


Protein Setup

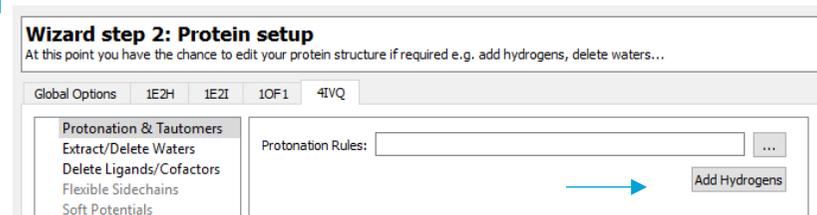
In the remainder of this section we will describe the steps required to prepare 4IVQ for docking:

- In the *Molecule Explorer* (off to the left-hand side of the Hermes interface) click on the “>” adjacent to 4IVQ and underneath *All Entries*. Right-click on *Ligands* and select **Delete** from the pull-down menu.
- Similarly, to delete Chain B, click on “>” adjacent to 4IVQ and then on “>” adjacent *Chains*. The two different chains (A and B) are shown; right-click on *B* and select **Delete** from the pull-down menu.
- Return to the *GOLD Setup* window and click on the 4IVQ adjacent to the *Global Options* tab. From within the 4IVQ tab, add hydrogen atoms to the protein by selecting **Add Hydrogens** from the first *Protonation & Tautomers* option in the Wizard.
- Still in 4IVQ tab, move to the next option by clicking on **Extract/Delete Waters**. From within this dialogue it is possible to specify water molecules that mediate protein-ligand interactions (active water), and to delete those that are not required. Since we don't want to extract any waters from this structure, click **Delete Remaining Waters**. When prompted *Are you sure you want to delete all waters?* click **OK**. You will be informed that 203 water molecules have been deleted.
- Click **Next** to proceed to *Define the binding site* tab of *GOLD Setup* window.

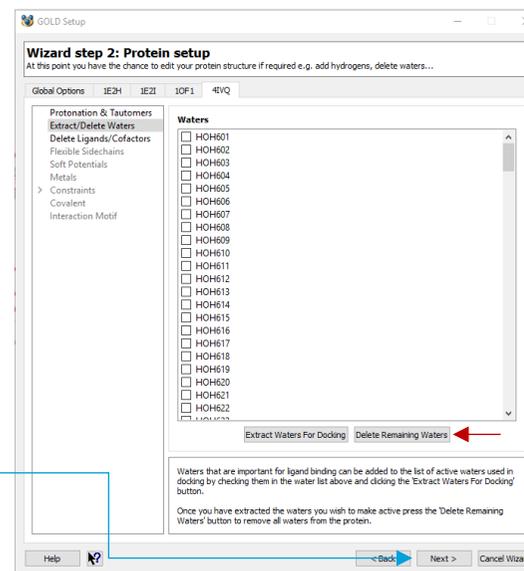
10



12



13



14

Exploring the Ligand for Ensemble Docking

Defining the Binding Site

Since the binding site definition for an ensemble must be a position suitable for all proteins, it is not possible to define the binding site from an atom or a list of atoms or residues. It is only possible to define the binding site from a point in space or from a ligand.

1. Load the reference ligand file *1E2K_ligand.mol2* in Hermes by clicking on the main menu option **File** and then **Open** from the resultant pull-down menu. This loads the ligand in the 3D view and makes the ligand available for binding site definition.
2. While on *Define the binding site* step, in the GOLD Wizard, click to activate the *Ligand* radio button. Select *A_1:TCM500, 1E2K_ligand* to determine the binding site. Leave the default all atoms within 6 Å of the ligand for the binding site definition. It can help here to switch off the display H-atoms using the **Show hydrogens** tick box in the top-level menu of Hermes. Carbon atoms outside of the binding site will turn purple. Click **Next** to proceed.
3. At this point you are giving the option to load a configuration file template. These templates can be used to load recommended settings for a number of different types of docking protocols (see [GOLD User Guide](#)). In this example, we will specify all docking settings manually. Click **Next** to proceed to the *Select ligands* step in GOLD wizard.

2

Wizard step 3: Define the binding site
The binding site can be defined by several different ways: an atom, a point or a reference ligand. Atoms can be selected in the visualiser.

Global Options: 1E2H 1E2I 1OF1 4IVQ

Wizard steps:
1. Select a protein
2. Protein setup
3. **Define the binding site**
4. Configuration template
5. Select ligands
6. Choose a fitness function
7. GA search options
8. Finish

Atom - select an atom in the visualiser or enter an atom index
View

Point - select atoms to define a centroid or edit XYZ
X: Y: Z: View Reset

Ligand
A_1:TCM500, 1E2K_ligand

List of atoms or residues
Filename: ... View

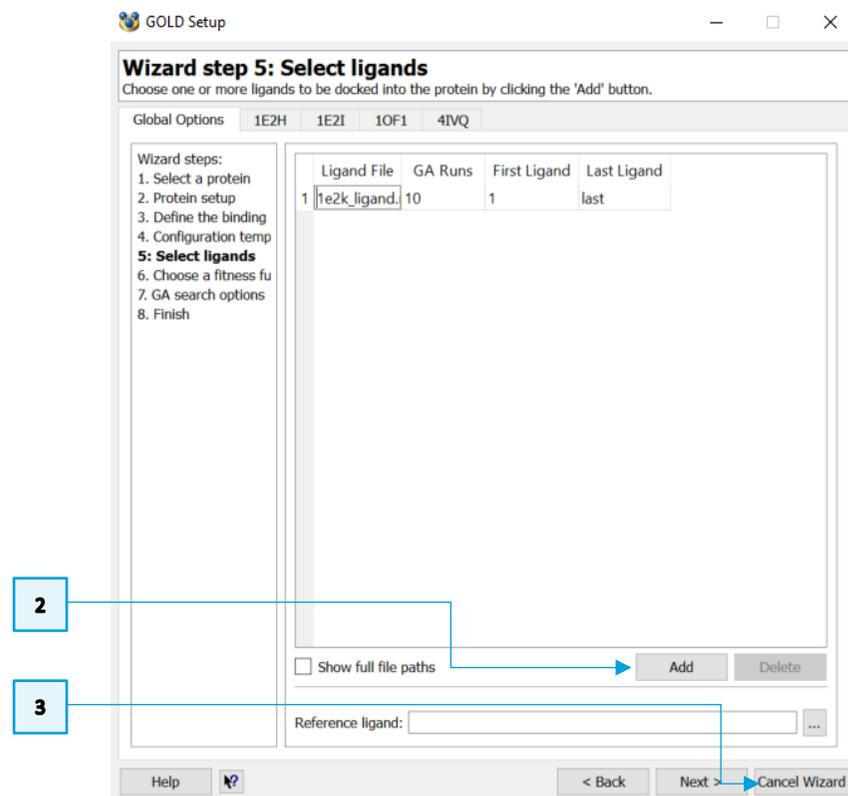
Select all atoms within Å
 Generate a cavity atoms file from the selection Refine Selection
 Detect cavity - restrict atom selection to solvent-accessible surface
 Force all H bond donors/acceptors to be treated as solvent accessible
Add Definition as a Selection

Help ? < Back **Next >** Cancel Wizard

Selecting Ligand(s) for Docking

To proceed with the GOLD setup, we must specify the ligand that we want to dock in the ensemble of TK proteins.

1. The ligand is provided in the tutorial folder (*1E2K_ligand.mol2*). As with the protein file, all hydrogen atoms must be present in the ligand file. We have already added the hydrogen atoms to the ligand.
2. Specify the ligand by clicking the **Add** button at the bottom of the *GOLD Setup* window. Navigate to the folder to which you copied the tutorial files, select *1E2K_ligand.mol2* then click **Open**. The *1E2K_ligand.mol2* is now listed under *Ligand File*.
3. *GOLD Wizard* provides the key steps for docking; however, more advanced options (*i.e. Configure Waters*) are available outside the main Wizard. To access these advanced options, click **Cancel Wizard**.



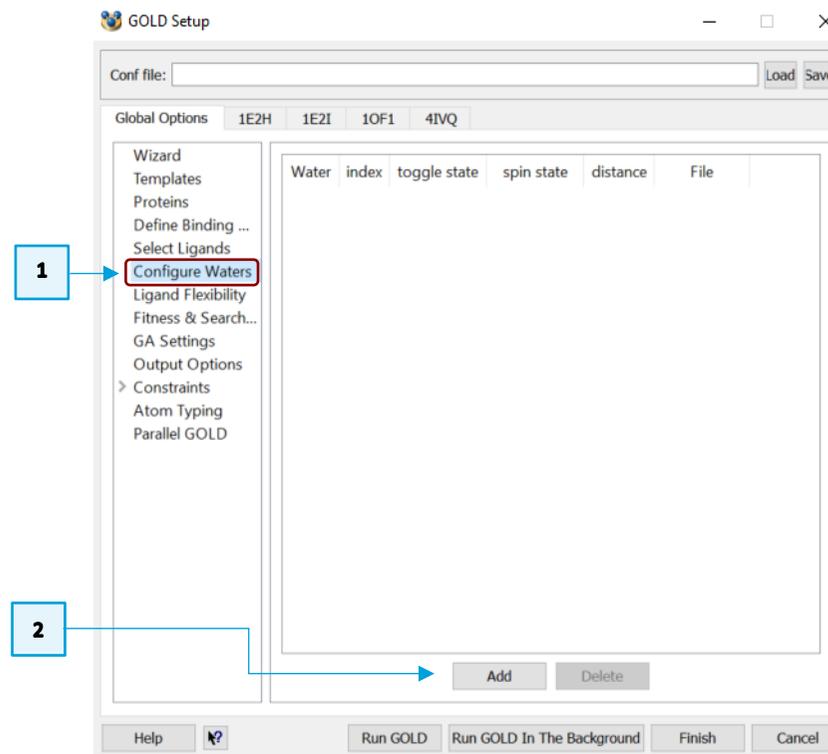
Setting Water in the Binding Site (Active Waters)

Before proceeding with the docking setup, we must define the active waters that we want to be considered during the ensemble docking.

Three active water molecules have been detected: two waters that coordinate H-bonds between the thymidine ring and Arg176 of TK_{HSV1}, and a third water molecule that coordinates the hydrogen bonds between the nucleobase of the ligand and the side chain of Gln125. This water molecule can compete with the thymidine ring of MCT to form direct hydrogen bonds with Gln125.

The active waters must be provided in separate files, one water molecule per file. You can find these files in the tutorial folder.

1. Pick **Configure Waters** from the list of available options in *GOLD Setup* window.
2. The dialogue is empty, so we need to specify our water molecules by reading in the water files. To do this click on the **Add** button, select the three water files then click **Open**.
3. The water molecules will be listed in the *Configure Waters* dialogue. By default, each water molecule in *Configure Waters* list will be retained in the binding site during docking and will be allowed to spin and toggle to optimise the position of the molecule and the orientation of the hydrogens. These settings can be customised for specific water molecules within the *Waters* dialogue in the *GOLD Setup* window.
 - *Toggle* state leaves GOLD to decide whether the water should be present or absent (bound or displaced by the ligand) during the docking.
 - *On* sets the water to be always present in the binding site and allows the hydrogen positions to vary during docking in order to maximise the



3

Water	index	toggle state	spin state	distance
1 water_1	1	toggle ▼	spin ▼	0
2 water_2	1	toggle ▼	spin ▼	0
3 water_3	1	toggle ▼	spin ▼	0

hydrogen bonding score both from interactions with the protein and the ligand.

- The *Off* water state option allows a water to be removed from consideration during docking.

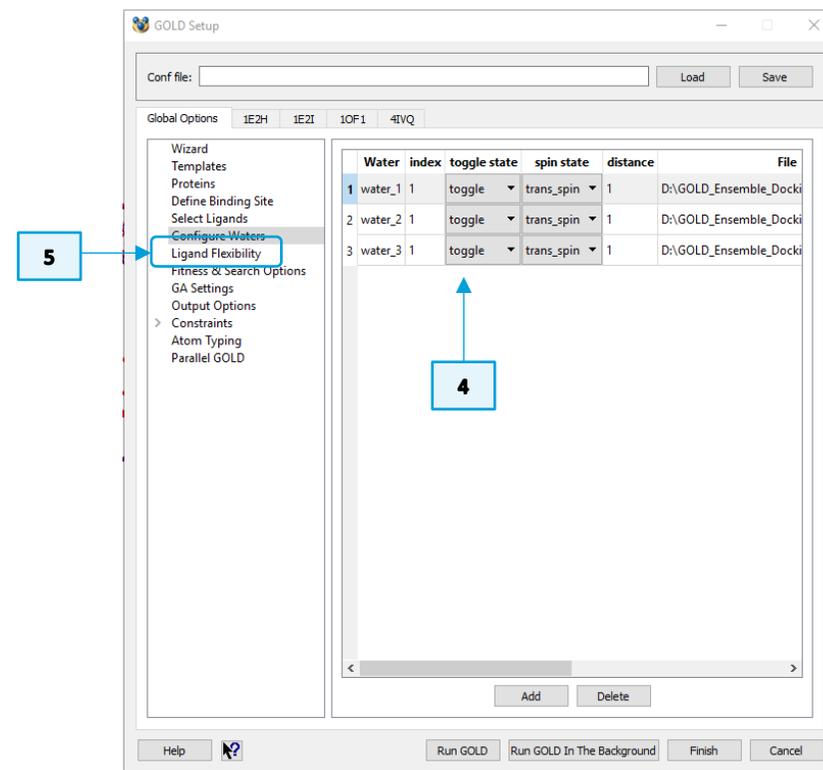
Leave the **toggle state** as *toggle* for this tutorial. This means GOLD will decide whether the waters should be present or absent.

- The orientation of the waters can be also changed.
 - Activating the *spin* option makes GOLD automatically optimise the orientation of the hydrogen atoms.
 - If you activate the *trans_spin* option and input a translation value into the distance dialogue, this will make GOLD spin and translate the water molecule to optimise the orientation of the hydrogen atoms as well as the water molecule's position within a defined radius. Note that the distance value must be between 0 and 2 Å.
 - Activating the *fix* option makes GOLD using the orientation specified in the input file.

Set the **spin state** to *trans_spin* from the dropdown menu. Set the **distance** to 1 Å, by double-clicking in the box and typing "1". This means that the waters are allowed to translate up to 1 Å.

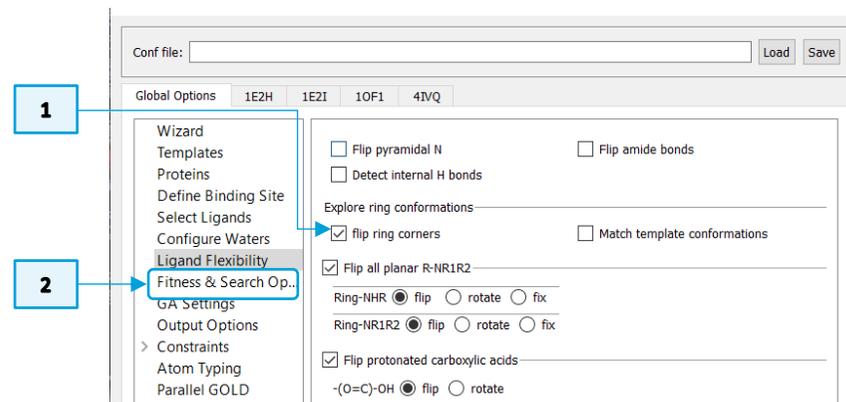
After docking, a summary of which waters were retained or displaced and their contribution to the fitness score can be found in the *Analysis of active water placements* section of the `gold_1E2K_ligand_m1.log` file.

- Click on *Ligand Flexibility* to continue with the GOLD Setup.



Setting Ligand Flexibility

1. Activate the **flip ring corners** tick-box to allow GOLD to perform a limited conformational search of cyclic systems by allowing free corners of the rings in the ligand to flip above and below the plane of their neighbouring atoms.
2. Click on *Fitness & Search Options* to continue with the GOLD Setup.



Setting Docking Parameters

Selecting a Fitness Function

1. During the docking run the solutions found by GOLD are scored according to a fitness function. Ensure that the default *CHEMPLP* scoring function is selected in *Fitness & Search Options* dialogue.

2. By default, **Allow early termination** check box is switched on. Switch it off by deactivating the tick box next to *Allow early termination*. This will ensure that as many solutions as possible are explored.

3. Click on *GA Settings* dialogue.

Specifying GA Settings

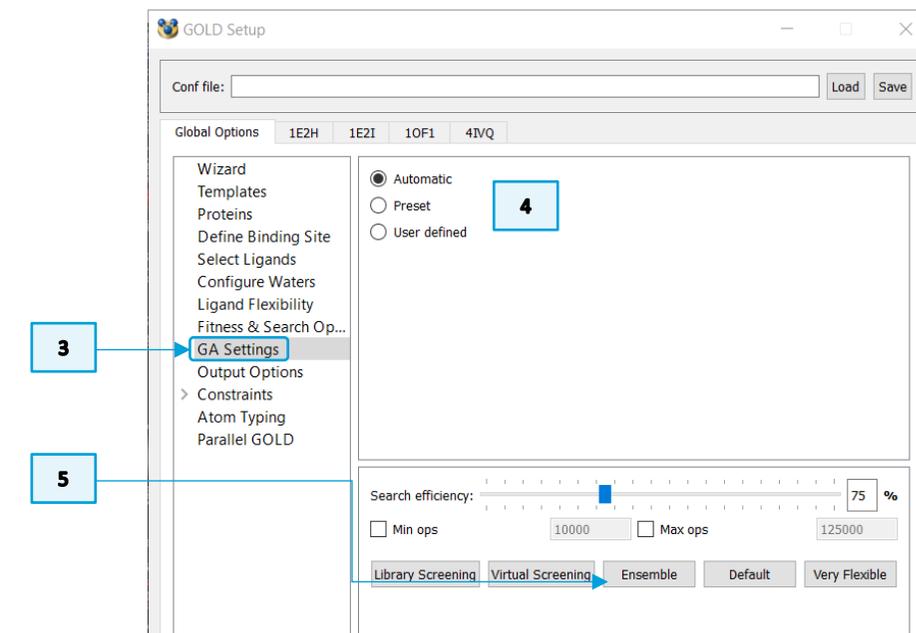
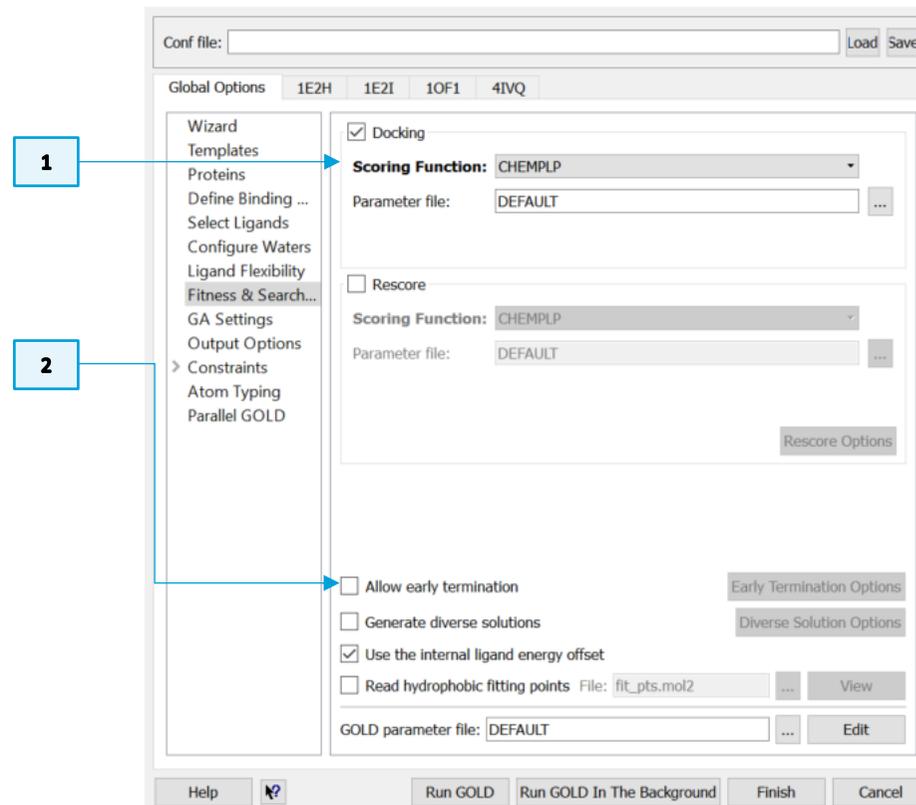
GOLD optimises the fitness score using genetic algorithm (GA). A number of parameters control the precise operation of the genetic algorithm. These settings are encapsulated into three speeds:

- Slow (most accurate): this equates to 100,000 GA operations
- Medium: 50,000 operations
- Fast (least accurate): 10,000 operations

There is a trade-off between speed and reliability. The fewer options, the faster the docking, but the search space will be less explored.

4. Enable automatic GA setting by clicking the **Automatic** radio button and ensure the *Search efficiency* is set to 100%. This will make GOLD automatically calculate an optimal number of operations for a given ligand, thereby making the most efficient use of the search time.

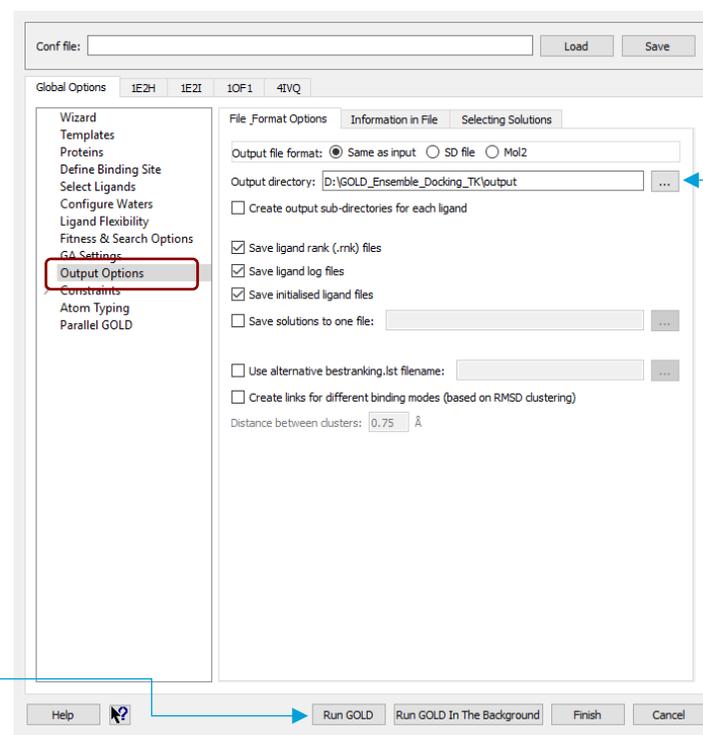
5. Click on the **Ensemble** button. This will set the search efficiency at 75% and it is recommended for ensemble docking.



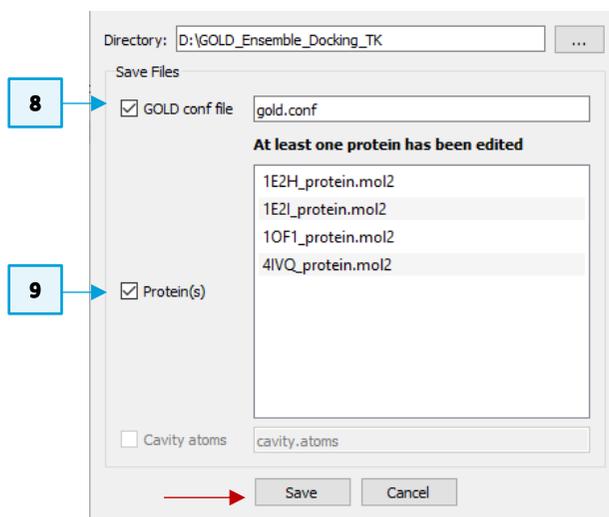
Run the Ensemble Docking

6. Before starting the run, select *Output Options*. Click on the ... button next to *Output directory* and specify a directory to which you have permission to write; this is where the GOLD output files will be written.
7. We have now finished setting up our docking. Click the **Run GOLD** button at the bottom of the GOLD interface. You will be presented with a *Finish GOLD Configuration* window containing *Save Files* options.
8. Ensure the *GOLD conf* file tick-box is activated and rename the *.conf* file as *gold_ensemble_TK.conf*.
9. Ensure that Protein(s) tick-box is activated. We want to save the edited 4IVQ structure. Note that all proteins will be saved, including the not edited ones. Click **Save** to start the docking.

6



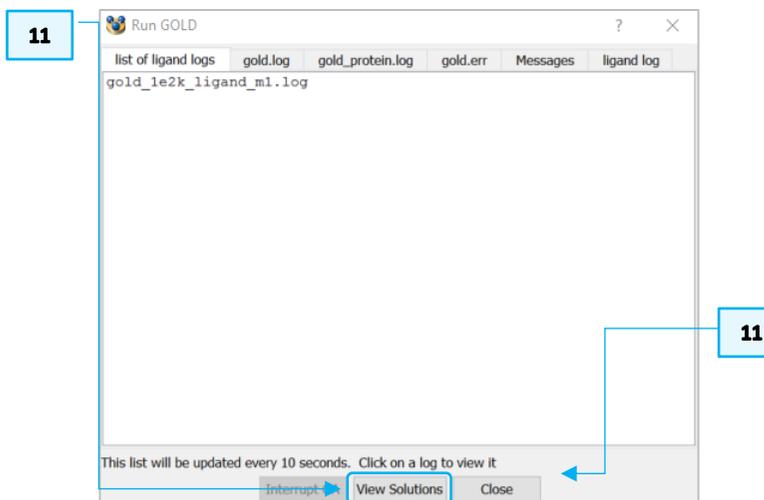
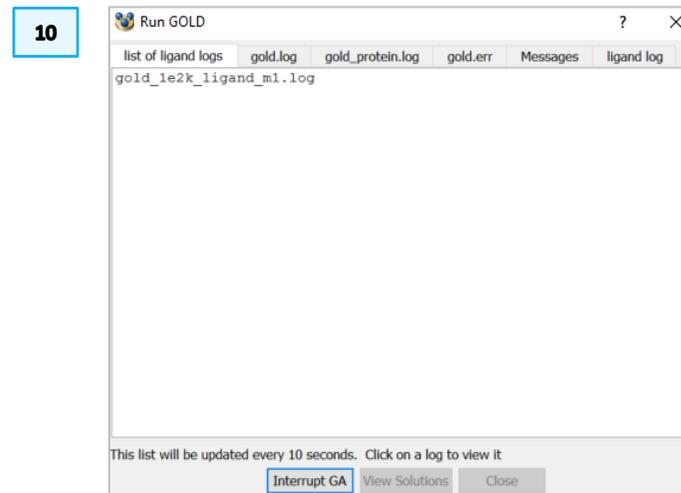
7



8

9

10. As the job progresses output will be displayed in several tabs in the *Run GOLD* window.
11. Once the job is complete, start by clearing all initial files in Hermes by going to the Hermes menu **File** then **Close All Files**. Then, return to the *Run GOLD* window and load the docking results into Hermes by clicking on the **View Solutions** button. We have finished with the *Run GOLD* window now, so close the window by clicking on the **Close** button. In the *GOLD Setup* window, click on **Cancel** button to close this window as it is no longer needed.



Analysing Docking Results

Visualising Docking Results

1. Return to the Hermes 3D view and look at the *Docking Solutions* tab in *Molecule Explorer*. Use the Up and Down arrows on your keyboard to change between docking solutions. Let us think about what results we can expect:

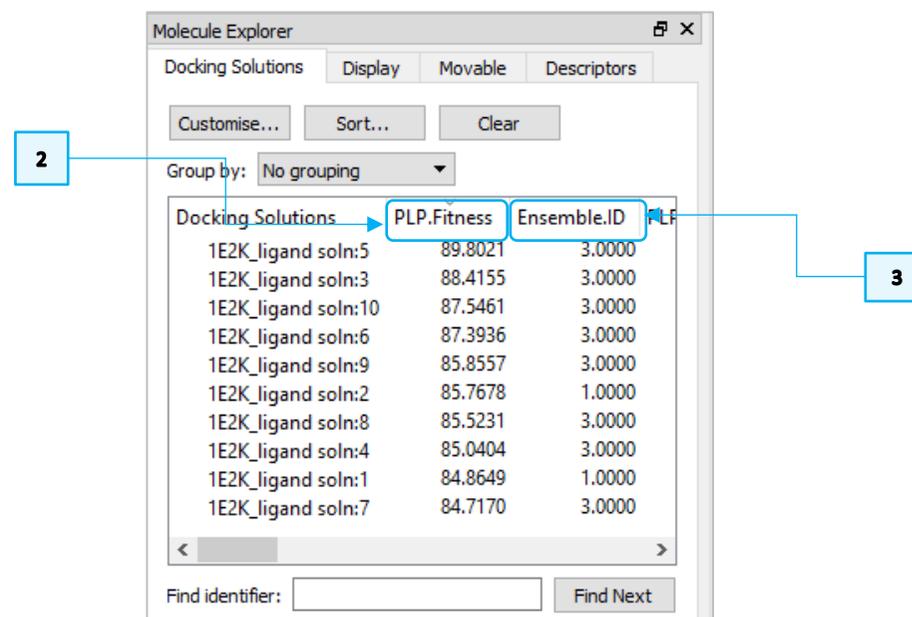
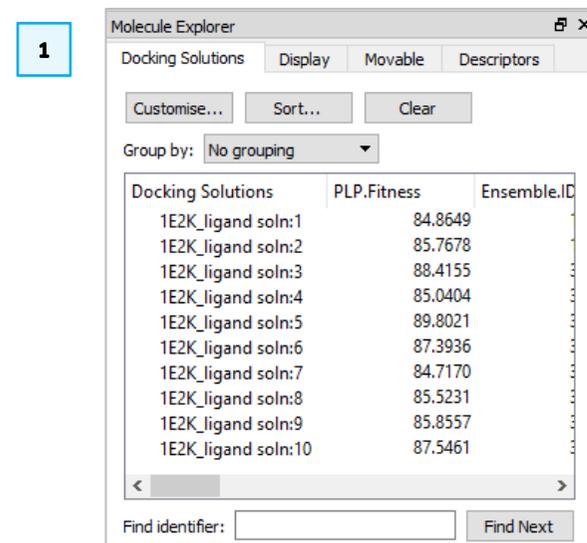
We chose to dock one ligand into four proteins. Starting from a superimposed set of protein structures, GOLD evolves a separate population of individuals (representing ligand conformations) for each protein structure that is part of the ensemble. The best ligand conformation found in any of the ensemble structures is returned. For example, if for a given GA run a ligand gets the scores 10 in protein 1, 20 in protein 2 and 15 in protein 3, protein 2 will be selected.

2. The docking solutions are given in their docked order with their corresponding fitness score listed under the column headed *PLP.Fitness*. If required, the solutions can be ordered by clicking on this *PLP.Fitness* header to determine which is the highest scoring.

Please note: Due to the non-deterministic nature of GOLD your results may vary from those described in this tutorial.

3. We have obtained 10 docking solutions as this is the default number of how many times our ligand was docked. The protein that the solution corresponds to may be one of four, identifiable by the ensemble index number (1-4).

The initialised protein is given a filename of the type *gold_protein_<ensemble_index>.mol2*, where the numbers correspond to the order in which the proteins are loaded. This index number is given in the docking solutions pane in Hermes as the column headed *Ensemble.ID*, next to the fitness score so you can see to which protein each solution corresponds.



4. GOLD gives best docking results for proteins 3 and, to a lesser extent, protein 1: 1OF1 and 1E2H, respectively.
5. The fact that protein 1OF1 (showed in magenta) gives the best docking results is not surprising considering that its co-crystallised ligand is the most structurally similar to MCT and so is its binding mode (shown in green).
6. By inspecting the binding mode, we can confirm that MCT forms two hydrogen bonds with the sidechain of Gln125 which in the 1OF1 model displaced one of the three water molecules.
7. The second-best scored model corresponds to 1E2H (showed in cyan), where MCT binds in a different way to the sidechain of Gln125 (with a single H-bond), and also displaces one of the three water molecules.

5

Molecule Explorer

Docking Solutions Display Movable Descriptors

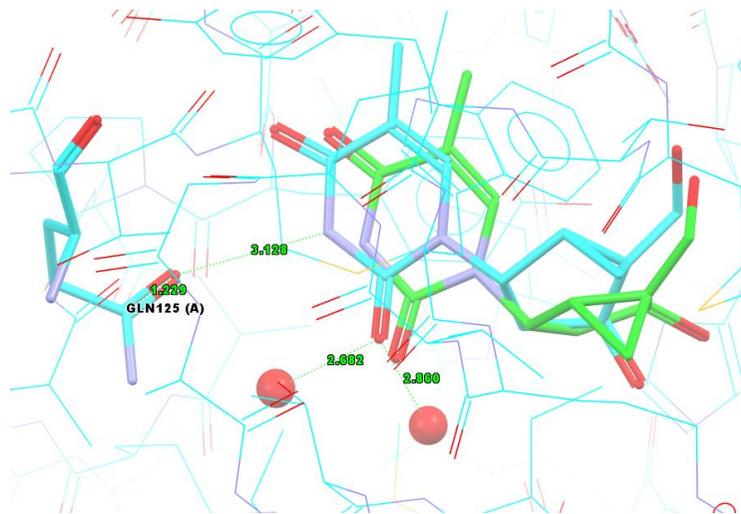
Customise... Sort... Clear

Group by: No grouping

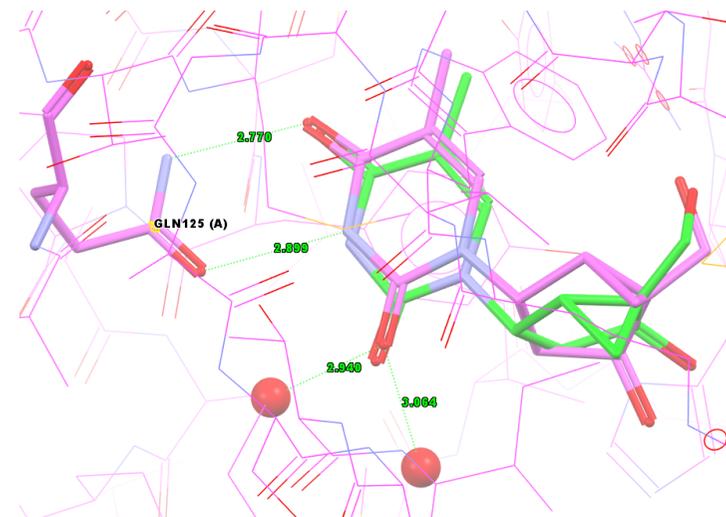
Docking Solutions	PLP.Fitness	Ensemble.ID	PLP
1E2K_ligand soln:5	89.8021	3.0000	←
1E2K_ligand soln:3	88.4155	3.0000	
1E2K_ligand soln:10	87.5461	3.0000	
1E2K_ligand soln:6	87.3936	3.0000	
1E2K_ligand soln:9	85.8557	3.0000	
1E2K_ligand soln:2	85.7678	1.0000	←
1E2K_ligand soln:8	85.5231	3.0000	
1E2K_ligand soln:4	85.0404	3.0000	
1E2K_ligand soln:1	84.8649	1.0000	
1E2K_ligand soln:7	84.7170	3.0000	

Find identifier: Find Next

7



6



Conclusions

- The crystallographically observed conformation of the 1E2K_ligand can be compared to the poses found when docking this ligand to the ensemble of TK_{HSV1}.
- The best ranking pose obtained in protein model 3 (i.e. 1OF1) reproduced the crystallographically observed conformation of the ligand.
- By allowing water molecules to rotate, translate and turn on and off GOLD sampled different protein models, identifying the correct ligand binding mode in the different protein conformations.