# Exercise_7 Molecular Dynamics Simulation

Andrea Scaioli

June 2025

## 1 Introduction

The goal of this work is to compute the binding free energy $\Delta G_{\text{bind}}$ of a biomolecular complex by performing umbrella sampling simulations, followed by analysis using the Weighted Histogram Analysis Method (WHAM). In particular, we study the dissociation of an outer protofilament from the A$\beta$(1–42) amyloid fibril—an important system in the context of neurodegenerative diseases—using classical molecular dynamics (MD) simulations.

Standard MD simulations often struggle to sample rare events such as binding or unbinding transitions due to the presence of high free energy barriers. To overcome this limitation, **umbrella sampling** is employed. Umbrella sampling enhances sampling efficiency by applying harmonic biasing potentials to restrain the system along a predefined reaction coordinate, typically called a collective variable. In this case, the collective variable is the center-of-mass (COM) distance between the protofilament and the rest of the fibril.

Multiple simulations, each centered around a different value of the reaction coordinate, are carried out using these harmonic biases. Each simulation (or "umbrella window") preferentially samples a narrow region of the reaction coordinate space. By ensuring sufficient overlap between the sampled regions, one can reconstruct the unbiased free energy profile along the coordinate.

To extract the free energy surface (FES) from the biased data, the **Weighted Histogram Analysis Method (WHAM)** is employed. WHAM combines the histograms obtained from each umbrella window, taking into account the biasing potentials and sampling weights, to iteratively estimate the unbiased probability distribution. The free energy as a function of the collective variable is then recovered using the relation

$$G(i) = -k_B T \ln \left( \frac{\langle P(i) \rangle}{\langle P_{\text{max}} \rangle} \right),$$

where $\langle P(i) \rangle$ is the estimated unbiased probability for bin $i$ and $\langle P_{\text{max}} \rangle$ is the maximum probability observed. The binding free energy $\Delta G_{\text{bind}}$ is obtained as the difference in free energy between the bound and unbound states along this profile.

This method allows for the accurate estimation of binding affinities, even for strongly interacting systems, and provides detailed insight into the free energy landscape governing biomolecular association and dissociation processes.

## 2 Methods

The molecular dynamics simulations were performed using the GROMACS software suite, following a standard umbrella sampling workflow. The system under investigation consists of an A$\beta$(1–42) amyloid fibril, and the goal was to compute the binding free energy $\Delta G_{\text{bind}}$ associated with the separation of one of its outer protofilaments.

### System Preparation

The first step involved preparing the simulation system. The protein structure (PDB ID: 2BEG) was solvated in a triclinic box filled with SPC water molecules, ensuring sufficient space along the pulling axis

to accommodate the dissociation process. Ions were added to neutralize the system and mimic physiological ionic strength. The topology and coordinate files were generated using standard GROMACS commands.

## Energy Minimization and Equilibration

After constructing the system, energy minimization was performed to relax any steric clashes. This was followed by equilibration under NPT conditions to allow the solvent and ions to equilibrate around the protein. The resulting structure provided a stable starting point for the pulling simulations.

## Index File and Group Definition

An index file was created using `gmx make_ndx` to define custom groups corresponding to the two protofilaments involved in the binding/unbinding process. These groups were named `Chain_A` and `Chain_B`, representing the parts of the amyloid fibril to be separated.

## Pulling Simulation

A pulling simulation was conducted to generate initial configurations for umbrella sampling. This was automated via the bash script `02.simprep.sh`, which extracts snapshots along the center-of-mass (COM) distance coordinate. Harmonic restraints were applied to pull one protofilament away from the fibril in the $z$-direction. The script also outputs the relevant GRO files and distance data needed for the subsequent umbrella simulations.

```
#!/bin/bash
# This line tells the system to execute the script using the Bash shell.

# === STEP 1: Prepare the input binary file for MD run ===
gmx grompp -f md_pull.mdp -c npt.gro -p topol.top -r npt.gro -n index.ndx -t npt.cpt -o pull.tpr -maxwarn 3
# gmx grompp: Preprocesses input files for the MD run
# -f md_pull.mdp     : MD parameter file for pulling simulation
# -c npt.gro         : Input structure (from NPT equilibration)
# -p topol.top       : Topology file
# -r npt.gro         : Reference structure (usually same as -c)
# -n index.ndx       : Index file containing custom groups
# -t npt.cpt         : Continuation file (checkpoint)
# -o pull.tpr        : Output run input file
# -maxwarn 3         : Allow up to 3 warnings before stopping

# === STEP 2: Run the pulling MD simulation ===
gmx mdrun -ntomp 4 -deffnm pull -pf pullf.xvg -px pullx.xvg
# gmx mdrun: Runs the simulation
# -ntomp 4           : Use 4 OpenMP threads
# -deffnm pull       : Default filename prefix (will produce pull.xtc, pull.edr, etc.)
# -pf pullf.xvg      : Output pulling force vs time
# -px pullx.xvg      : Output position of pulled group vs time

# === STEP 3: Measure distance between groups during MD ===
gmx pairdist -f pull.xtc -s pull.tpr -n index.ndx -sel "com of group Chain_A" -ref "com of group Chain_B" -o dist.x
# gmx pairdist: Computes distances between pairs of atoms/groups
# -f pull.xtc        : Trajectory file
# -s pull.tpr        : Input structure with topology
# -n index.ndx       : Index file for group selection
# -sel ...           : Select center-of-mass of group Chain_A
# -ref ...           : Reference: center-of-mass of group Chain_B
# -o dist.xvg        : Output file with distances

# === STEP 4: Extract frames at distance increments of 0.2 nm ===
sed '/^\@/d' dist.xvg |\
```

```
# Remove lines starting with '@' (plotting metadata)
sed '/^\#/d' |\
# Remove lines starting with '#' (comments)
awk '$2 > i+0.2{a[k++]=$1;i=i+0.2}END{for(j=0;j<501;j++)if(a[j] != 0.000 && j > 0 || j < 1 )printf("%7.3f%1s",a[j],
# awk:
# For each line, if the distance (column 2) increases by more than 0.2 nm from the last stored,
# save the time/frame (column 1) to array a[].
# Then print the first 501 non-zero time points, one per line, with 3 decimal precision.
# Output is a list of frames (in ps) where the COM distance increased by ~0.2 nm.

# === STEP 5: Extract coordinate snapshots at selected frames ===
for frame in $(cat important_frames.txt)
do
    printf "0\n" | gmx trjconv -s pull.tpr -f pull.xtc -o conf${frame%%.*}.gro -b $frame -e $frame
    # gmx trjconv: Extracts a single frame from the trajectory
    # -s pull.tpr       : Input structure
    # -f pull.xtc       : Trajectory
    # -o confXX.gro     : Output frame file, named using the frame time (before decimal)
    # -b £frame         : Begin at this frame (time in ps)
    # -e £frame         : End at this frame (same as begin, so only one frame)
    # printf "0\n"      : Selects the 'System' group automatically (group 0)
done

```

### Umbrella Sampling Simulations

Each extracted snapshot served as the starting structure for an individual umbrella sampling simulation (referred to as an umbrella window). A harmonic bias potential was applied in each window to restrain the system at a specific COM distance. The simulations were performed in two steps: a short NPT equilibration followed by a 500 ps production run. The spacing between windows was approximately 0.2 nm to ensure sufficient overlap between sampling distributions.

### Free Energy Analysis

After completing all umbrella sampling simulations, the results were analyzed using the `gmx wham` utility. This tool implements the Weighted Histogram Analysis Method (WHAM) to reconstruct the unbiased free energy surface along the COM distance coordinate. The final binding free energy $\Delta G_{\text{bind}}$ was extracted as the difference between the free energy of the bound and unbound states on the profile.

## 3 Results

# Results

### Pulling and Umbrella Sampling Trajectories

Figure 1 illustrates the key observables monitored during the pulling simulation that generated the initial configurations for umbrella sampling. Panel (a) shows the increase in center-of-mass (COM) distance between the two protofilaments as a function of simulation time, confirming that the pulling was effective in gradually separating the outer filament from the rest of the amyloid fibril. Panel (b) tracks the absolute position of the pulled protofilament's center of mass, while panel (c) displays the corresponding pulling force over time. The force profile is consistent with the expected response from a harmonic potential, with peak forces appearing as the protofilament resists separation due to intermolecular interactions.

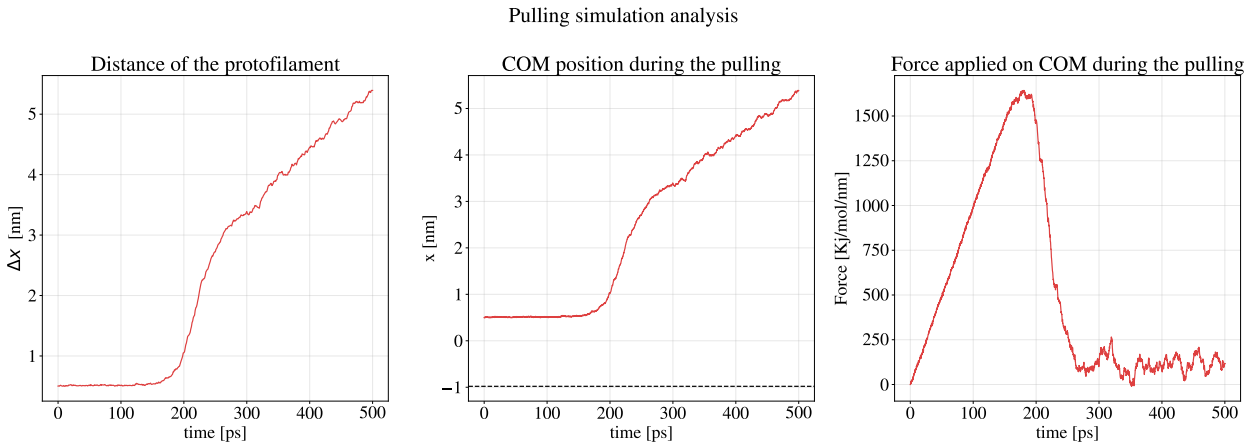


Figure 1: (a) Center-of-mass distance between the protofilaments over time, (b) COM position of the pulled filament, and (c) force applied during pulling simulation. These results were obtained from the initial pulling run used to generate starting configurations for umbrella sampling.

## Free Energy Surface and Binding Energy Estimate

Following the umbrella sampling simulations, the Weighted Histogram Analysis Method (WHAM) was used to compute the free energy surface (FES) along the COM distance coordinate. The results are shown in Figure 2, where the obtained FES profile is compared with the literature reference. While the overall shape and trend of the computed FES are qualitatively consistent with expectations, the limited sampling in some regions led to discontinuities and artefacts in the profile.

These issues are a consequence of a subset of umbrella sampling windows failing during simulation, leading to insufficient histogram overlap in specific regions of the reaction coordinate. As a result, the estimated binding free energy $\Delta G_{\text{bind}}$ is lower in magnitude compared to the reference value of $-50.5$ kcal/mol. Nevertheless, the windows that completed successfully yielded physically meaningful results, and the partial FES still provides a useful qualitative picture of the binding process.

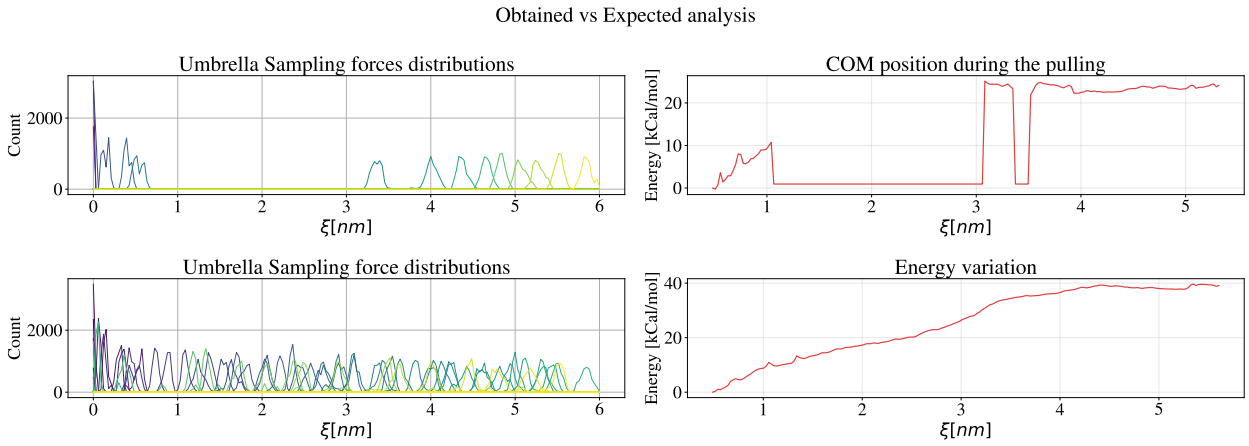


Figure 2: Comparison between the computed free energy surface (blue) and the literature reference profile (gray dashed line). The mismatch is primarily due to missing umbrella windows caused by simulation failures.

To improve the accuracy of the free energy estimate and resolve the gaps in the FES, targeted post-sampling simulations in the under-sampled regions are recommended. These additional simulations would enhance the histogram overlap and lead to a more reliable reconstruction of the full free energy profile.