Molecular Docking and Molecular Dynamics Simulations of Ligands Against HIV-1 Protease

Adam Barla
BS7107
NTU, Singapore
n2308836j@e.ntu.edu.sg

I. Introduction

Molecular docking is a computational technique used to predict the binding mode of a ligand to a protein target. It is a crucial step in drug discovery, as it can help identify potential drug candidates. In this report, I describe the computational methods used to perform molecular docking and molecular dynamics simulations on a dataset of ligands against a protein target. The ligands were docked against the HIV-1 protease variant G48T/L89M protein seen in Figure 2.

HIV-1 protease is an enzyme that plays a crucial role in the replication of the human immunodeficiency virus (HIV). It cleaves the newly synthesized polyproteins into mature protein components of an HIV virion – the infectious form of a virus outside the host cell.

The docking was performed using `QuickVina` and the ligand with the best fit was selected for molecular dynamics simulation done using `GROMACS`. The simulation was carried out a second time, this time with Saquinavir (seen in Figure 1), a known inhibitor of the HIV-1 protease.

Both ligands were compared in terms of their Root-Mean-Square Deviation (RMSD) and other properties to determine the quality of the fit and the potential of the new ligand as a drug candidate.

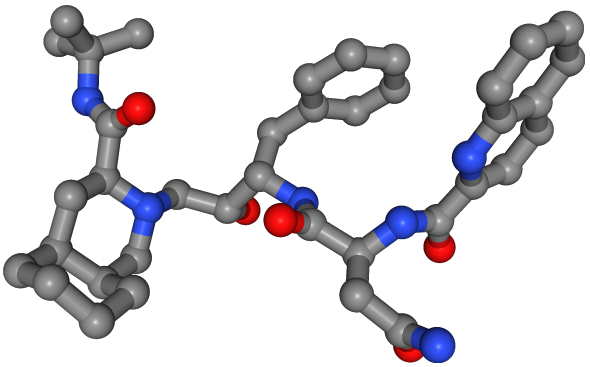

Figure 1: Saquinavir

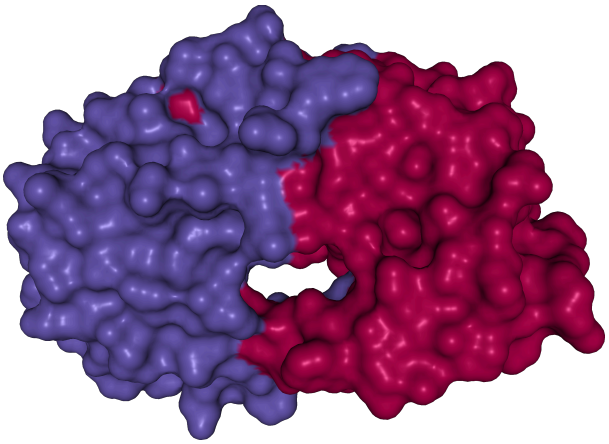

Figure 2: HIV-1 protease variant G48T/L89M

II. Methods

Before performing MD I selected the ligand with the best fit from the docking results. I used a dataset of 2116 ligands to dock against the HIV-1 protease. All files and scripts used in this project can be found on `github` at `adambarla/HIV-protease-binding`.

A. Molecular Docking

I downloaded the X-ray crystal structure of HIV-1 protease variant G48T/L89M in complex with Saquinavir from `rcsb` in a single PDB file. Before separating the ligand from the protein, I accessed the middle atoms of the Saquinavir which was positioned inside the Protease from the start. I looked at the distances to surrounding protein which can be seen in Figure 8 in the appendix. Based on these distances, I selected an atom that was the nearest to the surrounding protein. The coordinates of this atom were used as the center of the bounding box for the docking simulation. Parameters for the docking were set as follows:

```
center_x = 21.877        size_x = 20
center_y = -0.510        size_y = 20
center_z = 11.108        size_z = 20
```

I then split the molecule into two files `saquinavir.pdb` and `hiv1.pdb` and prepared them using `prepare_receptor` and `prepare_ligand` from the ADFR Suite.

B. *Selecting a ligand with the best fit*

I used a `python` code, which can be found in the `notebook.ipynb`, to parse the `qvina` output files and extract the most negative binding affinity score for each ligand out of 2116 tested. This involved identifying lines containing the `"REMARK VINA RESULT"` and capturing the first numerical value (the binding affinity score).

The top 10 best ligands selected can be seen in the Table 1. The two top ones (`fda_553` and `fda_554`) are just variations of Saquinavir with some additional hydrogen atoms. I decided to select the ligand with the third lowest binding affinity score named `fda_1700` and it can be seen in Figure 3. I was able to find the ligand in the ZINC database by converting it to SMILES format with the following command:

```
obabel -ipdbqt fda_pdbqt/fda_1700.pdbqt \
-osmi -O fda_1700.smi
```

The ligand `fda_1700` corresponds to the one with the zinc id ZINC001560410173 in the database.

The ligand's binding affinity score was −11.1 kcal/mol which is slightly less than Saquinavir's −11.4 kcal/mol.

C. *Molecular dynamics*

In this section, I outline the sequence of computational steps taken to prepare a molecular system for dynamic simulation using `GROMACS`.

Initially, I converted the ligand's structural data from `PDBQT` to `PDB` format with *Open Babel*, separating molecules and adding hydrogen atoms. Next, `Antechamber` was used to generate a `MOL2` file, calculating `AM1-BCC` charges and setting the net charge and multiplicity. The `parmchk2` command then created a force field modification file for missing parameters. With `tleap`, I prepared the system using the `ff14SB` force field, incorporating ligand parameters. The final step involved the `parmed_amber2gmx.py` script (see `scripts` folder) to convert `AMBER` files to `GROMACS` format. Following the successful setup of the ligand, I applied the `tleap` and `parmed_amber2gmx.py` steps to prepare the protein component.

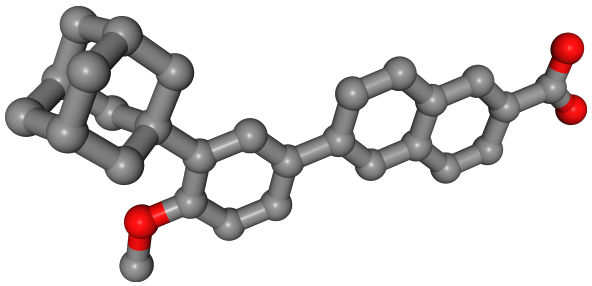

Figure 3: Ligand name `fda_1700` in the dataset. Ligand's ZINC ID is ZINC001560410173. It has reached the third lowest binding affinity score of −11.1 kcal/mol out of 2116 ligands tested.

Ligand	Model									min
	1	2	3	4	5	6	7	8	9	
fda_553	-11.4	-10.4	-10.4	-9.7	-9.6	-9.5	-9.3	-9.3	-8.9	**-11.4**
fda_554	-11.4	-10.8	-10.4	-10.1	-10.0	-9.9	-9.9	-9.5	-9.0	**-11.4**
fda_1700	-11.1	-10.2	-10.1	-9.9	-9.9	-9.8	-9.7	-9.6	-9.6	**-11.1**
fda_1755	-11.0	-10.8	-10.7	-10.4	-10.1	-10.0	-10.0	-10.0	-9.9	**-11.0**
fda_871	-11.0	-11.0	-10.8	-10.7	-10.6	-10.1	-9.9	-9.8	-9.8	**-11.0**
fda_872	-11.0	-11.0	-10.8	-10.6	-10.5	-10.5	-10.1	-9.6	-9.5	**-11.0**
fda_1829	-10.9	-10.9	-10.6	-10.6	-10.2	-10.2	-10.1	-10.1	-10.0	**-10.9**
fda_95	-10.8	-10.4	-10.4	-9.8	-9.7	-9.6	-9.3	-9.3	-8.7	**-10.8**
fda_161	-10.8	-9.7	-9.5	-9.5	-9.5	-9.2	-9.2	-9.1	-8.9	**-10.8**
fda_160	-10.7	-9.9	-9.7	-9.6	-9.6	-9.5	-9.4	-9.2	-8.8	**-10.7**

Table 1: Top 10 ligands with the lowest minimal (out of all models tested) binding affinity score

1) *Solvation of the system:* I then combined the ligand and protein systems into a single `GROMACS` input file `sys.gro` while making sure that the information about atom types and molecules was correct.

I created a box around the system using the `editconf` command with a buffer of 1.0 nm around the centered protein-ligand complex.

I solvated the system with water molecules around the molecule in the bounding box creating a `wat.gro` file. I used `configs/spc903.gro` as the solvent configuration. The topology file `sys.top` was updated with the number of solvent molecules added.

Next preprocessing step was to compile the system topology from `sys.top`, the solvated configuration from `wat.gro` with `gmx grompp` using minimization parameters from `configs/mini.mdp`, to generate a portable binary run file `bions.tpr`.

Then I neutralize the system with `gmx genion` command which replaces solvent molecules with ions to reach a desired ionic concentration of 0.15 mol/l. The topology `sys.top` was again updated with the ion information, and the output configuration is saved as `ions.gro`.

I also added water molecules to the system together with Na and Cl ions to neutralize the system.

2) *Energy minimization:*

After solvation, I combined the `ions.gro` and `sys.top` files to create a portable binary run file `mini.tpr` with `gmx grompp` using the `mini.mdp` parameters. I performed energy minimization of the system using the `gmx mdrun` command with the `mini.tpr` created by the previous step.

3) *Running the simulation:*

I created an index file `index.ndx` with `gmx make_ndx`.

Using `configs/pr.mdp` parameters, I created a portable binary run file `pr.tpr` with `gmx grompp` to run the position restrained simulation which was done with `gmx mdrun`.

Using `configs/md.mdp` parameters, I created a portable binary run file `md001.tpr` with `gmx grompp` to run the production simulation which was done with `gmx mdrun`. Finally, I ran the production simulation with `gmx mdrun` using the `configs/md.mdp` parameters.

The simulation was run for 5000000 steps totaling 10 ns and it was repeated with Saquinavir as the ligand.

III. Results

I analyzed the results of the molecular dynamics simulations to compare the ligands in terms of their stability and conformational changes over time.

A. Root-Mean-Square Deviation (RMSD)

RMSD is used to measure the average distance between atoms of superimposed structures. It is a commonly used metric to assess the conformational stability and structural changes of a macromolecule (protein) over time during a simulation.

Lower RMSD values indicate higher structural stability. During the initial phase of the simulation, the RMSD will typically increase rapidly as the system departs from the starting structure. Once equilibrated, the RMSD will plateau, indicating that the system is sampling around an average structure.

I calculated the RMSD (root mean squared deviation) to check how far the peptide structure changes from the initial one.

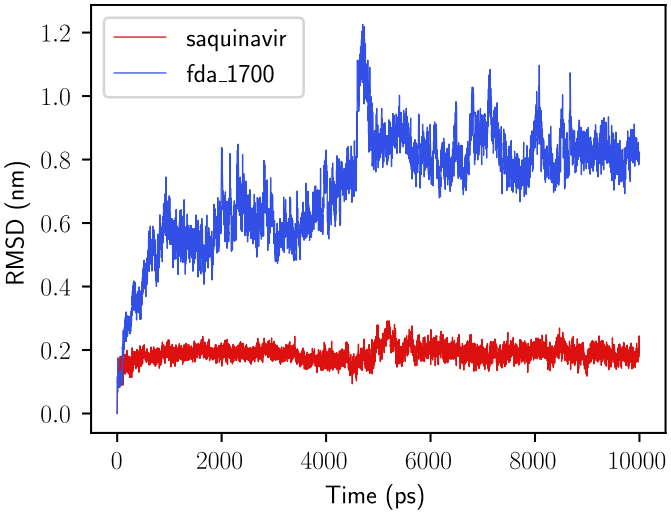

Figure 4: Root-Mean-Square Deviation (RMSD) of a ligand after performing a least squares fit to a protein over time in a molecular dynamics simulation. The figure shows a comparison of the RMSD of Saquinavir and the ligand `fda_1700`.

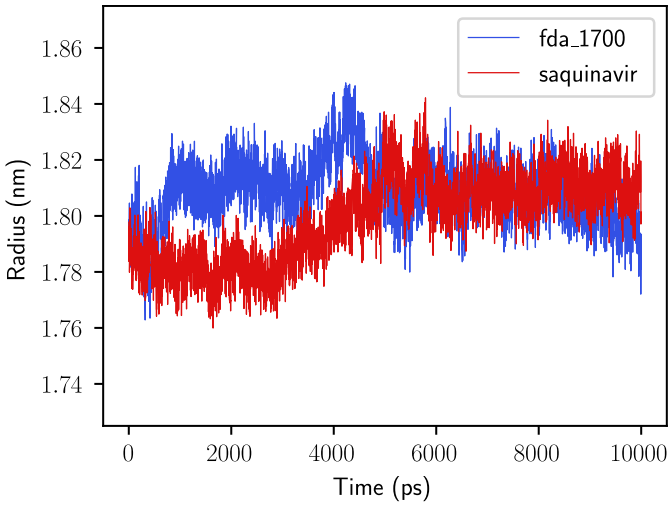

Figure 5: The radius of gyration (R_g) of the protein when simulated with the ligand `fda_1700` and Saquinavir over time.

The comparison of RMSD for the ligand `fda_1700` and Saquinavir over time can be seen in Figure 4.

B. Radius of Gyration (R_g)

The radius of gyration is a measure that describes the distribution of components (such as atoms) around the center of mass of a molecule. It gives us an idea of the molecule's "compactness" and can inform us about its three-dimensional structure. R_g is calculated using the positions of all the atoms in the molecule, weighted by their masses. Mathematically, it is defined as the root-mean-square distance of the system's parts from its center of mass.

The comparison of R_g of the protein when simulated with the ligand `fda_1700` and Saquinavir over time can be seen in Figure 5.

C. Root-Mean-Square Fluctuation (RMSF)

Root Mean Square Fluctuation (RMSF) measures the deviation of positions of a selection of atoms (usually the backbone atoms) over time from a reference position (often the time-averaged position). It's used to understand the flexibility and dynamics of a molecule or molecular complex within a simulation. Calculating RMSF per residue gives insight into which parts of the protein are more flexible or rigid during the simulation. High RMSF values indicate regions with high flexibility, while low RMSF values indicate more rigid, stable regions. This can help identify flexible loops, stable cores, or regions that undergo conformational changes in response to ligand binding or other factors.

The comparison of RMSF for the protein when simulated with the ligand `fda_1700` and Saquinavir over time can be seen in Figure 9 and Figure 10 in the appendix.

The most mobile regions for protein with `fda_1700` are residues 41, 53, 51, and 50. While for protein with Saquinavir,

the most mobile regions are residues 41, 17, 69, 18, and 16. Positions of these residues can be seen in the Figure 6.

D. Radial Distribution Function (RDF)

The Radial Distribution Function (RDF) is a measure of the probability of finding a particle at a distance r from a reference particle. Function $g(r)$ measures how density varies as a function of distance from a reference particle. If $g(r) = 1$, it means that particles are distributed at that distance in a completely random, homogeneous manner, as expected in an ideal gas or the bulk phase of a liquid.

A hydration shell is a layer of water molecules that surrounds a solute when it's dissolved in water. This interaction of the protein surface with the surrounding water is often referred to as protein hydration and is fundamental to the activity of the protein. Solvation shell water molecules can also influence the molecular design of protein binders or inhibitors.

I calculated the RDF for the protein with `fda_1700` and Saquinavir to compare the distribution of water molecules around the heavy atoms of the protein (not including hydrogen atoms). The RDF plots with the distribution of water molecules around the protein with `fda_1700` and Saquinavir and highlighted peaks can be seen in Figure 7.

IV. Discussion

A. RMSD

Looking at Figure 4, the `fda_1700` simulation shows significant initial conformational changes, as evidenced by the initial rise in RMSD. It stabilizes somewhat but continues to fluctuate throughout the simulation, suggesting this structure is more flexible or undergoes conformational changes during the simulation. The `saquinavir` line shows much lower RMSD values, which remain relatively stable throughout the simulation. This suggests that `saquinavir` maintains a more stable conformation compared to `fda_1700` during the same simulation timeframe.

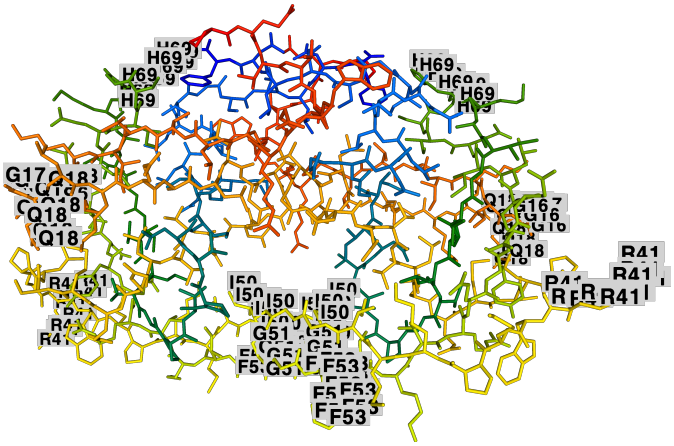

Figure 6: HIV-1 protease variant G48T/L89M with residues that have the highest RMSF values marked.

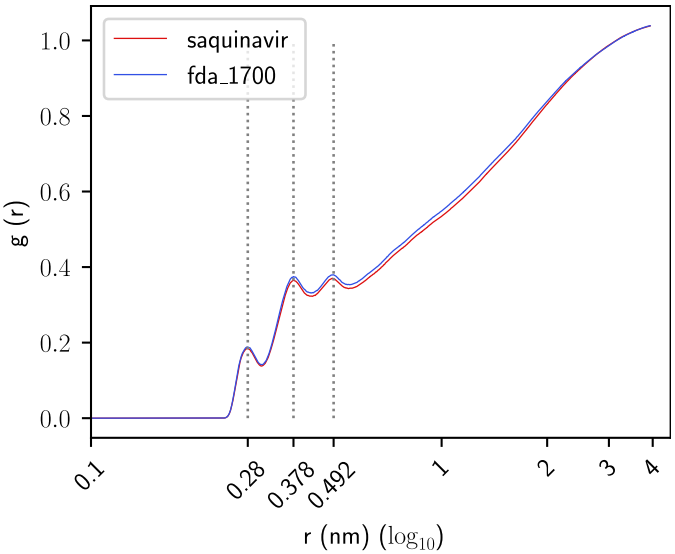

Figure 7: The Radial Distribution Function (RDF) graph shows how the density of water molecules varies as a function of distance from the protein complexed with two different ligands: `fda_1700` and Saquinavir. The x-axis is the distance from the protein in nanometers on a logarithmic scale, and the y-axis is the $g(r)$, a measure of density relative to bulk water. Vertical lines highlight peaks in the RDF plot.

B. Radius of Gyration

Comparison of R_G in Figure 5 shows that the protein with the `fda_1700` ligand starts with a larger R_g, which might indicate it initially adopts a more expanded conformation that is less stable or allows greater flexibility in the protein structure than when bound to Saquinavir. However, towards the end, the drop in R_g for `fda_1700` below Saquinavir's could indicate that the protein-ligand complex has reached a stable conformation after undergoing necessary conformational adjustments.

C. RMSF

The RMSF plots in Figure 9 and Figure 10 show that the protein with `fda_1700` has more pronounced fluctuations, with some very high peaks, indicating regions of significant flexibility. This might suggest that the fda_1700 ligand causes some regions of the protein to be more dynamic.

The RMSF values are generally lower for the protein with Saquinavir, implying the protein is overall less flexible with this ligand. The lower peaks suggest more rigidity or a stable conformation.

Interestingly, the residues with the highest RMSF values differ between the two simulations. From the Figure 6, we can see that the most mobile regions for protein with Saquinavir are on the surface of the protein, while for protein with `fda_1700`, they are more towards the core of the protein (near the binding site). This could mean that the `fda_1700` ligand doesn't bind as tightly to the protein as Saquinavir, leading to more flexibility in the core regions.

D. RDF

The RDF plots in Figure 7 show the distribution of water molecules around the protein with `fda_1700` and Saquinavir.

For both ligands, there's a notable increase in water density at certain distances from the protein, which could indicate preferred distances where water molecules are more likely to be found due to interactions with the protein-ligand complex. Peaks and troughs represent areas of higher and lower water molecule density, respectively.

The lines for "fda_1700" and "saquinavir" track closely together, suggesting similar hydration patterns for both ligands when bound to the protein. Differences in the lines could indicate differences in how water molecules interact with and organize around the different ligands. If Saquinavir fills the binding "hole" in the protein better than `fda_1700`, it may displace more water molecules from the cavity, resulting in a lower RDF, which we observe in the plot. This could indicate a tighter binding of Saquinavir to the protein compared to `fda_1700`.

V. Conclusion

In this report, I described the computational methods used to perform molecular docking and molecular dynamics simulations of ligands against the HIV-1 protease variant G48T/L89M. I selected the ligand `fda_1700` with the lowest binding affinity score (aside from Saquinavir variants) from the docking results and compared it to Saquinavir in terms of RMSD, R_g, RMSF, and RDF.

The results suggest that the protein with `fda_1700` undergoes more conformational changes and has higher flexibility compared to Saquinavir. The higher RMSD, R_g, and RMSF values for `fda_1700` indicate that it may not bind as tightly to the protein as Saquinavir, leading to more flexibility in the core regions of the protein. The RDF plots show similar hydration patterns for both ligands, but differences in water density around the protein could indicate differences in how water molecules interact with the protein-ligand complex.

Overall, the results suggest that Saquinavir may be a more stable and tightly bound ligand compared to `fda_1700`, which was expected given Saquinavir's known inhibitory activity against the HIV-1 protease.

I. Appendix

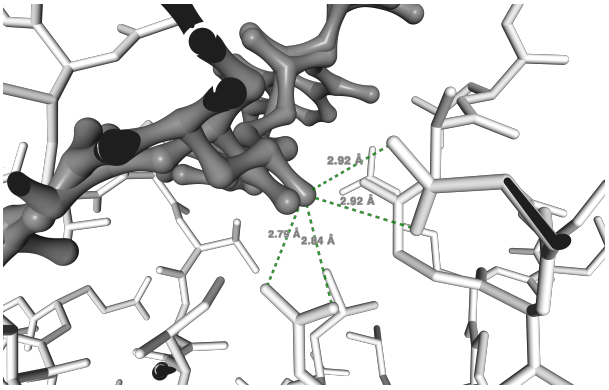

Figure 8: Distances between the center atom `A/ROC100/O2A` of the Saquinavir and surrounding atoms of the Protease.

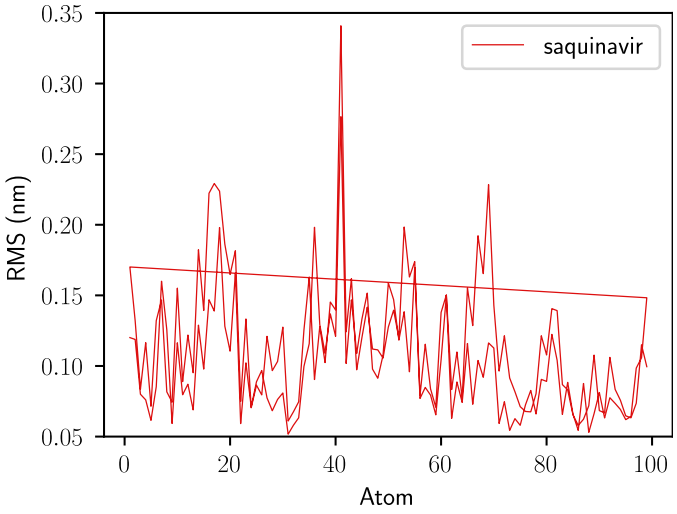

Figure 9: Root-Mean-Square Fluctuation (RMSF) of the protein when simulated with the Saquinavir over time.

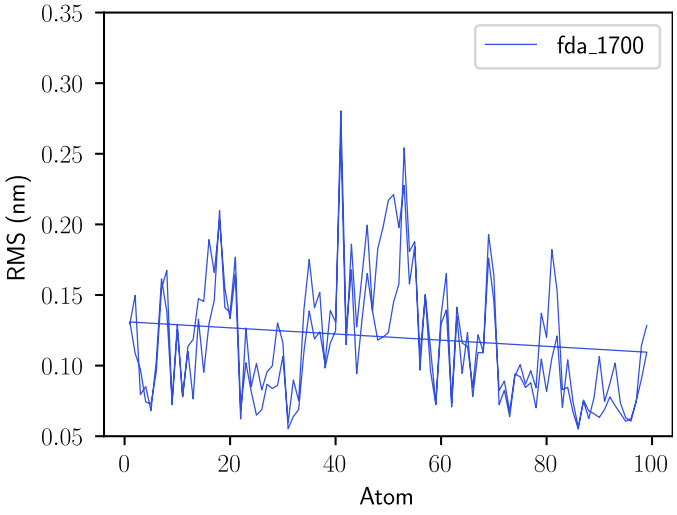

Figure 10: Root-Mean-Square Fluctuation (RMSF) of the protein when simulated with the ligand `fda_1700` over time.