CICLOP v2.0 Tutorial

A Concicse Guide to studying water dynamics within protein cavities

November 24, 2024

Introduction

CICLOP v2.0 is a robust Python-based molecular dynamics trajectory analysis tool tailored for identifying and characterizing protein cavities (internal cavities) and the water contained within them. Here, we provide a concise guide on how to use CICLOP to characterize water within protein cavities

Dependencies

CICLOP is built on python3.11. The following are the dependencies required to run CICLOP

1. numpy, scipy, math, pandas, sys, os, time
2. MDAnalysis
3. tqdm
4. multiprocessing
5. shapely
6. gc
7. argparse
8. concurrent.futures
9. collections

Installation

To install and run CICLOP, execute the following commands.

```
chmod +x CICLOP.py
sudo cp CICLOP.py /usr/local/bin/CICLOP
```

Now you are ready to go!

Basic Commands

The following command can simply list all functionalities available in CICLOP

```
1 CICLOP -h
```

Cavity Lining and Water Detection

Given a solvated structure file, one can detect the cavity lining atoms and the water contained within the cavity using the following command.

```
CICLOP -w tip3p -f 3los_Tutorial.pdb -o Output.pdb
```

The flag $-w$ is a mandatory flag for all analyses. The argument provided to $-w$ is required to identify the nature of the water model, i.e., the number of atoms in the water model. The choice of the model does not really matter, i.e., you may provide $-w$ *tip3p* even if your system has been solvated with SPC/E or for any three-point model for that matter.

Cavity Lining and Water Detection

When running the above command, you will be prompted to choose a group for output. The selected group will be written out to the output file. In the supported formats, the B-Factor columns will be modified so that inner surface atoms will have a B-Factor of 9999.

One can now visualize or select the cavity lining atoms using any of the numerous available packages, such as Chimera, PyMol, and MDAnalysis.

Cavity Lining and Water Detection

Suppose one desires to identify the cavity along with the water within the cavity at some desired distance from the cavity lining, i.e., water within $1 - 5\mathring{A}$ or $3 - 10\mathring{A}$ from the cavity lining. In that case, either of the following commands are to be run.

```
1 CICLOP -w tip3p -f 3los_Tutorial.pdb -layer_range -o Output.
    pdb
2 CICLOP -w tip3p -f 3los_Tutorial.pdb -layer_range -ls 3 -le
    10 -o Output.pdb
```

Cavity Lining and Water Detection

1. The use of the $-ls$ and $-le$ flags will set the bounds for the distances from the cavity lining between which you wish to identify water molecules. If not specified, you will be prompted to provide these values once cavity water detection has been completed.
2. The same procedure will be used for selecting ensembles of water for further analyses.

Cavity Charge Profile

1. The charge profile of the cavity lining residues along the protein axis can be discerned for a single frame or for multiple frames using the following commands.
2. For a single frame,

```
CICLOP -w tip3p -charge_plot -s 3los_Tutorial.tpr -f 3los_Solv.pdb -o 3los_Charge.pdb -to_csv 3los_Charge.csv
```

3. For a time-averaged plot

```
CICLOP -w tip3p -charge_plot -s 3los_Tutorial.tpr -traj 3los_Tutorial.xtc -t0 0 -tf 50 -step 0.5 -o 3los_Charge.pdb -to_csv 3los_Charge.csv -axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102
```

Cavity Charge Profiles

1. The inputs provided to the $-t0$, $-tf$, and $-step$ flags are the initial and final time points and the time steps between frames, respectively, all in picoseconds. These parameters will decide the number of frames over which any time-dependent analysis is carried out.
2. The flags $-axis\_x/y/z$ are used to define the axis vectors of the protein. CICLOP, by default, discerns the axis of the protein for alignment to the $Z-Axis$. These flags circumvents the axis discernment, making the analysis significantly faster.
3. For the above trick to work, one must first align the protein trajectory to a particular frame, preferably the energy minimized structure, which is common practice before proceeding for further analysis.

Cavity Charge Profiles

1. Once alignment is done, one can safely choose the axis of the energy-minimized structure to be the axis for all frames. The axis vector can be found within the log file produced after running any of the single-frame analyses on the single-frame structure
2. The flag $-to\_csv$ is used to write out the produced data to CSV(Comma Separated Values) files with a name of your choice. If not given, a CSV file will still be produced with a default name of the form **CICLOP_Chosen_Analysis-YYYY-MM-DD-HH-MM-SS**. (e.g.: CICLOP_Charge_Profile-2024-9-15-18-47-58.csv)

Cavity Charge Profiles: Results

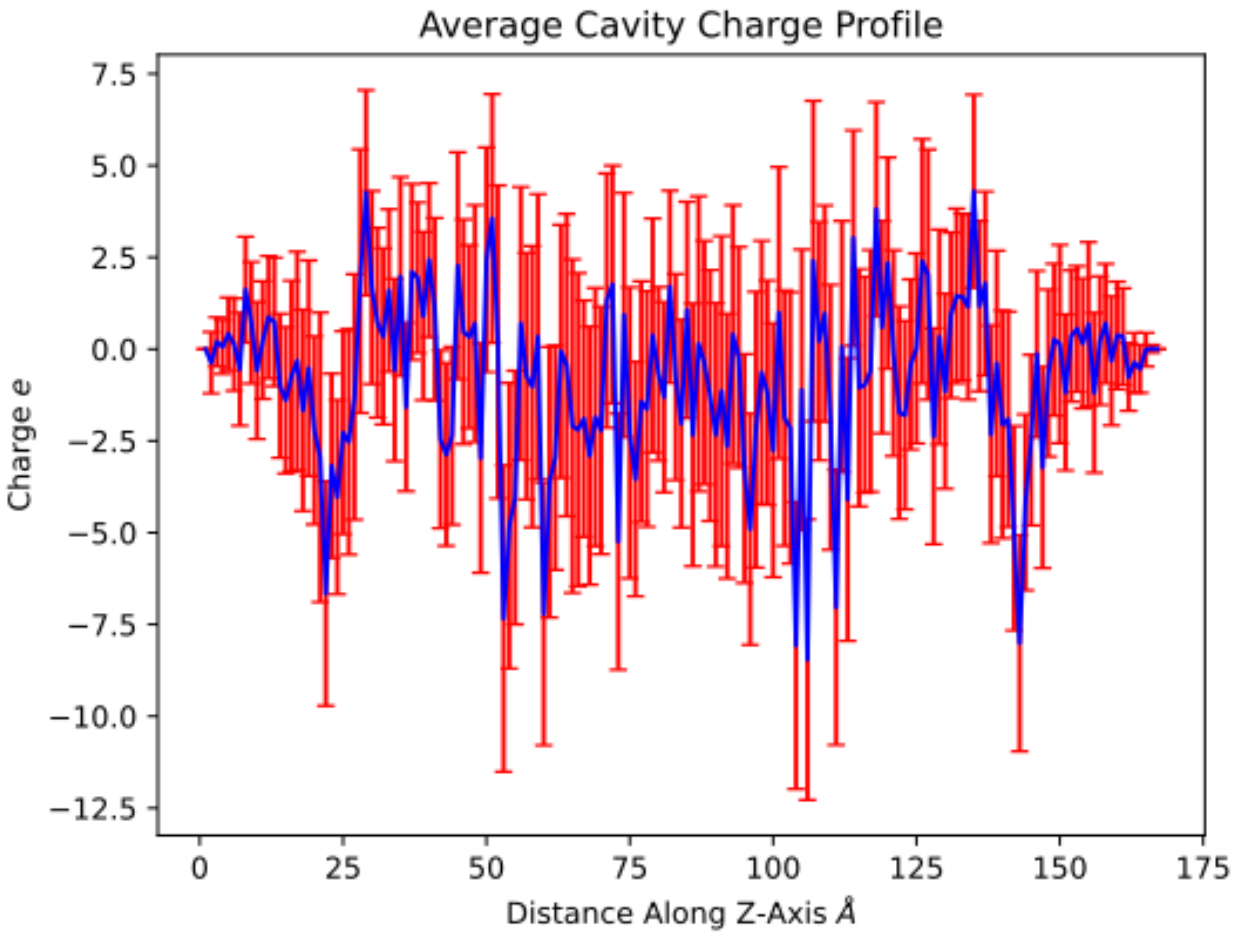

Figure: Time Averaged Charge Profile

Cavity Charge Profiles: Results

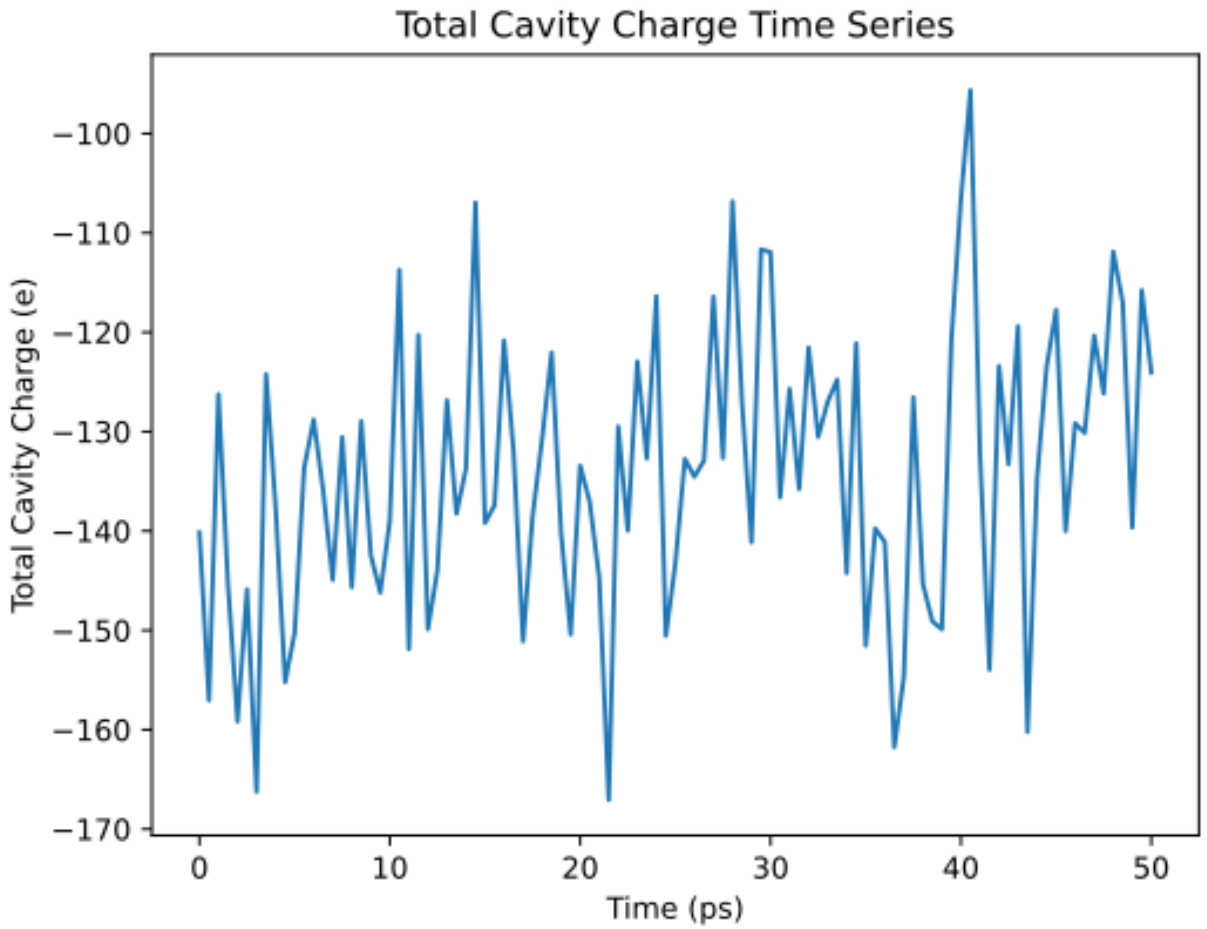

Figure: Total Cavity Charge Time-Series

Cavity Charge Profile: Results

1. One can see in the plots that the cavity charges are predominantly negative, and the maximum magnitude of negative charge is towards the central region of the protein cavity.
2. The charge time series plots show that the net charge in the cavity is negative.

Cavity Radius and Volume Profiles

1. One can also calculate the radius and volume profiles of the protein cavity along the protein axis, both for a single frame and multiple frames.
2. For a single frame,

```
CICLOP -w tip3p -rad_profile -f 3los_Tutorial.pdb -o 3
    los_Rad.png -to_csv 3los_Rad.csv
```

3. For a time-averaged plot,

```
CICLOP -w tip3p -rad_profile -s 3los_Tutorial.tpr -traj
     3los_Tutorial.xtc -t0 0 -tf 50 -step 0.5 -o 3
    los_Rad.png -to_csv 3los_Rad.csv -axis_x 0.0052 -
    axis_y 13.999 -axis_z 0.0102
```

4. For volume profiles, one can simply replace the *−rad_profile* flag with the *−vol_profile* flag.

Cavity Radius and Volume Profile: Results

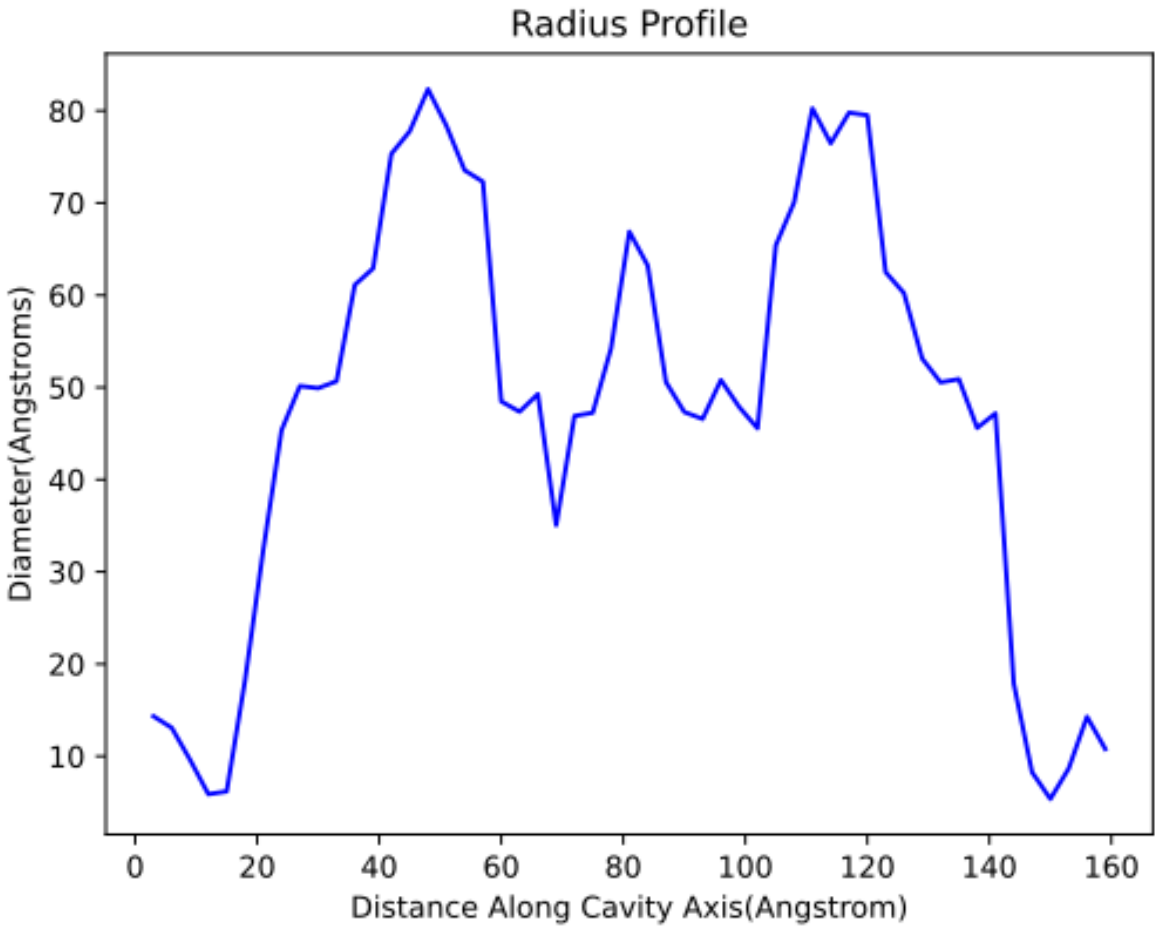

Figure: Cavity Radius Profile

Cavity Radius and Volume Profile: Results

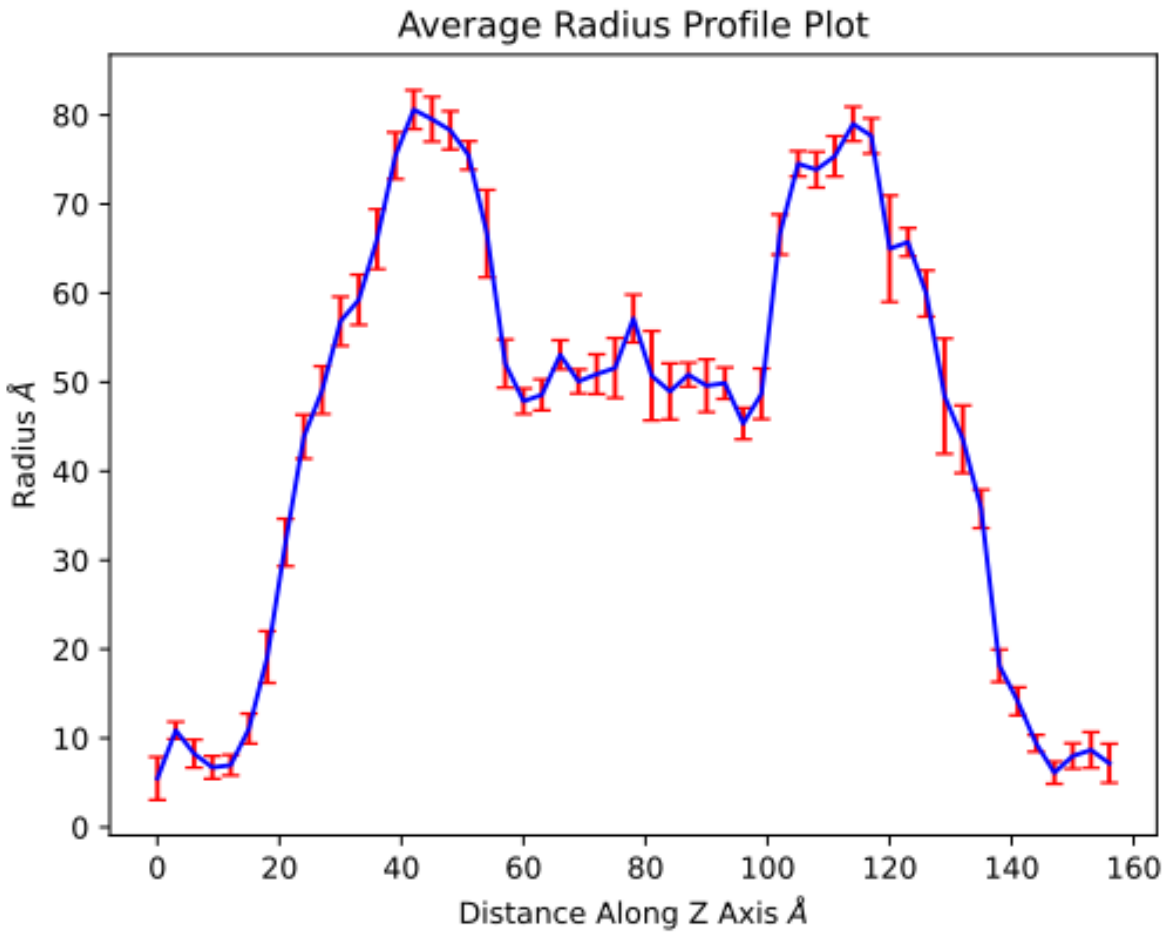

Figure: Average Cavity Radius Profile

Water Density Profiles

1. The variation in water density as a function of distance from the cavity wall can be calculated for both a single frame and multiple frames of a trajectory using the following commands.
2. For a single frame,

```
CICLOP -w tip3p -f 3los_Tutorial.pdb -dens_plot -o 3
    los_Dens_Plot.png -to_csv 3los_Dens_Plot.csv
```

3. For a time-averaged plot,

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -dens_plot -t0 0 -tf 50 -step 0.5 -
    o 3los_Dens_Plot.png -to_csv 3los_Dens_Plot.csv -
    axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102
```

Water Density Plots: Results

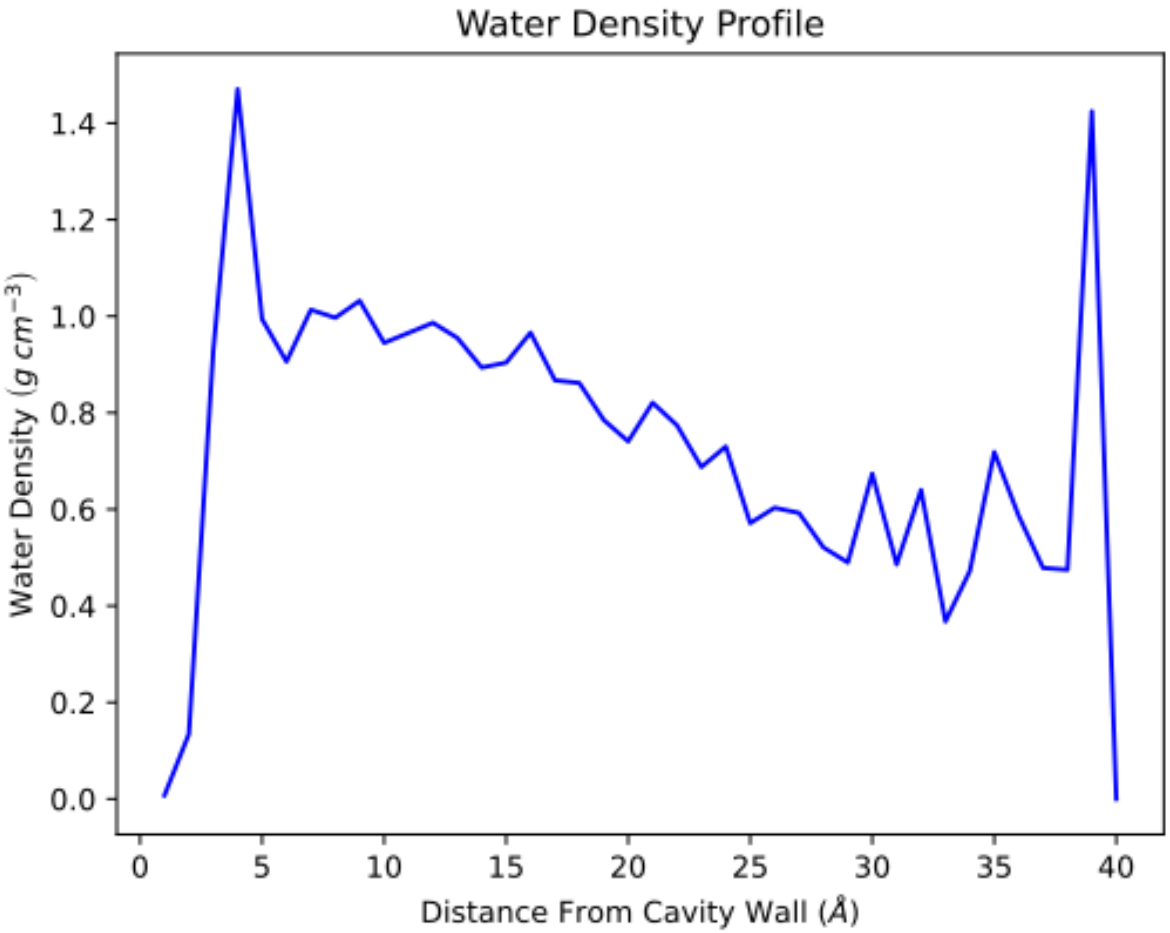

Figure: Cavity Water Density Profile

Water Density Plots: Results

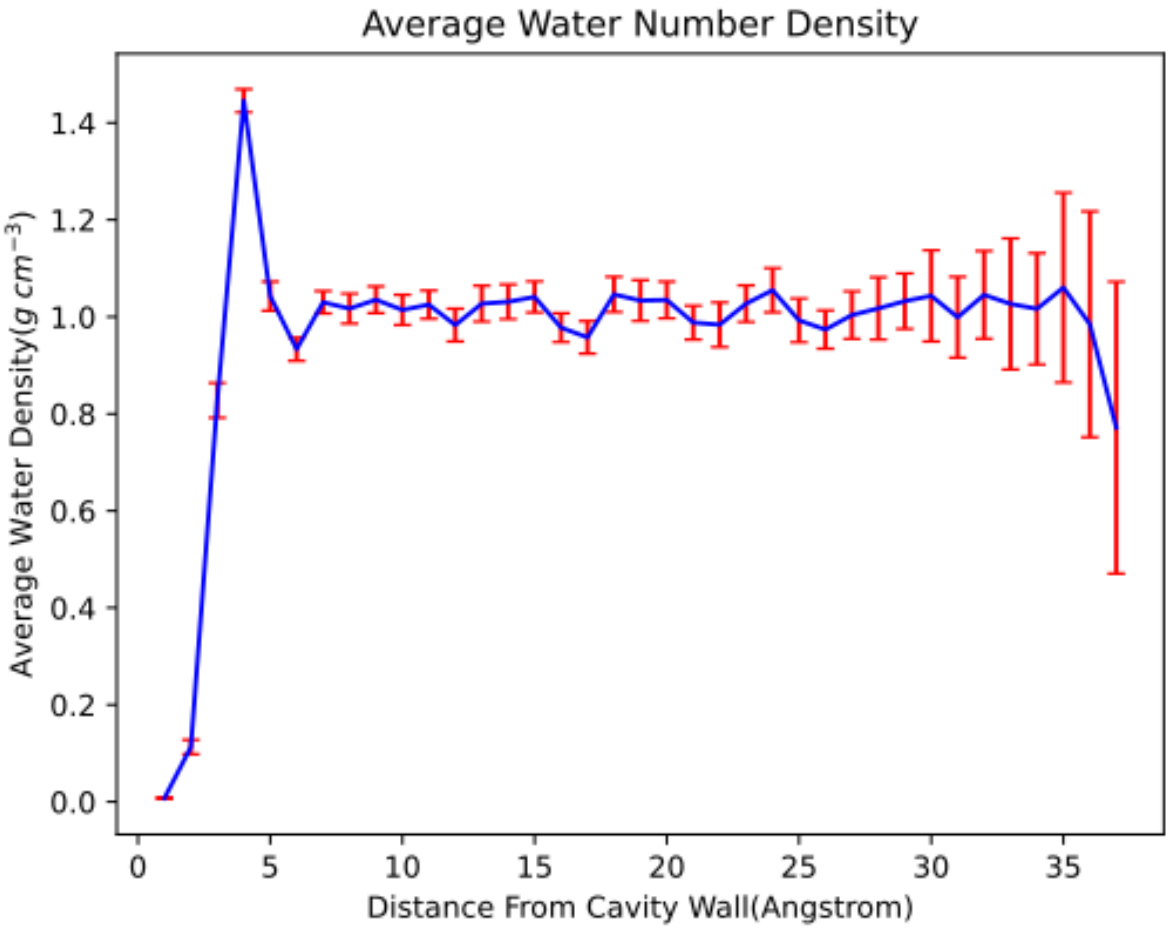

Figure: Average Cavity Water Density Profile

Water Dynamics

1. We provide the functionality to calculate multiple water-dynamical quantities within protein cavities, namely
    1. Water Residence Times
    2. Diffusion Coefficient
    3. Rotational Relaxation Times
    4. Water Movement Propensity
2. Care should be taken while preparing systems for these analyses. Due to the rather fast dynamics of water, to get accurate results, the input trajectories are preferred to be made with a 1 fs integration step with a saving frequency ranging from $1 - 10$ fs. The trajectories provided with this tutorial are only for representative purposes and have a saving frequency of 0.5 ps, which is not preferred for the following analyses.

Water Residence Times

1. Residence time calculations give us on average the amount of time spent by water molecules **continuously** within a determined region.
2. The word "continuously" should be taken seriously, as in this calculation, we consider continuous occupancy to be water molecules that stay within the defined region for every time step between the endpoints t_1 and t_2. A water molecule that leaves the specified region at a time t and again returns at a time $t + dt$ is not considered to be a continuous occupant.

Water Residence Times

Residence times are calculated by defining a region within the protein cavity by providing the distance bounds from the cavity wall using the $-ls$ and $-le$ flags. For example, one would use the following command to calculate the residence time of water in a region $1 - 5\mathring{A}$ from the cavity wall.

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3los_Tutorial.xtc
    -res_time -ls 1 -le 5 -t0 0 -tf 50 -step 0.5 -o 3
    los_Res.png -to_csv 3los_Res.csv -axis_x 0.0052 -axis_y
    13.999 -axis_z 0.0102
```

Water Residence Times: Results

The residence time obtained for the selected ensemble of water is $\tau_{Res}^{1-5} = 20.53ps$

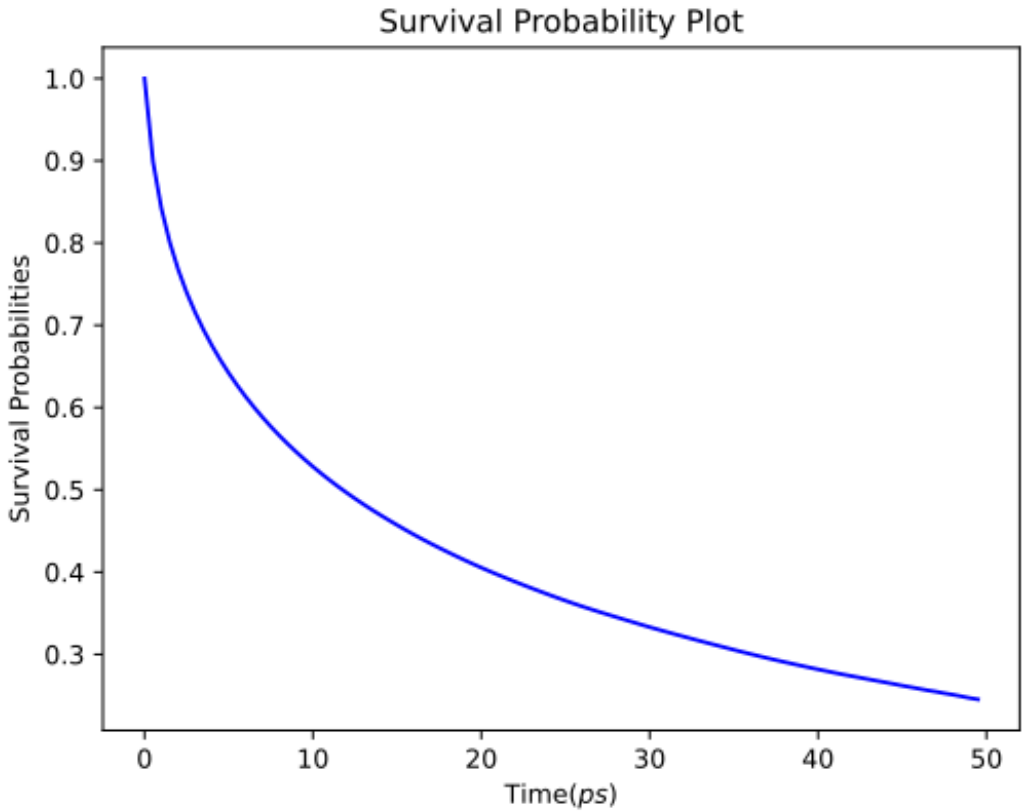

Figure: Survival Probability of Water b/w $1 - 5\mathring{A}$ from the Cavity Wall

Water Residence Times

Alternatively, if one wishes to calculate the residence time for the entirety of the water within the cavity, i.e., if one wishes to calculate the average amount of time that water stays within the protein cavity, one would use the *−all_water* flag to select the entire set of cavity water.

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3los_Tutorial.xtc
    -res_time -all_water -t0 0 -tf 50 -step 0.5 -o 3los_Res
    .png -to_csv 3los_Res.csv -axis_x 0.0052 -axis_y 13.999
    -axis_z 0.0102
```

Water Residence Times

1. The calculation of residence times is quite time-consuming as it involves finding the water molecules within a defined region in the protein cavity over thousands of frames. Running this command on a single node will take a very long time to complete.
2. Hence, to speed up the process, CICLOP supports multiprocessing, which has been implemented using ProcessPoolExecutor. To run your command with multiprocessing, modify the above command as

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -res_time -all_water -t0 0 -tf 50 -
    step 0.5 -o 3los_Res.png -to_csv 3los_Res.csv -
    axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102 -
    multi_process -frame_storage 15
```

Water Residence Times

1. Here, $-multi\_process$ implements the calculations via multiprocessing, and $-frame\_storage$ is used to define the number of nodes you wish to provide. Here, we are parallelly computing 15 frames at a time, and you can set this to any reasonable number based on the number of nodes available to you. The default value of $-frame\_storage$ is 10.
2. One must be careful while multiprocessing as the RAM usage will shoot up significantly.
3. It is to be noted that multiprocessing has also been implemented in all previously mentioned functions that employ multi-frame analyses, such as time-averaged radius, charge, volume, and water density profiles.

Water Diffusion Coefficient

1. The diffusion coefficient of water within a particular region within the cavity can be calculated by using the following command

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -cav_diff -ls 1 -le 5 -t0 0 -tf 10
    -step 0.5 -o 3los_Diff.png -to_csv 3los_Diff.csv -
    axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102
```

2. Here, the choice of the time period over which the calculations is to be done should be based on the value of residence time for water molecules in that particular region.
3. At time steps after the particular ensemble residence time, the water molecules, on average,, have left the region of interest and no longer accurately describe your desired region of water.
4. As before, $-all\_water$ can be used to select all the water within the cavity.

Water Diffusion Coefficient

1. Here, we use Einstein's relationship to find the diffusion coefficient by fitting a straight line to the average squared distance vs time plot.
2. To avoid fitting to the ballistic regime, by default, we fit the data by ignoring the first and last 10% of the data.
3. If one wishes to alter this range, one can make use of the $-fit\_limit$ flag to determine the fraction of data to be overlooked from the initial and final time points.
4. For example, if one wishes not to use the data corresponding to the initial and final 15% of the data, one can use the command

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -cav_diff -ls 1 -le 5 -t0 0 -tf 10
    -step 0.5 -fit_limit 0.15 -o 3los_Diff.png -to_csv 3
    los_Diff.csv -axis_x 0.0052 -axis_y 13.999 -axis_z
    0.0102
```

Water Diffusion Coefficient: Results

The value of the diffusion constant for the selected ensemble is $D^{1-5} = 0.15\ \mathring{A}^2\ ps^{-1}$

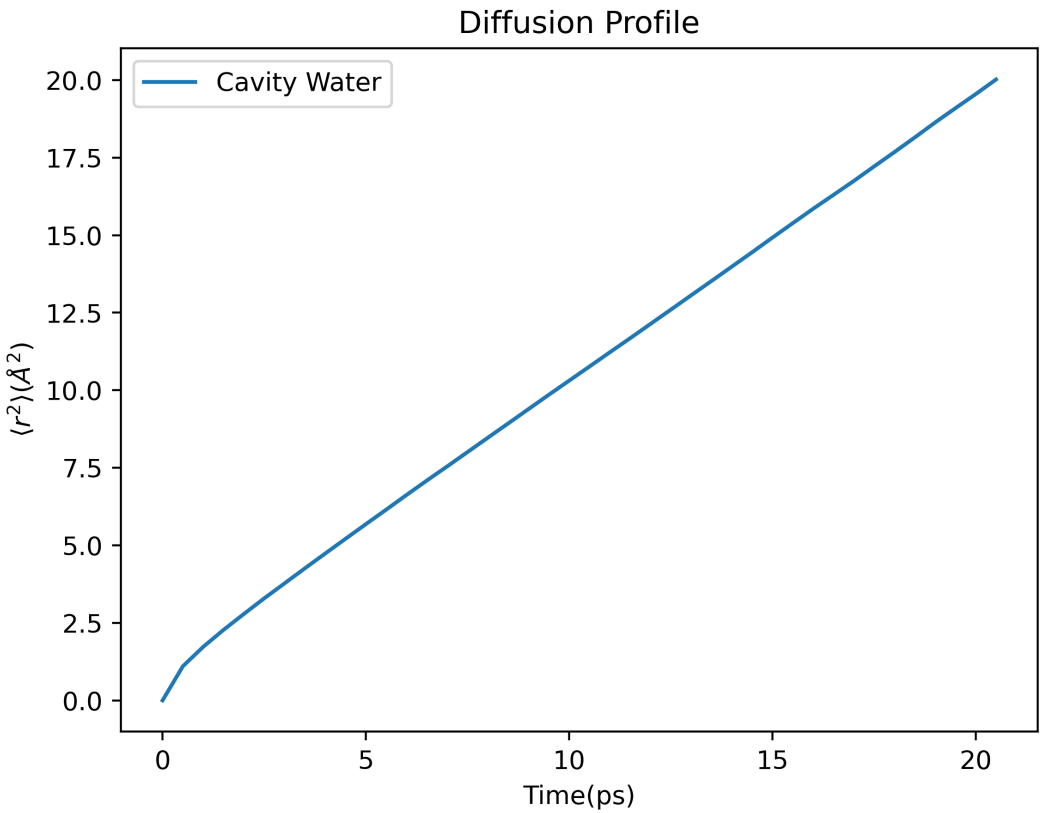

Figure: Diffusion Profile of Water b/w $1 - 5\mathring{A}$ from the Cavity Wall

Rotational Relaxation Times

1. The rotational reorientational relaxation times of water molecules can be found using the command,

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -rel_plot -ls 1 -le 5 -t0 0 -tf 10
    -step 0.5 -o 3los_Rel.png -to_csv 3los_Rel.csv -
    axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102
```

2. Here, as mentioned above, the analysis should only be performed within the residence time of the ensemble of water under consideration.
3. Once again, to select all the water molecules within the cavity, one may use the $-all\_water$ flag.

Rotational Relaxation Times: Results

The relaxation time of water in the selected ensemble is $\tau^{1-5} = 8.44\ ps$

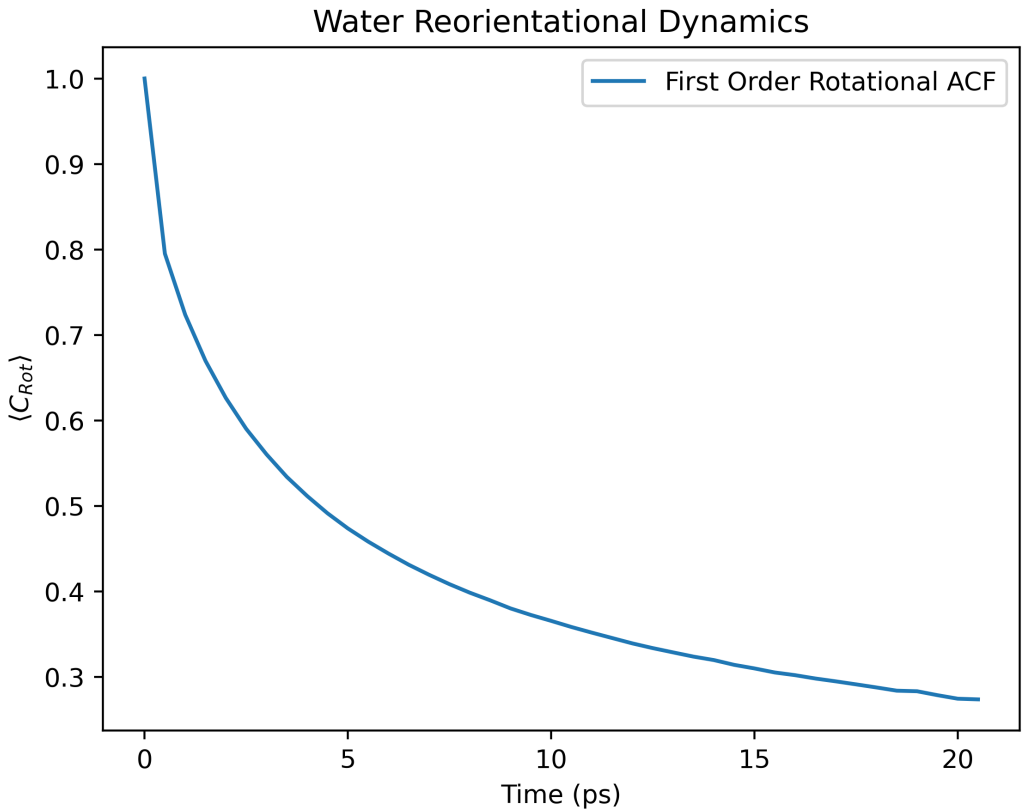

Figure: Orientational Relaxation of Water b/w $1 - 5\mathring{A}$ from the Cavity Wall

Water Movement Propensity

1. We can study the propensity of water to move to different regions from an initial region over time.
2. This analysis tracks the movement of water molecules from an initially identified region over time, i.e., what fraction of water moves towards the cavity wall, what fraction moves towards the center of the cavity, what fraction moves out of the cavity, and what fraction remains where it was initially.
3. The command to run this analysis is,

```
CICLOP -w tip3p -s 3los_Tutorial.tpr -traj 3
    los_Tutorial.xtc -water_prop -ls 6 -le 10 -t0 0 -tf
    7 -step 0.5 -o 3los_Prop.png -to_csv 3los_Prop.csv -
    axis_x 0.0052 -axis_y 13.999 -axis_z 0.0102
```

4. Here, one may make use of the $-multi\_process$ flag to speed up analysis.

Water Movement Propensity: Results

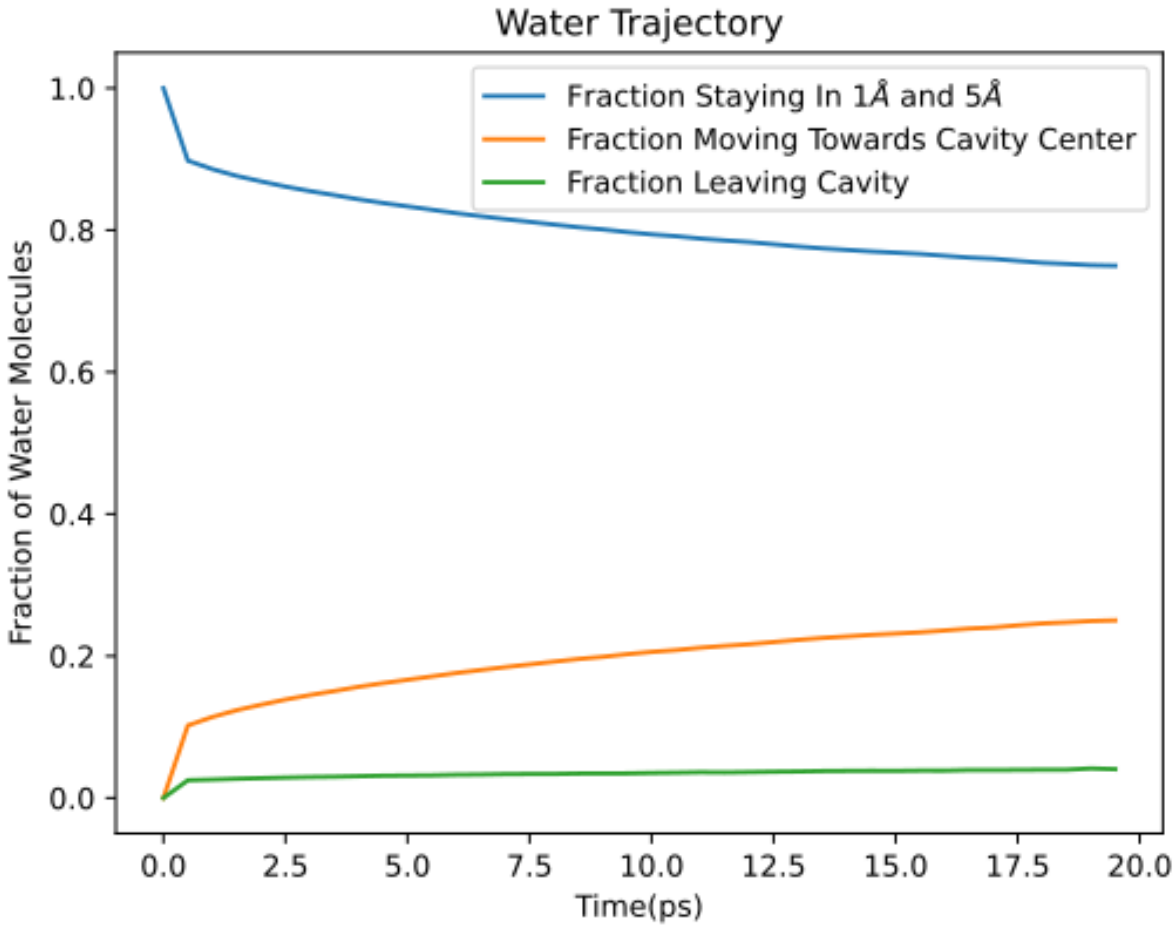

Figure: Water Movement Propensity for Water b/w $1 - 5\mathring{A}$ from the Cavity Wall

Supplementary Flags: *−select*

1. CICLOP uses MDAnalysis for atom selection purposes. Hence, if one wishes to select a particular set of protein atoms for cavity water discernment, one can use the flag *−select "MDAnalysis selection string"*. This allows one to use the MDAnalysis selection language to select regions of interest.
2. One can use this flag to select non-protein groups as well. If one wishes to find the cavity water within a liposome, you can use the *−select* flag for selecting your liposome lipids for cavity identification.
3. Default setting for the *−select* flag is the string *"protein"*.

Supplementary Flags: *−sol_select*

1. CICLOP uses the water oxygen atoms as a reference to carry out all analyses. By default, this selection is made assuming the solvent residue name is "*SOL*", and the water oxygen has a name "*OW*".
2. Since this may not always be the case, the *−sol_select* flag allows one to select water oxygen atoms that are named with different nomenclature.
3. For example, if the residue name is "*TIP3*" and the name is "*OH2*", one can use the flag as *−sol_select* "*resname TIP3 and name OH2*".
4. **The selection must always be for the water oxygen and not the whole molecule.**

Supplementary Flags: *–fit_group*

1. By default, CICLOP uses the protein backbone atoms, "*CA*" or "*CB*", i.e., protein backbone carbon atoms, to produce the best-fit axis.
2. If one wishes to select a different group for alignment, one can use the *–fit_group* followed by their desired selection in MDAnalysis selection language to make this selection.
3. Eg: *–fit_group "name ROH and resname CHOL"*
4. The default selection is "*name CA or name CB*"

Notes : Input Files

1. Analysis of dynamical quantities associated with water requires one to create trajectories that have $1fs$ integration time steps with a saving frequency of $1 - 10fs$ with a total simulation time of $25 - 100ps$. This is absolutely necessary for properly interpretable results due to the rather fast dynamics of water.
2. Structural features such as cavity radius, volume, charge, and water density profiles can be discerned with standard trajectories having $2fs$ integration timesteps and saving frequencies ranging from $10 - 100ps$.
3. Charge profile generation requires the provision of structure files containing both coordinate and atomic partial charge information.

Notes : Output Files

1. CICLOP makes use of MDAnalysis to write out files from selected atom groups. Hence, all MDAnalysis-supported structure files are allowed as output.
2. B-factor/tempfactor modification is possible only if input files contain B-factor/tempfactor information. CICLOP provides B-factor/tempfactor modified outputs only if both the input and output files are in the **PDB** format.

Notes : Selections

1. The *−select*, *−sol_select*, and *fit_group* flags take inputs in the form of MDAnalysis selection strings. One can freely select the protein residues to be selected as structure.
2. *−sol_select* should be used in the selection of **water oxygen atoms**, and not whole water molecules unless the water model is coarse-grained. The default selection assumes a residue name of SOL and the water oxygen to be named OW.
3. CICLOP does not support analysis of implicit solvent simulations.
4. CICLOP supports analysis of coarse-grained MD simulations as well.
5. One must use the option *CG* with the *−w* flag to enable the selection of water beads in the simulation.