# Molecular Dynamics Analysis of the Conformational Flexibility of Loops in the SARS-CoV2 Spike Protein

by

Cassandra Wong

A research project
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Waterloo, Ontario, Canada, 2022

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

**Abstract**

The SARS-CoV-2 virus has been identified as the causative agent of the COVID-19 disease that led to a global pandemic. Considerable effort has been spent in solving the spike (S) protein structure, which plays a critical role in viral infection. We focus on the loop regions of the S protein receptor-binding domain (RBD) that experiences a higher predicted disorder propensity compared to regular secondary structures. Given its ability to hold multiple stable conformations, protein loops are more difficult to model as traditional loop modeling methods tend to be designed for predicting a single conformation. In this work, we show that the conformational space of flexible loop regions can be better described by combining loop modeling protocols and molecular dynamics (MD) simulations. It is applied to the S protein RBD structural template to explore the structural variability of Loop 3. Root-mean-squared deviation (RMSD) and cluster analysis of the resulting conformations were conducted to determine whether the generated decoys can be accurately grouped based on the energy state they belong to.

**Acknowledgements**

I would like to thank Dr. Samuel Wong for his continued support and guidance throughout the research project. I would also like to thank Dr. Mu Zhu for his precious time reading this paper as a second reader. Lastly, I would like to thank the Waterloo Math community and the CM friends I have made over the course of my Master's program.

## Dedication

This is dedicated to my friends and family for supporting me throughout my Masters's degree.

# Table of Contents

**Author's Declaration** ii

**Abstract** iii

**Acknowledgements** iv

**Dedication** v

**List of Figures** viii

**List of Tables** ix

**1 Introduction** 1

1.1 Overview of SARS-CoV-2 and S Protein . . . . . 1

1.2 Loop targets of the SARS-CoV-2 S protein . . . . . 2

1.3 Related Work . . . . . 3

1.4 Goal of the Project . . . . . 5

**2 Materials and Methods** 6

2.1 Preparation of Initial Structural Models . . . . . 6

2.2 Loop modeling methods . . . . . 7

2.3 MD Simulation . . . . . 8

2.4 Analysis of MD Trajectory Output . . . 9
2.4.1 Visualization of the Conformations . . . 9
2.4.2 Plotting the System Settings . . . 9
2.4.3 RMSD Analysis . . . 9
2.4.4 Cluster Analysis . . . 9

**3 Results and Discussion** **11**
3.1 Loop Modeling Decoys . . . 11
3.2 MD Simulations . . . 11
3.3 Output Analysis . . . 12
3.3.1 Visualization of Structures . . . 12
3.3.2 Temperature and Energy Data Analysis . . . 13
3.3.3 RMSD Analysis . . . 17
3.4 Cluster Analysis of the RBD-decoys hybrid conformation from MD trajectories 21

**4 Conclusion** **24**

**References** **26**

**APPENDICES** **29**

**A Codebase** **30**
A.1 Prepare PDB Files with pdb4amber . . . 30
A.2 Shell Script to Prepare Simulation Components . . . 30
A.3 Energy Minization and Equilibrium Input Files . . . 31
A.3.1 Minimization Input, '01_Min.in' . . . 31
A.3.2 Heating Input, '02_Heat.in' . . . 31
A.3.3 Production Input, '03_Prod.in' . . . 32

# List of Figures

1.1 The schematic representation of Monte Carlo (in blue) and Molecular Dynamics methods sampling (in red). By Qx8314 - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=94107414 . . . . . 4

3.1 Visualization of RBD (top) and Loop 3 (bottom) of 7dddA Decoys 1 to 5. 12

3.2 Visualization of RBD (top) and Loop 3 (bottom) of 6x0mB Decoys 1 to 5. 13

3.3 Change in Temperature of Decoy 1 in 7dddA During MD Simulation. . . . 14

3.4 Change in Energy of Decoy 1 in 6x0mB During MD Simulation. . . . . 14

3.5 Coordinate RSMD Analysis of Decoy 1 to Decoy 5 in 7dddA, along with their respective visualizations at peaks obtained via VMD. . . . . 19

3.6 Coordinate RSMD Analysis of Decoy 1 to Decoy 5 in 6x0mB, along with their respective visualizations at peaks obtained via VMD. . . . . 20

# List of Tables


3.1 Energy data at the final conformation of PDB: 7dddA. . . . . 15
3.2 Energy data at the final conformation of PDB: 6x0mB. . . . . 15
3.3 RMSD of the Final Conformation of 7dddA for Decoy 1 to 5. . . . . 15
3.4 RMSD of the Final Conformation of 6x0mB for Decoy 1 to 5. . . . . 16
3.5 Cluster Analysis Results of Frames from 7dddA and 6x0mB, $\epsilon = 1.9$ Å . . 23
3.6 Cluster Analysis Results of Frames from 7dddA and 6x0mB, $\epsilon = 1.0$ Å . . 23

# Chapter 1

# Introduction

## 1.1 Overview of SARS-CoV-2 and S Protein

The COVID-19 disease caused by a strain of the novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has rapidly spread worldwide, leading to a global pandemic. The first known case was detected in Wuhan, China in December 2019 [21]. SARS-CoV-2 is a single-stranded RNA virus, which has a higher mutation rate leading to a greater genetic diversity compared to DNA virus [5]. This leads to resistance against both vaccine and infection-induced immunity. Since the beginning of the global pandemic, a number of prominent variants have been observed, notably Alpha, Beta, Delta, Gamma, Epsilon, and Omicron [17].

SARS-CoV-2 encodes a total of 29 proteins, including the following structure proteins: the spike protein (S), the envelope protein (E), the membrane glycoprotein (N), and the nucleocapsid protein (N). In particular, the membrane-anchored S protein is a class I membrane fusion protein that plays a critical role in facilitating the penetration process into host cells by recognizing and binding to human angiotensin-converting enzyme 2 (ACE2) through its receptor binding domain (RBD) [14]. Throughout this process, the S protein adapts to multiple stable conformations. Currently, the vaccines and antibody-based therapeutic agents being administered worldwide work by facilitating the human body to produce a modified version of the S protein, which then creates neutralizing antibodies to disturb the attachment process between the S protein RBD and the human host-cell ACE2 receptor, thereby fighting against SARS-CoV-2 [13].

## 1.2 Loop targets of the SARS-CoV-2 S protein

Given the urgency to characterize the S protein to develop potential therapeutic options against the disease, its structure has been solved many times, using various experimental methods such as X-ray crystallography, cryo-electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR). As of September 29, 2022, there were 1162 3-D protein structure data found in the RCSB Protein Data Bank (PDB) (https://www.rcsb.org) for the SARS-CoV-2 S protein and RBD. These structural data come from a collection of experimental data, including standalone S protein, S protein in complex with antibodies and various forms of ACE2, and mutated strains of the S protein [2].

Identifying an abundance of static snapshots of the protein conformations is necessary for capturing the true range of its dynamic movement to develop therapeutics that can target the range of conformations. A notable example of its conformational dynamics is the ability of the S protein to be in an open, receptor-accessible state, and closed receptor-inaccessible state, depending on whether its RBD is in an up or down position, respectively [14]. Thus, the S protein in SARS-CoV-2 provides a rich source of structural data.

This research project focuses on the loop conformations of the S protein. In a protein, amino acid residues form regions of regular secondary structures like $\alpha$-helices and $\beta$-sheets. Protein loops are the flexible regions that join these structure elements together and do not have regular, easily observable patterns. More mutations happen in loops in comparison to the more conserved $\alpha$-helices and $\beta$-sheets, resulting in a higher predicted disorder propensity [18]. Protein loops that have multiple stable conformations are known to be difficult to model compared to rigid or inflexible structures [1][11][19]. Yet, modeling the loop region is necessary in order to develop downstream therapeutic applications. On top of playing a vital role in connecting regular secondary structures, protein loops are also important for other biological processes such as protein-protein interactions, and recognition sites. In the case of SARS-CoV-2, an extended loop of the S protein receptor-binding domain (RBD) interacts directly with the loops of ACE2 [20].

## 1.3 Related Work

Recent work concerning the COVID-19 spike protein has implemented loop modeling methods and molecular dynamics (MD) simulation. In Wong et al.'s paper, various loop modeling methods were implemented in loop targets within the SARS-CoV-2 S protein. Traditional ab initio loop modeling methods such as Rosetta's next-generation KIC (NGK) algorithm, the PETALS algorithm, and the DiSGro algorithm incorporate sampling-based techniques to explore the conformational space with the guidance of energy or scoring steps (i.e., finding the most likely conformation with the lowest potential energy) without directly making use of any structure templates of known loop conformations. Sphinx is another loop modeling method that is hybridized to take in loop structure fragments from sequence alignment and finish off the loop construction by ab initio sampling [19]. Monte Carlo simulation methods calculate the thermodynamical statistical probabilities of acceptance or rejection of moves by applying random perturbations to the system to generate an ensemble of representative configurations under specific thermodynamic conditions for a complex system [12]. Accuracy is usually measured by computing the root-mean-squared deviation (RMSD) of the residues of the predicted loop conformation versus the corresponding PDB structural data. These loop modeling methods have not been very successful in predicting loops with multiple conformations from a single structural template as they tend to be designed for predicting a single conformation [19].

On the other hand, molecular dynamics (MD) simulation has been utilized to probe the conformational space of the SARS-CoV-2 RBD region and obtained promising results in William et al.'s work [18]. MD simulates molecular movements by solving Newton's equations of motion for the molecules, resulting in the temporal evolution of the coordinates and the state of a given macromolecular structure, or a "trajectory" [12]. The authors have identified four loop regions within the RBD and found loop 3 (residues 475-487) and loop 4 (residues 495-506) to be the most flexible regions within the S protein RBD region as they show the largest root-mean-squared fluctuation (RMSF) values. Both loops are located directly adjacent to the binding interface with ACE2 and act as a stabilizing contact with ACE2 during the binding process. This suggests that MD could be a useful tool more generally, that could be used together with loop modeling methods to provide insight into loops with multiple stable conformations.

The main differences between Monte Carlo and MD simulation are that the MD simulation method is a deterministic technique that describes the time evolution of a system of particles, whereas the Monte Carlo simulation method is not deterministic as particles

are generated randomly to represent a target probability distribution consistent with the desired state of the system. In addition, the Monte Carlo method does not provide information regarding the temporal evolution of a system [12]. The schematic representation illustrating the difference in sampling the system's potential energy surface using Molecular Dynamics versus Monte Carlo simulation methods is shown below.

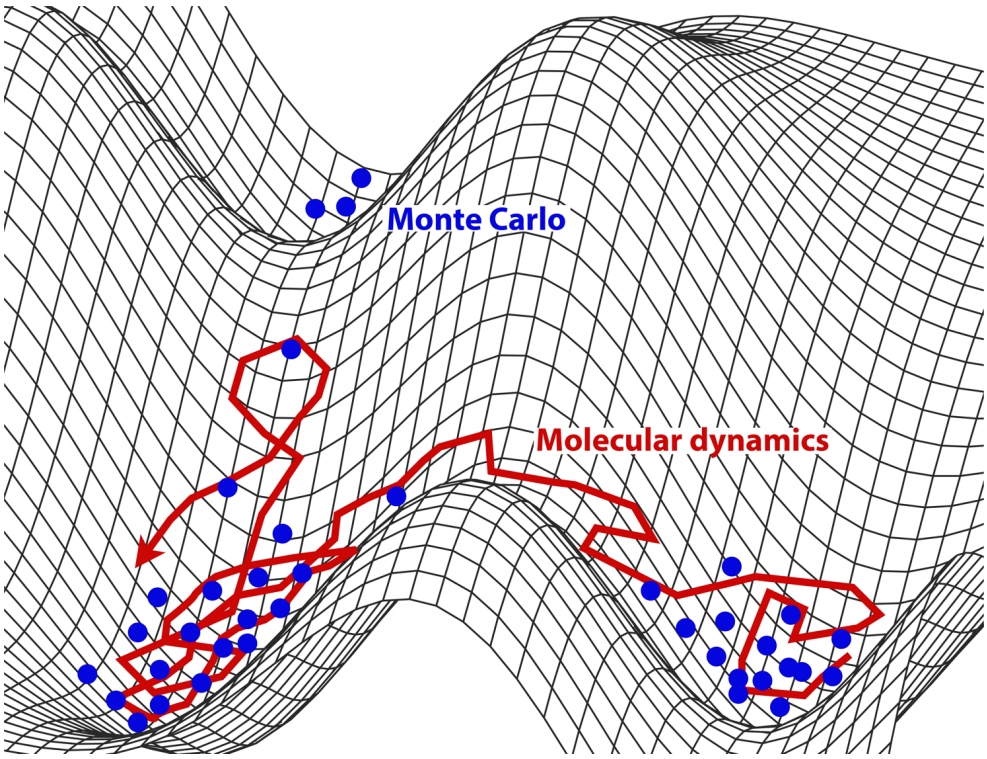



Figure 1.1: The schematic representation of Monte Carlo (in blue) and Molecular Dynamics methods sampling (in red). By Qx8314 - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=94107414

Monte Carlo and MD simulation methods can be combined such that the Monte Carlo method is used to generate promising random conformations from the energy surface, these can then be used to initialize an MD simulation for further exploration over time. Other recent studies have combined Monte Carlo and MD simulations to estimate the absolute binding free energy in the use case of drug design [8] and conducted a systematic study of protein loop structure predictions using MD simulations [6]. These papers suggest that this idea could be promising in related contexts.

## 1.4 Goal of the Project

This research project focuses on the loop regions within the SARS-CoV-2 Spike protein. The goal of the research project is to better describe the conformational space of flexible loops, by leveraging some combination of loop modeling methods and MD tools, and appropriately analyzing the output data from these methods.

# Chapter 2

# Materials and Methods

## 2.1 Preparation of Initial Structural Models

The initial structures of the SARS-CoV-2 Spike (S) protein were obtained in PDB format from the RCSB website (https://rcsb.org) on September 29, 2022, through the SARS-CoV-2 Resources page, then accessed all PDB structures for spike proteins and receptor binding domains [2]. We manually extracted all S protein structures that satisfy the following conditions: the S protein is not bound to any other macromolecules; the sequence length of the S protein is greater than 1000. This is to ensure that the selected full-length S protein structures are not affected by conformational changes from interactions with other entities. As a result, a total of 273 S protein structures were extracted, or 839 individual chains. The structures were categorized based on mutation, with the most prevalent variant being Delta at 12.8% (35), D614G at 12.45% (34), and Omicron at 11.7% (32). Wild-type S protein is at 8.8% (24).

The notable mutations that each variant contains are listed below [7]. Note that the nomenclature system for protein mutation is read as the first amino acid code changing to the second amino acid code at the numbered site (i.e., L452R is read as leucine (L) changing to arginine (R) at position number 452), and "del" refers to the amino acids flanking the deletion site.

- Delta: L452R, T478K, P681R
- Omicron: P641H, N440K, N501Y, S477N, many others

- Alpha: 69–70del, N501Y, P681H
- Gamma: K417T, E484K, N501Y
- Beta: K417N, E484K, N501Y

We are interested in the receptor-binding domain (RBD) region of the S protein, which corresponds to residues 333-526. We choose to focus on the wild-type variant structure because their loop structures are well-studied, such that there is more information on the wild-type variant's identified loop regions in comparison with other variants. From the initial PDB file obtained from the RCSB website, the RBD portion was extracted and isolated for subsequent loop modeling and MD simulation.

## 2.2 Loop modeling methods

Given the importance Loop 3 of the SARS-CoV-2 RBD has in the binding process with ACE2, it was chosen as our region of interest. It is located in residues 475-487. William et al.'s work has demonstrated its dynamic flexibility using MD simulation [18]. Wong and Liu's research has found the following PDB chains: 7dddA and 6xm0B to represent two distinct loop conformations observed in the PDB for this particular loop region [19]. The coordinates of 6x0mB were rotated to match the coordinates of 7dddA such that their loop regions were aligned.

There are several ab initio methods that explore the conformational space of a loop with the guidance of an energy or scoring function, including Rosetta's next-generation KIC (NGK) algorithm, the DiSGro algorithm, and the PETALS algorithm. The NGK algorithm was chosen as the loop modeling method in this project, however, other methods could also have been used. The NGK algorithm improves on a previous kinematic closure method, which consists of local conformational sampling and Monte Carlo minimization steps performed over two (coarse and full-atom) stages [15]. The program outputs the lowest energy loop structure found in each run. The program was iterated 500 times to obtain the desired ensemble of decoys. Hierarchical clustering was performed to summarize the total of 500 decoys into 10 specific ones to ensure that the MD starting points would be substantively distinct, where the medoid of each cluster acts as its representative. The top 5 decoys were then used to substitute the corresponding residue positions in the RBD PDB files from Section 2.1. This substitution provides context to model the change in loop conformation over time during molecular dynamics (MD) simulation.

## 2.3 MD Simulation

To perform molecular dynamics (MD) simulation, the AMBER 2021 package was used [3]. PDB files were analyzed and cleaned using pdb4amber, as well as removed all of the water molecules and added hydrogen atoms. The topology and coordinate files were built using tleap. The structure was treated with the ff19SB force field. The protein was immersed in a periodic solvent box, specifically, a truncated octahedron box of OPC water with at least 20.0 Angstroms buffering distance between the periodic box wall and the protein. Each system was neutralized by adding 2 $Cl^-$ counter ions. The nonbonded cutoff distance was 9.0 Angstroms.

Energy minimization of the system was performed, with 1000 cycles of steepest descent minimization, then 1000 cycles of conjugate gradient minimization. 2 fs integration step was used, with the output file being printed every 100 steps throughout the simulation process. Equilibration was then executed, where the system was heated from 0 to 300 K over 100 ps. After reaching the desired 300 K, the temperature was set to remain constant for another 40 ps with Langevin dynamics and a Langevin thermostat collision frequency of 5.0 $ps^{-1}$. The SHAKE algorithm was enabled to constrain all bonds involving hydrogen. The pressure was set to remain constant at 1 atm with a relaxation time of 2.0 ps. The production runs for the RBD lasted for 1 ns. The MD process totals 5,700 frames per decoy.

## 2.4 Analysis of MD Trajectory Output

### 2.4.1 Visualization of the Conformations

Visualization of the 3-D structures was carried out with VMD using trajectory outputs, with the colouring method set as "protein", and the drawing method set as "NewCartoon" [9].

### 2.4.2 Plotting the System Settings

The perl script in AMBER 2021 package was used to automate the data extraction process using output files from heating and production [3]. Graph plotting of the temperature and energy data was performed with Grace to ensure the simulation protocol did not have potential issues such as bad starting structure, unsuitable time steps, or wrong parameters [16].

### 2.4.3 RMSD Analysis

The root-mean-square deviation (RMSD) between its corresponding backbone atoms (N, $C_\alpha$, C) was used to assess structural changes in 3-D protein structures throughout the MD simulation. CPPTRAJ implemented in AMBER 2021 is used, and the trajectory inputs were the heating and production output files [3]. The mass-weighted RMSD used all the backbone atoms corresponding to the Loop 3 region. The first frame of the trajectory was used as a reference structure to calculate deviation. Grace was then used to plot the output of RMSD versus frame number [16].

### 2.4.4 Cluster Analysis

Cluster analysis with the average-linkage hierarchical agglomerative method was performed on the Loop 3 region for decoy 1 to decoy 5, corresponding to residues 475 to 487. Since each decoy was generated from the loop modeling method as per section 2.2, they all contained their own respective topology files. While CPPTRAJ in AMBER 2021 provided the tools to perform cluster analysis, it was unable to conduct clustering for conformations containing multiple topology files due to the underlying way coordinates data sets are handled [3]. Hence, cluster analysis was done manually.

First, an ensemble PDB file was created by CPPTRAJ in AMBER 2021. It was set to read and extract frames using an offset of 10 to reduce a total of 5000 frames to 500 frames [3]. Only the backbone atoms (N, $C_\alpha$, C) for the Loop 3 regions were selected. The program, PDBParser in Bio.PDB was used to read and parse the ensemble PDB file in its separated chain [4]. Its coordinates were saved as a 3x39 matrix, where row (3) refers to each of the X, Y, and Z coordinates, and column (39) refers to the 3 backbone atoms for each of the 13 residues. We created a Python script to calculate all pairwise RMSDs between the extracted frames (i.e., a total of 5000 frames from 10 decoys, with 500 frames per decoy) in both 7dddA and 6x0mB. Scikit-learn was used to run an average-linkage agglomerative clustering using the resulting distance matrix [10]. Numerous epsilon ($\epsilon$) cutoffs were evaluated and found that 1.0 and 1.9 Angstroms provide the most insights (i.e., $\epsilon > 1.9$ results in one cluster, and $\epsilon < 1.0$ results in more than 27 clusters). Hence, these epsilon ($\epsilon$) cutoffs were used.

# Chapter 3

# Results and Discussion

## 3.1 Loop Modeling Decoys

The NGK algorithm was successfully used to generate 500 new conformations for each of the Loop 3 region of the SARS-CoV-2-RBD, 7dddA and 6x0mB at residues 475 to 487. Cluster analysis was performed to gather 10 clusters, and the medoid of each cluster was used as its representative. The top 5 decoys were selected to substitute the initial RBD PDBs, resulting in 5 individual RBD PDB files with substituted Loop 3 decoys. See Section 2.2 for more details regarding the loop modeling method.

## 3.2 MD Simulations

As per Williams et al., [18], Loop 3 was found to be the most flexible loop within the SARS-CoV2-RBD. MD simulation of the RBD was performed to observe how its conformation and dynamics were altered over time. The MD simulation consisted of three main steps. First, the system was heated from 0K to 300K for 100 ps. Then, the temperature remained constant at 300K for 40 ps. Lastly, production time lasted for 1 ns, totalling 1140 ps for the whole simulation. The frame output was recorded every 100 steps. Since the process of heating and production resulted in 70,000 steps and 500,000 steps, the output becomes 700 and 5,000 frames, totalling 5,700 frames for the whole simulation. See section 2.3 for more details regarding MD settings.

Other possible experimental settings were explored. Firstly, protein position restrain was

in place during the heating process to allow the solvent to equilibrate around the protein without disturbing the protein structure. In addition, the production length was increased to 2 ns for to run the simulation for longer, biologically relevant timescales. Ultimately, they were not implemented in the study as they did not affect the output significantly.

## 3.3 Output Analysis

After performing loop modeling and MD simulation on the five decoys of 7dddA and 6x0mB, the output data from these methods were appropriately analyzed. The following analysis tools were performed: visualization of the 3-D structures, plotting the temperature of the energy data from output files, as well as RMSD analysis.

### 3.3.1 Visualization of Structures

The conformations of the RBD portion, as well as the extracted Loop 3 region in all 5 decoys of 7dddA and 6x0mB were visualized using VMD. The final frame of the simulation was imported. Based on the observation in both Figure 3.1 and Figure 3.2, it is obvious that Loop 3 in the SARS-CoV-2-RBD can take on a vast range of conformations.

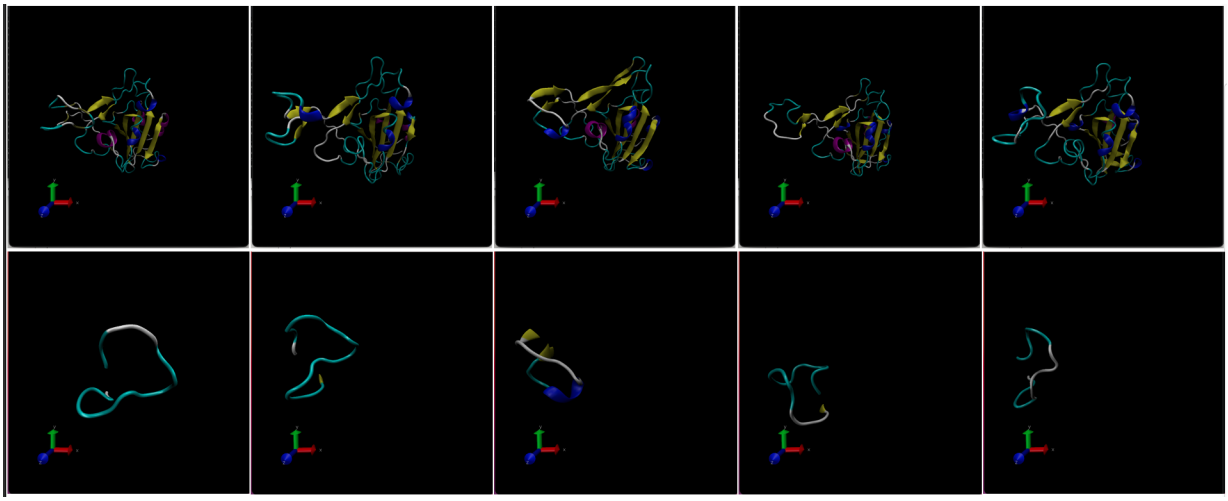


Figure 3.1: Visualization of RBD (top) and Loop 3 (bottom) of 7dddA Decoys 1 to 5.

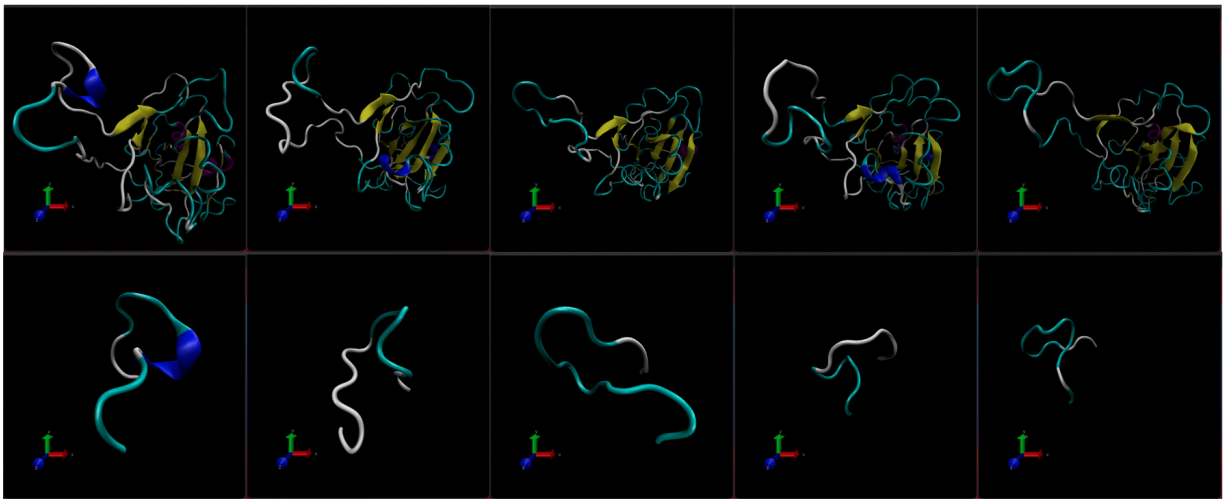


Figure 3.2: Visualization of RBD (top) and Loop 3 (bottom) of 6x0mB Decoys 1 to 5.

### 3.3.2 Temperature and Energy Data Analysis

The temperature of the system throughout the MD simulation is plotted to ensure the temperature rose smoothly and equilibrated at 300 K. The plot of decoy 1 in 7dddA is shown below in Figure 3.3. We see that the heating phase for the first 100 ps illustrates a gradual increase from 0 K to 300 K and no sudden jumps, as expected. The temperature during the rest of the simulation time from 100 ps to 1140 ps remains constant at 300 K.

The energy data is then visualized to confirm the rise in energy is smooth and proportional to the temperature of the system. The temperature plot of decoy 1 in 7dddA is shown below in Figure 3.4. The change in kinetic energy and potential energy from the first 100 ps is found to be from -472,612.26 KCal/Mol to -286,812.75 KCal/Mol, and 1,881.29 KCal/Mol to 43,173.12 KCal/Mol, respectively. The initial rise in the kinetic energy and potential energy is shown to be directly proportional to the temperature plot from Fig. 3.3. The total energy levels off at equilibrated values for the remainder of the simulation from 100 ps to 1140 ps. Note that the energy output is of the entire RBD rather than of the loop region. During the heating process for the first 100 ps, the increase in temperature provides energy to allow the structure to move to another conformation, then stabilizes at a new energy state when the temperature remains constant during production.

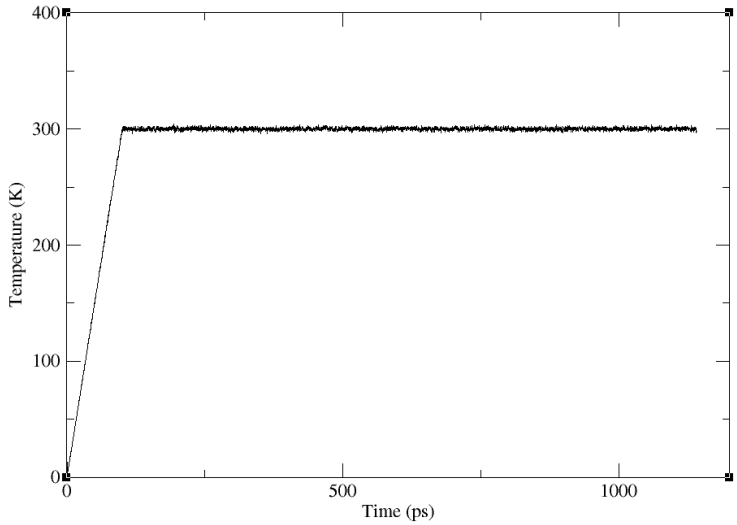



Figure 3.3: Change in Temperature of Decoy 1 in 7dddA During MD Simulation.

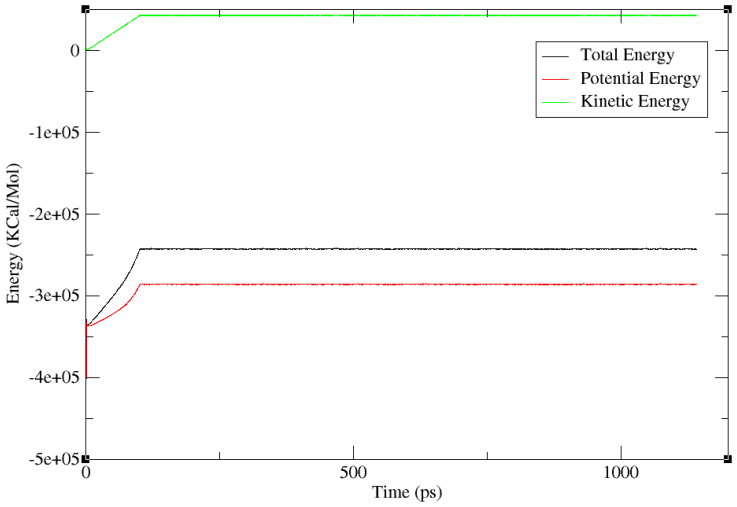



Figure 3.4: Change in Energy of Decoy 1 in 6x0mB During MD Simulation.

The energy data at the final frame of the MD simulation for both PDB: 7dddA and 6x0mB are summarized in Table 3.1 and Table 3.2 below. We focus on the potential energy value as it is used to locate the structure with the lowest energy to determine what the folded conformation of the structure is.

Table 3.1: Energy data at the final conformation of PDB: 7dddA.

| (kcal/mol) | Decoy 1 | Decoy 2 | Decoy 3 | Decoy 4 | Decoy 5 |
|---|---|---|---|---|---|
| Total Energy | -243,110.25 | -273,667.89 | -247,122.38 | -268,941.62 | -249,358.50 |
| Kinetic Energy | 42,983.45 | 48,763.73 | 44,058.91 | 47,802.32 | 44,603.72 |
| Potential Energy | -286,093.70 | -322,431.62 | -291,181.29 | -316,743.94 | -293,962.22 |

Table 3.2: Energy data at the final conformation of PDB: 6x0mB.

| (kcal/mol) | Decoy 1 | Decoy 2 | Decoy 3 | Decoy 4 | Decoy 5 |
|---|---|---|---|---|---|
| Total Energy | -291,051.61 | -301,253.91 | -339,251.95 | -330,984.17 | -316,120.52 |
| Kinetic Energy | 51,632.76 | 52,980.99 | 59,963.88 | 57,870.34 | 56,192.17 |
| Potential Energy | -342,684.37 | -354,234.90 | -399,215.83 | -388,854.50 | -372,312.69 |

The large discrepancy in the energy state at the final conformation between all 5 decoys for both 7dddA and 6x0mB shows that all decoys ended up in different regions of the energy surface. From Table 3.1 and 3.2, we observe that decoy 2 of 7dddA and decoy 3 of 6x0mB have the lowest potential energy. To determine whether these structures are the conformations that are most likely to be observed in nature, we compute the RMSD between the 5 decoys and the original loop conformation in the PDB.

Table 3.3: RMSD of the Final Conformation of 7dddA for Decoy 1 to 5.

| (Angstrom) | Decoy 1 | Decoy 2 | Decoy 3 | Decoy 4 | Decoy 5 |
|---|---|---|---|---|---|
| RMSD | 3.1412 | 1.9377 | 1.6487 | 1.9320 | 1.1852 |

Table 3.4: RMSD of the Final Conformation of 6x0mB for Decoy 1 to 5.

| (Angstrom) | Decoy 1 | Decoy 2 | Decoy 3 | Decoy 4 | Decoy 5 |
|---|---|---|---|---|---|
| RMSD | 2.3470 | 2.5579 | 2.1803 | 2.2531 | 2.1471 |

The conformations with the lowest potential energy are expected to have the lowest RMSD value to the original loop conformations as per our hypothesis. However, Table 3.3 and 3.4 indicate otherwise, as Decoy 3 in 7dddA and Decoy 5 in 6x0mB have the lowest RMSD values. A possible explanation is that the energy output is of the entire RBD rather than of the loop region, and low potential energy may indicate favoritism of the structure of the RBD instead of the Loop region.

### 3.3.3 RMSD Analysis

Coordinate root-mean-squared deviation (RMSD) metrics are used to assess the change in conformation to a reference set of coordinates.

RMSD is defined as follows:

$$RMSD = \sqrt{\frac{\sum_{i=0}^{N} \left[ m_i * (X_i - Y_i)^2 \right]}{M}}$$

where N represents the number of atoms, $m_i$ refers to the mass of atom $i$, $X_i$ and $Y_i$ are the coordinate vector for target atom $i$ and reference atom $i$, respectively, and $M$ refers to the total mass.

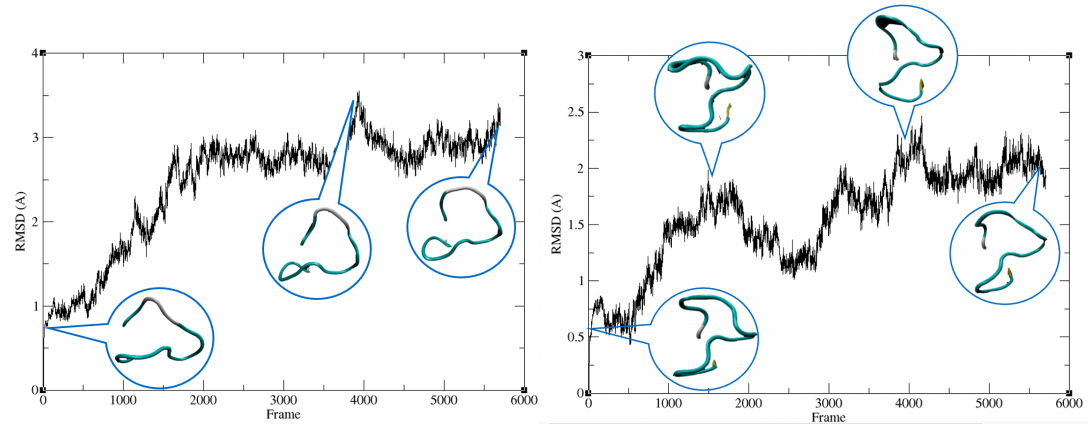
All the frames obtained from the heating and production processes of the MD simulation are imported, and the coordinates of the Loop 3 region, which is part of the large ACE2 binding interface, are used for the RMSD analysis. See section 2.4 for more information regarding RMSD runs and settings. After calculating the RMSD with respect to the reference structure (loop structure of the starting decoy), the output is plotted below in Figure 3.5 and 3.6. The X-axis refers to the frame number and the Y-axis is the coordinate RMSD to the original loop conformation (in Angstroms). For visualization purposes, the first and last frames, as well as frames of certain peaks in the RMSD plots are labeled with their 3-D structures obtained from VMD.

Our RMSD plots in Figure 3.5 and 3.6 show that decoy 1 to decoy 5 in 7dddA and 6x0mB drift away from the initial structure over time, and most of the RMSD plots plateau around frame number 3000. There is a huge conformational shift early on in the simulation as the temperature of the system increases for the first 700 frames from 0 K to 300 K. Given the vastly different visualized images of the first and last frames, the trajectories may have folded up to an alternative structure that is different from the native structure as during MD simulation, the structure is continuously unfolding and refolding as the system attempts to locate a more energy-stable structure. The plateau portion of the RMSD plots suggests that the systems have folded up to a stable equilibrium conformation.

There are peak regions in the RMSD plot for some of the decoys, such as Figure 3.5a, 3.5b, 3.6a, and 3.6c. As the system is exploring the structure's conformational space, it may detect stable energy states that are geometrically farther (resulting in a higher RMSD) from the starting state and are briefly visited by the MD trajectories. We notice that in

Figure 3.6c, the peak regions around frame 2000 and frame 4000 have a similar RMSD at 2.25, yet their visualized structures look different.

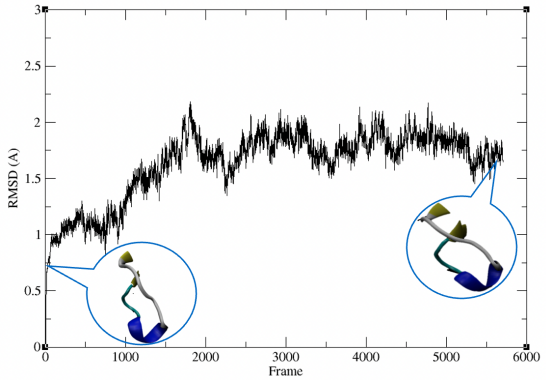
Another interesting observation is that the RMSD values of 7dddA show a larger fluctuation compared to 6x0mB, where their RMSD values range from 1.18 to 3.14 for 7dddA after MD simulation, with decoy 5 (Fig. 3.5e) having the lowest RMSD value and decoy 1 (Fig. 3.5a) having the largest RMSD value. Whereas the RMSD values are between 2.15 and 2.56 for 6x0mB, with decoy 5 (Fig. 3.6e) having the lowest RMSD value and decoy 2 (Fig. 3.6b) having the largest RMSD value. This suggests that the Loop 3 region from decoys 1 to 5 of 7dddA is more conformationally flexible than 6x0mB. Perhaps the two template structures of SARS-CoV-2 RBD differ in certain localized areas, resulting in the contrast seen in the flexibility in their respective Loop 3.



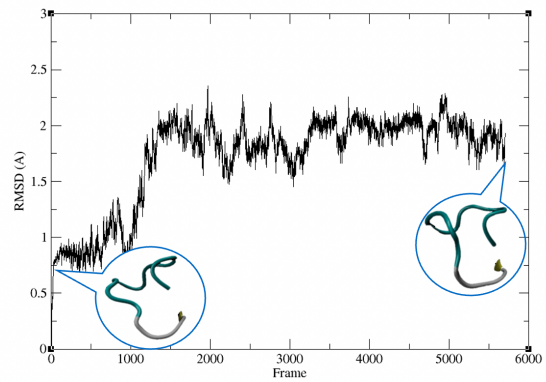

(a) Decoy 1



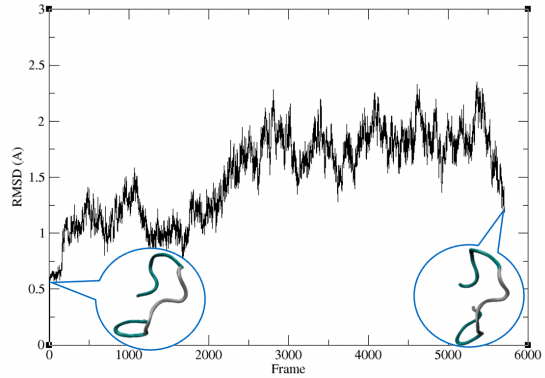

(b) Decoy 2




(c) Decoy 3

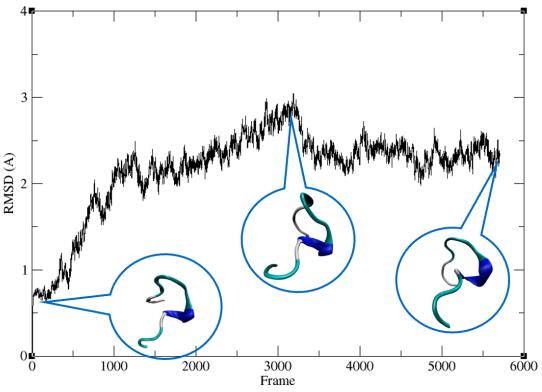



(d) Decoy 4

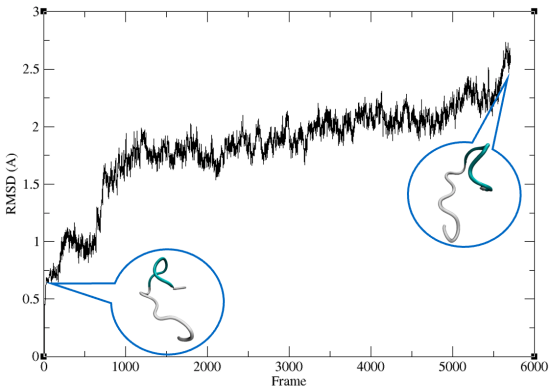



(e) Decoy 5

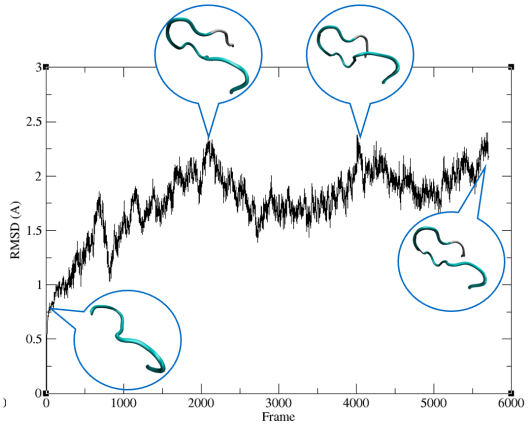
Figure 3.5: Coordinate RSMD Analysis of Decoy 1 to Decoy 5 in 7dddA, along with their respective visualizations at peaks obtained via VMD.



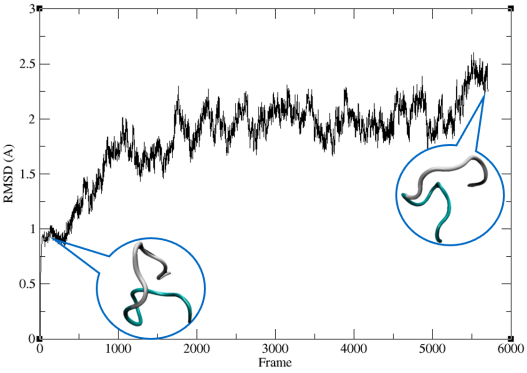

(a) Decoy 1



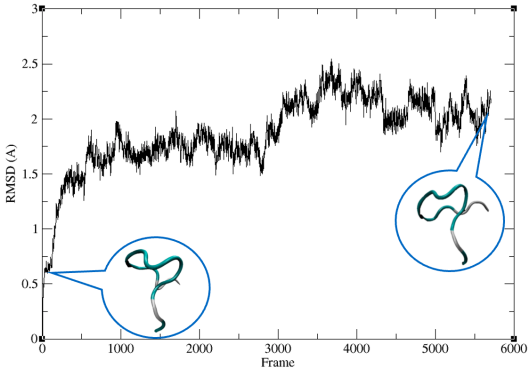

(b) Decoy 2




(c) Decoy 3




(d) Decoy 4




(e) Decoy 5

Figure 3.6: Coordinate RSMD Analysis of Decoy 1 to Decoy 5 in 6x0mB, along with their respective visualizations at peaks obtained via VMD.

## 3.4 Cluster Analysis of the RBD-decoys hybrid conformation from MD trajectories

To further our understanding of the conformational states of Loop 3 in SARS-CoV2-RBD binding interfaces that were obtained during the MD simulations, cluster analysis was performed using a hierarchical agglomerative (bottom-up) algorithm. See section 2.4 for more details regarding the cluster analysis settings.

With an epsilon ($\epsilon$) cutoff of 1.9 Å, the trajectories in decoys of 7dddA and 6x0mB were separated into six clusters. Whereas with an epsilon ($\epsilon$) cutoff of 1.0 Å, the trajectories were separated into nine clusters. The average of the pairwise RMSD similarity metric within the cluster compared to all other members of that cluster is calculated, along with its standard deviation. Clusters with a low RMSD represent a group of conformations that are similar to one another, while in contrast, clusters with a large RMSD correspond to a group of conformations that are vastly different from each other. The results are summarized in Table 3.5 and 3.6 below.

We want to determine whether the combination of loop modeling tools and MD simulation can transverse energy regions and predict conformations that belong to various regions from a single starting loop structure. If that is the case, then clusters containing frames from various sources would be expected. Table 3.5 and 3.6 show that the resulting clusters have a relatively low heterogeneity, such that all the frames of a particular decoy belonged to a cluster, rather than having frames generated from a specific decoy belonging to numerous distinct clusters. We observe that clusters that contain structures from more than one decoy have a significantly higher RMSD (i.e., least similar to one another) than clusters that only contain structures from one decoy.

Overall, while the combination of loop modeling and MD simulation was able to explore the conformational space within the local region, structures generated from each decoy after MD simulation appears to be distinct as they belong to the same cluster, and are unable to jump from one energy state to another as per cluster analysis results. Also, while the coordinates of 7dddA and 6x0mB are rotated to match up their Loop 3 regions, and are expected to hold the same loop conformation as per Wong et al.'s work [19], the trajectories produced from them did not merge at all, as shown in the cluster analysis.

With clusters containing structures from multiple starting decoys, we can assume that they may belong to a local energy region with close proximities. Cluster 1 from 7dddA

contains structures generated from decoy 1, decoy 2, and decoy 5, with an RMSD of 2.008 (see 3.5). Referencing back to Figure 3.5a, 3.5b, and 3.5e, we can see that their last frames contain a similar structure. Also, cluster 2 from 3.5 contains structures generated from decoy 1 and decoy 5, with an RMSD of 1.372, and cluster two includes decoy 2 and decoy 3, with an RMSD of 1.937. While in cluster 2, decoy 1 and decoy 5 don't look visually related based on Figure 3.6a and 3.6e, in cluster 3, the last frame of Figure 3.6c appears to have a similar structure when it is rotated 90 degrees counterclockwise as Figure 3.6b.

In conclusion, the combination of loop modeling and MD simulation could only be expected to gain information regarding the conformational space within the decoy's local energy region, rather than exploring all the possible structures from different energy regions with just a single initial structure.

Table 3.5: Cluster Analysis Results of Frames from 7dddA and 6x0mB, $\epsilon = 1.9$ Å

| Cluster | Cardinality | RMSD (Mean) | RMSD (St Dev) | Structures in Cluster |
|---|---|---|---|---|
| 1 | 1500 | 2.008 | 0.955 | 7dddA Decoy 1, Decoy 2, Decoy 5 |
| 2 | 1000 | 1.372 | 0.698 | 6x0mB Decoy 1, Decoy 5 |
| 3 | 1000 | 1.937 | 1.248 | 6x0mB Decoy 2, Decoy 3 |
| 4 | 500 | 0.57 | 0.157 | 6x0mB Decoy 4 |
| 5 | 500 | 0.691 | 0.222 | 7dddA Decoy 3 |
| 6 | 500 | 0.671 | 0.267 | 7dddA Decoy 4 |

Table 3.6: Cluster Analysis Results of Frames from 7dddA and 6x0mB, $\epsilon = 1.0$ Å

| Cluster | Cardinality | RMSD (Mean) | RMSD (St Dev) | Structures in Cluster |
|---|---|---|---|---|
| 1 | 1000 | 1.345 | 0.535 | 7dddA Decoy 1, Decoy 2 |
| 2 | 500 | 0.641 | 0.163 | 6x0mB Decoy 1 |
| 3 | 500 | 0.681 | 0.197 | 6x0mB Decoy 2 |
| 4 | 500 | 0.746 | 0.231 | 6x0mB Decoy 3 |
| 5 | 500 | 0.57 | 0.157 | 6x0mB Decoy 4 |
| 6 | 500 | 0.761 | 0.214 | 6x0mB Decoy 5 |
| 7 | 500 | 0.691 | 0.222 | 7dddA Decoy 3 |
| 8 | 500 | 0.671 | 0.267 | 7dddA Decoy 4 |
| 9 | 500 | 0.945 | 0.363 | 7dddA Decoy 5 |

# Chapter 4

# Conclusion

Loops within a protein not only serve as the join for secondary structure elements such as $\alpha$-helices and $\beta$-sheets in a protein, but also play a role in other significant biological functions. These regions experience a higher level of disorder in comparison to regular secondary structures. Hence, accurately modelling loops is a relevant problem in the field of structural biology. Loop regions in the SARS-CoV-2 S protein at residues 475–487, or "Loop 3" was found to be involved in binding with ACE2 as per Williams et al. [18], which is a significant viral infection process. Loop 3 was found to be highly flexible while in its unbound state, suggesting the possibility of a conformation that inhibits binding with ACE2. Therefore, Loop 3 was chosen to be the focus of this research project to discover a wider range of conformational states that the development of therapeutics could target prior to the binding between S protein and ACE2.

In this research project, we presented the approach to combining loop modelling methods and MD simulation tools to tackle the challenges associated with describing the conformational space of flexible loop regions in a protein structure. We extracted the RBD regions within SARS-CoV-2, examined their structural variability based on the available structures in the PDB, and applied loop modelling methods to obtain all the possible conformations. Then, the generated loop structures were grouped into 5 clusters or decoys. MD simulation was performed with minimization, equilibrium, and production phases. A total of 5,700 frames per decoy was observed to have substantive structural variability and be able to adopt multiple distinct conformations according to the RMSD analysis with the starting structure. The potential energy data reveals that decoy 2 of 7dddA and decoy 3 of 6x0mB have the lowest potential energy. This observation suggests that they are most likely to exist in the natural state.

Cluster analysis was performed with the hierarchical agglomerative (bottom-up) algorithm. It was found that all the conformations generated from each decoy after MD simulation belonged to the same cluster. This suggests that they were unable to jump between energy states. While the combination of loop modeling and MD simulation could be leveraged to discover the structural space within the local energy region, it could only be expected to gain information regarding the conformational space within the decoy's local energy region, rather than exploring all the possible structures from different energy regions with just a single initial structure.

Overall, the S-protein sequence and structure dataset in the form of PDB is a rich source of information that could be leveraged to predict its conformational space. The combination of loop modeling methods and MD simulation tools requires more fine-tuning to explore a wider range of possible conformations at different energy regions using only a single initial structure as input. Future studies could attempt to implement the loop modeling and MD simulation combination on other known loops, such as "Loop 4" in the S protein, corresponding to residue position 495-506. Loop 4 is also involved in binding with ACE2 in the host cell [18]. Another improvement of the study is to extend the production time to biological relevant time scales to microseconds.

# References


[1] Amélie Barozet, Marc Bianciotto, Marc Vaisset, Thierry Simeon, Hervé Minoux, and Juan Cortés. Protein loops with multiple meta-stable conformations: a challenge for sampling and scoring methods. *Proteins: Structure, Function, and Bioinformatics*, 89(2):218–231, 2021.

[2] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[3] David A Case, H Metin Aktulga, Kellon Belfon, Ido Ben-Shalom, Scott R Brozell, David S Cerutti, Thomas E Cheatham III, Vinícius Wilian D Cruzeiro, Tom A Darden, Robert E Duke, et al. *Amber 2021*. University of California, San Francisco, 2021.

[4] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[5] Siobain Duffy. Why are rna virus mutation rates so damn high? *PLoS biology*, 16(8):e3000003, 2018.

[6] Jia-Jie Feng, Jia-Nan Chen, Wei Kang, and Yun-Dong Wu. Accurate structure prediction for protein loops based on molecular dynamics simulations with rsff2c. *Journal of Chemical Theory and Computation*, 17(7):4614–4628, 2021.

[7] Centers for Disease Control, Prevention, et al. Sars-cov-2 variant classifications and definitions. 2021.

[8] Joan F Gilabert, Oriol Gracia Carmona, Anders Hogner, and Victor Guallar. Combining monte carlo and molecular dynamics simulations for enhanced binding free

energy estimation through markov state models. *Journal of Chemical Information and Modeling*, 60(11):5529–5539, 2020.

[9] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[10] Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer, 2016.

[11] Claire Marks, Jiye Shi, and Charlotte M Deane. Predicting loop conformational ensembles. *Bioinformatics*, 34(6):949–956, 2018.

[12] Eric Paquet and Herna L Viktor. Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *BioMed research international*, 2015, 2015.

[13] Herb F Sewell, Raymond M Agius, Denise Kendrick, and Marcia Stewart. Covid-19 vaccines: delivering protective immunity, 2020.

[14] Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition by sars-cov-2. *Nature*, 581(7807):221–224, 2020.

[15] Amelie Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PloS one*, 8(5):e63090, 2013.

[16] PJ Turner. Xmgrace, version 5.1. 19. *Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR*, 2, 2005.

[17] Saudi Arabia WHO. Who coronavirus disease (covid-19) dashboard. *World Health Organization*, 2020.

[18] Jonathan K Williams, Baifan Wang, Andrew Sam, Cody L Hoop, David A Case, and Jean Baum. Molecular dynamics analysis of a flexible loop at the binding interface of the sars-cov-2 spike protein receptor-binding domain. *Proteins: Structure, Function, and Bioinformatics*, 90(5):1044–1053, 2022.

[19] Samuel WK Wong and Zongjun Liu. Conformational variability of loops in the sars-cov-2 spike protein. *Proteins: Structure, Function, and Bioinformatics*, 90(3):691–703, 2022.

[20] Renhong Yan, Yuanyuan Zhang, Yaning Li, Lu Xia, Yingying Guo, and Qiang Zhou. Structural basis for the recognition of sars-cov-2 by full-length human ace2. *Science*, 367(6485):1444–1448, 2020.

[21] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 2020.

# APPENDICES

# Appendix A

# Codebase

## A.1 Prepare PDB Files with pdb4amber

```
pdb4amber 7d_decoy10.pdb > 7d_decoy10.amber.pdb --dry --reduce
```

## A.2 Shell Script to Prepare Simulation Components

```
tleap % Start the program
source leaprc.protein.ff14SB
source leaprc.water.opc
seq = loadPdb "7d_decoy10.amber.pdb"
solvateoct seq OPCBOX 20.0
addIons seq Cl-2‘ or ‘addIons seq Na+ 1
saveamberparm seq 7dAdecoy10.parm7 7dAdecoy10.rst7
quit
```

## A.3 Energy Minization and Equilibrium Input Files

### A.3.1 Minimization Input, '01_Min.in'

```
Minimize
 &cntrl
  imin=1,
  ntx=1,
  irest=0,
  maxcyc=2000,
  ncyc=1000,
  ntpr=100,
  ntwx=0,
  cut=9.0,
 /
```

### A.3.2 Heating Input, '02_Heat.in'

```
Heat
 &cntrl
  imin=0, ntx=1, irest=0,
  nstlim=70000, dt=0.002,
  ntf=2, ntc=2,
  tempi=0.0, temp0=300.0,
  ntpr=100, ntwx=100,
  cut=9.0,
  ntb=2, ntp=1, ntt=3,
  taup = 2.0,
  gamma_ln=5.0,
  nmropt=1,
 /
&wt type='TEMP0', istep1=0, istep2=50000, value1=0.0, value2=300.0 /
&wt type='TEMP0', istep1=50001, istep2=70000, value1=300.0, value2=300.0 /
&wt type='END' /
```

### A.3.3 Production Input, '03_Prod.in'

```
Production
 &cntrl
  imin=0,
  ntx=5,
  irest=1,
  nstlim=1000000,
  dt=0.002,
  ntf=2,
  ntc=2,
  temp0=300.0,
  ntpr=100,
  ntwx=100,
  cut=9.0,
  ntb=2,
  ntp=1,
  ntt=3,
  gamma_ln=5.0,
  ig=-1,
 /
```