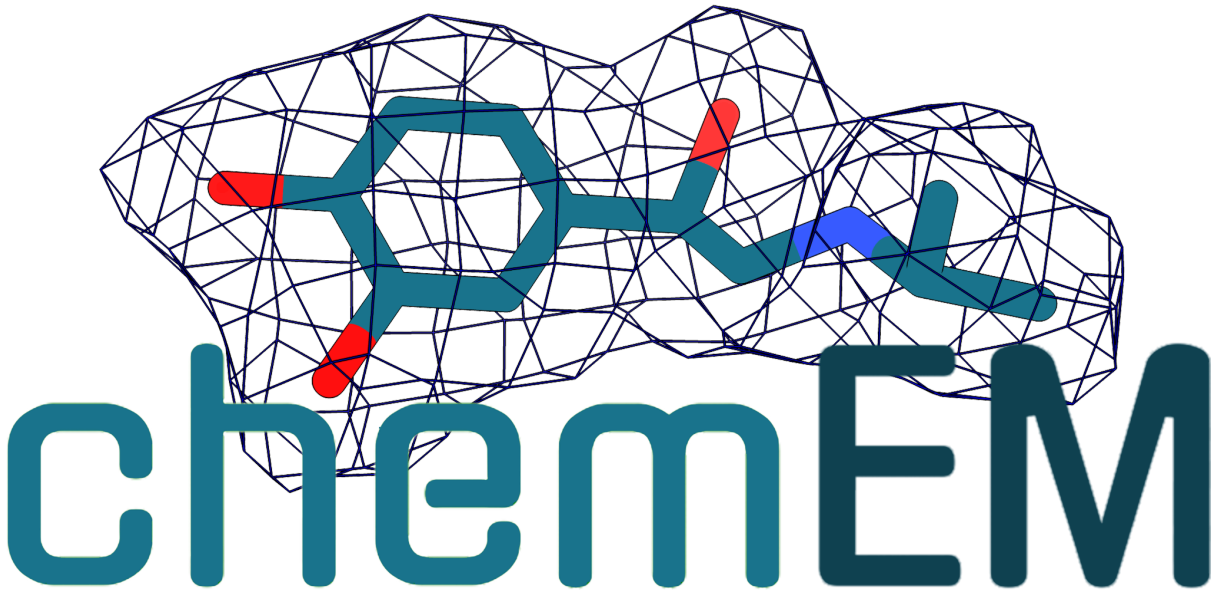
chemEM

ChemEM – Documentation v0.3.0

1.0 Introduction..... 2
1.1 Contact Information..... 4
1.2 What's New in v0.3..... 4
2.0 Getting Started..... 5
2.1 Installation..... 5
2.2 Quick Start Guide..... 6
2.3 The configuration file..... 6
2.3.1 Required data..... 6
2.3.2 Enable Stages..... 8
2.3.3 Pre-processing parameters..... 10
2.3.4 Fitting parameters..... 11
2.3.5 Post-processing parameters..... 13
2.4 Protonation, Chirality and Rings..... 15
2.4.1 Protonation..... 15
2.4.2 Rings..... 16
2.4.3 Chirality..... 16
2.5 Force Fields..... 16
2.6 Surface residues definition..... 17
3.0 Output files..... 17
3.1 Pre-processing..... 17
3.2 Fitting..... 18
3.3 Post-processing..... 19
3.4 Protonation..... 19
4.0 Examples and tips..... 20
4.1 Defining a centroid..... 20
4.2 Improving a difference map..... 21
4.2.1 Auto split point..... 21
4.2.2 Box size..... 24
4.2.3 Thresholds..... 25
4.3 Docking with no cryo-EM map..... 27
4.4 Rescoring ligands..... 27
4.5 Multi-ligand docking..... 28
5.0 References..... 30

1.0 Introduction

ChemEM (Sweeney et al. 2024) docks small molecules to cryo-EM maps using a two stage approach (Figure 1). In Stage 1 an approximate fit of a ligand into the protein structure is generated using a difference density map between the protein-ligand complex map and a simulated protein map.

An initial fit is achieved by molecular docking using an empirical scoring function integrated with the MI score. This ensures that the generated conformation is both physico-chemically acceptable and well-fitted to the difference map. In this stage, the difference map acts as a constraint for the docking algorithm, reducing the conformational space to be searched.

In Stage 2 (Figure 1), candidate conformations are refined into the full cryo-EM density map using a flexible fitting approach with the AMBER forcefield-14SB19 protein parameters and the SAGE OpenForceField 2.0.020 parameters for small molecules. This stage fine-tunes the initial fits generated in Stage 1, whilst simultaneously refining the fit of binding-site atoms.

ChemEM is a software platform specifically designed to simplify and automate the process of fit-ting small molecules into cryoEM maps. To begin using the software, users must provide four key elements: a fitted protein structure in PDB format, a ligand file in SDF, Mol2, or SMILES string format, a cryoEM density map in MRC format, and the centre point of the binding site to expedite the cal-culation of difference maps (Figure 1).

Following the provision of these inputs, the soft-ware's pipeline is initiated, requiring no further in-tervention from the user. However, in cases where certain structures necessitate more comprehensive assessment, ChemEM provides the flexibility for users to examine the preprocessed difference maps before progressing to the docking phase.

In complex scenarios, such as when two or more ligands are being fitted simultaneously to the same binding site, ChemEM allows users to manually assign disconnected densities to a specific ligand input. Furthermore, if high-quality dif-ference maps cannot be generated using the default parameters, users have the option to adjust the number of protein atoms and the density of the map that are considered in the difference map cal-culations. This offers an effective method to opti-mise the map for fitting.

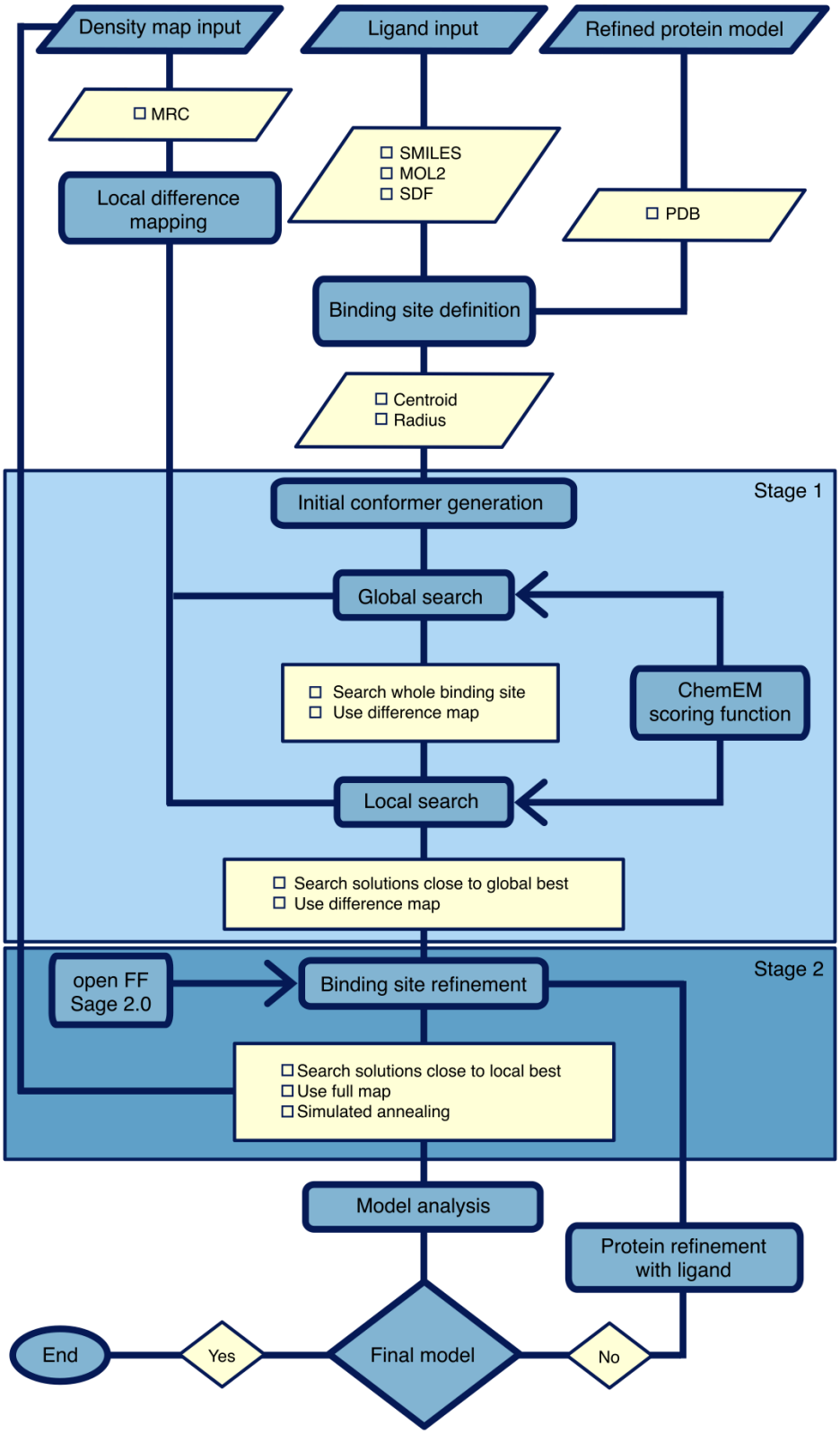

Figure 1. Schematic overview of the ChemEM methodology for docking small molecules, also in combination with cryo-EM maps.

1.1 Contact Information

For bug reports, feedback, or feature suggestions, please contact:

- **Aaron Sweeney**
    - **Email: aaron.sweeney@cssb-hamburg.de**

For inquiries regarding licensing, commercial or otherwise, please contact:

- **Maya Topf**
    - **Email: maya.topf@cssb-hamburg.de**

1.2 What's New in v0.3

Updates in ChemEM v0.3 included more elegant and informative error handling, along with introducing some new functionalities.

– user supplied difference maps

ChemEM now provides a mechanism for submitting user supplied difference maps for docking.

– protonation state assignment

The protonation state of ligand smiles strings is now assigned using di-Morpholite (Ropp et al. 2019).

– ions introduced

We have introduced the ions Ca, Cl, Co, Cu, Fe, K, Mg, Mn, Na, Ni, Zn to the algorithm. For now ions remain stationary during the fitting and post-processing stages but are taken into account in scoring functions if they are located within the binding sites of ligands being fitted.

–Hold atoms

The hold-fragments feature allows users to post-process solutions generated by ChemEM while restraining specific atoms.

– System ligands

The add system ligands feature enables the integration of ligands into the system whose positions and conformations are pre-determined. These ligands remain stationary during the fitting and post-processing stages but are taken into account in scoring functions if they are located within the binding sites of ligands being fitted.

– Surface residue definition

ChemEM now defines surface residues for docking using a SASA based method with the Shrake & Rupley algorithm (Shrake and Rupley 1973) calculated at the residue level.

–docking radius parameter

Binding site residue definition has changed in v0.3 (see section 2.3.4)

– Force Fields

We now include a way to different Amber force fields for ion parameters and implicit solvent models.

– Tests

Please note that the **test function has been deprecated** as of ChemEM version 0.0.2. We now require users to download the latest test data and example configuration files.

For your convenience, these files can be accessed and downloaded from the following URL:
gitlab.com/topf-lab/chemem_test_data/-/archive/main/chemem_test_data-main.zip
Alternatively, if you are familiar with Git, you can clone the repository using the following command in your terminal or command prompt:
git clone https://gitlab.com/topf-lab/chemem_test_data.git

2.0 Getting Started

2.1 Installation

ChemEM is currently distributed via anaconda. To begin using ChemEM please download anaconda or mini-conda (https://www.anaconda.com/download).

In the terminal Initiate a new python environment with anaconda to run ChemEM:

```
conda create --name <env_name> python=3.11
```

!?Note – currently ChemEM v0.0.3 is only available with python 3.11

Activate the new environment with

```
conda activate <env_name>
```

Install mamba package manager with:

```
conda install -c conda-forge mamba
```

Next install the ChemEM and it's dependencies with the command:

```
mamba install -c chemem -c conda-forge chemem
```

2.2 Quick Start Guide

Once the installation is complete and you should be able to use ChemEM from the command line by specifying a configuration file with the necessary data like so:

```
chemem <configuration_file>
```

To test the installation you can download the test data with the command:

```
git clone https://gitlab.com/topf-lab/chemem_test_data.git
```

2.3 The configuration file

The ChemEM software requires the use of a configuration file as an input for fitting small molecules to cryoEM data, molecular docking, rescoring, and post-processing.

2.3.1 Required data

The minimum require data to run ChemEM is shown below:

```
#ChemEM config file
protein = ~/test_data/7jjo_protein.pdb
ligand = CC(C)NC[C@@H](c1ccc(c(c1)O)O)O
centroid = (134.087, 133.507, 174.180)
output = ~/some_directory
densmap = ~/test_data/7jjo.mrc
resolution = 2.6
```

– Protein
The path to a protein file in PDB format (as of ChemEM v0.1.0, no water molecules or should be present).

– Ligand
The ligand to be fit as a SMILEs string. Alternatively the ligand can be given as a .sdf or .mol2 file:

```
ligand = ~/test_data/7jjo_ligand.sdf
ligand = ~/test_data/7jjo_ligand.mol2
```

Additionally, for rescoring ligands can be specified using the *ligands_from_dir* option with a directory path containing ligands in .mol2 or .sdf file format:

```
Ligands_from_dir = ~/test_data/
```

– Centroid
The centroid of the binding site the ligand should be fit to.

– Output
The directory to write output files.

The above are all the necessary files needed to run docking, re-scoring or post-processing without Cryo-EM data.

To include cryo-EM data two further options are needed.

– Densmap
The path to the cryoEM density map in .mrc format.

– Resolution
The resolution of the density map

In addition to the densmap and resolution data, the map contour value may also be specified:

```
map_contour = 3.0
```

If this option is not given the map_cotour will be calculated automatically by ChemEM.

– Difference map
The user can supply a difference map to be used in docking calculations, by specifying the map file:

```
differece_map = <difference map path>
```

!? – Note : if you supply a user defined difference map ChemEM will ignore all commands to create one including the auto-split commands.

2.3.2 Enable Stages

The next part of the configuration file that is required is the enabled stages section:

```
#enabled stages
pre_process = 1
pre_process_split_density = 0
auto_split_point = 0
auto_split_zone = 0
fitting = 1
dock_only = 0
post_process = 0
rescore = 0
```

The enabled stages section allows the user to specify what functions ChemEM should run with the given data. To enable a stage set the value to 1 to disable a stage set the value to 1.

– Pre Processing

The preprocessing stage sets up ChemEM for docking, fitting, post processing or rescoring, given the input data. Generally, this option should always be enabled unless loading a ChemEM object file.

For docking, rescoring or post processing it is enough just to enable the pre_process stage.

!Tip - the preprocessing stage is generally quick, and it may be a good idea to run with just this stage and evaluate the difference map used for fitting visually first for example with UCSF Chimera.

However, for fitting to CryoEM data the preprocessing option calculates a difference map to be used (output file = ~/preprocessing/difference_map_<difference map

id>.mrc). These maps can contain small regions of non-cryoEM density that you may want to remove.

To do this there are three options available: *pre_process_split_density, auto_split_point, and auto_split_zone*.

– Pre Process Split Density

This option splits the difference map into regions of disconnected density; each region is labelled as a unique integer (output file = ~/preprocessing/difference_map_D0_split_density_< mask id>.mrc). The labelled masks can be applied manually (see pre-processing parameters section 2.3.3) or automatically using one or more (see Multi ligand section 4.5) given centroid points.

To include relevant densities automatically ChemEM has two methods.

– Auto split point

The *auto_split_point* option calculates the centre point of disconnected densities in the mask and takes the density that is the closest to the specified binding site centroid (see Required data section 2.3.1).

–Auto split zone

The *auto_split_zone* option takes the disconnected densities that have their centre points within a given radius of the specified binding site centroid (see Required data section 2.3.1). The given radius can be specified with the pre-processing option *auto_split_radius* (see pre-processing parameters section 2.3.3)

– Fitting

The fitting option runs stage 1 of the ChemEM/ChemDock algorithm.

– Dock only

If this option is enabled ChemEM will run without the cryoEM data only using the ChemDock score.

– Post process

The post process option runs stage 2 of the ChemEM/ChemDock algorithm. Taking the top scoring solutions from the fitting process and molecular dynamics running

simulated annealing. The number of solutions to refine can be specified using the *post_process_num_solutions* option (see post-processing parameters section 2.3.3).

– Rescore

The rescore option will rescore one or more solutions and output a rescore.txt file in the top level directory of the output. Solutions can be specified one by one in the config file with multiple *ligand* lines or using *ligands_from_dir* parameter (see section 2.3.1) .

2.3.3 Pre-processing parameters

Below are the parameters that can be included in the config file to alter the included binding site atoms or the creation of the difference map for stage 1 of the algorithm.

```
segment_dimensions = (30, 30, 30)
label_threshold_sigma = 3
label_threshold = 0.5
auto_split_radius = 6.0
```

– Segment dimensions

Controls the size of the segment used. E.g. (30, 30, 30) specifies a box around the centroid with edges in x,y and z dimensions of 30 Å.

– Label threshold sigma

This option specifies the standard deviations to use for contouring the difference map when defining disconnected densities.

– Label threshold

This option specifies the minimum density cutoff value to use for contouring the difference map when defining disconnected densities.

– Auto split radius

When the auto_split_zone option is used, a distance limit from the binding site's centre is set. Any disconnected densities within this distance will be incorporated into the difference map.

2.3.4 Fitting parameters

For the fitting function three parameters can be changed that affect the scoring during docking:

```
mi_weight =  50
global_k = 50
docking_radius = [15.0, 15.0, 15.0] | 15.0
```

– Mutual information (MI) weight

The MI weight parameter controls the weighting given to the MI (W^{MI}) score when fitting to cryoEM data according to the equation:

$$ChemEM\ score\ =\ (MI\ *W^{MI})\ +\ ChemDock\ score$$

– Global K

The value of Global K controls the weight of the map potential used for minimising solutions with OpenFF (N.B this applies in both the fitting and post-processing stages). A lower value puts less weight on the map potential according to the equation:

$$Eq\ 9.\ E^{map}\ =\ -\ global\_k\ *\ D(x,y,z)$$

– Docking radius

Defines the binding site residues used in the docking calculations. In previous versions, this was defined as a radius, represented by a single float. However, in ChemEM v0.3, the binding site is defined as a box, using a given centroid and box dimensions. From the residues within this box, those important for the docking score are identified using a SASA-based method (see section 2.6). The box size should approximate the size of the segment used to calculate the difference map. If this parameter is not supplied, it defaults to the segment dimensions. A separate 'docking radius' parameter is included in cases where the two values need to be unlinked, e.g., if the segment size needed to calculate the difference map was unusually small or large. Additionally, for backward compatibility, a single float value can still be supplied; however, this will be converted into a box size by the algorithm. For example, if a value of 15.0 is supplied, this would be converted to a box with dimensions [15.0, 15.0, 15.0].

The following fitting options relate to how OpenForceField minimisation is carried out on docked solutions:

```
flexible_side_chains = 0
platform = OpenCL
solvent = False
```

– Flexible site chains

Defines whether to treat binding site side chains as flexible during the docking stage. Valid inputs are 0 (disabled, default) or 1 (enabled)

– Platform
Specifies how calculations with OpenMM/OpenFF should be calculated, valid inputs are 'OpenCl', 'CUDA' or 'CPU'. For more information regarding the differences in platforms see the OpenMM documentation (http://docs.openmm.org/latest/userguide/library/04_platform_specifics.html) .

– Solvent

If enabled implicit solvent will be used when minimising solutions with openFF.
Valid inputs are 0 (disabled, default) or 1 (enabled)**.**

Finally, parameters are available to control the Ant colony optimisation (ACO) algorithm during the fitting stage are:

```
n_ants = 20
N_cpu = 20
rho = None
max_iterations = 40
generate_diverse_solutions = 1.5
```

– *N* Ants

The number of ants to use corresponds to the number of solutions constructed at each iteration of the ACO algorithm.

– *N* CPUs

This defines the number of cpus to use during the ACO.

– Rho

This is the pheromone evaporation coefficient. It represents the rate at which the pheromone trail intensity decreases over time. This parameter is critical because it prevents the algorithm from converging too quickly on a suboptimal solution by allowing pheromone trails to fade, which encourages exploration of alternative paths.

– Max iterations

The maximum number of iterations of the ACO to run. By default this value is calculated based on the degrees-of-freedom included in the fitting i.e. the higher the number of rotatable bonds the greater the number of iterations needed.

– Generate diverse solutions

The *generate_diverse_solutions* parameter takes a value for an RMSD cutoff for which solutions are considered diverse. If this parameter is enabled the top solutions will be

2.3.5 Post-processing parameters

There parameters for the post-processing stage fall into two main categories, related to what is to be post-processed and how the post-processing should take place. Options for the former are:

```
#-----post processing parameters-----
#---general---
post_process_num_solutions = 15
refine_side_chains = 1
Post_process_radius = 15.0
```

– Post process num solutions

The number of top solutions from fitting to post process.

– Refine side chains

If turned on, the binding site side chains will be refined along with the ligand. Valid inputs are 0 (disabled) or 1 (enabled, default)**.**

– Post process radius

The maximum radius from the centroid at which residues side chains will be flexible.

Options that affect how simulated annealing steps take place are:

```
#---simulated annealing params---
cycles = 4
norm_temp = 300
top_temp = 315
temperature_step = 1
steps = 1000
heating_interval = 100
```

– Cycles

The number of '*heating*' and '*cooling'* cycles for each ligand solution. Where a solution file is generated after each cycle.

– Normalisation temperature

The lowest temperature that the system is cooled to.

– Top temperature

The highest temperature the system is raised to.

– Temperature step

The step size when raising or cooling the system temperature determines how quickly the temperature decreases (where 1 step is equal to 1 femtosecond). A smaller step size means the temperature decreases slowly, allowing for a more thorough exploration of the solution space, which can potentially lead to finding a better minimum.

– Steps

The number of steps to run the simulation when the system reaches the top and normalisation temperatures

– Heating interval

The number of steps to run at each temperature step when heating and cooling the system.

2.4 Protonation, Chirality and Rings.

2.4.1 Protonation

It is advised that the protonation states explicitly in the ligand smiles string. However, ChemEM by default will attempt to set the protonation states automatically. This protonation state assignment uses dimorphite-DL (Ropp et al. 2019). Dimorphite-DL assigns protonation states based on pH ranges. This can result in multiple plausible protonation states. There is some effort by ChemEM to choose a suitable protonation state for the pH range. However, it is advised to use the following protocol:

```
chemem.protonate <configuration_file | ligand smiles>
```

The command chemem.protonate will apply Dimorphite-DL protonation to your ligands. This functionality can be accessed in two ways:

There are two options that control the results of the Protonation:

By providing the path to your configuration file: This method allows you to specify detailed settings and multiple ligands in a single command.

By supplying a ligand SMILES directly as an argument: This is a quick way to protonate a single ligand without the need for a configuration file.

When running *chemem.protonate*, you have several options that can influence the outcome:

```
pH = [6.4, 7.4]
pKa_prec = 1.0
```

– pH
The pH at which to apply the protonation.

– pKa_prec
The pKa precision factor to use when estimating moiety pKa ranges.

Once you have decided on a final protonation state for your ligand, you can include the ligand SMILES string in your configuration file and disable the automatic protonation feature. This can be done by setting the protonation option to 0 in your configuration file. Doing so ensures that ChemEM uses the protonation state you have specified without making any automatic adjustments.

```
protonation = 0
```

2.4.2 Rings

Ring information when reading ligands from smiles can sometimes be lost. ChemEM uses RDKit to assign ring information before calculations. If this functionality is affecting you're calculations it can be turned off in the configuration file by including the line:

```
rings = 0
```

2.4.3 Chirality

To run molecular dynamics calculations in ChemEM, the ligand chirality information must be set for all chiral centres in a molecule. It is recommended to include this information within the supplied SMILES string. However, ChemEM will attempt to assign any missing chirality data from the structure generated from the supplied SMILES string or molecule file. To turn off this functionality, include the following line in your configuration file:

```
chirality = 0
```

2.5 Force Fields

Various force fields are available for ion parameters and implicit solvent. By default the following Amber14 force fields are included.

```
Protein = amber14/protein.ff14SB.xml
Ions = amber14/tip3pfb.xml
Implicit solvent = amber14/tip3pfb.xml
```

Specific Ion or implicit force fields can be added to the calculations by adding the following to your configuration file:

```
forcefield = amber14 #required when specifying force fields
forcefield = <ion/implicit solvent force fields>
```

Currently supported force fields for ions are:

```
amber14/tip3p.xml
amber14/tip3pfb.xml
amber14/tip4pew.xml
amber14/tip4pfb.xml
amber14/spce.xml
amber14/opc.xml
amber14/opc3.xml
```

Currently supported force fields for implicit solvent are:

```
implicit/hct.xml
implicit/obc1.xml
implicit/obc2.xml
implicit/gbn.xml
implicit/gbn2.xml
```

2.6 Surface residues definition

ChemEM now defines surface residues for docking using a SASA based method with the Shrake & Rupley algorithm (Shrake and Rupley 1973) calculated at the residue level. Residues are defined as surface if they have a SASA above the defined cutoff (default 5.0 Å). To change the cutoff use the following in your configuration file:

```
sasa_cutoff = 5.0
```

The residues included in the SASA calculation are defined with either the '*docking_radius'* or '*segment_dimensions'* parameters (See section 2.3.4).

3.0 Output files

In the configuration file, users can specify the output directory where ChemEM will save its resulting files. Upon execution, ChemEM organises the output into three distinct subdirectories, each named according to the stage of the algorithm that produced the files.

3.1 Pre-processing

Contains all the files generated during the initial preparation and setup phase of the algorithm. In this directory you will find files related to the difference mapping.

```
#----pre_processing_files-----
full_map_M0_segment_S0.mrc
difference_map_D0.mrc
difference_map_reverse_D0.mrc
difference_map_D0_split_density_X0.mrc
difference_map_D0_split_density_auto_point_0_X0.mrc
difference_map_D0_split_density_auto_zone_0_X0.mrc
```

– Full map segment

The file named full_map_M0_segment_S0.mrc holds the specific region extracted from the full map identified as M0, which is used in the segment S0.

– Difference map

The *difference_map_D0.mrc* file, designated by ID D0, represents the initial difference map generated by comparing the segment to the full map.

– Reverse difference map

The file *difference_map_reverse_D0.mrc,* is the reverse difference map created by subtracting the experimental map from the simulated apo map.

– Difference density masks

The file *'difference_map_D0_split_density_X0.mrc'* contains the difference map with labelled masks. If the '*pre_process_split_density'* is enabled in the enabled stages of the config file, this file is generated. Each disconnected density above the label threshold is given an integer ID. This file can help when deciding to split density by zone or by point in removing non-ligand density from the difference map.

–Auto split point

The file '*difference_map_D0_split_density_auto_point_0_X0.mrc'* contains the difference map output if the '*auto_split_point'* option is enabled. This option applies a density mask for the labelled density closest to the binding site centroid.

–Auto split zone

The file '*difference_map_D0_split_density_auto_zone_0_X0.mrc'* contains the difference map output if the '*auto_split_zone'* option is enabled. This option applies a density mask for all disconnected densities within a given radius of the binding site centroid.

3.2 Fitting

Holds files that are the result of the ChemEM algorithm's fitting process (stage 1).

```
Ligand_0_refined.sdf
Ligand_1_refined.sdf
Ligand_2_refined.sdf
...
PDB
results.txt
times.txt
```

– Ligands

Ligands are output in *.sdf* format and numbered with an integer.

– Atomic models

Within Fitting there is a folder titled '*PDB'* that holds the ligand solutions fit to protein models.

— ChemEM scores

The '*results.txt'* file contains the ChemEM scores for each ligand, organised from the best to worst scoring solutions.

– Runtime

The 'times.txt' file holds information on the total runtime of docking.

3.3 Post-processing

All the solutions for all files that are post-processed are found in this folder.

```
Ligand_0_cycle_1_ligand_1.sdf
Ligand_0_cycle_1.pdb
Ligand_0_cycle_2_ligand_1.sdf
Ligand_0_cycle_2.pdb
...
results.txt
times.txt
```

– Ligands

Ligand solutions are found in .sdf format. The file name corresponds with the ligand ID from the fitting stage, also indicated is the simulated annealing cycle, and the number of the ligand within the binding site (for a single ligand this will always be ligand_1).

– Atomic models

Atomic models containing fit ligand and protein are given in *.pdb* format.

— ChemEM scores

The '*results.txt'* file contains the ChemEM scores for each ligand, organised from the best to worst scoring solutions.

– Runtime

The 'times.txt' file holds information on the total runtime of docking.

3.4 Protonation

The output of running chemem.protonate will depend on the data supplied. If a SMILES string is supplied, the identified protonation states will be printed to the console. When a

configuration file is supplied, the output will be saved to the file protonation.txt at the top level or in the specified output directory.

The fill includes the input smiles and pH range followed by the possible protonation states found for each input ligand.

```
Input SMILES: CC(C)NC[C@@H](c1ccc(c(c1)O)O)O
Found states at pH range 6.4 - 8.4:
      CC(C)[NH2+]C[C@H](O)c1ccc([O-])c([O-])c1
      CC(C)NC[C@H](O)c1ccc(O)c(O)c1
      CC(C)[NH2+]C[C@H](O)c1ccc(O)c(O)c1
      CC(C)[NH2+]C[C@H](O)c1ccc(O)c([O-])c1
      CC(C)NC[C@H](O)c1ccc([O-])c(O)c1
      CC(C)NC[C@H](O)c1ccc([O-])c([O-])c1
      CC(C)NC[C@H](O)c1ccc(O)c([O-])c1
      CC(C)[NH2+]C[C@H](O)c1ccc([O-])c(O)c1
```

4.0 Examples and tips

4.1 Defining a centroid

The simplest way to define a binding site is to use a molecular viewer such as UCSF Chimera/ChimeraX, Coot or pyMol.

Select a few residues around the assumed binding site and calculate the centroid of this, the centroid does not need to be exact (Figure 2).

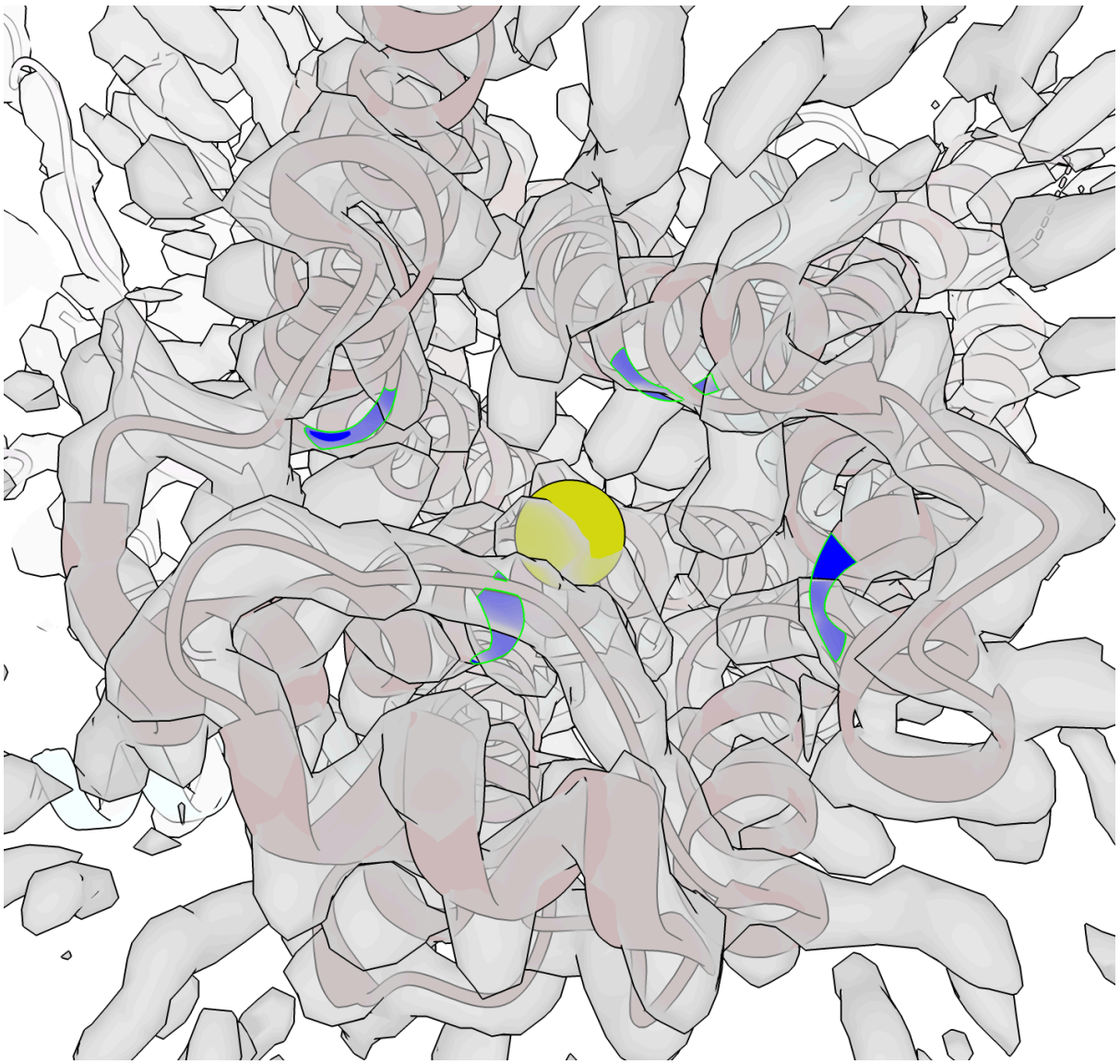

Figure 2. An approximation of a binding site centroid (yellow ball) in ChimeraX, made by calculating the centroid of four residues that line the binding site (blue residues green boarder).

4.2 Improving a difference map

4.2.1 Auto split point

```
#ChemEM config file
protein = ~/test_data/7jjo_protein.pdb
ligand = CC(C)NC[C@@H](c1ccc(c(c1)O)O)O
centroid = (134.087, 133.507, 174.180)
output = ~/some_directory
densmap = ~/test_data/7jjo.mrc
resolution = 2.6
```

```
#enabled stages
pre_process = 1
pre_process_split_density = 0
auto_split_point = 0
auto_split_zone = 0
fitting = 0
dock_only = 0
post_process = 0
rescore = 0
```

The above shows the config file used to generate an initial difference map for the protein model 7jjo deposited in the PDB (https://www.rcsb.org/structure/7JJO).

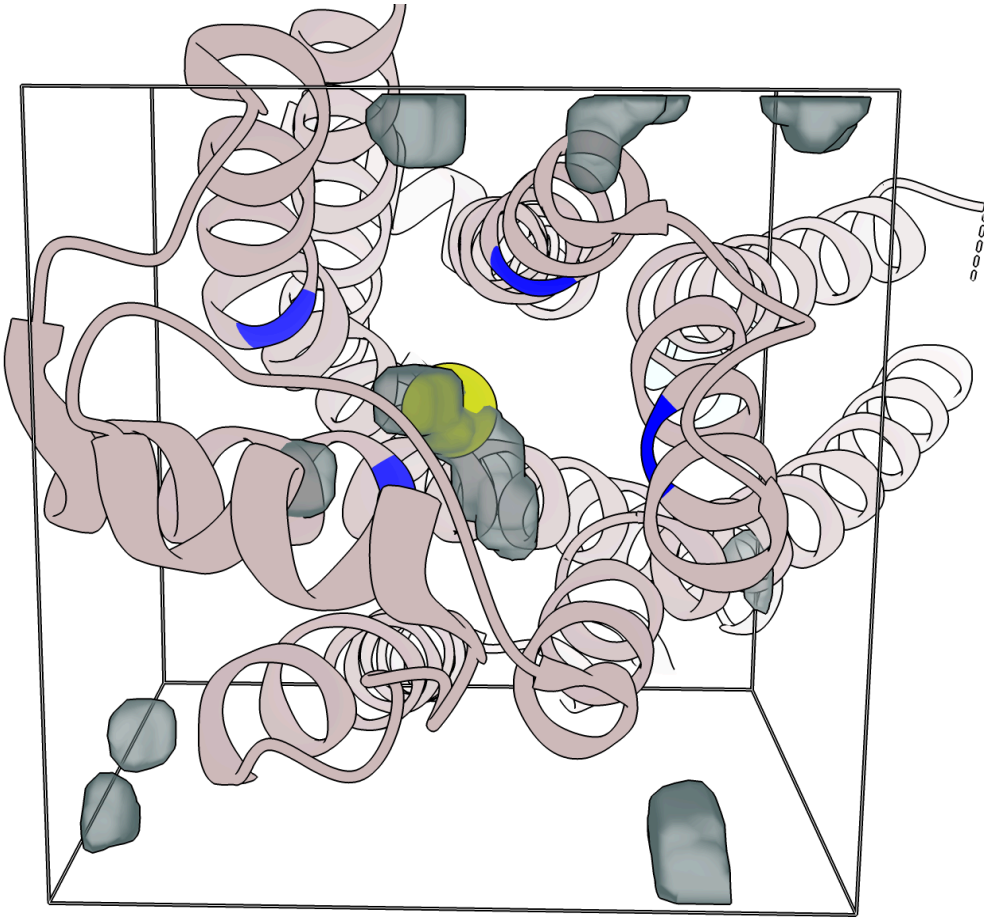

Figure 3. The generated difference map (blue/grey translucent density) in the segment defined by the centroid (yellow ball) and box size (enclosing black box).

This example shows how to check and improve a difference map. A typical initial difference map is shown in Figure 2. In this map the assumed ligand density is within the binding site, with small regions of disconnected density present in the periphery. This density is close to

the protein and should not affect the results of ChemEM, however can easily be removed by enabling the '*auto_split_point'* option with:

```
...
auto_split_point = 1
...
```

This removes the unwanted non-ligand density from the difference map (Figure 4)

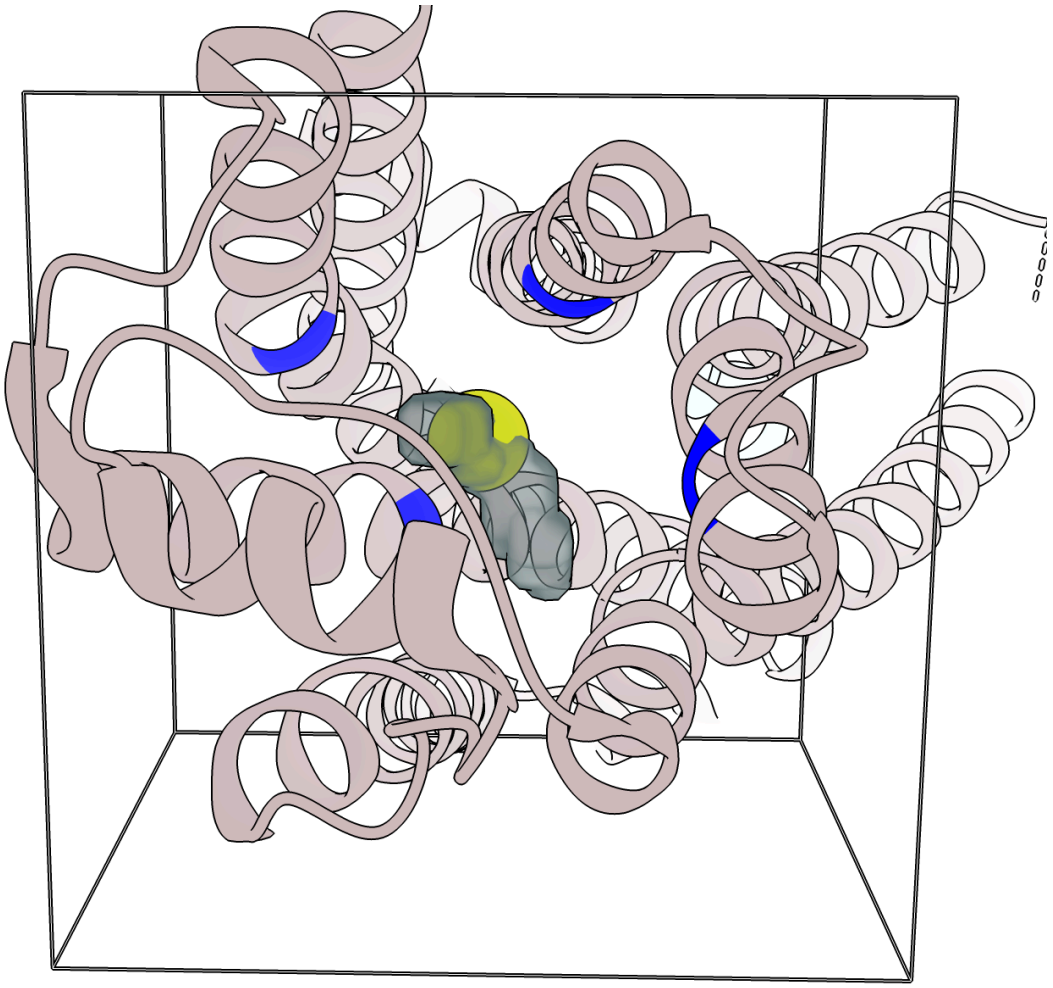

Figure 4. The difference map using the *'auto_split_zone'* option (blue/grey translucent density) in the segment defined by the centroid (yellow ball) and box size (enclosing black box).

4.2.2 Box size

This example shows how changing the '*segment_dimesions'* option can help improve the quality of the difference map. Below is the configuration file used to generate a difference map using the deposited protein model and map for the PDB entry 6NYY (https://www.rcsb.org/structure/6nyy):

```
#enable stages
preprocess = 1
pre_process_split_density = 1
auto_split_point = 1
auto_split_zone = 0
fitting = 0
dock_only = 0
post_process = 0

#data and working directory
protein = ~/6nyy_protein.pdb
ligand =
c1nc(c2c(n1)n(cn2)[C@H]3[C@@H]([C@@H]([C@H](O3)CO[P@](=O)(O)O[P@@](=O)(N
P(=O)(O)O)O)O)O)N
densmap = ~/6nyy.mrc
resolution = 3.0
centroid = (36.016, 70.61, 85.53)
output = ~/6nyy

#pre-processing options
segment_dimensions = (20, 20, 20)
```

Using a box size of 20 $Å^3$ yields a relatively good difference map (Figure 5, left). However this can be improved upon by increasing the segment dimensions to 25 $Å^3$ (Figure 5, right). By changing the segment dimension line in the configuration file:

```
#pre-processing options
segment_dimensions = (25, 25, 25)
```

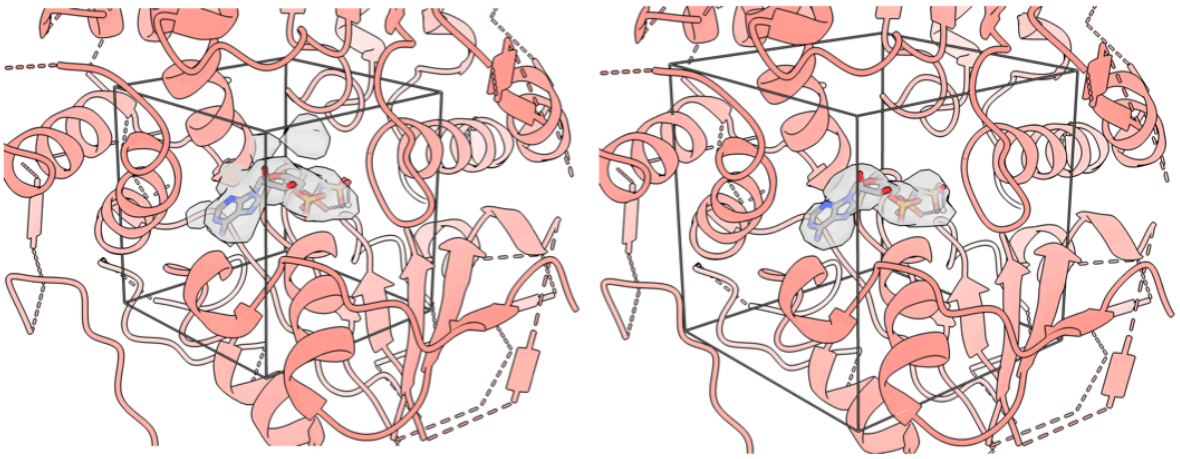

Figure 5. The left image shows the difference map generated with the default box size of (20 Å^3). The image on the right shows the difference map generated using a box size of (25 Å^3).

4.2.3 Thresholds

This example shows how to use the *'label_threshold'* option to improve a difference map. The configuration file below calculates a density difference map using the deposited map and protein model from the PDB entry 6X1A (https://www.rcsb.org/structure/6x1a):

```
#ChemEM config file

#enable stages
preprocess = 1
pre_process_split_density = 1
auto_split_point = 1
auto_split_zone = 0
fitting = 0
dock_only = 0
postprocess = 0

#data and working directory
protein =~/6x1a_protein.pdb
ligand =
c1cc(nc(c1)OCc2ccc(cc2F)C#N)C3CCN(CC3)Cc4nc5ccc(cc5n4C[C@@H]6CCO6)C(=O)O
densmap = ~/6x1a.mrc
resolution = 2.5
#map_contour = 3.0
centroid = (131.34, 116.77, 155.03)
output = ~/6x1a
```

Using the above configuration file yields a difference map with clear protein density included, even when using '*auto_split_point*' (Figure 6A). By inspecting the density in Chimera it is clear that at lower thresholds the density is connected (Figure 6B). At a threshold of 0.2 the density is no longer connected. We can apply this threshold to difference map operations by including the following line in the config file:

```
label_threshold = 0.2
```

Re-running the difference map now removes the unwanted density (Figure 6C).

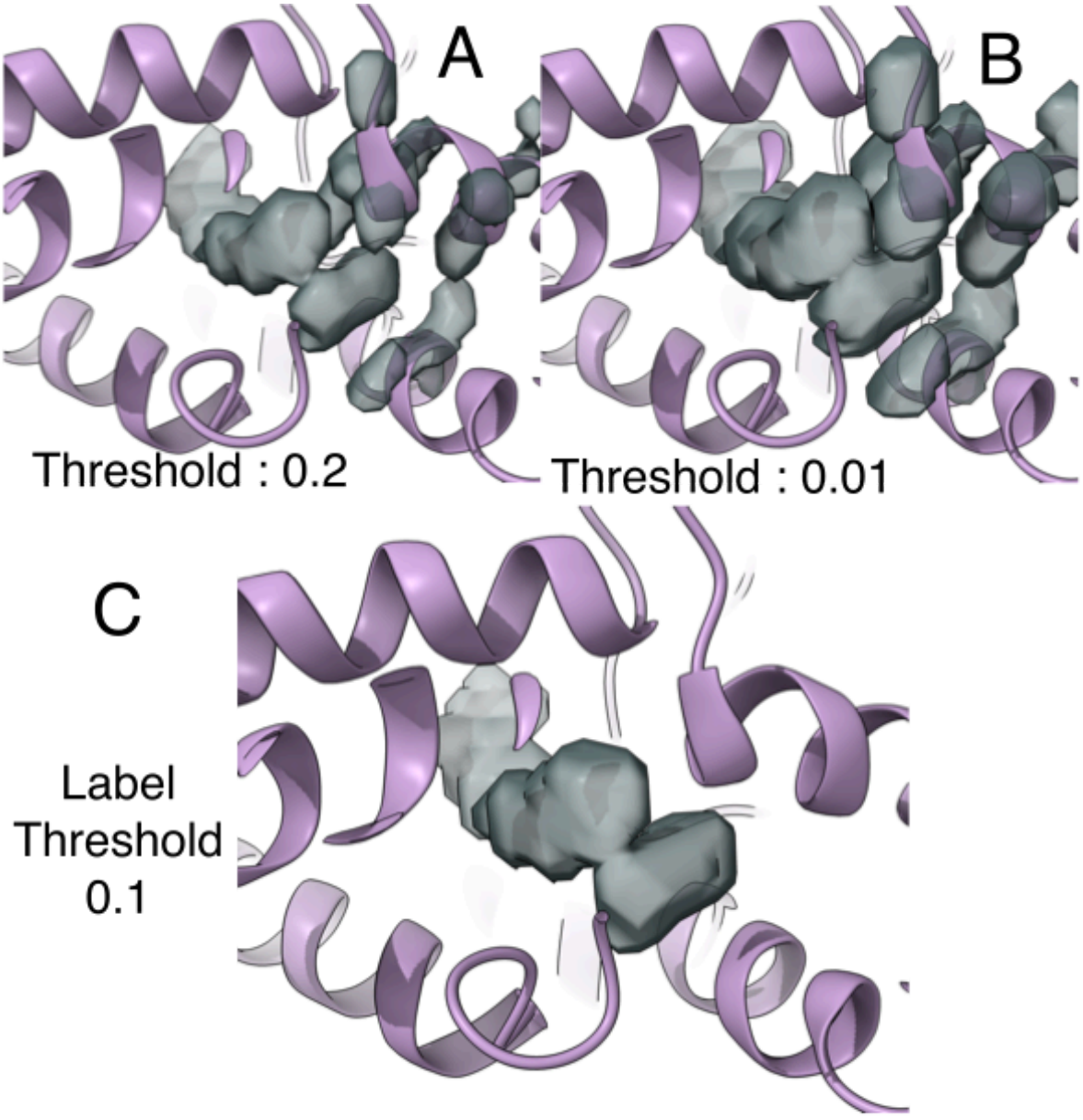

Figure 6. Density difference maps created using ChemEM.

4.3 Docking with no cryo-EM map

Docking without the cryo-EM map can be achieved by enabling the '*dock_only*' function in the enabled stages section of the configuration file. The pre-processing stage must still be enabled to set up the segment for docking. The enable the fitting/and or post-processing stages as needed.

```
#ChemEM config file
protein = ~/test_data/7jjo_protein.pdb
ligand = CC(C)NC[C@@H](c1ccc(c(c1)O)O)O
centroid = (134.087, 133.507, 174.180)
output = ~/some_directory
densmap = ~/test_data/7jjo.mrc
resolution = 2.6

#enabled stages
pre_process = 1
pre_process_split_density = 0
auto_split_point = 0
auto_split_zone = 0
fitting = 1
dock_only = 1
post_process = 0
rescore = 0
```

4.4 Rescoring ligands

To rescore ligands with the ChemDock or ChemEM scores, the preprocessing and rescore options must be enabled. To discount the Mutual information score in the rescoring enable the dock only stage.

Ligands can be uploaded individually to the configuration file like so:

```
ligand = ~/test_data/Ligand_1.sdf
ligand = ~/test_data/Ligand_2.mol2
```

Additionally, for rescoring ligands can be specified using the *ligands_from_dir* option with a directory path containing ligands in .mol2 or .sdf file format:

```
Ligands_from_dir = ~/test_data/
```

The output of rescoring will be in the top level directory specified in the configuration file under the '*output*' option in a file called '*rescore.txt*'.

4.5 Multi-ligand docking

ChemEM is able to fit multiple ligands simultaneously into a single binding site, to set up a configuration file for this there are a few changes that need to be included in the configuration file.

```
#ChemEM config file

#enable stages
pre_process = 1
pre_process_split_density = 1
auto_split_point = 1
auto_split_zone = 0
fitting = 1
dock_only = 0
post_process = 0
rescore = 0

#data and working directory
protein = ~/test_data/6tti_protein.pdb
ligand = CC(=O)Nc1cc(no1)C(F)(F)F
ligand = CC(=O)Nc1cc(no1)C(F)(F)F
ligand = CS(=O)C
ligand = CS(=O)C
densmap = ~/maps/6tti.mrc
resolution = 2.5
map_contour = 0.07
centroid = ( 43.908,  54.415,  49.755)
output = ~/multi_fit
```

Firstly all ligands must be entered individually. It is possible to use a single centroid for all four ligands as in the example case (Figure 7), where the difference map shows four conspicuous peaks in the density that lie within the binding site.

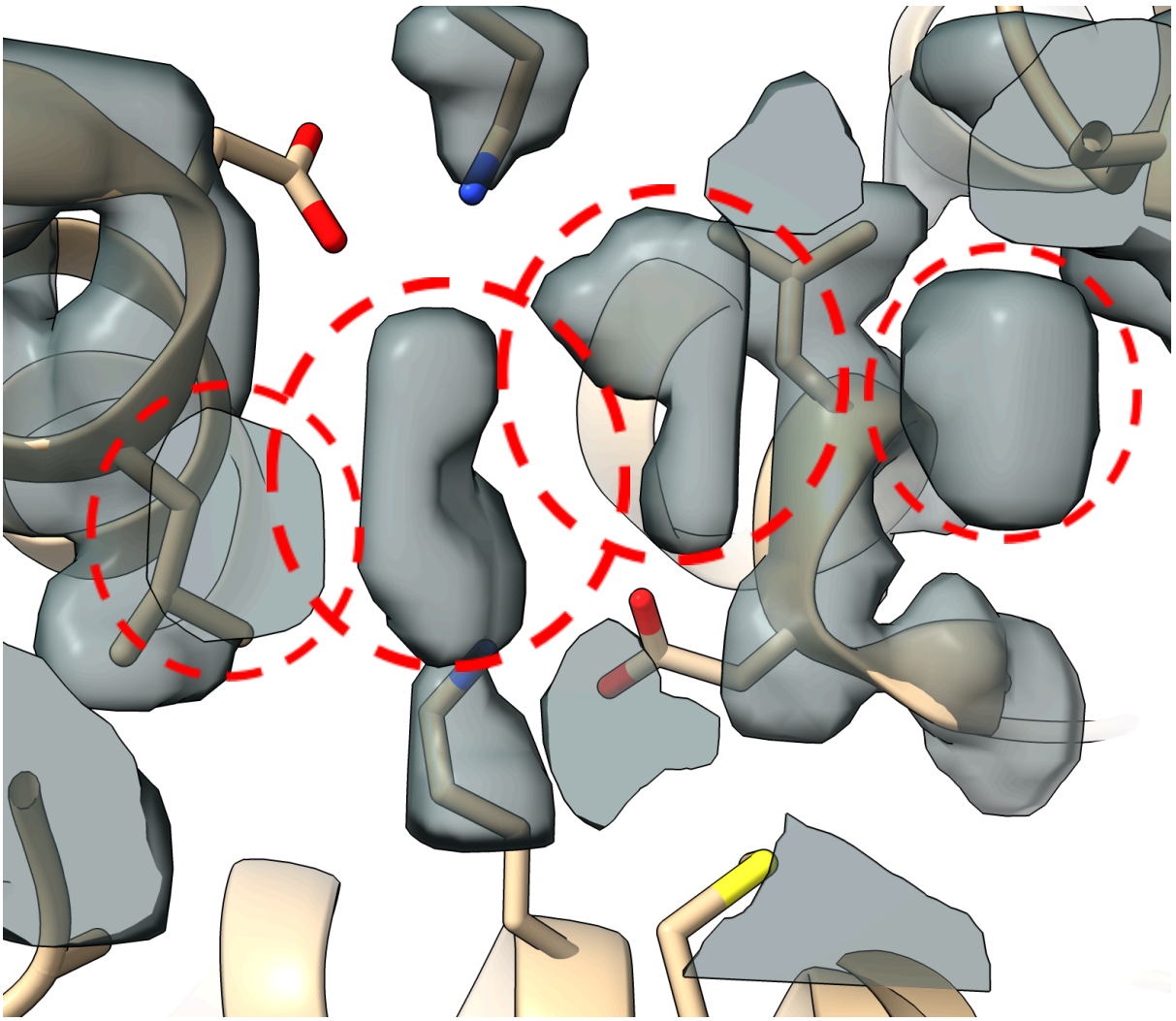

Figure 7. The density map generated using a single centroid.

However, it is also possible and may lead to more accurate solutions to enter multiple centroids like so:

```
...
centroid = ( 43.908,  54.415,  49.755)
centroid = ( 39.049,  58.586,  49.760)
centroid = ( 40.625,  62.434,  52.641)
centroid = ( 42.391,  50.558,  52.642)
...
```

The order of the ligands entered and the centroids must correspond.

This will create four individual difference maps for use during fitting (Figure 8).

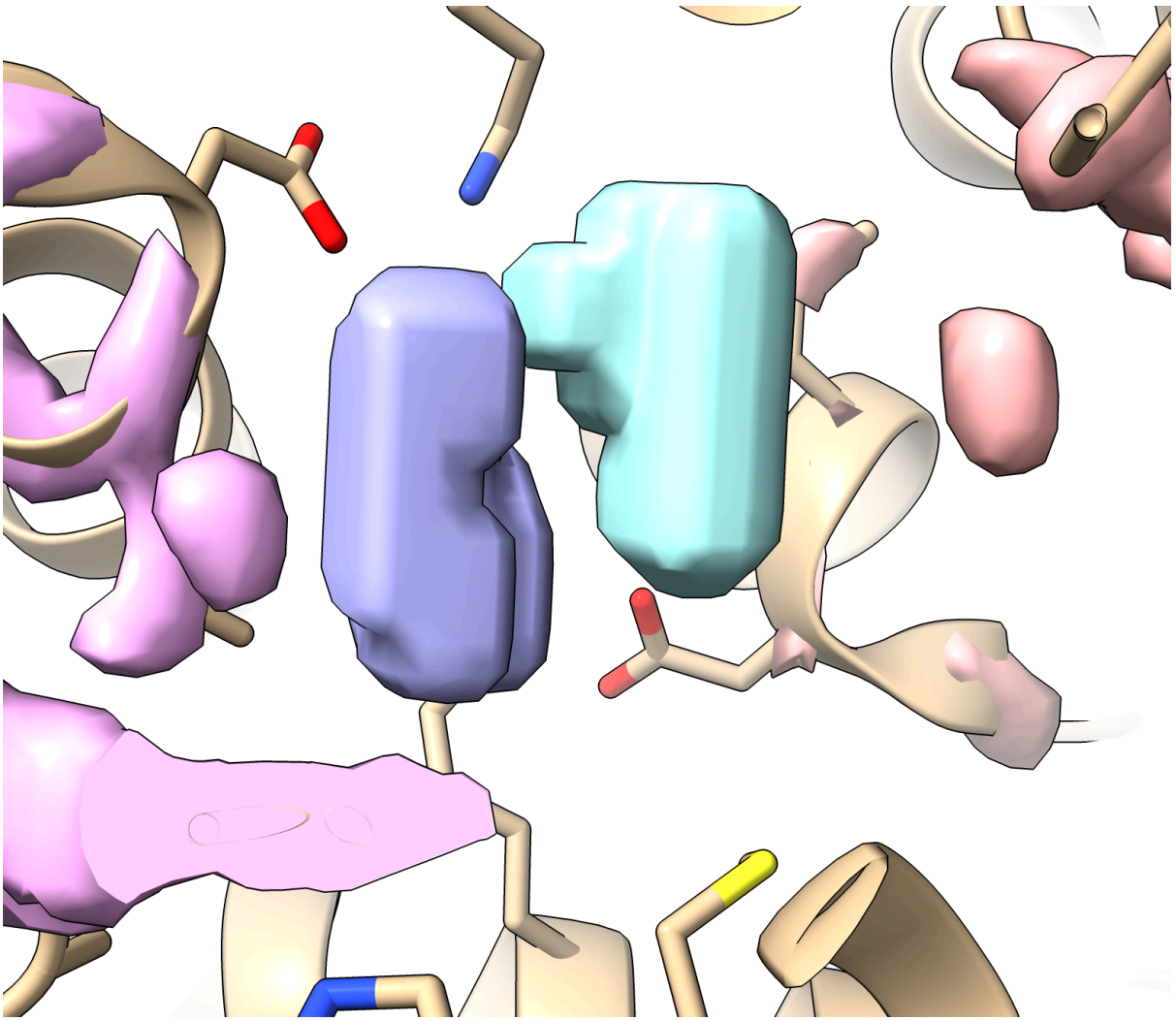

Figure 8. Difference density map generated using multiple centroids. Each color is an individual density map.

Using the multi-fitting function in this way will mean that solutions for a given ligand will only be explored around the density map for which it is assigned.

5.0 References

Ropp, Patrick J., Jesse C. Kaminsky, Sara Yablonski, and Jacob D. Durrant. 2019. "Dimorphite-DL: An Open-Source Program for Enumerating the Ionization States of Drug-like Small Molecules." *Journal of Cheminformatics* 11 (1): 14.

Shrake, A., and J. A. Rupley. 1973. "Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin." *Journal of Molecular Biology* 79 (2): 351–71.

Sweeney, Aaron, Thomas Mulvaney, Mauro Maiorca, and Maya Topf. 2024. "ChemEM: Flexible Docking of Small Molecules in Cryo-EM Structures." *Journal of Medicinal Chemistry* 67 (1): 199–212.