Molecular Dynamics Exercise VIII: Self-supervised Protein Simulation with Convergence and Principal Components Analysis

Tristan Alexander Mauck

December 23, 2024

This report deals with the simulation of the protein HIV-protease (5YOK) solvated in water. The R_{gyr} and RMSD(t) were calculated and analysed for convergence using block standard error analysis. Finally, a principal component analysis of the protein's movement was conducted. All simulations were performed using GROMACS.

1 Introduction and Procedure

For the simulation of the protein, the .pdb-file of HIV protease (5YOK) from the RCSB PDP protein data bank was used [3]. The non-standard groups included in original file which are not paramterized in the force field were taken out using the PyMol software. A visualization by VMD of the protein with the non-standard groups removed is displayed below.

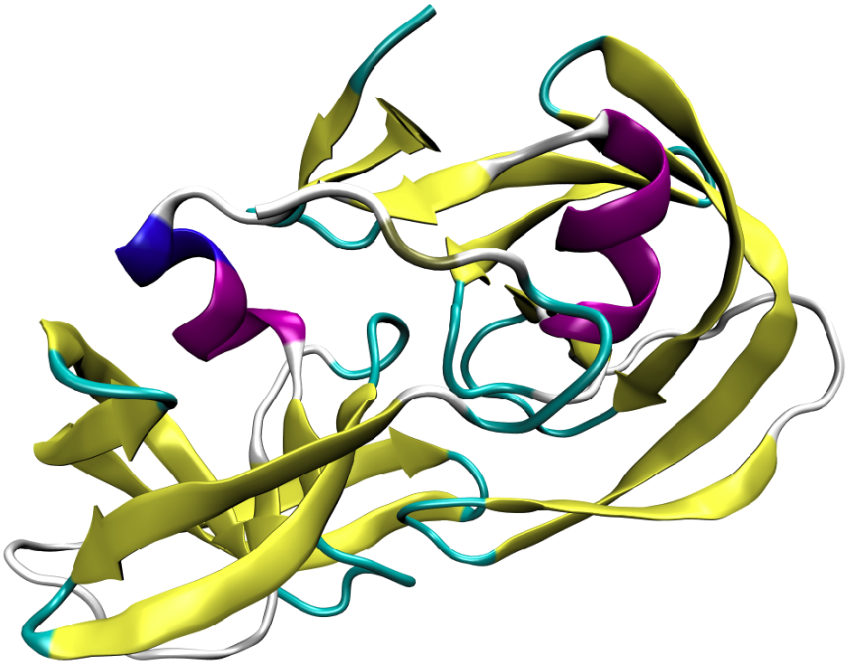

Figure 1: The protein 5YOK after the removal of the non-standard groups.

For the conversion from the .pdb file to a .gro file, the AMBER99SB-ILDN force field from 2010 and the recommended TIP3P 3-site model for water was used. The protein was the placed into a solvated box (TIP3P water model) of dimensions $7.00\,\text{nm} \times 7.00\,\text{nm} \times 7.00\,\text{nm}$ and 9 CL^-

ions were added to the solvent to ensure charge neutrality. This box size was chosen as the length of the protein along its longest axis is roughly 5.50 nm and the cutoff distances for the forces were set to 1.0 nm. In this way, it is ensured that the protein does not interact with its image. In all simulations the Van der Waals forces were modelled using the Lennart-Jones Potential and the electrostatic forces were computed using the fourth order Particle-Mesh-Ewald method. Furthermore, all bonds to hydrogen atoms were constrained to their equilibrium length with the help of the fourth order LINCS algorithm. The short range neighbour search was performed using the Verlet cutoff scheme with a range of 0.8 nm (Probably not ideal choice).

After an energy minimization using the steepest descent algorithm, the following runs were conducted for this report with $T_0 = 300$ K and $P_0 = 1.0$ bar:

Run	Δt [ps]	Steps	Thermostat	Barostat
NVT Equilibration	0.002	10 000	Berendsen	-
NPT Equilibration	0.002	100 000	Berendsen	Berendsen
Production Run	0.002	2 500 000	Nose-Hoover	Parinello-Rahman

Table 1: Time steps, thermostats and barostats used for the simulations conducted.

All runs were conducted using the GROMACS 'md' integrator, an implementation of the leap frog algorithm [4].

2 Simulation Results and Discussion

2.1 Equilibration Runs

As mentioned above three equilibration runs were conducted before the conduction run. The results are shown below:

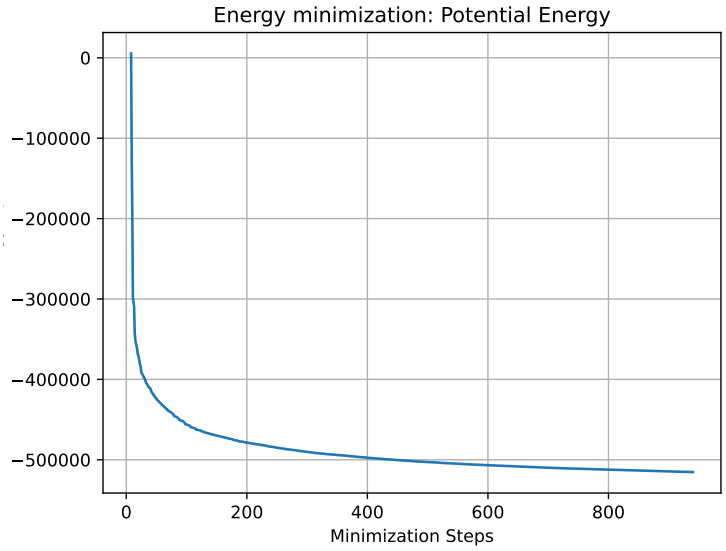

Figure 2: Equilibration of potential energy during the optimization run.

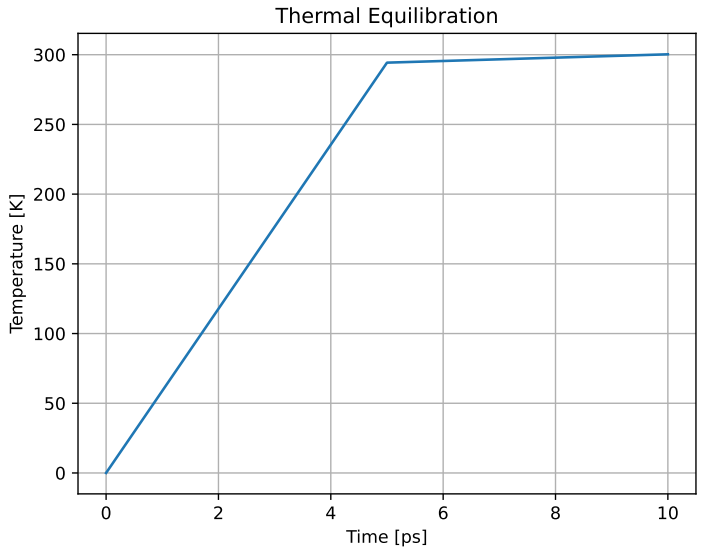

Figure 3: Temperature equilibration during NVT equilibration run.

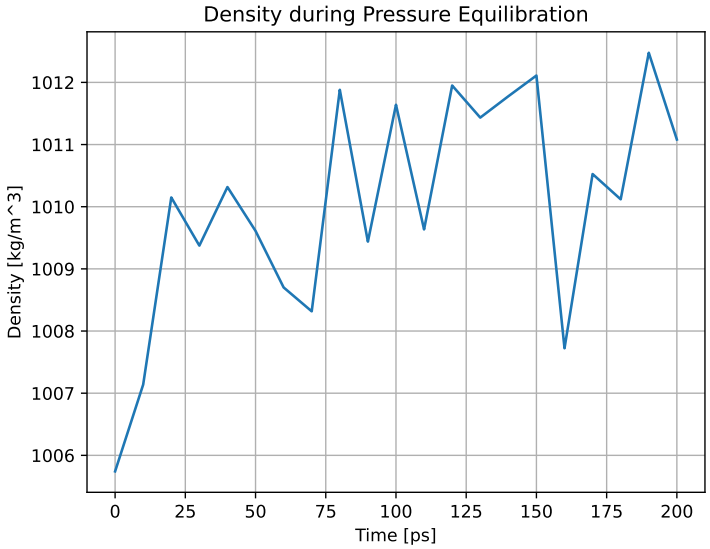

Figure 4: Development of the density during pressure equilibration.

As can be seen the potential energy was minimized during the energy minimization and the temperature reached the desired 300 K after roughly half the simulated 10 ps. The density development during the pressure equilibration is shown in figure 4. It shows fluctuations, which however are smaller then 1% of the average value. Based on this it is concluded that the simulation was equilibrated before the production run.

2.2 RMSD and R_{gyr} calculation

The RMSD and the R_{gyr} of the 5YOG protein were found from the production run and are shown below:

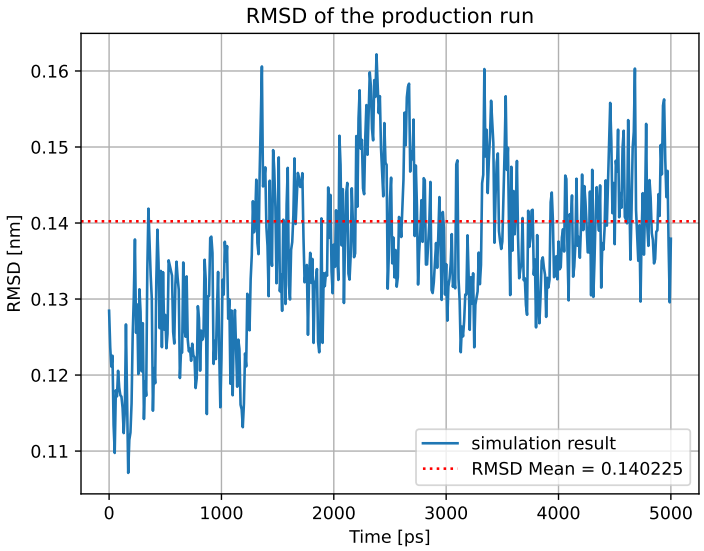

Figure 5: RMSD for the trajectory of the production run.

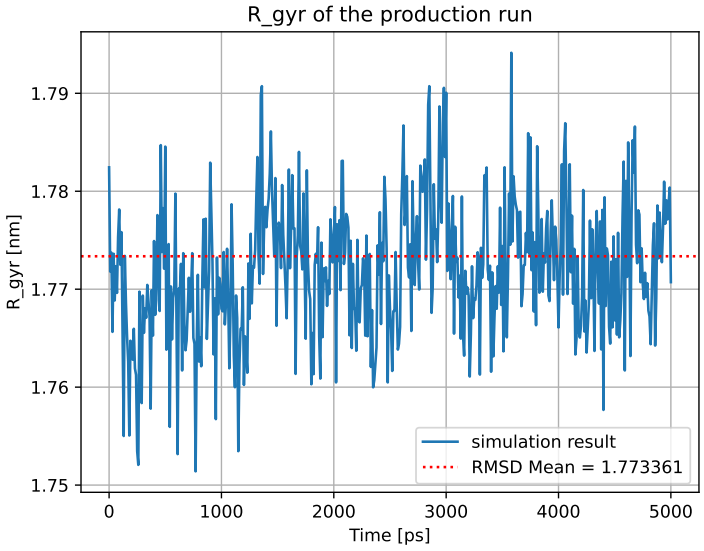

Figure 6: R_{gyr} for the trajectory of the production run.

One spots that both quantities seem to need roughly 1000 ps to reach their final values. The averaged quantities are RMSD = 0.140 nm and R_{gyr} = 1.773 nm. To further investigate the convergence of the two quantities a block standard error analysis was conducted for both. From this analysis one can learn whether a simulation is conserved and further samples will not change the result significantly.

Looking at the figures 7 and 8, one finds that the sampling was in both cases not yet fully sufficient. If enough samples have been sampled for a given observable and n exceeds the correlation time, the curve should become a horizontal line. Clearly, the RMSD result is far from this. The R_{gyr} result seems to to be better converged, but also here the convergence could be better. Hence, more than the used 2.500.000 steps are necessary for a 'full' convergence.

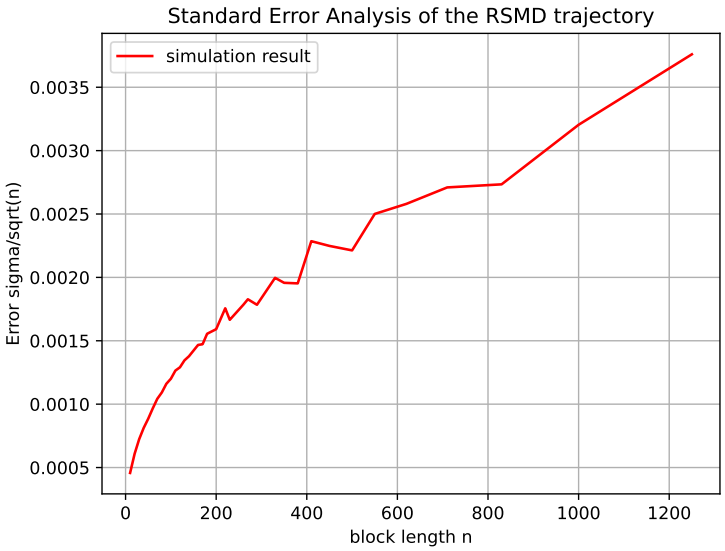

Figure 7: Standard Error Analysis of the RMSD trajectory. Notice the linear behaviour for large n.

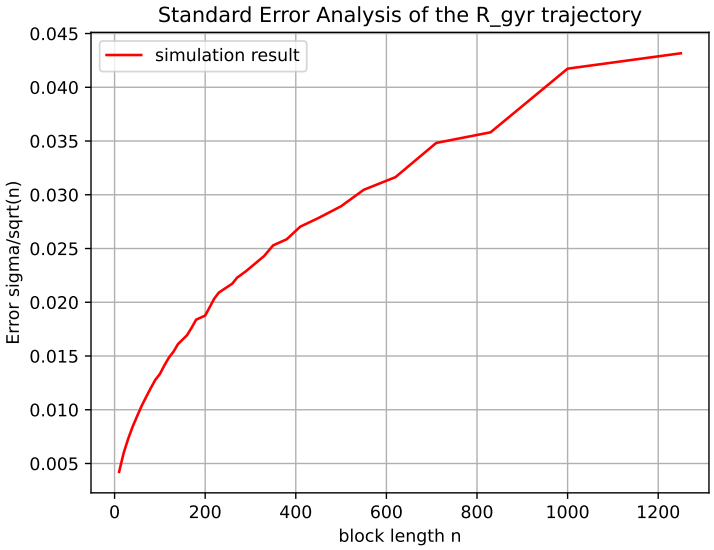

Figure 8: Standard Error Analysis of the RMSD trajectory. Here the curve shows a flattening for large n.

2.3 PCA Analysis

Finally, a principal component analysis is used to discover the most important motions of the molecule.

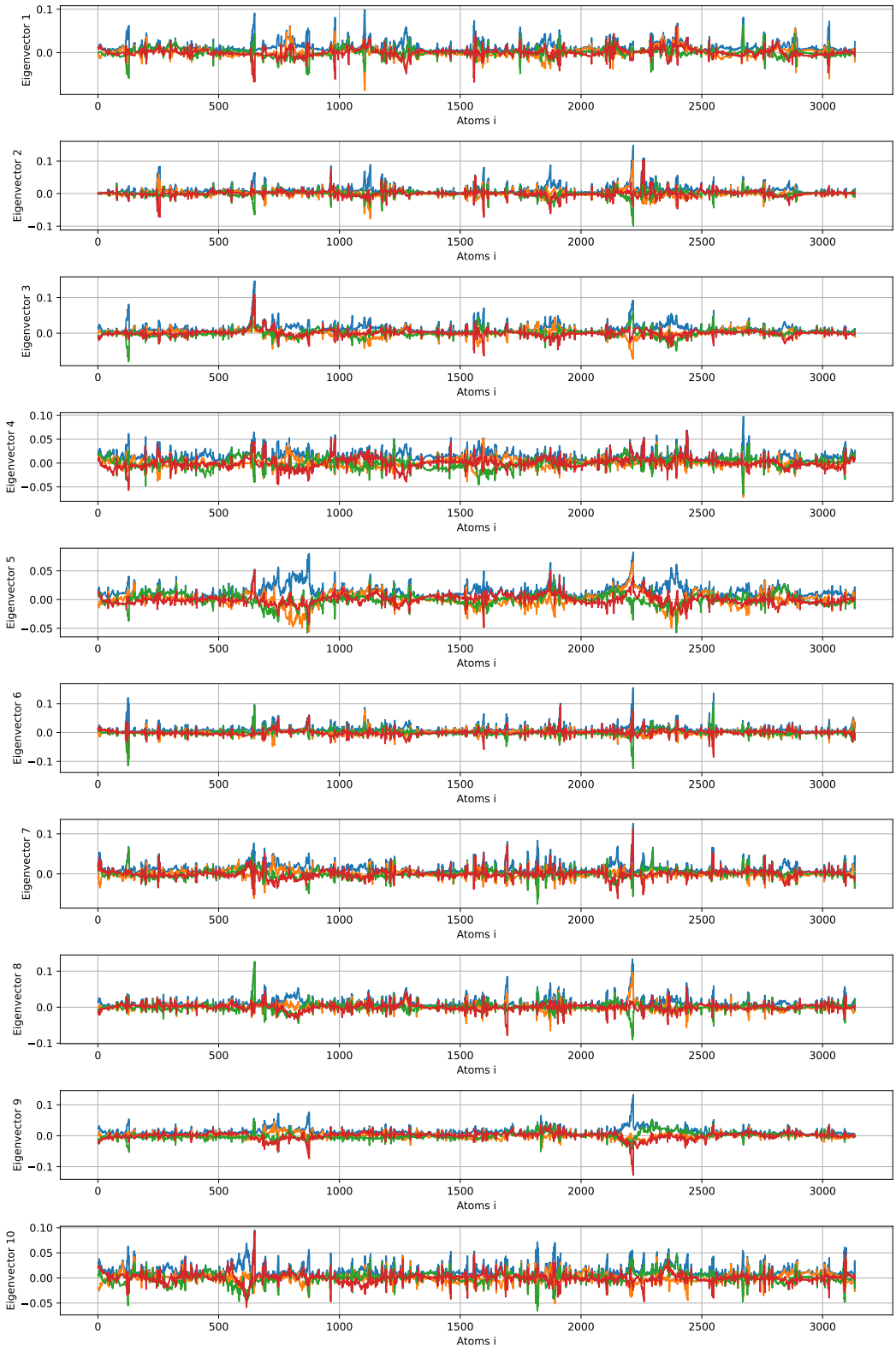

Figure 9: Components of the ten principal components. The x axis shows the atoms (1-3134). The blue line shows the total contribution of one atom to the principal component while the orange, green and red curves show the contributions of the X-,Y, and Z-components of the respective atoms.

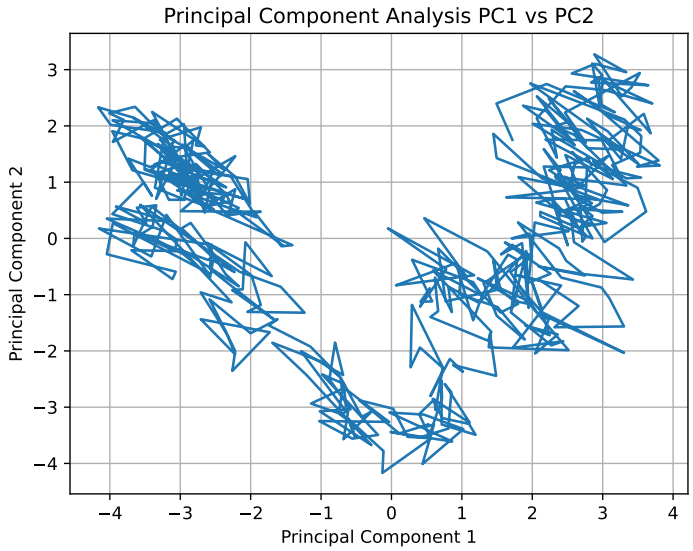

Figure 10: Total value of principal component value 1 plotted against total value of principal component 2 over the production trajectory.

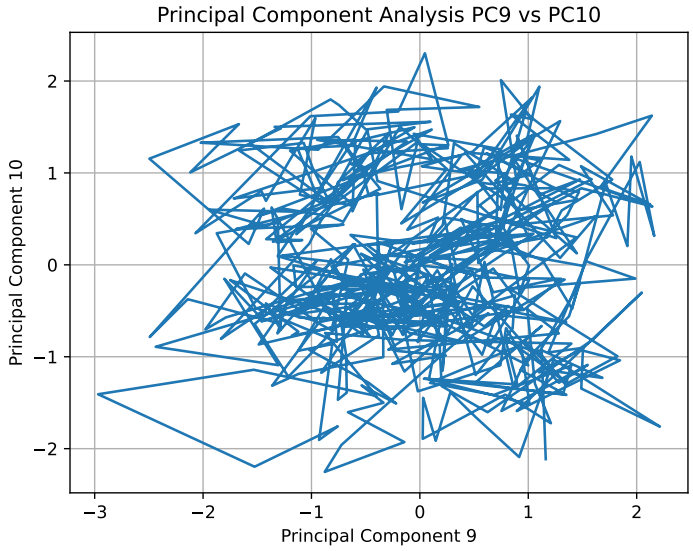

Figure 11: Total value of principal component value 9 plotted against total value of principal component 10 over the production trajectory.

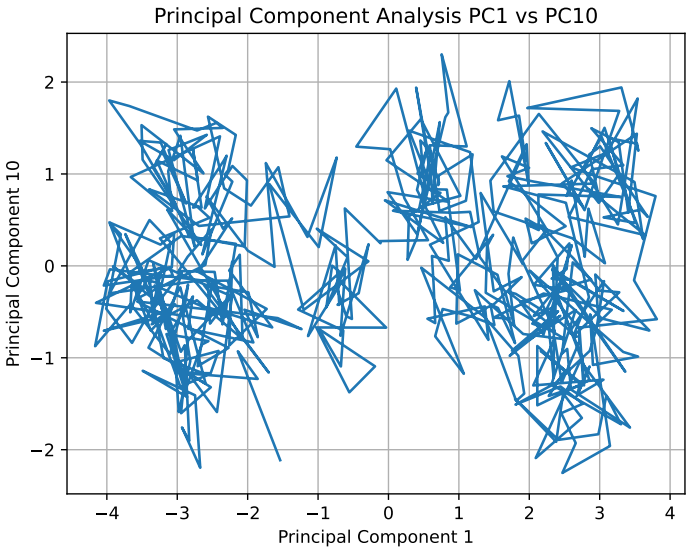

Figure 12: Total value of principal component value 1 plotted against total value of principal component 10 over the production trajectory.

Firstly, the contributions of every atom to the ten largest principal values are shown (figure 9). As expected from a PCA analysis there is no distinct pattern visible. Most atoms only contribute a small amount to the eigenvector with a few being more prominent. This shows that the most important motions observed in the simulations were collective motions of the protein rather than isolated motions. Also, there is no distinctive difference in the components of the ten principal values. Again, this was to be expected as they encode fairly complicated collective motions.

Furthermore, the amplitudes of the principal values with respect to each other were analyzed (figure 10,11 and 12). The first and also expected observation is that the amplitudes of the first and second principal value are larger than the amplitudes of the ninth and tenth value. This is essentially by construction and means that the first principal values encode the most significant motions of the molecule. While in the plots of PC1 vs PC10 and PC9 vs PC10, the space seems to have been extensively sampled, it seems from the plot of PC1 vs PC2 that only a certain trajectory (V shape) was sampled. This could indicate a lack of sampling or mean that only certain combinations of the first two principal modes are possible.

Lastly, the extreme amplitude configurations of the first principal mode were visualized using VMD software (see pictures below). This motion, being the most prominent one in the simulation, is fairly small. It is a small torsion of the upper part of the molecule against the lower part of the molecule. The most significant part is a twisting of the opposing beta sheet loops in the foreground of the images.

The simulation was successfully prepared and conducted. The desired RSMD and R_{gyr} were calculated. BES and PCA analysis however indicate that the number of steps and hence the sampling of the configurations space in the production run were insufficient. Due to this further runs with more steps should be performed.

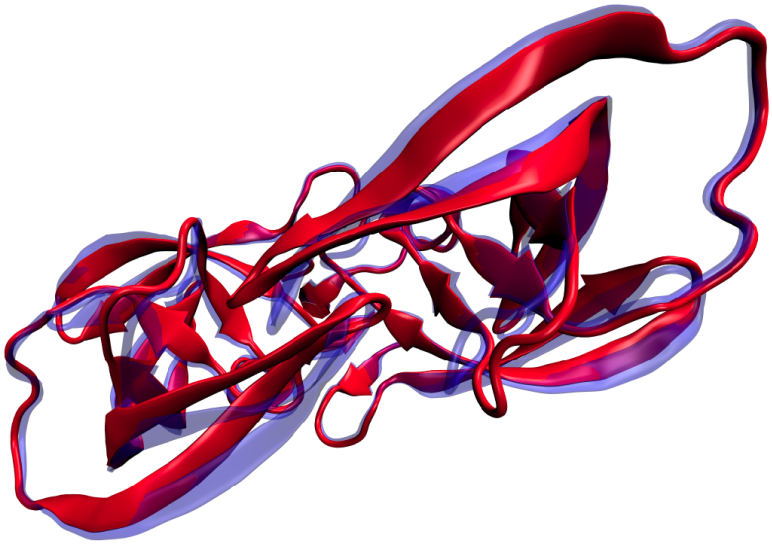

Figure 13: Maximal amplitude of the largest principal value.

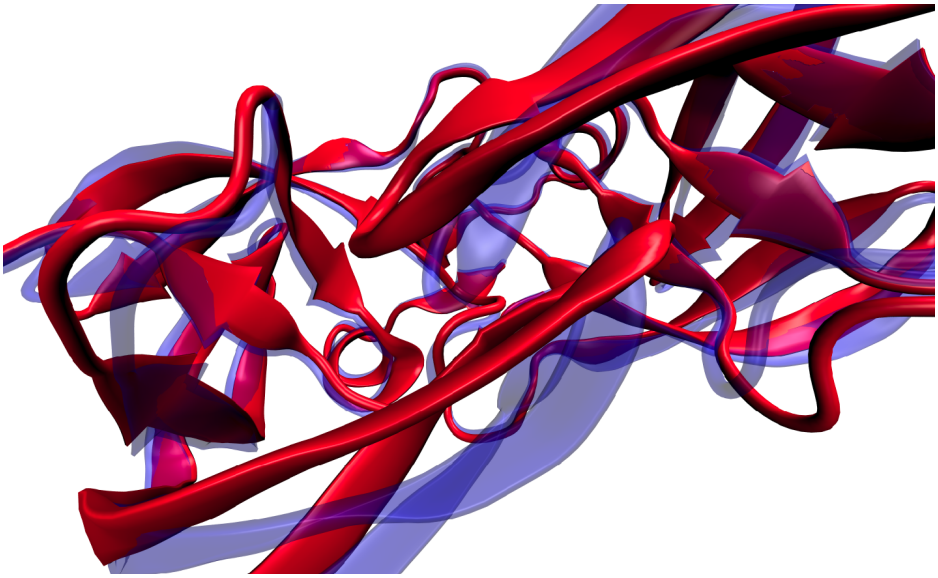

Figure 14: Enlarged version of the figure above showing the twisting of the beta loops in more detail.

3 Bibliography

[1] Zacharias, Martin et Al. Molecular Dynamics: From Basics to Applications Lecture Notes; 2023
[2] Vollmers, Luis; Zacharias, Martin; Reif, Maria; Molecular Dynamics: Exercise 8 - Self-Supervised Protein Simulation with Convergence and Principal Component Analysis; 2023

[3] RCSB PDB Protein Data Bank, Structure of HIV protease 5YOK, `https://www.rcsb.org/structure/5YOK` [accessed 25.06.2023, 14:44]

[4] GROMACS online manual: .mdp options `https://manual.gromacs.org/current/user-guide/mdp-options.html` [accessed 25.06.2023, 15:45]