Exercise_8 Molecular Dynamics Simulation

Andrea Scaioli

June 2025

1 Introduction

In this exercise, we studied the structural dynamics of the HIV protease protein through a combination of statistical and dimensionality reduction techniques. Our analysis focused on two complementary approaches: Blocked Standard Error (BSE) analysis and Principal Component Analysis (PCA).

The BSE method was employed to assess the statistical reliability and convergence of observables extracted from the molecular dynamics (MD) trajectory. By examining how the standard error evolves as a function of block size, we were able to identify the degree of correlation between sampled configurations and estimate the error in time-averaged quantities more accurately.

To explore the dominant collective motions of the protein, we applied PCA to the atomic positional fluctuations. This technique transforms the high-dimensional trajectory data into a new basis set of orthogonal principal components, ordered by the variance they explain. The first few components often capture large-scale, biologically relevant motions of the protein, while the remaining components represent smaller, less significant fluctuations. This allowed us to gain insights into the essential dynamical behavior of HIV protease and to evaluate the sampling quality of the simulation.

Through these analyses, we obtained a comprehensive understanding of the statistical quality and dominant motions present in the MD simulation of HIV protease, shedding light on its dynamic features and potential functional states.

2 Methods

The molecular dynamics (MD) simulations of the HIV protease protein were performed using GROMACS. The structure was obtained from the Protein Data Bank (PDB ID: 5YOK), and the system was prepared through the following steps.

System Preparation

The topology was generated using the `amber99sb-ildn` force field and the TIP3P water model:

```
gmx pdb2gmx -f 5YOK.pdb -o processed.gro -water tip3p -ignh -ff amber99sb-ildn
```

The protein was placed in a dodecahedral box with a minimum distance of 1.0 nm from the box edge, and the box was solvated with SPC216 water molecules. Ions were added to neutralize the system and reproduce physiological conditions (0.15 M NaCl), replacing solvent molecules:

```
gmx editconf -f processed.gro -o newbox.gro -c -d 1.0 -bt dodecahedron
gmx solvate -cp newbox.gro -cs spc216.gro -o solvated.gro -p topol.top
gmx grompp -f ions.mdp -c solvated.gro -p topol.top -o ions.tpr -maxwarn 1
gmx genion -s ions.tpr -o solv_ions.gro -p topol.top -pname NA -nname CL -neutral -conc 0.15
```

Energy Minimization and Equilibration

The system was minimized using the steepest descent algorithm for up to 50,000 steps with a cutoff scheme of Verlet and PME electrostatics:

```
gmx grompp -f minim.mdp -c solv_ions.gro -p topol.top -o em.tpr
gmx mdrun -v -deffnm em
```

Subsequently, the system underwent NVT and NPT equilibration phases. The NVT equilibration was performed at 300 K using the V-rescale thermostat with positional restraints on the protein:

```
gmx grompp -f nvt.mdp -c em.gro -r em.gro -p topol.top -o nvt.tpr
gmx mdrun -v -deffnm nvt
```

The NPT equilibration used the Parrinello-Rahman barostat to maintain a pressure of 1 bar:

```
gmx grompp -f npt.mdp -c nvt.gro -r nvt.gro -p topol.top -o npt.tpr
gmx mdrun -v -deffnm npt
```

Preproduction and Production Runs

A preproduction MD run of 1 ns (500,000 steps) was executed with no position restraints, using the Nose-Hoover thermostat and Parrinello-Rahman barostat:

```
gmx grompp -f preprod.mdp -c npt.gro -r npt.gro -p topol.top -o md_pre.tpr
gmx mdrun -v -deffnm md_pre
```

The production run was extended to 10 ns using identical parameters:

```
gmx grompp -f prod.mdp -c md_pre.gro -r md_pre.gro -p topol.top -o md.tpr
gmx mdrun -v -deffnm md
```

Trajectory Analysis

The resulting trajectory was post-processed by centering the protein and removing periodic boundary artifacts:

```
gmx trjconv -s md.tpr -f md.xtc -o prod_center.xtc -center -pbc mol
```

Standard observables such as the radius of gyration and RMSD were computed:

```
gmx gyrate -f prod_center.xtc -s md.tpr -o rgyr.xvg
gmx rms -s md.tpr -f prod_center.xtc -o rmsd.xvg -tu ns
```

Principal Component Analysis

Principal Component Analysis (PCA) was performed to extract the main modes of motion. First, the covariance matrix of the positional fluctuations was constructed and diagonalized:

```
gmx covar -f prod_center.xtc -s md.tpr -o eigenval.xvg -v eigenvec.trr -xpma covar_matrix.xpm
```

Then, the trajectory was projected along the first 10 eigenvectors:

```
gmx_mpi anaeig -v eigenvec.trr -f prod_center.xtc -s md.tpr -proj eigproj1_10.xvg -comp eigcomp1_10.xvg
```

This enabled us to visualize the conformational sampling in reduced-dimensional space and assess convergence and sampling quality based on the dominant principal components.

System Preparation

The molecular structure of the HIV protease (PDB ID: `5YOK`) was used as the initial model. The topology was generated with the `amber99sb-ildn` force field and the `TIP3P` water model using the `gmx pdb2gmx` tool. A dodecahedral simulation box was defined with a minimum distance of 1.0 nm between the protein and the box edge. The system was solvated using `gmx solvate` with the `spc216` water model.

To neutralize the system and reach a physiological salt concentration (0.15 M), sodium and chloride ions were added using `gmx genion`, replacing water molecules. This step followed the preprocessing step with `gmx grompp` using the `ions.mdp` file, where the integrator was set to `steep` and the electrostatics treated with the Particle Mesh Ewald (PME) method with a real-space cutoff of 1.0 nm for both Coulomb and van der Waals interactions.

Energy Minimization and Equilibration

Energy minimization was carried out with the steepest descent algorithm until convergence or a maximum of 50,000 steps. The minimization used the `minim.mdp` settings, maintaining the PME method and Verlet cutoff scheme.

Following minimization, two equilibration phases were conducted. The NVT equilibration was run at 300 K using the V-rescale thermostat with a coupling time constant $\tau_t = 0.1$ ps and position restraints applied on the protein heavy atoms. This was followed by NPT equilibration using the Parrinello-Rahman barostat at 1 atm, with $\tau_p = 1.0$ ps and the same temperature coupling as in the NVT phase. The timestep in the NPT phase was set to 2 fs, and the simulation was run for 50,000 steps.

Production Simulation

A short preproduction run of 1 ns was first carried out using the `preprod.mdp` file to allow relaxation without restraints. This simulation employed the Nose-Hoover thermostat (with $\tau_t = 1.0$ ps) and Parrinello-Rahman barostat (with $\tau_p = 2.0$ ps), applying isotropic pressure coupling. The constraint algorithm was applied to all hydrogen bonds, and dispersion corrections were enabled.

Subsequently, a production run was launched under the same settings using the `prod.mdp` file. The trajectory was centered and made whole using `gmx trjconv` with the `-center -pbc mol` options.

Trajectory Analysis

Post-simulation analysis included the computation of the radius of gyration (Rg) using `gmx gyrate` and the root mean square deviation (RMSD) with `gmx rms`, both applied to the protein backbone over the production trajectory.

To investigate the essential dynamics of the protein, Principal Component Analysis (PCA) was performed. The covariance matrix of atomic positional fluctuations was computed using `gmx covar`, and the eigenvectors and eigenvalues were extracted. Projection of the trajectory onto the first ten principal components was carried out using `gmx anaeig`, generating the files used for further analysis and visualization of conformational sampling.

Results

Convergence and Blocked Standard Error Analysis

The root-mean-square deviation (RMSD) and radius of gyration (R_g) were monitored throughout the trajectory to assess structural stability. As shown in Figure 1, both metrics plateau after an initial equilibration phase, indicating a reasonably stable simulation.

To quantify the statistical quality of the sampling, a Blocked Standard Error (BSE) analysis was performed (Figure 2). Although the BSE does not converge to a perfectly flat line—an ideal only reached in infinite sampling—it flattens sufficiently to suggest that the observable of interest is reasonably converged. However, further sampling would be beneficial to reduce statistical uncertainty and improve the reliability of ensemble averages.

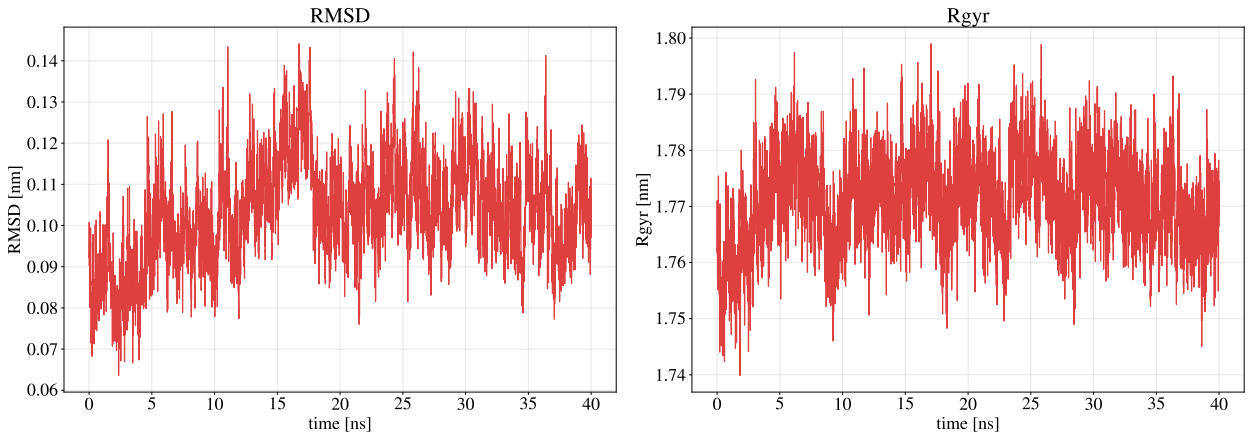

Figure 1: RMSD (left) and radius of gyration (right) as a function of time. Both indicate structural stability after initial equilibration.

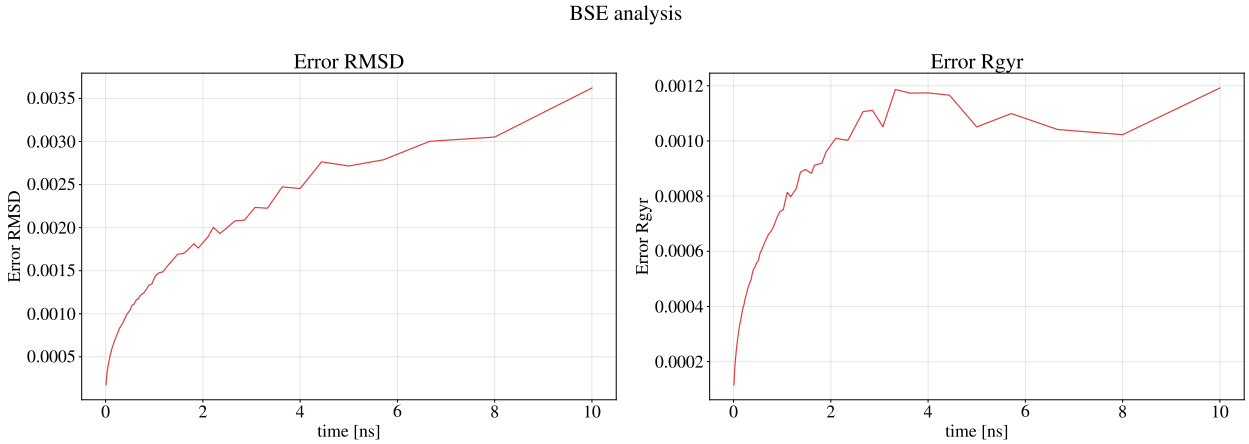

Figure 2: Blocked Standard Error (BSE) plot for the chosen observable. A flattening trend indicates acceptable convergence.

Principal Component Analysis

Principal Component Analysis (PCA) was performed to investigate the collective motions sampled during the MD simulation. Projections onto the first ten eigenvectors reveal that motion along PC1 and PC2 is well sampled and clustered (Figure 3), indicating a dominant low-dimensional conformational landscape. In contrast, projections along PC1 vs. PC10 and PC9 vs. PC10 (Figure 4) appear scattered, highlighting the decreasing importance and sampling quality of higher-order principal components.

These plots support the use of PC1 vs. PC2 as a meaningful representation of the conformational ensemble, as they capture the majority of collective motion while remaining sufficiently converged for qualitative analysis.

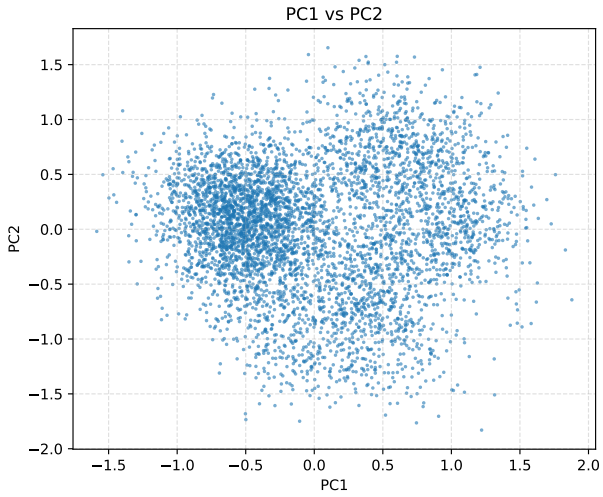

Figure 3: Projection of the trajectory onto PC1 and PC2. Clear clustering suggests good sampling along the dominant collective modes.

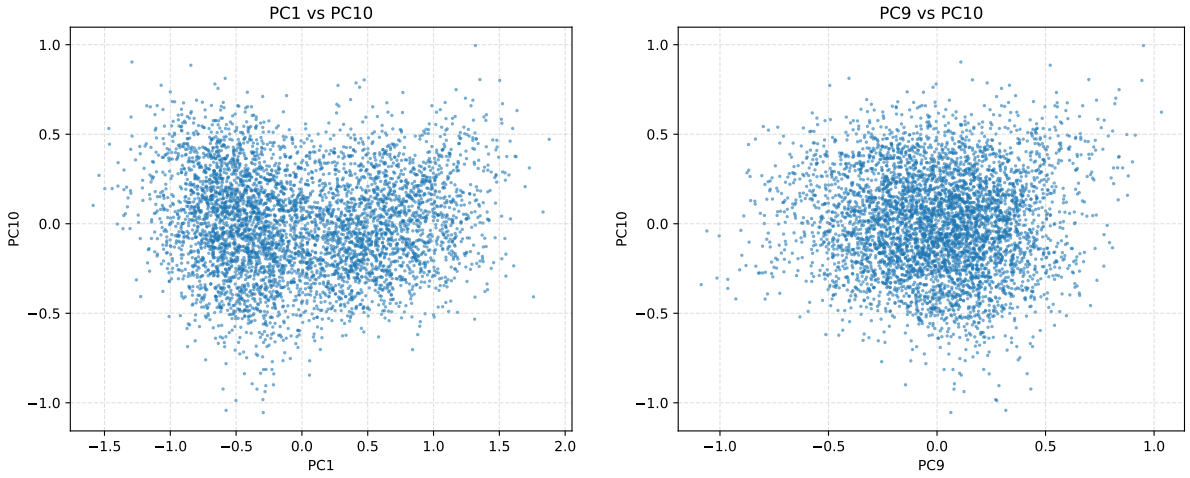

Figure 4: Projections onto PC1 vs. PC10 (left) and PC9 vs. PC10 (right) show scattered distributions, indicating less relevant, poorly sampled motions.

Comparison of Average and Extreme Structures and Analysis of Eigenvector Contributions

To better understand the structural variability observed in the simulation, we compared the average structure (‘average.pdb‘) with an extreme conformation (‘extreme.pdb‘) extracted from the trajectory. Figure 5 shows the overlay of these two structures, highlighting the regions of maximum deviation.

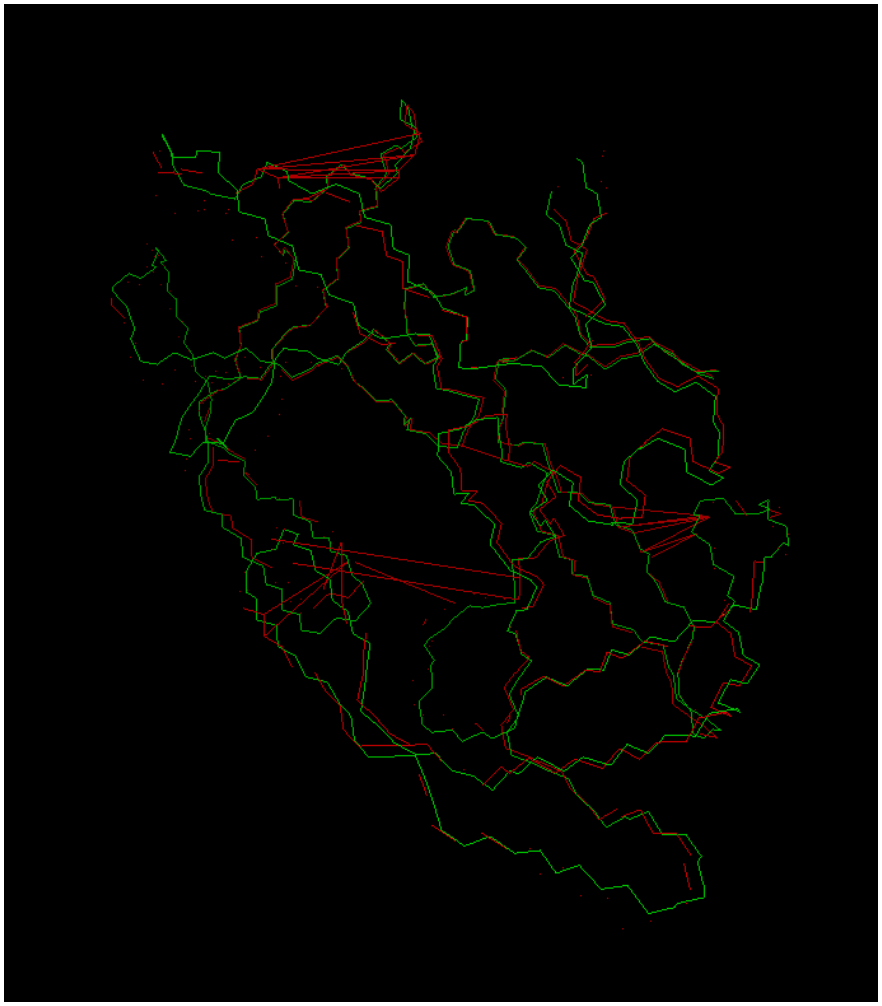

Figure 5: Overlay of the average structure (green) and an extreme conformation (red) of the HIV protease protein. Regions with significant deviations indicate flexible or highly dynamic parts of the protein.

Furthermore, to identify which atoms contribute most to the observed motions, we plotted the components of the principal eigenvectors as a function of atom index (Figure 6). The peaks in these plots correspond to atoms that have a greater influence on the respective eigenvector mode, often indicating key residues or flexible loops involved in the protein's dynamics.

This analysis allows us to rationalize structural changes and focus on functionally relevant regions that are dynamically significant.

End-State Structural Analysis

The final structure of the HIV protease protein was visualized to examine the binding site and overall conformation. Figure 8 shows the end-state snapshot, which maintains the characteristic dimer structure and preserves the geometry of the active site. This structural consistency reinforces the interpretation of

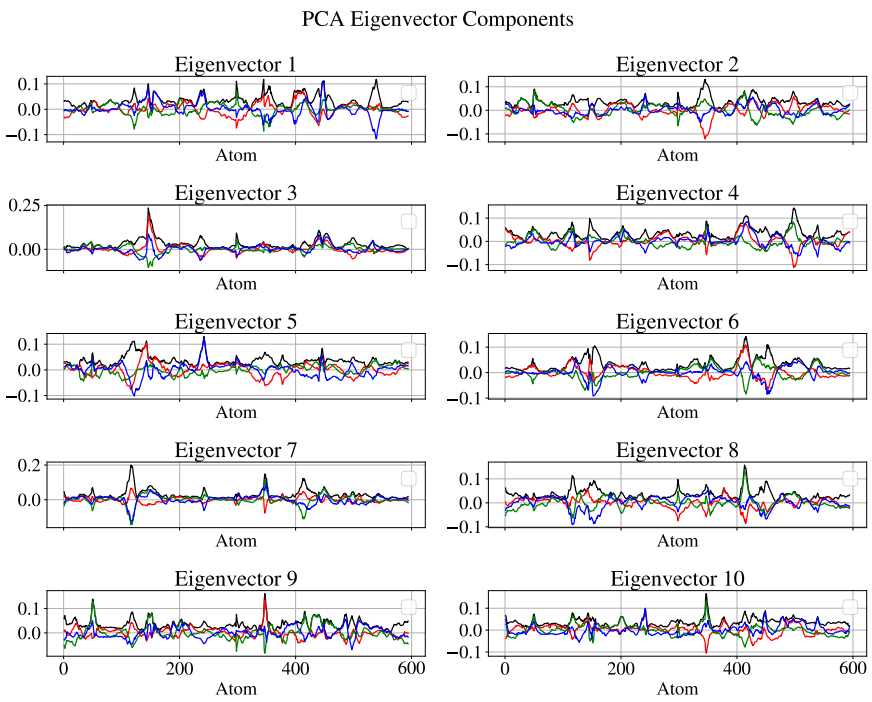

Figure 6: Projection of selected eigenvectors along atom indices. Peaks indicate atoms with major contributions to the principal modes of motion.

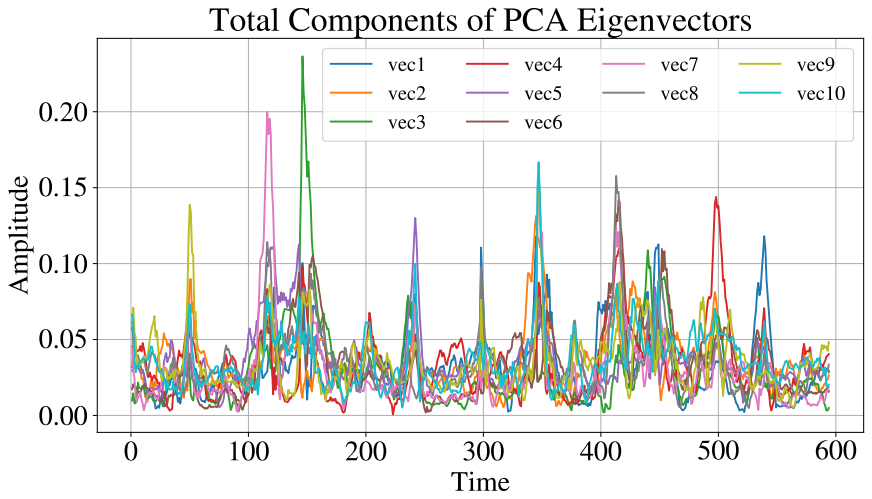

Figure 7: Total components overlapping of the 10 eigenspaces

convergence from RMSD, R_g, and PCA.

When choosing a representative view of the structure, care was taken to highlight relevant functional features such as the flap regions and the binding pocket. This visual context supports a qualitative assessment of structural stability and sampling sufficiency over the course of the simulation.

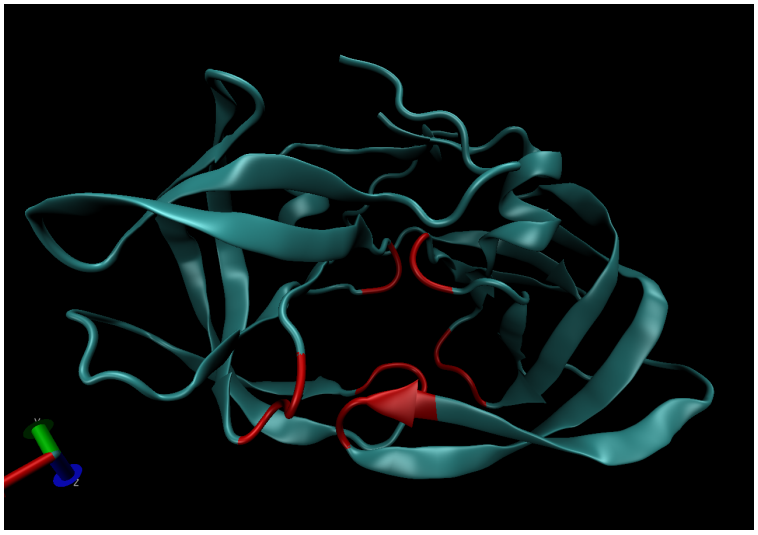

Figure 8: Final frame of the simulation, showing the stable conformation of the HIV protease with intact binding site in red.