Structural bioinformatics

Assignment 3

Lianne van Ruijven

Question 1

The technique used in the May et al. (2014) paper for calculating the PMF is constraint force integration.

→ Describe the essential difference between constraint force integration and umbrella sampling.

* The F mean (which is a function of the reaction coordinate), is calculated between a window of the reaction coordinate, when using umbrella sampling. For constraint force integration, the Fmean is calculated for specific values of the reaction coordinate.

→ Name two main assumptions that need to hold for these methods to work correctly (only one is explicitly mentioned in the paper).

* constant N (number of particles), V (volume), E (energy))
* Assumes equilibrium

# Starting up

The 1VET.pdb file contains coordinates.. how do you get the kinetic energy from a pdb file?

## Getting the necessary files

* For the 1VET structure: there is a chain A and chain B!
* Je kan zien welke delen elkaar raken.. is volgens mij de betastrands die dicht bij elkaar liggen

## Generating the topology

In this part:

* the pdb file is checked to be complete.. so are all atoms present in all amino acids.. The file 1vet\_add.pdb has alle the atoms that would be missing, added.
* Then the potential energie from the coordinates is determined.

done

## Setting up the system

Set up the size of the system.. so putting the system in a box of 1.2 nm around the molecule and it is centered in the box.

Conf.gro contains our molecule centered in a box

done

## Energy minimizing the system

We want to minimize the energy because… try to stabilize the structure. 🡪 The goal of energy Minimization is to find a set of coordinates representing the minimum energy conformation for the given structure.

* Output file Em/fit.pdb contains both structures, means: 1VET before and after minimizing the energy. Difference is super small, you cannot really see any difference. The super long loop in the protein connects the two chains.. this happens because the difference between the two chains is not that clear.. so this is an artifact. Chromacs doesn’t know about different chains..
* 0.098 is wat Fabienne krijgt. Dit is best wel laag.. dus komt wel overeen met de 3D model in chimerea.. dat de rmsd > 0 komt waarschijnlijk door de lange loop in het midden.

done

# Umbrella sampling

## Coarse-graining

Oke done, is just explenation

## Distance

Distance from c1 to c2 is 23.535

For the umbrella sampling:

* minimal is 23.535 armstrong = ongeveer 2.3 nm
* max ?? paper 4.44 as max

## Equilibrations

Ik heb range 2.2:4.5 gekozen

python3 ../src/Pull.py -f 1VET-em.pdb -c A:B -d 2.2:4.5:0.1

Lager dan 1.7 gaat niet zo lekker, omdat dan de twee proteïnes te dicht bij elkaar komen.

Question 2

Em\_sol\_posre 🡪 potential energy minimum. So there is a really low temperature. Because there is nothing going on there

MD\_sol\_posre 🡪 adding temperature by adding velocitys. Now we equilibirate the whole system. Bu tprotein can’t move only water.

Md\_sol\_eq 🡪 same as step above, but now protein can move as well. Just with a smaller time. This also makes sure that your results that your are analyzing, do not contain the part where you simulation is still adapting to the water mollecules. This is namely not a correct result which you want to use.

We are sampling with the umbrella sampling. So each sample is a starting configuration in a box filled with water for a MD simulation. Why do we run short simulations to sample?

* so the water can adjust to interacting with the proteins
* so we know the starting configuration is in equilibrium
* we can set the simulation parameters (NVT) in which we run the equilibrium simulation.
* so that the system relaxes sufficiently in the target ensemble to allow the production run to be commenced. Relaxes means having it in a energy minimum.

The difference between EM and MD is:

* The goal of energy Minimization is to find a set of coordinates representing the minimum energy conformation for the given structure . This energy conformation is the potential energy. So in EM, only the bonded (bonds, angles, dihedrals and impropers) and non-bonded (coulomb, van der waals) forces can be adjusted. This won’t have large effects on the secondary structure of a protein and the changes won’t be large.
* MD takes potential (coordinates) and kinetic (velocity) energy into account. An MD simulation without constraints can totally unfold and fold a protein, so has a large effect..

Anton: when you are starting, you are transferring your structure in a different structure multiple times. From only coordinates, to water environment. In the water environment, you don’t want to have a disruption in the simulation because of a wrong environment energy.

## Production simulations

Why did some distances fail?

2.2 and 2 and 3.3 failed for Fabienne.. you can see this at that there is no folder for this distances.

Anton doesn’t really know.. it also happens in the paper.. maybe this happens because the proteins are in a box.. maybe at these distances the interaction is on the other side of the box… 2.2

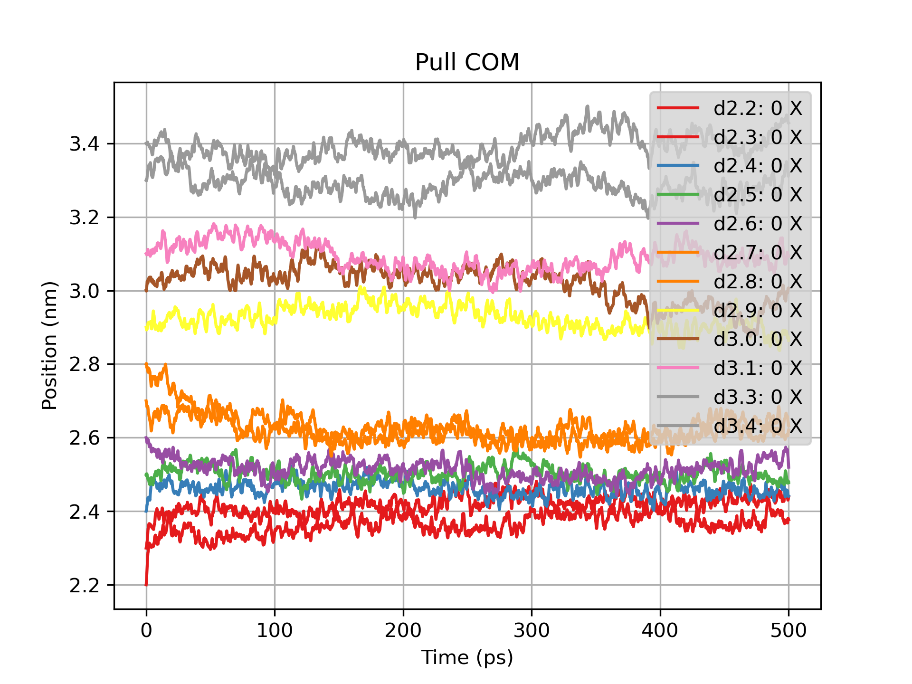
DONE!

# Analyzing the results

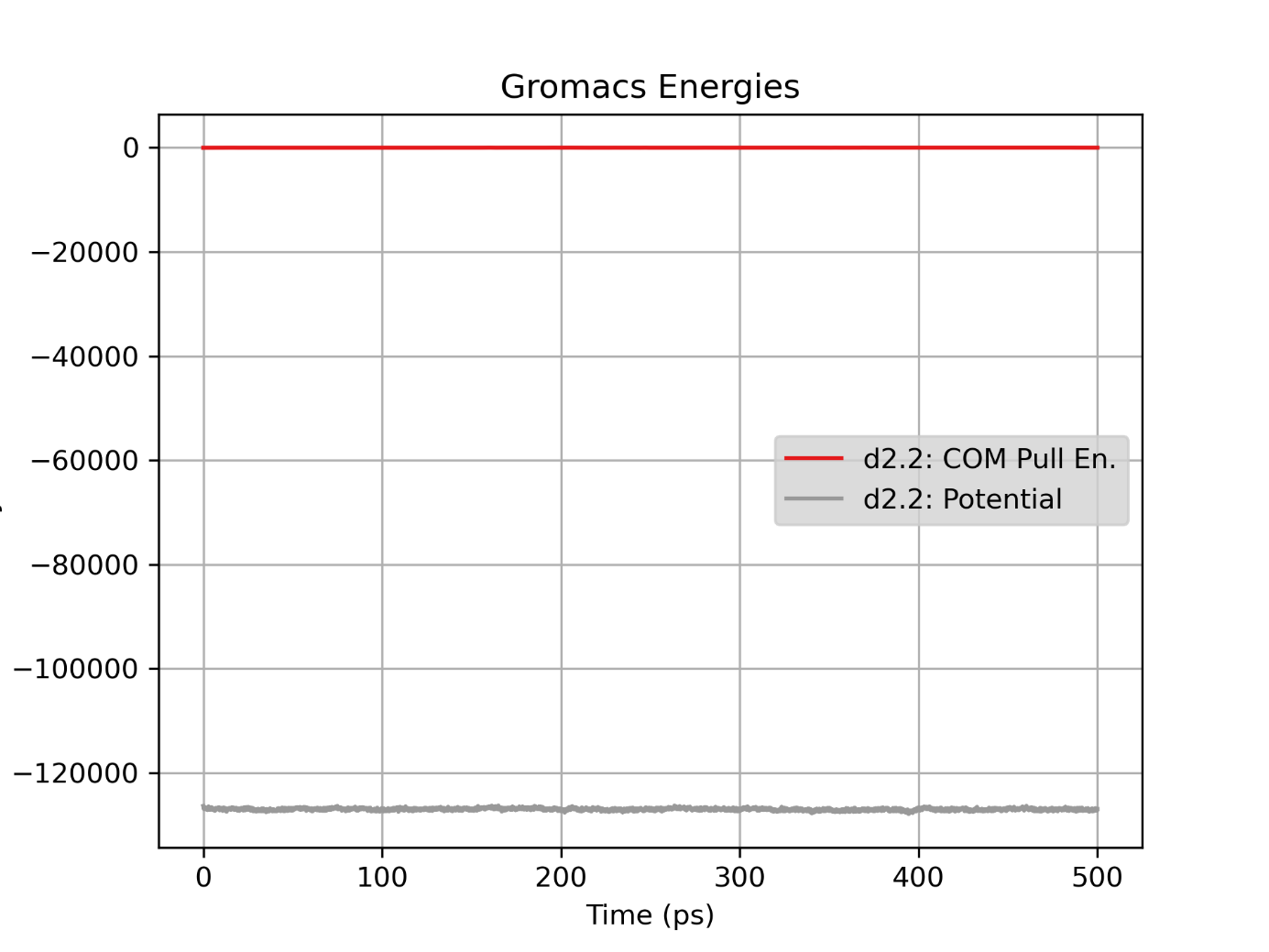
## Analysis according to umbrella sampling

### Extracting energies and distances

Center of mass distances as function of time:

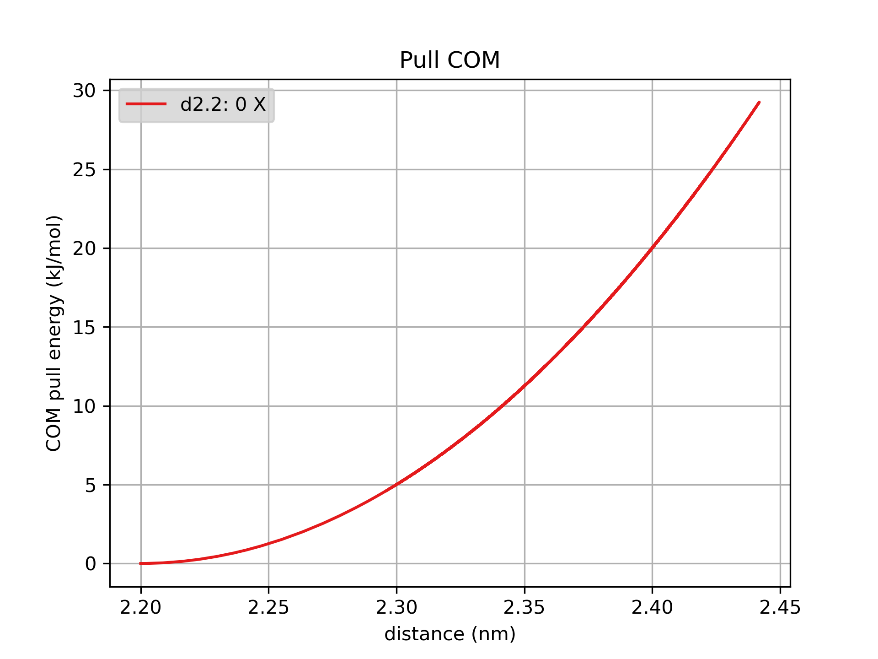


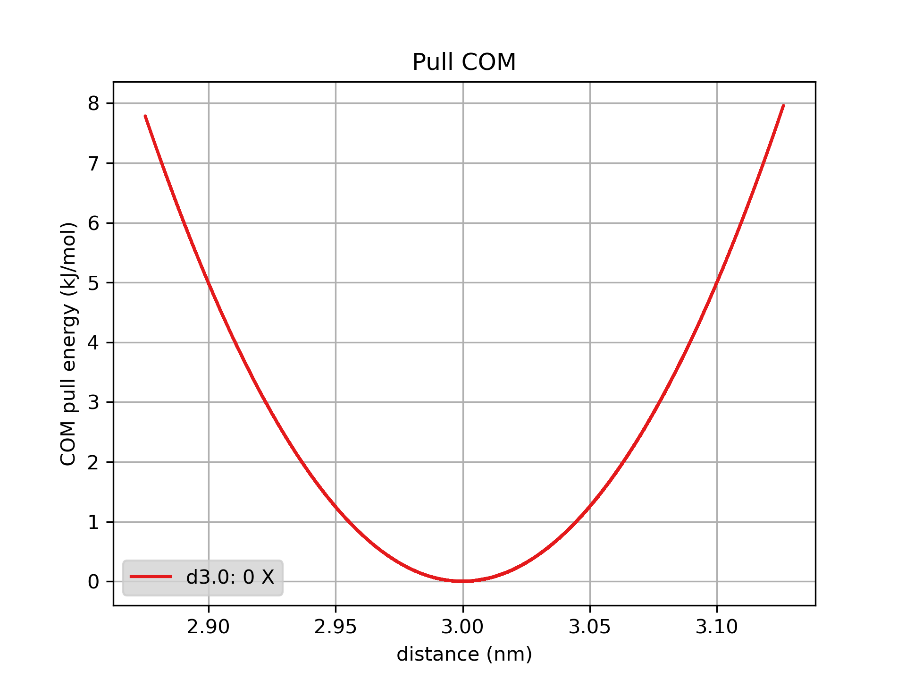
Energy potential of sample with distance 2.2. COM pull En = umbrella potential energy and Potential = total potential energy.



When considering the shape of the umbrella potential (Q3-5, and section 5.1.2), you may want to re-read section 15.8.1 and .2 in the book "Umbrella Sampling in MC" and "... using quadratic potentials", and the figure and box on the following two pages.

Question 3





D\_0 is the starting distance of the umbrella sample and d is the current distance in the simulation. The COM pull energy has a relation with the distance between d\_0 and d (Eumbrella = k\_umbrella(d-d0)^2 ). This means, that when the distance between d0 and 1 increases, the COM pull energy also increases. And this means that when d0 lays at a distance in which the free energy function is on a slope, the simulation will move away from d0 and towards the free energy minimum in the umbrella region. This is the simulation shown in the figure with the d2.2 plot. Now let say the d0 lays in a free energy minimum or far away from a free energy minimum, like on a platue, like in the figure with plot for d3.0. The d will move around the free energy landscape during the production simulation, but because the d0 lays in a free energy minimum or far away from it, the probability of d being similar to d0 is bigger. The simulation is not drawn away from d\_0. So the COM pull energy will be smaller around the free energy minimum (which is around 3 nm distance in plot for d3.3).

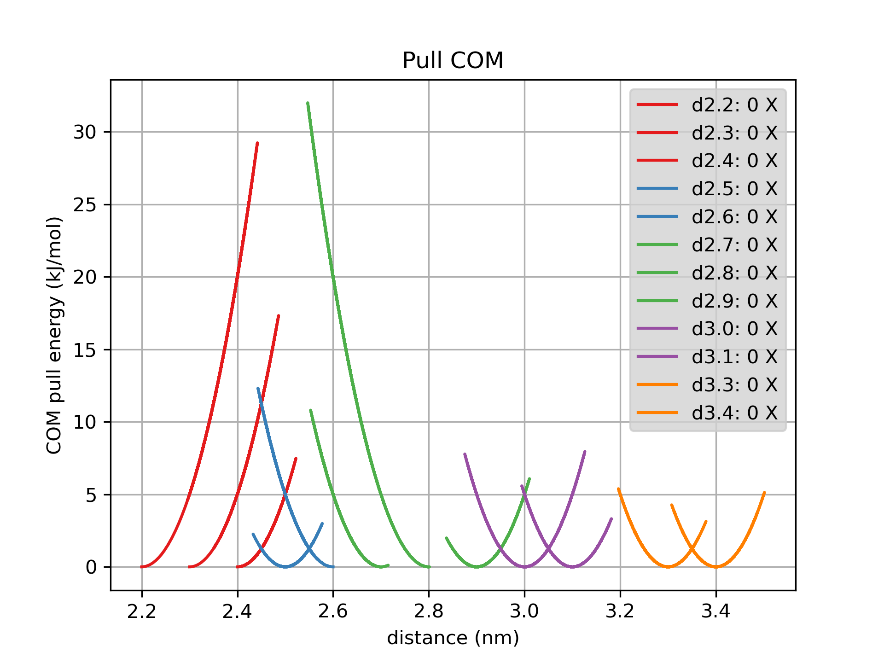
For an infinitely long sample (so with an infinit long simulation time) you would excpect a similar plot. The plot which shows the COM pull energy for d2.2 may increase to a larger maximum COM pull energy. But because of the umbralla sampling, the simulation has a distance restrain. This means (for plot d2.2) the following counts: 0 >=(d\_0 – d)^2 >= (d\_0 – d\_free energy minimum ) ^2.

Command line: belangrijk dat je eerst dist.xvg en dan energy.xvg als input gebruikt:

(structbio) -bash-4.4$ python3 PlotCOM\_vs\_distance.py -xvg\_f d2.2/md\_sol\_prod/dist.xvg d2.2/md\_sol\_prod/energy.xvg -o comvsdist\_d2\_2\_test

Question 4

python3 Merge\_files.py -xvg\_f d?.?\*/md\_sol\_prod/dist.xvg d?.?\*/md\_sol\_prod/energy.xvg -o plot\_all



The free energy minimum lies around a distance of 2.5 nm. This is based on the fact that the plots for d2.2, d2.3, d2.4 are not parabolic but have a upward slope, which means the simulation moves away from the starting distance (d\_0) and moves towards a larger distance, which lays around the free energy minimum. Because the COM pull energy increases when d\_0 – d increases, the COM pull energy increases in this plot. The plots for d2.5, d2.6 and d2.7 have a downward slope. This means as well that during the simulation, the d moves away from the d\_0, but because the distance at which the free energy lays is smaller than the d\_0, the COM pull energy gets smaller at a higher d.

Question 5

(structbio) -bash-4.4$ python3 Average\_potential.py -xvg\_f d?.?\*/md\_sol\_prod/dist.xvg d?.?\*/md\_sol\_prod/energy.xvg -a -o plot\_average\_potential

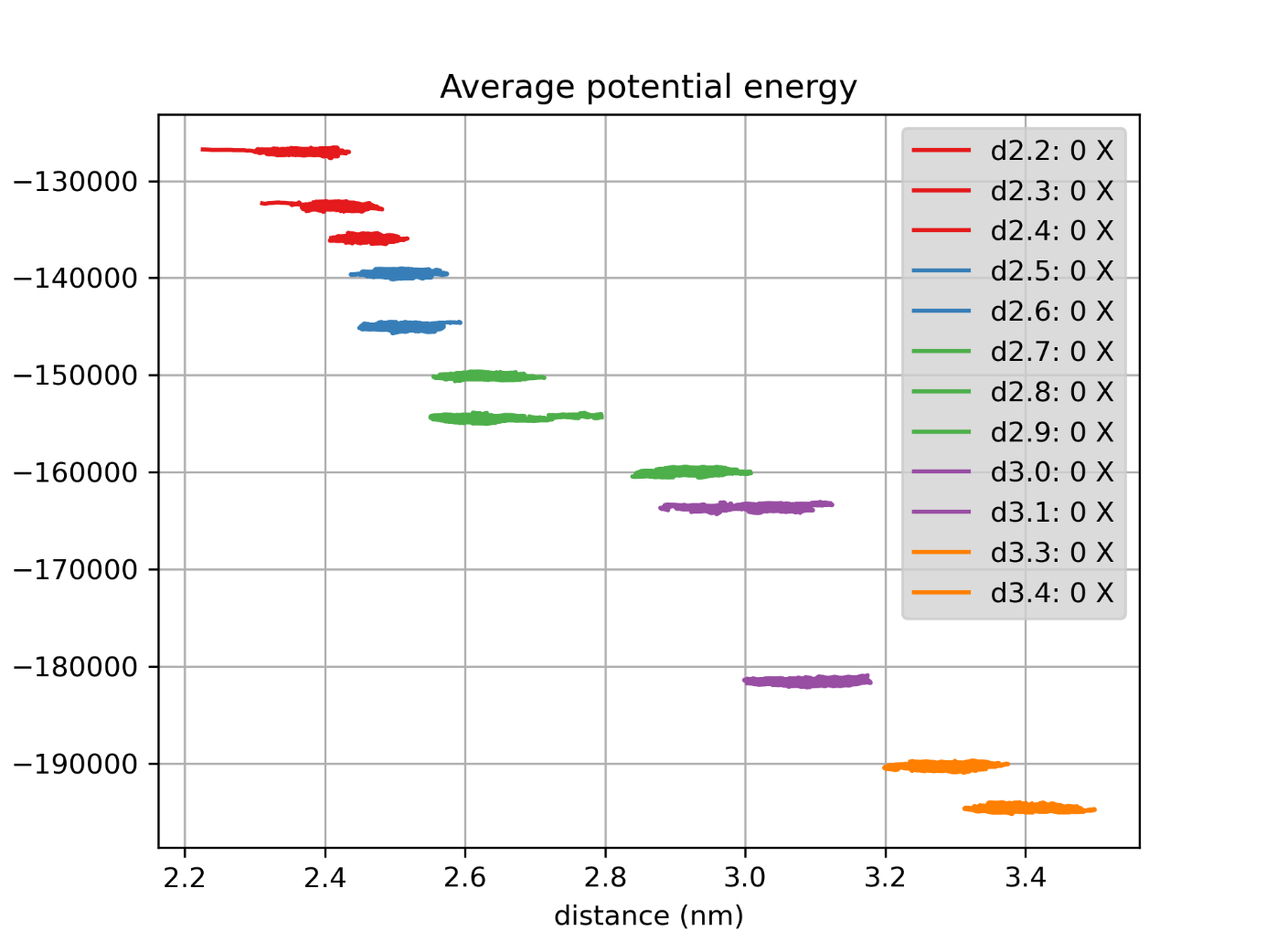
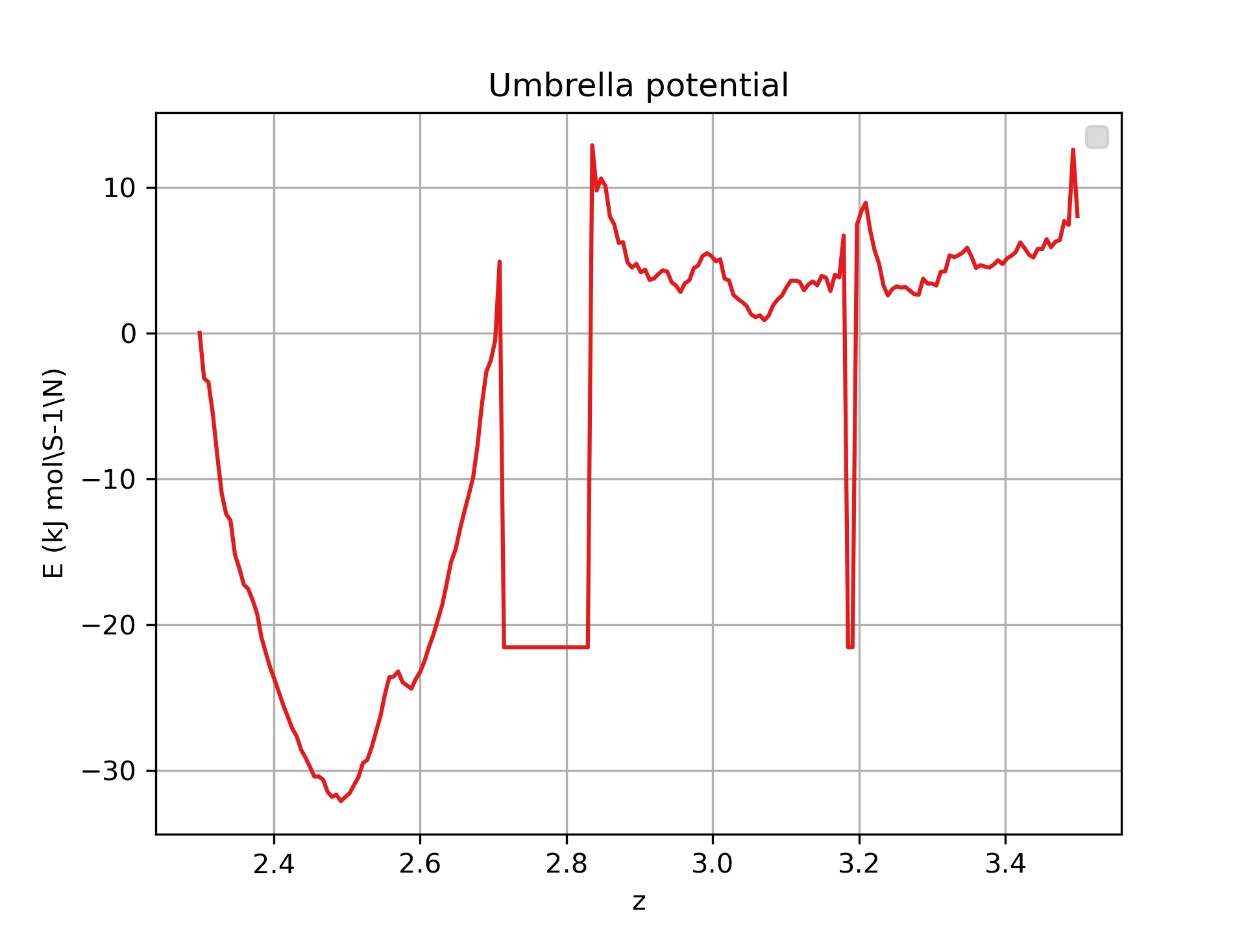
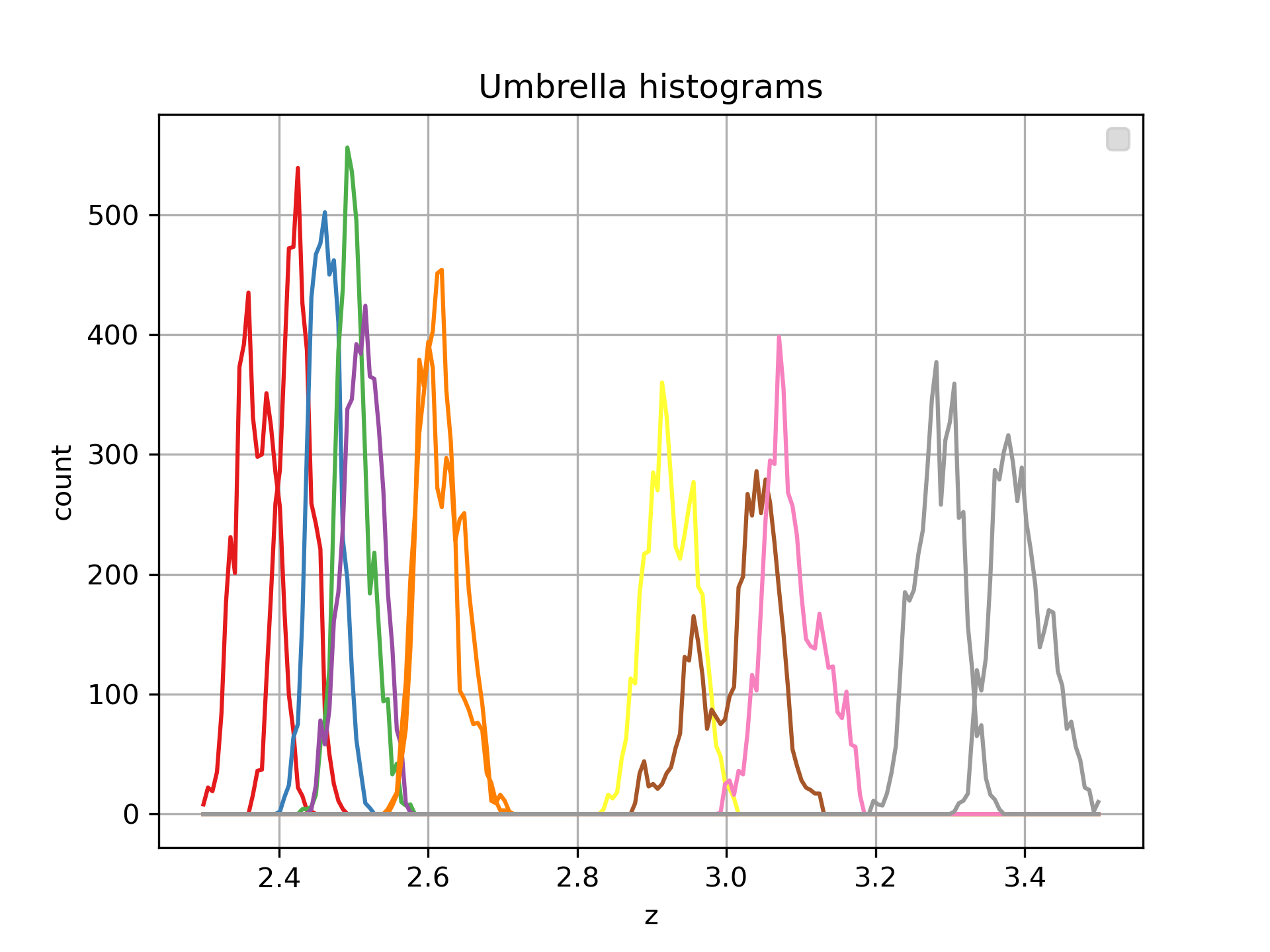


Figure: average potential energy during 12 simulations (d2.2 till d3.4). The average potential energy decreases.

### Unbiasing the umbrella potential

## Automated analysis using WHAM





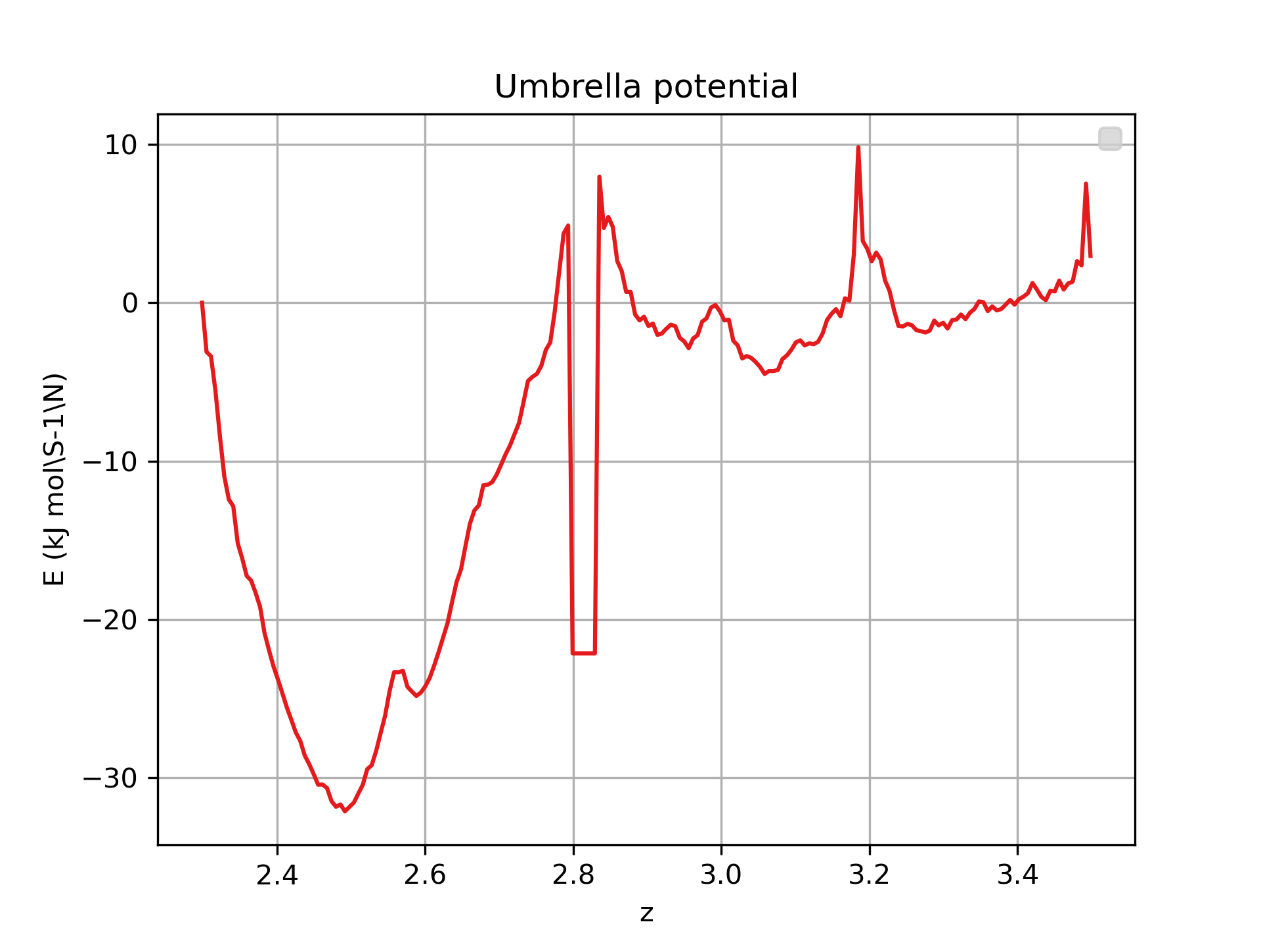
Question 6

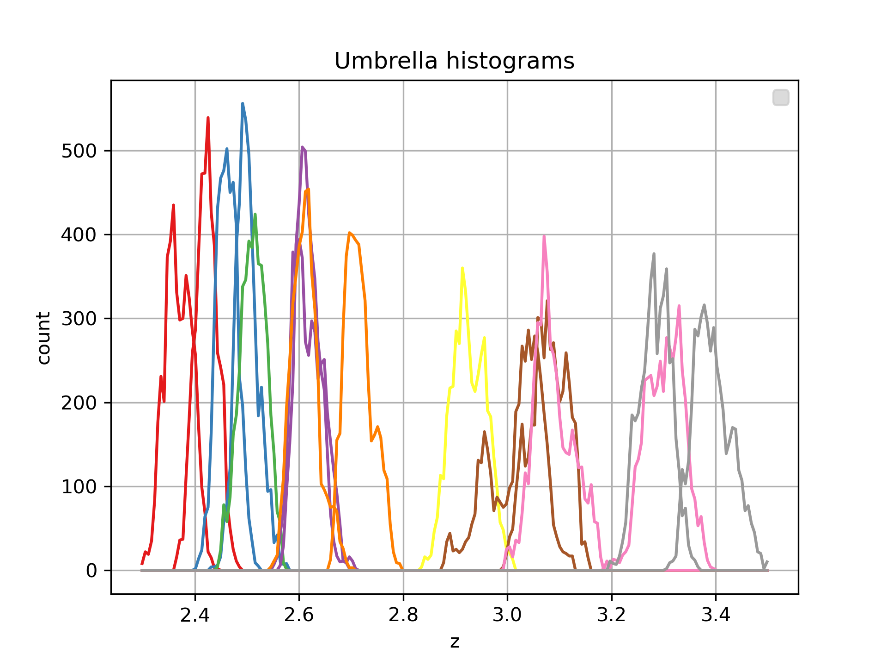
* In the histogram there is a gab between distances 2.7 and 2.8. This means there are no samples from this range and this results in the PMF plot, at the straight drop of the line and flat horizontally line between approximately 2.7 and 2.8. The reason for this, is that d2.8 lays on the top of the downwards slope and makes the simulation move to a smaller distance, closer to the free energy minimum.
* In the histogram you also see smaller and higher paraboles between 2.4 and 2.6 than between 2.8 and 3.4. This is because, as shown in the PMF plot, there is a energy free minimum at 2.5. So the number of time the simulation ends up between 2.4 and 2.6 is higher. Between 2.8 and 3.4, the simulation does not have a attraction to a certain distance because there is an energy plateau. This results in the paraboles be more wider and lower.
* The overlap between 2.4 and 2.6 is condence. This is because, regardless the starting distance, all the simulations in this eara are drawn to the same minimum free energy around 2.5.

Question 7

I choose to combine my output with the output of Fabienne Kick. So she gave me her simulation results for d2.7, 2.8, 3.1 and 3.3. This improved my PMF plot. It decreased the gap around 2.8 and removed the gap around 3.2.

When comparing my improved results with two others that used the first solution (so thy rerun the production simulations with different starting points), they don’t have any gaps anymore. This can be because they changed the starting position.





Question 8

No, only the sampling methods are different. Determining the PMF is similar, so in case of ideal sampling, the graphs should be the same. Comparing my graph with the one in the paper.

Question 9

steric clashes, physical contact, hydrophobic effect, water-mediated interactions

Downward slope: 2 proteins are in a bound state. But not the most favorable state. Eather the proteins are to close to each other, or the distance between the 2 proteins is similar as in the free energy minimum but the formation of the proteins is different so the COM distances gets smaller. Steric clash.

Global minimum: Bound state of the 2 proteins. Most stable state so highest probability to find this state.

Upward slope: Larger distances between the proteins, energy is increasing. There is water mollecules between the proteins. But it stilly wants to go back to the global minimum.

Plateau: The two proteins are unbound totally. There is no energy pulling the two proteins towards each other. Hydrophob

Question 10

The MARTINI force-field does not explicitly describe backbone hydrogen bonds. → Where may backbone hydrogen bonds be particularly important for this complex? Refer back to the structural properties of the interface you looked at in the beginning of the practical (section 3.1).