A Dynamic Look Into Hydrogen Bonding in Proteins

A thesis

submitted in fulfilment

of the requirements for the degree

of

Masters of Chemistry (Research)

at

The University of Waikato

by

TROY VAN TIEL

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2022

Contents

Acknowledgments

I am grateful to have had Associate Professor Jo Lane as my supervisor and for his support and guidance to allow me to pursue this research. I am also grateful for the environment that was created at the University of Waikato for me to complete this research in. I also want to extend my unending gratitude to my family and friends who have provided their love and support throughout this year of research

.

Abstract

This thesis aims to highlight the importance of analysing the dynamic nature of hydrogen bonds and their temporal variation. We also evaluate whether distance alone can and should be used as the main estimator for hydrogen bond strength. Molecular dynamics simulations of three common protein building blocks were conducted in triplicate to provide temporal data for our investigation. This data was then run through a newly developed program called Hybond, which analyses a given hydrogen bonding interaction as a function of time based on the reduced electron density gradient. The results found were promising and showed that the strength of a given hydrogen bond changes significantly throughout each simulation trajectory. This demonstrates the dynamic nature of these hydrogen bonds, and we provide a workflow to efficiently characterise individual interactions as a function of time. We found that distance correlates well with kinetic energy density however there are points that do not fit this correlation. In these cases, hydrogen bond strength cannot be estimated through distance alone. Other properties that Hybond outputs have also shown promising correlations with kinetic energy density, and these can be used in place of distance to estimate the strength of the interaction. Finally, we investigated some of the challenges that arise when trying to determine the time-averaged strength of a given hydrogen bond and how different averaging approaches might be more appropriate than others, depending on the underlying science question that is being asked.

List of Figures

List of Tables

DEFINTIONS USED IN THIS THESIS

- **NCI** – Non-Covalent Interactions, there are different classifications of interactions such as, electrostatic, Π-effects, van der Waals forces and hydrophobic effects. Hydrogen bonds are classified as electrostatic and are the focus of this thesis.

- **HB** – Hydrogen Bond, common electrostatic interaction that contributes greatly to protein structure and the functionality of the protein.
- **MD** – Molecular Dynamics, normally a simulation that takes atoms moving into account.
- **RDG** – Reduced electron Density Gradient, calculated from the density and its first derivative is used to describe the deviation from a homogenous electron distribution. (1) (2)
- **DFT** – Density Functional Theory, an electronic structure method that produces an electron density, which can be further analysed.
- **NAMD** – Nanoscale Molecular Dynamics program, the software that was used to run most of the simulations in this paper.
- **.DCD/.dcd file** – The file type of the trajectory output used by Hybond to get each frames data. (Binary file)
- **.PDB/.pdb file** – The file type that holds the atom and position data of the protein.
- **.PSF/ .psf file** – The file type that holds the structural information of the protein.
- **AUC** – The calculation done to find the Area Under the Curve of the averaging methods

CHAPTER ONE

1.1 Proteins and Hydrogen Bonding

Proteins are responsible for many different functions in living organisms, from catalysis of metabolic reactions (3) to DNA replication (4). The structure of a protein is determined by its sequence of amino acids, with each amino acid having a different side chain group. The combination of amino acids affects how the protein folds and what overall 3D structure it will adopt. There are different levels of structure inside a protein, these are listed below and shown in Figure 1.1.1:

1. Primary Structure – The sequence of amino acids that make up the protein
2. Secondary Structure – These are repeating units in the protein structure that are held together by non-covalent interactions. Common structures formed are the alpha helix and beta sheet.
3. Tertiary Structure – The makeup of secondary structures that are stabilised by many different covalent and non-covalent interactions including hydrogen bonding, salt bridges and disulfide bonds.
4. Quaternary Structure – How multiple separate tertiary structures fit together to function as single overall protein complex.

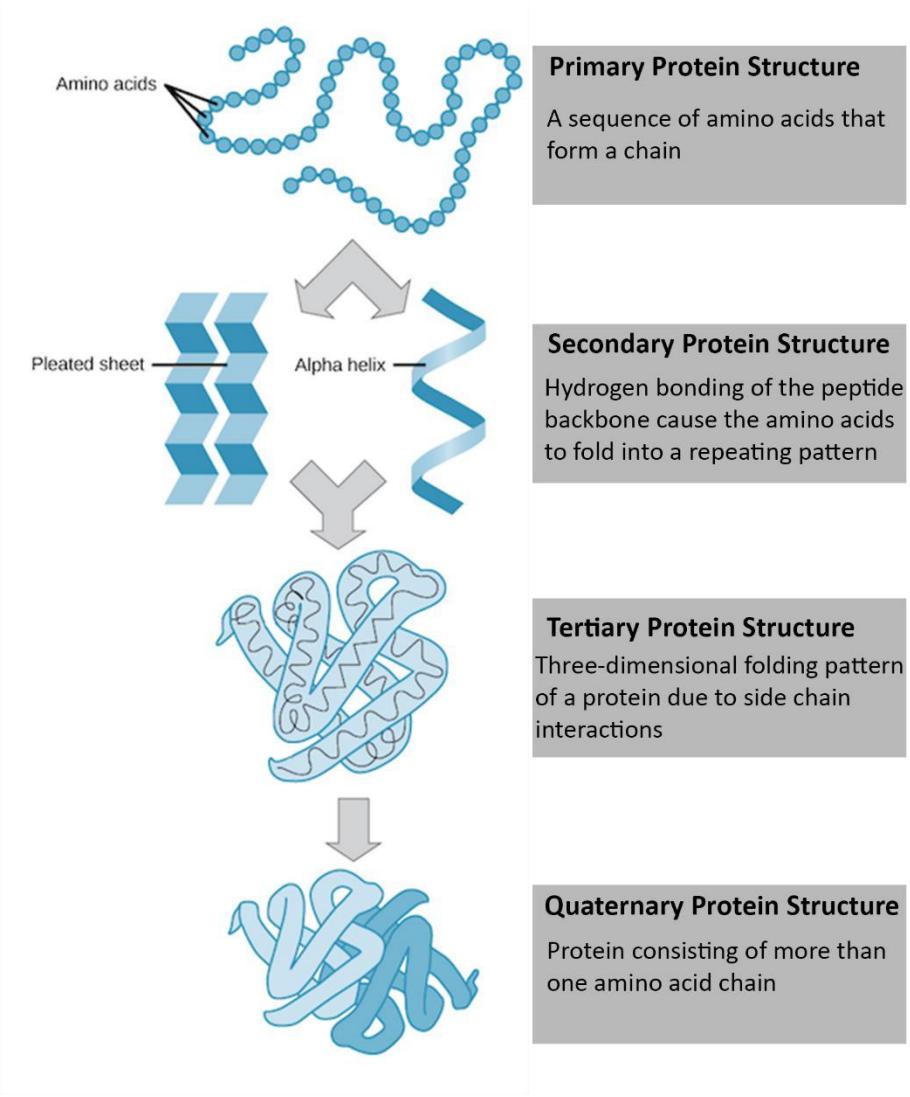


Figure 1: Different Levels of Structure in Proteins

Figure 1.1.1. Different Levels of Structure in Proteins

Proteins are inherently dynamic in nature, changing their structure and hence position of the atoms, as they complete a variety of different functions. (5) This means that proteins operate on a dynamic basis and the non-covalent interactions present are ever-changing, along with these structural changes. Hydrogen bonds are considered to be the most important noncovalent interactions in proteins, and are the main contribution to the secondary structure of proteins, and by extension, the tertiary structure as well (6). Hydrogen bonds are also responsible for how the active sites of the protein interact with other molecules. The breaking and forming of these bonds (7) are also a main factor on how these proteins behave in nature.

1.2 Hydrogen Bonds and their Types

Hydrogen bonds occur between a hydrogen atom that is attached to an electronegative atom X and interacts with another electronegative atom Y. The main strength of hydrogen bonding is dependent on how electronegative the X and Y atoms are, with more electronegative atoms leading to stronger hydrogen bonds. The distance and angle of the hydrogen bond also influence the energy. A full formal IUPAC definition (8) can be found for hydrogen bonding, which outlines some other factors that affect these interactions.

For many years, hydrogen bonds have been simply defined by geometric parameters thought to be energetically ideal, including the distance between the donor and acceptor groups and the corresponding angles, which are thought to be energetically ideal. The hydrogen bond distance is normally defined between the two electronegative atoms involved in the bond (rather than being defined in terms of the hydrogen and acceptor atoms), as these are, in practice, easier to accurately detect experimentally. Hydrogen bonds in water range from 2.7 Å to 3.3 Å, with 3.0 Å being the average bond length (9). The hydrogen bond angle is defined with the H atom in the centre and the acceptor and donor electronegative atoms on either side. Hydrogen bonds inside proteins can have slightly

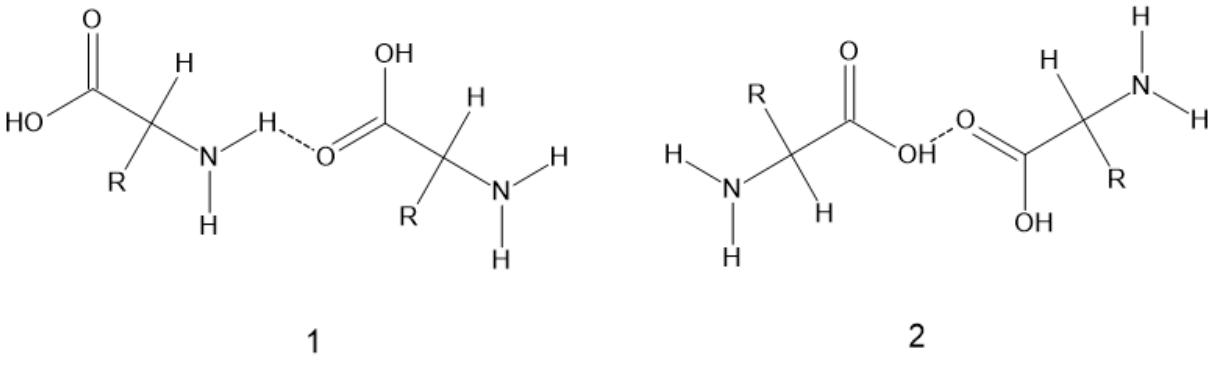


Figure 1.2.1: Two main environments of hydrogen bonding present in proteins, (1) Involves a nitrogen in the hydrogen bond, (2) involves just oxygen in the hydrogen bonding interaction.

more variation in the mean value due to the chains of the protein bending and stretching, which can push atoms closer together resulting in shorter hydrogen bonds (10).

There are two main types of hydrogen bonds that exist in proteins, which are shown in figure 1.2.1. The NH---O hydrogen bond is the most prevalent, due to the high number of NH and C=O groups associated with the peptide bonds that make up the protein backbone. Protein chains are made from multiple amino acids that have undergone a dehydration reaction, the reaction mechanism for the formation of these polypeptides is shown in Figure 1.2.2. During the formation of the polypeptide chains, the OH group along with one hydrogen from an NH_2 group are removed in the form of water leaving only one OH group at the end of the chain. Consequently, there are relatively few OH donors available in proteins, with these groups

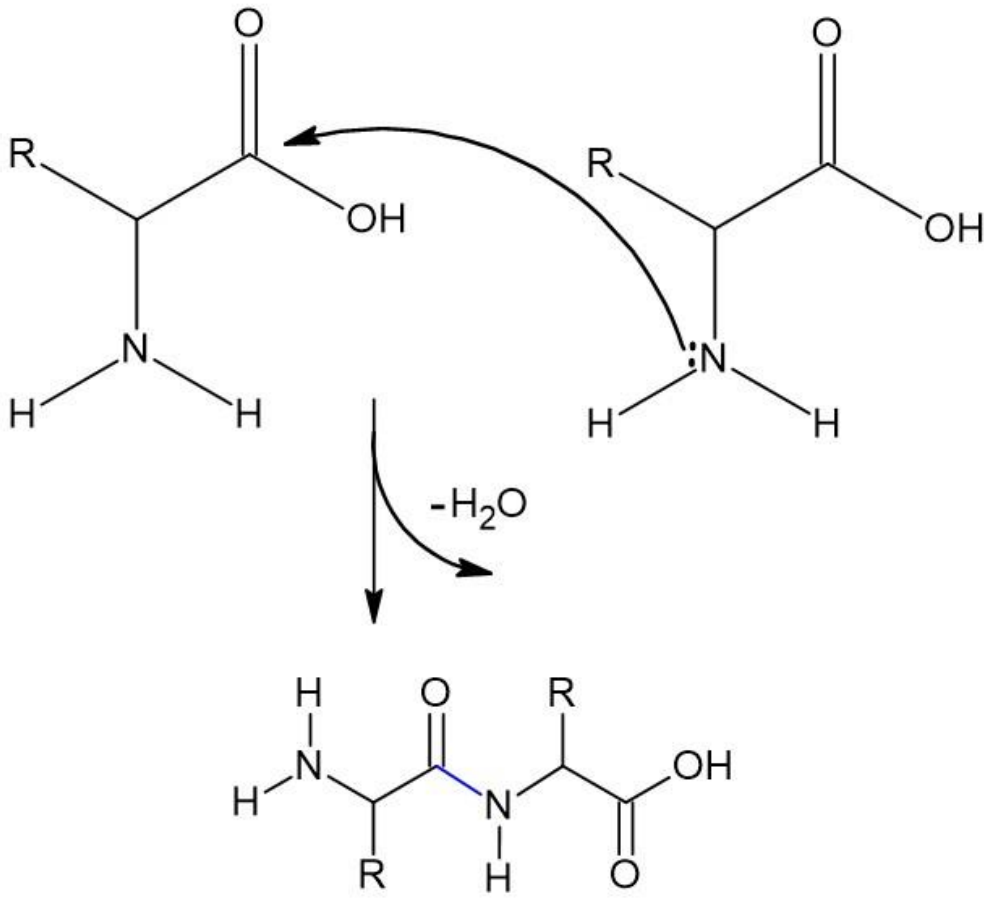


Figure 1.2.2: Mechanism for the dehydration of two alanine molecules to form the start of a polypeptide chain

mainly being present on the sidechains of a limited number of polar and acidic amino acids.

This is the reason that there are so few O-H --- O (2 in Figure 1.2.1) interactions in proteins. As such, we restricted our investigation of hydrogen bonds in this thesis to the NH---O type.

1.2.1 Hydrogen Bonding in Protein Secondary Structures

There are many different types of hydrogen bonding in proteins, which play different roles, including maintaining the structure within the protein and creating reactive sites within the protein where docking of substrates and other molecules can occur. The main secondary structures that arise from hydrogen bonding in proteins are listed below. While there are many more types of secondary structures in proteins, these ones represent a large portion of the secondary structures that have been observed.

- Alpha helix
- Beta sheet
- Coil (multiple helices)
- Turn (connection of Alpha helix or beta sheet)

These main groups can also exhibit some variations that offer a slightly different environment for the hydrogen bond. The 3-10 and Pi helix are both variations of the alpha helix. The alpha helix exists with an n to n+4 residue hydrogen bond. The 3-10 and Pi helices have an n to n+3(10 atoms between H-bond) and an n to n+5 hydrogen bond, respectively. These variations of the alpha helix were shown in Figure 1.2.3. The 3-10 and Pi helix are uncommon in protein structures due to being less favourable energetically. The 3-10 helix is seen more frequently towards the end of a helix chain (11), this is most likely due to the helix unravelling (as going from left to right in Figure 1.2.3) to give more compression of the backbone atoms. The Pi helix is a rarer case, however there are examples and certain protein structures do lend themselves to include this compact helix. It has been shown that previous

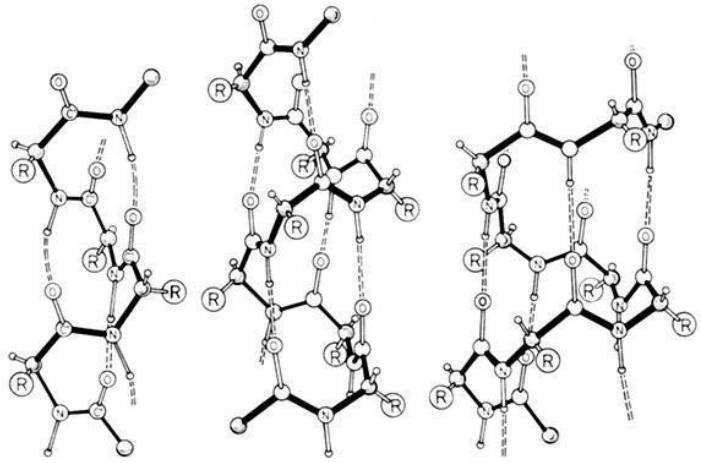


Figure 1.2.3: Different variations of alpha helix: 3-10(left), Alpha helix(middle), Pi helix(right)

algorithms used to search for these structures are not that sophisticated, meaning many of the variant structures are unable to be automatically detected. (12). These different structures of helix show a very important characteristic which is variation. The alpha helix appears in many optimized protein structures, however if this helix went through different states of folding to form either a 3-10 or Pi helix throughout the simulation, this could be an important phenomenon to track.

Another major secondary structure in many proteins is the beta sheet. These sheets can be arranged in three different ways to each other, parallel, mixed, and anti-parallel. They hydrogen bond with each other across each sheet, unlike the helices, which hydrogen bond within the same backbone chain. This variation of the beta sheet secondary structure is shown in Figure 1.2.4.

Bridging of hydrogen bonds can occur between any of the secondary structures, which makes for many hydrogen bonds that reside in many different environments (13). The energies of these bonds may differ significantly and show areas of proteins that are not held together as strongly as others. This variation in structure yet again shows how these structures could change throughout a simulation, especially if there is an overall folding or unfolding motion of the protein. Finding the trends and changes in these secondary structures could be very

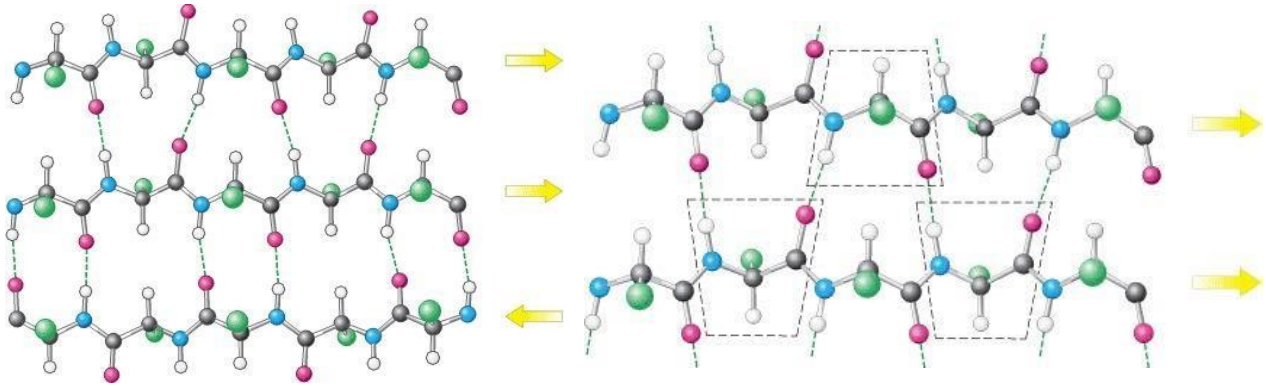


Figure 1.2.4: Beta sheet Bonding; Mixed/antiparallel (Left) Parallel (Right)

useful. While the bonds looked at in this paper have had the secondary structures they connect noted, a more detailed analysis with a larger data set would be needed in order to derive statistically significant differences between different secondary structure types.

1.3 Theory for Calculating Hydrogen Bond Strength

For simple dimers, the hydrogen bond strength can be calculated by optimizing the dimer and monomer structures and comparing the relative energies. This minimizes the energy in the structure so that the energy in the dimer can be compared to the energy of two of the monomers. This isolates the intermolecular bond that spans the two monomers which make up the dimer and thus the energy can be trivially calculated. In equation 1, we let the total energy of the hydrogen bonding interaction be E_{tot} and the energy of each monomer be E_{m1} and E_{m2} and the total energy of the dimer be E_{dim}.

$$E_{tot} = E_{dim} - (E_{m1} + E_{m2}) \qquad (1)$$

There are many different electronic structure methods that could be used for optimizing the geometry of the dimer/monomer, the most common are density functional theory (DFT) methods, such as B3LYP or ω-B97X-D.

For intramolecular hydrogen bonds or systems that cannot be easily separated into distinct fragments where the hydrogen bond is present or absent, it is much more difficult (if not impossible) to determine the hydrogen bond strength. Consequently, in many cases, the distance between the atoms involved in the hydrogen bond is used as a proxy measure of its strength. However, geometry alone is an imperfect estimate of hydrogen bond strength, hence there are ongoing efforts to understand these interactions using other approaches. (14)

One of these approaches involves investigating non-covalent interactions (including hydrogen bonds) by analysing the reduced electron density gradient (RDG), which is defined as (s)

$$s(\boldsymbol{r}) = \frac{1}{2(3\pi^2)^{\frac{1}{3}}} \cdot \frac{|\nabla\rho(\boldsymbol{r})|}{\rho(\boldsymbol{r})^{\frac{4}{3}}} \qquad (2)$$

Equation 2 is used to create an iso-surface in the region between the atoms involved in the hydrogen bond, where the interaction is defined by an iso-volume that can be visualised. An example of such iso-surface can be seen in Figure 1.3.1 (15). Iso-surfaces from our own software, Hybond, are similar and are shown later in this thesis.

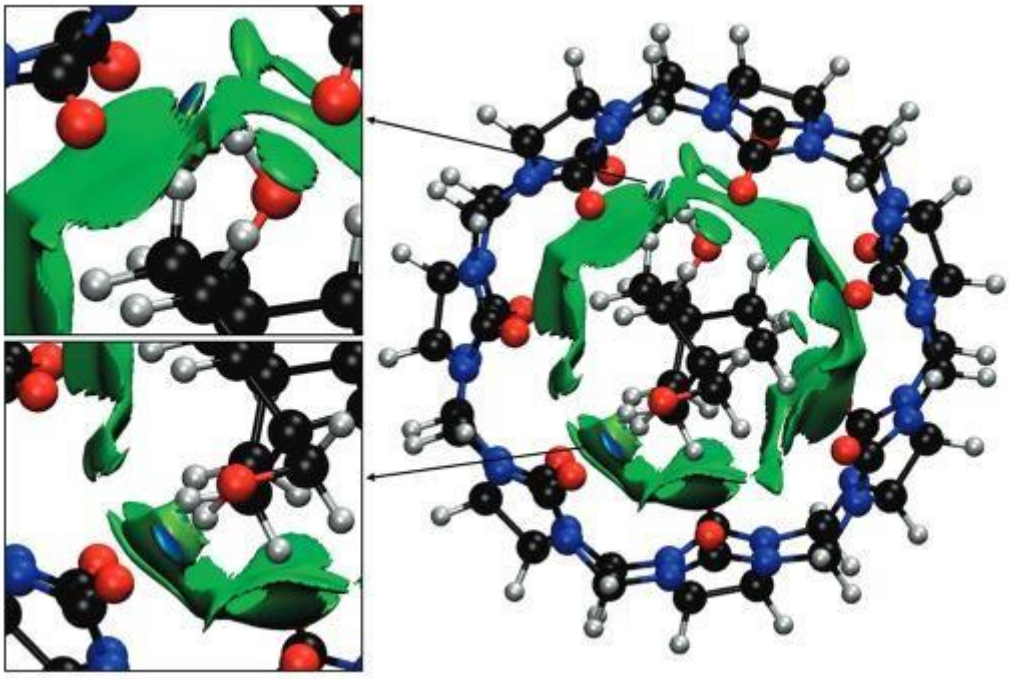


Figure 1.3.1: NCI isosurfaces in Cucurbit[7]uril-bicyclo[2,2,2]octane with a cut-off value of 0.5 a.u., Sourced from NCIPLOT journal article (15)

Within these iso-surfaces, we can investigate various properties of the electron density, such as the kinetic and potential energy density. The kinetic energy density has been used to show properties of hydrogen bonding accurately (16) (17). The kinetic energy density is used extensively in this thesis as this has been used as one of the more valuable parameters in the past. There are other parameters that the kinetic energy density can be compared with such as the Electron Localisation Function (ELF). ELF shows the measure of finding an electron in the nearby area of the interaction. Using ELF to determine what is happening (18) to the electrons in a hydrogen bond is not a novel idea but its use is uncommonly seen. Using these properties will hopefully highlight trends that are seen over the course of the simulation.

1.4 Existing Software for MD Simulations

Molecular Dynamics (MD) simulations have been used for many years to describe molecules from all the disciplines of chemistry. Of particular interest is the use of MD to simulate proteins and other large biological macromolecules. These include many non-covalent interactions, which contribute to the overall 3D structure and are involved with interacting with other molecules to form large, complex systems. These systems are of great importance as they are targets for big industries such as pharmaceuticals. First starting in the 1970's, MD simulations have gained much popularity in biochemistry to look at different properties of both reactions and structures. Modern day MD simulations are typically carried out on supercomputers that have access to a much more powerful set of processing hardware. There are many popular software packages that can set up and run simulations along with tools to describe the data that these simulations output.

Each of these software packages have their own strengths and weaknesses when running MD simulations. Each simulation software uses a set of forcefields to describe the atoms that are present within the molecules, and these are generally decided on by the software package being used. A common MD simulation package is the NAMD software. (19) NAMD is a

widely used general purpose molecular dynamics package that is commonly used for simulating proteins. The software package easily allows parallelisation and uses the common CHARMM or AMBER forcefields to describe the atoms and interactions that are occurring in the simulations. There are other interchangeable software packages available to do similar MD simulations, such as GROMACS (20)/GROMOS (21), CHARMM (22) and AMBER (23) .

There are existing tools (in the form of programs) for analysing MD simulations such as PyContact (24), RIPMD (25) and MDAnalysis (26) (27) but they do not investigate hydrogen bonding in much depth. They allow analysis of basic properties of hydrogen bonds such as the distance and angle of the bond, along with an overall count of the number of HBs in the protein. However, they only scratch the surface of the information that can be gathered about a hydrogen bond over the time of an MD simulation. The main focus for most of these programs is looking at the total number of NCIs in a molecule as a whole. This cannot give a detailed explanation of what is happening at a specific region of a protein or even closer, a single hydrogen bond. This need for programs that investigate specific regions is growing as a better understanding of hydrogen bonds is attained.

Recently, non-covalent interactions have been described by analysing the reduced electron density gradient. Programs like NCIPLOT do this already but are not optimised for identifying individual interactions and struggle to characterise all of the interactions in large protein systems due to the extreme computational resources needed. Our research group has been working on the answer to this problem. An in-house software package called Bonder (28) was developed to solve this problem, with the ability to find localised areas of interactions quickly, using a flood fill algorithm to construct the NCI iso-volume. This makes it trivial to separately analyse all individual interactions in a molecule at massively reduced computational cost. In this thesis, the functionality of Bonder was extended so that an

individual interaction could be analysed for each frame of a MD simulation trajectory to give a temporal lens on its strength.

1.5 Objectives

This thesis sets out to evaluate the strength of hydrogen bonds in some prototypical protein systems through a temporal lens using MD simulations. Our hypothesis is that the strength of a given hydrogen bond varies appreciably as a function of time due to environmental and structural changes within the protein. Furthermore, as MD simulations are stochastic in nature, we expect that different simulation trajectories should exhibit different hydrogen bond strength behaviour, although with sufficient sampling these should lead to the same timeaveraged values.

There are two underlying premises for this research. Firstly, that different hydrogen bonds exhibit different strengths, even when the same atom types are involved, due to the orientation and surrounding environment. Secondly, that a simple geometric analysis of the atoms involved in a hydrogen bond is insufficient to determine its strength alone.

This research should highlight the importance of analysing hydrogen bonds in a dynamic environment (as opposed to a static view of bonding). A workflow is developed so that similar research can be undertaken to suit any reaction or interaction involving non-covalent interactions, across the various disciplines of chemistry. Finally, the development of Hybond/Bonder will also be extended with this research, allowing similar and more extensive research to be conducted investigating the temporal lens of non-covalent interactions.

We initially planned on investigating the temporal nature of hydrogen bonds using a series of different proteins that each exhibit a range of secondary structures. However, after we completed some benchmarking simulations using the APOA1 protein, it became clear that this approach was not ideal due to the high computational cost and the added complexity of

needing to deconvolute potentially competing trends. Instead, after discussion with Prof Vic Arcus (29), we decided to run simulations on some common protein building blocks obtained from the CATH database. These protein building blocks were large enough to give representative secondary structure environments for hydrogen bonding but at much reduced computational cost.

The three protein building blocks chosen were the Rossmann Fold, which is made up of a mixture of beta sheets and alpha helices, the Jelly Roll, which is made up of mostly of beta sheets with one main alpha helix, and the Alpha-Alpha Barrel, which is made up of many alpha helices with a very small beta sheet off to one side. These three protein building blocks consist of less than 20,000 atoms (including the corresponding water box), which makes the computational cost more manageable. The structures can be seen in Figures 1.5.1, 1.5.2 and 1.5.3.

Our first protein building block is the Rossmann Fold. It is a very common structure in many proteins making up 20% of the known structures in the protein database. The Rossmann Fold is also one of the five main structural motifs that occur in proteins (30). This section of protein has a mixture of alpha helices and beta sheets which interchange with each link. Since the Rossmann fold is a well-documented and captured part of proteins it means there can be comparisons made to this data and follow-on research conducted. This along with the varying hydrogen environments going from alpha helix to beta sheet and back to alpha helix and so

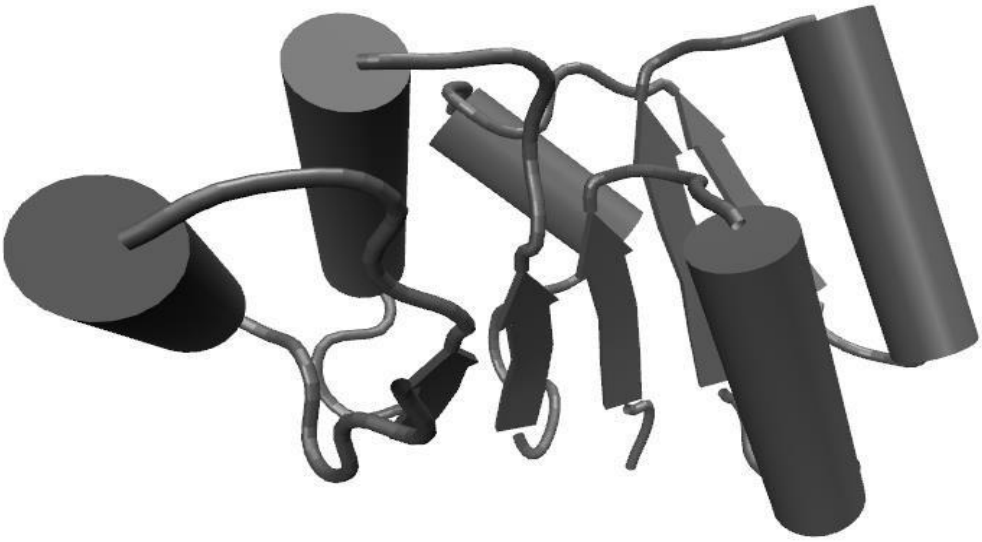


Figure 1.5.1: Structure of Rossmann Fold

forth, gave a good starting place for analysis and a multitude of different hydrogen bonds to select for in depth analysis. The structure of the Rossmann fold is shown in Figure 1.5.1.

The second protein building block is the alpha-alpha barrel. This portion of proteins is made up almost entirely of alpha helices. This formation of secondary structures gives many environments for hydrogen bonds in the chains as well as bridging between different helices. This structure is shown in Figure 1.5.2. The overall construction of the protein segment is an array of alpha helices that are arranged from top to bottom and form in a ring shape giving a barrel like structure. There are amino acid chains that are classified as turns between each helix. These turns also have hydrogen bonding present in them to stabilise the bend formation of the amino acids in the turn. While the alpha-alpha barrel is less common than that of others such as the TIM Barrel or alpha-beta barrels it is still a good example of a barrel formation that occurs in many proteins. This along with only having alpha helix secondary structures

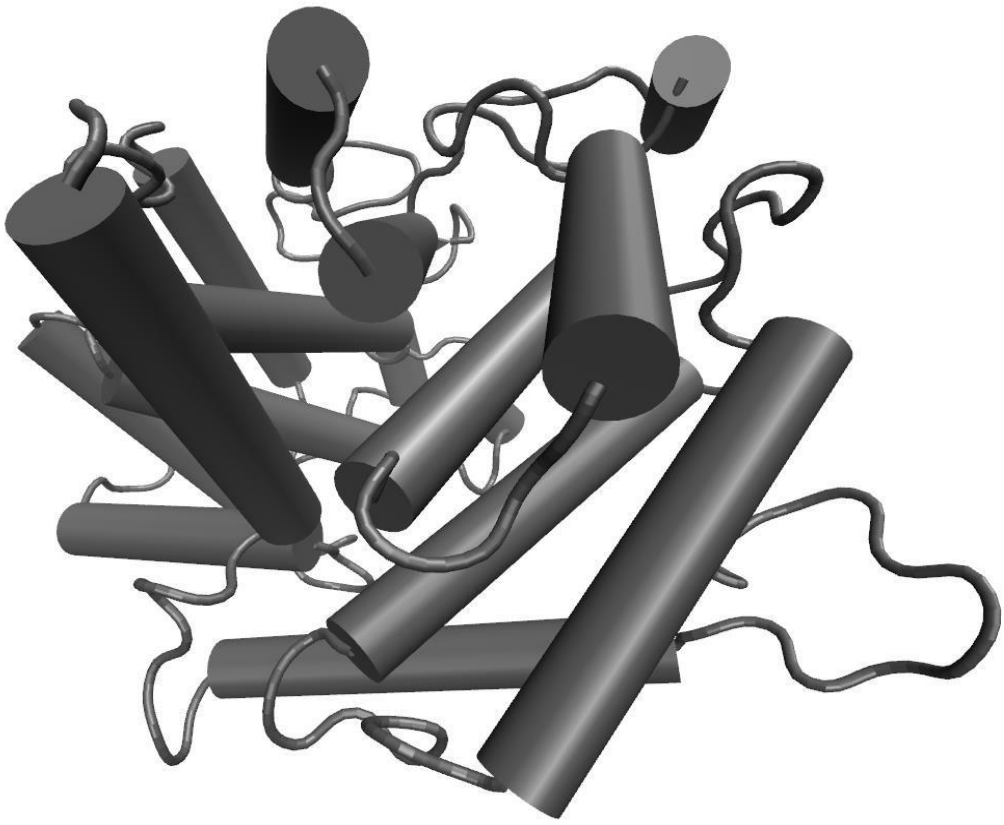


Figure 1.5.2:Structure of Alpha-alpha Barrel

allows an analysis of whether there is some energy difference between the hydrogen bonds in this environment compared to the other two protein environments.

The third, and final, protein building block is the Jelly Roll shown in Figure 1.5.3. This partial structure is prevalent in viruses and is the reason for many of the structures that have been discovered (31). Jelly Rolls are primarily made up of eight beta sheets. These beta sheets are arranged in two groups of four to make up a larger sheet. There are a few alpha helices in the protein which help to link the sheets and the Jelly Roll to other parts of larger proteins. The beta sheets in the centre will hold enough environments for hydrogen bonding to occur and allow analysis of a purely beta aligned interaction. This allows the third type of environment to be studied which is beta sheet rich. There is not expected to be much difference in energy between these three environments and there are many other factors that need to be considered when comparing the results.

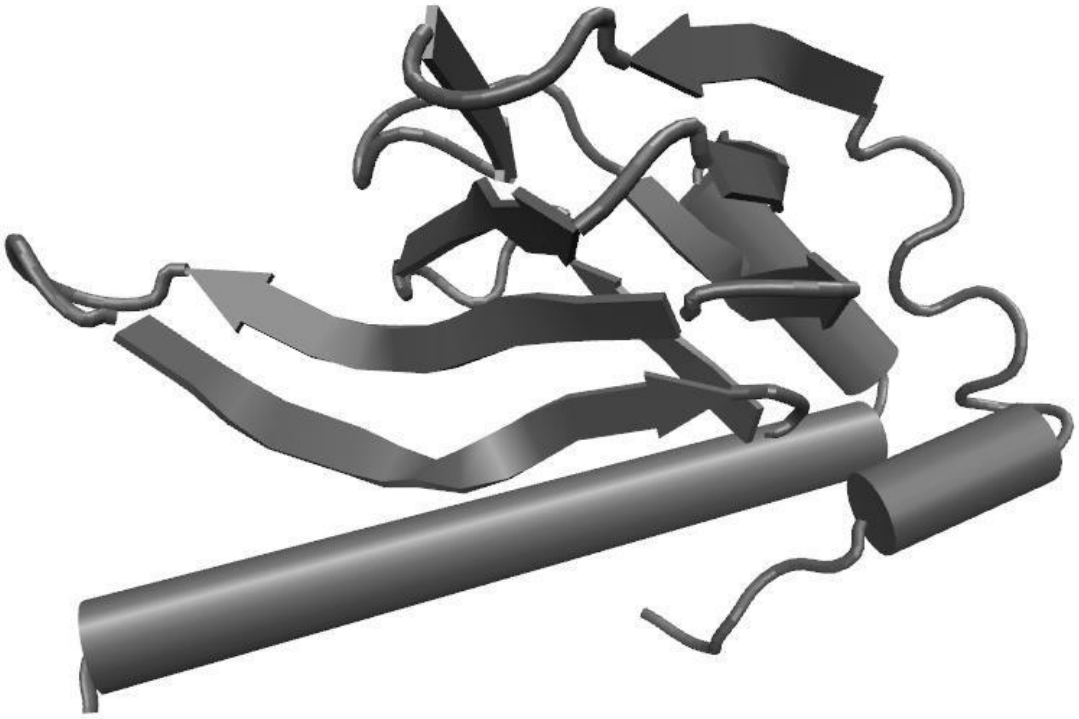


Figure 1.5.3: Structure of Jelly Roll

From the selected structures (CATH domains: 1rwhA01, 1ft9B01 and 5nurE00) each NH---O hydrogen bond is shown in blue and each OH---O hydrogen bond shown in red (Figures 1.5.4, 1.5.5 and 1.5.6). This shows that there were many environments per protein building block that the analysis could be conducted on.

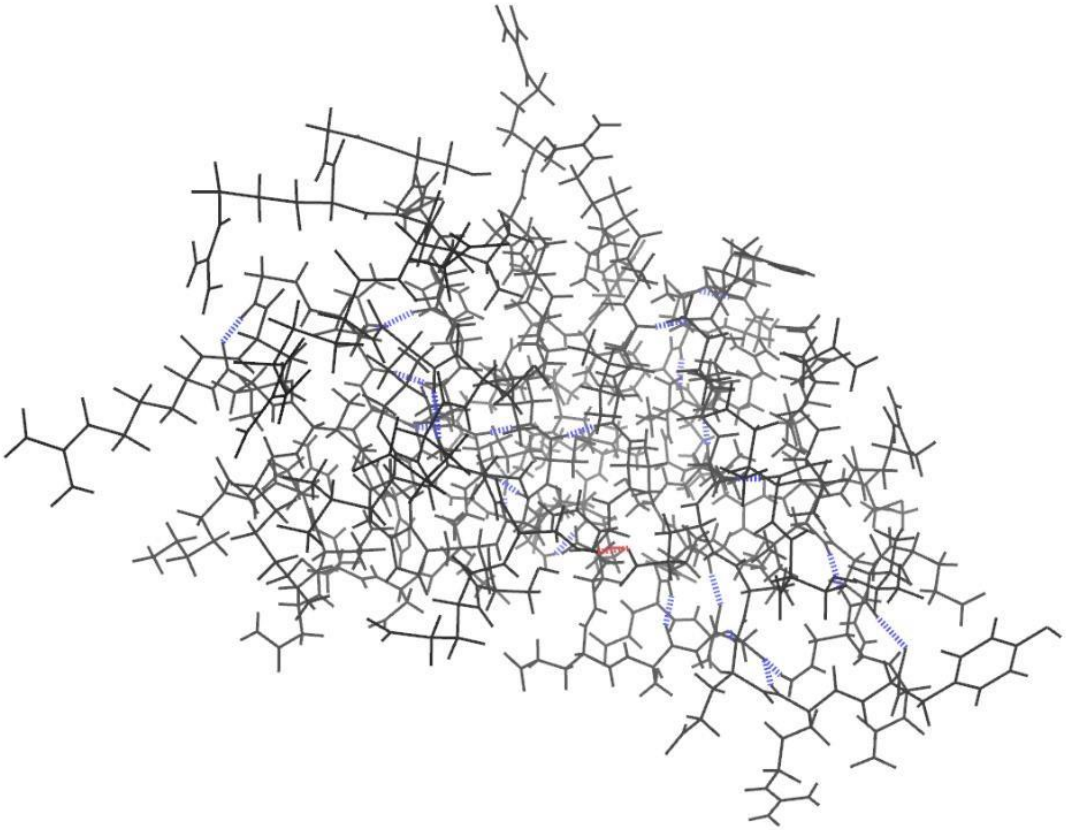


Figure 1.5.4: Rossmann Fold Hydrogen Bonding Environments

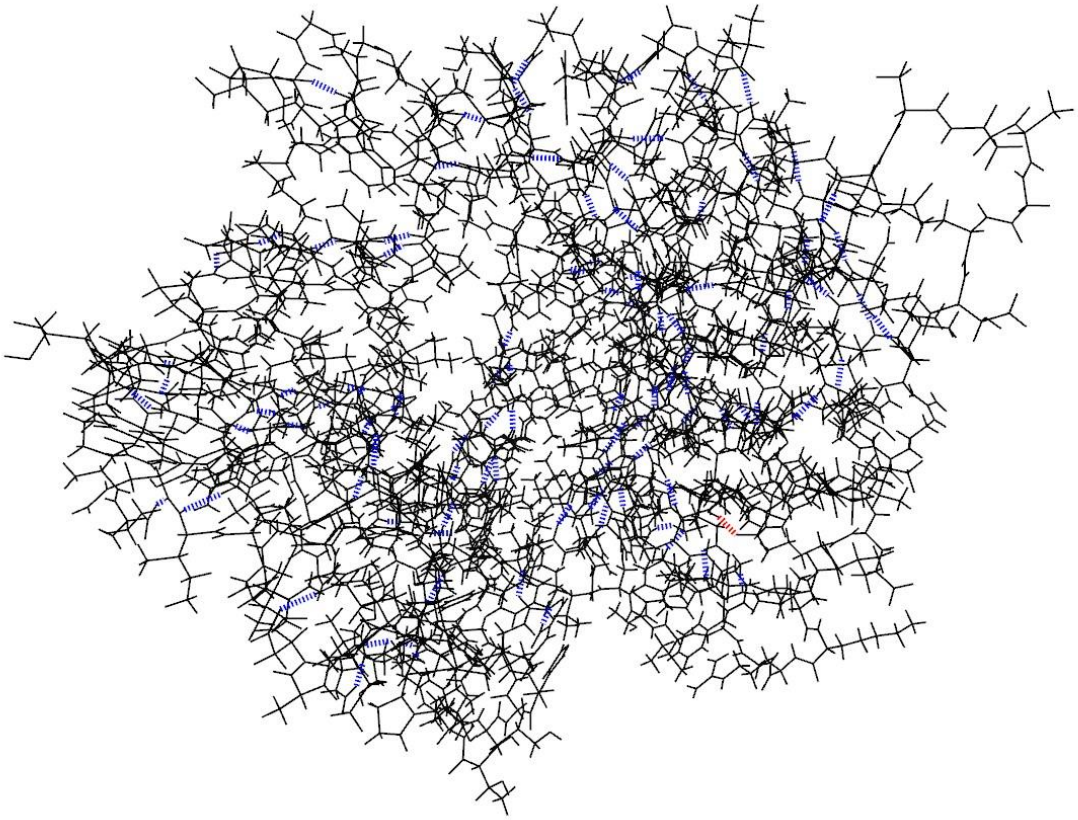


Figure 1.5.5: Alpha-Alpha Barrel Hydrogen Bonding Environments

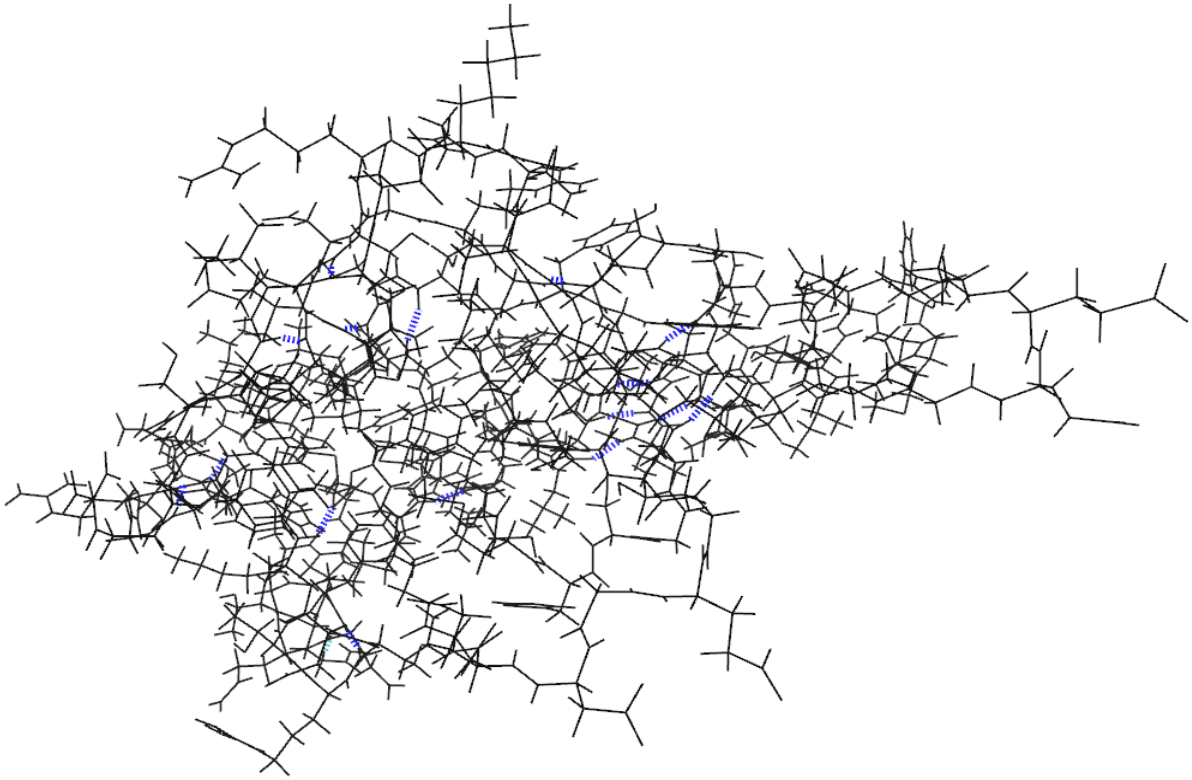


Figure 1.5.6: Jelly Roll Hydrogen Bonding Environments

CHAPTER 2

2.1 OVERALL WORKFLOW

To ensure that this research can be easily applied and extended to other systems, a modular workflow approach was taken, consisting of three main steps:

1. Prepare and run a molecular dynamics simulation for a given protein. We took our initial PDB structures from the CATH database and used NAMD for the simulations but any PDB file and variety of MD packages could be used.
2. Apply Non-Covalent Interactions (NCI) theory to each frame of the MD simulation. As no current software has this feature set, a new programme, Hybond, was written that automates the use of Bonder, which is an existing package that uses NCI theory to analyse static structures.
3. Analyse the NCI results for each frame of an MD simulation using some custom Java and R scripts.

A detailed breakdown of each step in the workflow is shown in Figure 2.1.1.

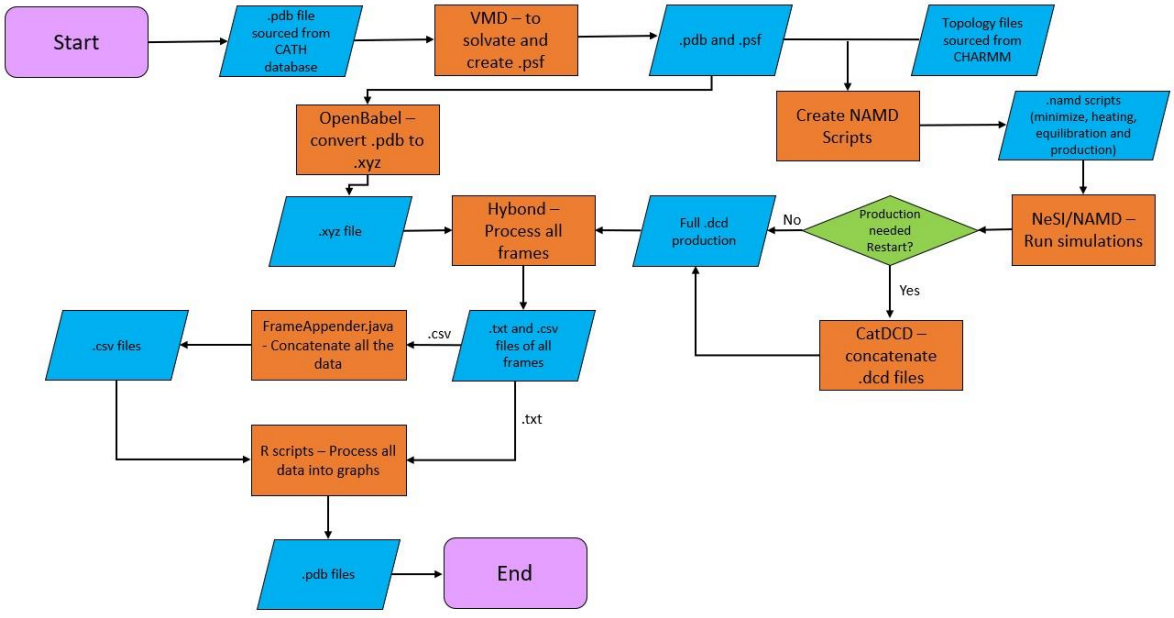


Figure 2.1.1: Overall Workflow, Blue Parallelograms are Inputs, Orange Rectangles are Processes and Green Diamonds are Decisions

2.1.1 Protein Preparation and MD simulations

The .pdb files can be obtained from the CATH database. The protein can be solvated at this stage. Although there are multiple ways of doing this VMD was used to solvate the proteins in a water box for this research. Ions are added at this stage to balance out the charge within the system. The atoms involved in a hydrogen bond can be selected within VMD to be studied and Hybond will use the index (atom positions) of these atoms. It must be noted that after the analysis has occurred the iso-surface can also be viewed in VMD. An example of this is shown in Appendix 1. This feature was mainly used as a double check for any unwanted interactions. After solvation the .pdb and .psf file are ready, OpenBabel can be used to acquire the .xyz file of the subject protein. The NAMD simulation files must be updated with the cell origin and the cell basis vectors. This was done using the maxmin.tcl script. Once this is done and the topology files have been loaded the minimization simulation can be run. Minimization removes energy from the system and lets the molecule rest in a low energy state. The files that are output from the minimization step can be used as the input for the heating step. This puts energy back into the system by heating the molecule in steps to 300K. This gives the molecule the correct energy to avoid any complications with energy inside the system. The output from the heating step is used for the equilibrium step. This step lets the molecule move around and find an average energy that is close to how the protein would be in nature. This prevents any unwanted twists or folds in the molecule that could force interactions to be of higher or lower energy than expected. Once equilibration is completed the production runs can be started. These production runs are sampled, and the trajectory is recorded in the format of a .dcd file. If the simulation needed to be restarted this would produce another .dcd file starting from where the other trajectory stopped. These were concatenated together using the CatDCD software (32).

2.2 Simulation Parameters

The main parameters that needed to be considered for the simulation were the timespan and how often the simulation would be sampled for data. These, along with the temperature and the pressure of the system were the front runners for parameter consideration. Most of the other settings were kept at the standard or the default for the NAMD software. The choice of ensemble for the system was also taken into consideration.

Simulations conducted were run for a total time of 500 ns. This length is now far from some of the state of the art simulation times of multiple milliseconds that started appearing the in the late 2000's (33) on frameworks specifically designed for this like Anton, and took off in following years (34) (35) (36). However, 500 ns is a good middle range for a simulation of this size due to the frequency of structural changes within the molecule. Hydrogen bonds have been found to vibrate on the time scale of picoseconds (37). The structural folding and bending of proteins have been expected to be on the time scale of microseconds to milliseconds (38). Running multi millisecond simulations was not feasible with the amount of computing power that was available during this project. The simulations that have been produced will show the variation in the vibration of the hydrogen while also allowing a smaller snapshot of the large movement of the protein overall. Longer simulations with a finer resolution for the step size would be ideal. However, with the computing power and storage accessibility, 500 ns simulations were more than reasonable. The simulations were run using NVT ensembles, with the required constants being the amount of substance (N), volume (V) and Temperature (T). NVT is required to control temperature and volume during the heating step, allowing the system to slowly heat to 300K. It was then decided that all the steps should be run using the NVT ensemble to avoid complications.

2.3 Hybond NCI Theory

Bonder (28) is an in house software package that was developed to analyse non covalent interactions in a molecule. Bonder draws a line between two chosen atoms and checks for an interaction. If this is successful Bonder then activates the flood fill algorithm (39) and constructs an iso-surface (using supplied cut-off values) to describe the interaction found. This iso-surface then has specific properties analysed for each cut-off value (if 0.3 was selected Bonder analyses 0.1,0.2 and 0.3 for the interaction) and output to files that correspond to the negative, positive and total portions[1] of the interaction. The negative and positive portions of the interaction correspond to the attractive and repulsive forces within the bond respectively.

Hybond has been written to be agnostic to the software used for the underlying MD simulation software, provided a .dcd output can be obtained (either directly or converted from some other trajectory file format). Hybond takes the binary .dcd output and turns it into separate frames (.xyz files) so that they can be input into Bonder. The only valid inputs for standard Bonder operation are .wfn (wavefunction) files (where the electron density has been pre-computed) and .xyz files (where the electron density will be calculated in Bonder using a promolecular approach based solely on atom positions). Hybond also takes indices of atoms for calculating proximity, distance, and angle. Hybond then acts as a wrapper for Bonder setting up the inputs that are normally needed for running Bonder. This allows a smooth analysis to occur for a complete simulation which extends the functionality of NCI analysis which up until this point, was limited to static structures. Hybond also calculates the distance and angle of the interaction using Euclidean Distance (40) mathematics and the arccos (41) rule respectively. It also uses a quick search of the surrounding atoms and records if they are

[1] The total output file is only made when both the positive and negative output files exist. Bonder given the right situation might only output the negative portion of the interaction as no positive portion was detected.

within a 5 Å range of the atoms that are involved in the interaction. Hybond has retained the standard output from the Bonder program and this results in many files being produced, roughly five times the amount of frames. This information is then processed using other programs and scripts to clean and graph the data.

2.4 Processing of NCI Results

The java program FrameAppender takes the output from each Bonder call that Hybond makes. For this analysis only the negative portion of the files was taken since Bonder was consistently outputting the negative portion of the interaction. This is done by setting a flag in the program arguments. The java program takes these files and separates the RDG cut-off values and the data associated with them from each file. For the analysis conducted in this thesis it produces separate files with all the information for the 0.1, 0.2 and 0.3 cut-off values as 0.3 was selected as the cut-off for each Bonder call that Hybond made. These are again output as .csv files so they can be easily modified and loaded into the R script in Rstudio.

The processed file along with the interaction distance and angle files are loaded into the R script and put in the same data frame. A type is added so that the R script can differentiate between different simulations or interactions. The last data column that is added to the data frame is the index/timestamp of each frame. Larger data frames are now constructed that either consist of the same interaction across all three simulations or all the interactions of one simulation. These are the data frames that are then graphed. Any averaging that occurred while making the graphs was a simple moving average and is explained as part of the results in Chapter 3.3 where an analysis of different averaging methods was conducted. Graphs were constructed that looked at a single interaction and others were made to compare two or more interactions. Later in the thesis, graphs were made where two well characterised interactions were taken, and the data cleaned so that the large majority of points lay within the trend.

Lines were then fitted to these trends to see what relationships there were in the data. This is further explained in Chapter 4.2.

2.5 Initial Testing to get a Representative Sample of Interactions

We applied the workflow outlined in section 2.1 to the three protein building blocks discussed in section 1.6. As each protein contains a very large number of hydrogen bonds, there were several factors that needed to be considered when picking the bonds for analysis. This was to ensure that the chosen interactions would give a viable, overall result that showed how the interactions behave in the protein over a temporal space. The first factor was identifying interactions using VMD that existed at both the start and the end of the simulation i.e. a simple visual inspection of whether the initial and final geometries were likely to accommodate a hydrogen bond. For each protein building block interactions one and two were picked from the start of the simulation whereas interactions three and four were picked from the end. This avoids having interactions that are only present during that specific folding motif of the molecule.

Later once some of the data had been analysed, we realised that one of the selected hydrogen bonds involved an NH_3 donor group. This meant that the hydrogen that was involved in the interaction was one of three equivalent hydrogens that, interchange during the MD simulation. As Hybond was originally written to analyse and track only one of these hydrogen atoms, when they switch places, Hybond was no longer analysing the interaction between the closest hydrogen to the acceptor group. The code for Hybond was modified to take into account if there is more than one hydrogen in a given interaction environment that could be involved in the hydrogen bond. This was done by calculating the distance of each hydrogen to the electronegative acceptor atom and taking the one with the shortest distance as the hydrogen that is participating in the interaction. This was later modified to take into

account environments with 2 oxygen atoms as these can behave in the same manner.

CHAPTER 3

3.1 Data Outcome and Validity

The outcome of the Hybond analysis depends on the environment of the interaction. There are 3 different outcomes that can be observed from the analysis on each frame. These are listed below.

- **Discrete Interaction**

  A discrete interaction is where the iso-surface that was constructed is well defined. This results in little interference from nearby interactions and atoms. Analysis continues as normal on these interactions and a valid set of data is obtained in the form of .csv files. As explained in chapter 2.3, the files that are output can be of positive and/or negative portions of the interaction. This produces data that can then be graphed to compare the strength of the interaction.

- **Non-Discrete Interaction**
- A non-discrete interaction is where the iso-surface of an interaction has grown too large, exceeding the threshold volume of 5 $Å^3$. This occurs when Hybond attempts to capture the target interaction but includes nearby interactions. The influence of nearby interactions results in an unrepresentative value of the positive and negative portions (and therefore total energy) of the target interaction. N/A is inserted into the files, instead of the influenced values from the analysis. This was done for ease of producing graphs since R has built-in functions handle an N/A parameter. It must be

noted that N/A does not describe the real nature of this interaction and is handled differently in Chapter 4.2. The energy of the target interaction is undefinable under these circumstances.

- **No Interaction**

  A no interaction outcome from the analysis of the frame is straight forward. This occurs when Bonder tests for interactions, along the line that is drawn between the atoms involved, and is unsuccessful resulting in no output files. Bonder then moves onto the next frame. In this case the post processing of the data fills in these points as 0's.

3.2 RDG Testing to Determine Cut-Off Value

3.2.1 Initial Testing of RDG Cut-Off

To determine the most appropriate RDG cut-off value to be used, one of the hydrogen bonds in the Rossmann Fold building block protein was systematically investigated. There are three main factors that were used to determine whether a certain cut-off value is suitable or not. These factors are the number of frames that are terminated due to large volumes, the runtime of the program, and the number of points that are analysed inside the iso-surface of the interaction. The cut-off value greatly affects these three parameters as with a larger volume to define the interaction there is more chance of other interactions being caught in the same volume, which results in a terminated frame. Having other interactions inside of the isosurface means that all the integrated properties will not be representative of just the one hydrogen bond interaction that we are primarily interested in. A larger cut-off also will affect the number of points that are being analysed in the iso-surface, which effects the numerical accuracy of the integrated properties of the interaction. With the increase in number of points there is also an increase in the runtime, as more analysis needs to be conducted by Bonder.

The results from this preliminary test are shown in Table 3.2.1 and an overview of the runtimes and points per frame are shown in Figures 3.2.1 and 3.2.2.

Table 3.2.1: All RDG Cut-off Values Tested on One Interaction

RDG Cut-off Value	Terminated Frames	Percentage of Valid Frames	Total Time (Minutes)	Average Number of Points[2]
0.1	29	97.1%	14.1	39
0.2	8	99.2%	29.6	350
0.3	9	99.1%	94.1	1760
0.4	62	93.8%	288.4	5806
0.5	562	43.8%	513.3	10287

Over the range of the cut-off values there are multiple trends. The first observed is the number of terminated frames that were output from the Hybond analysis. Apart from the higher starting number at 0.1 the general trend is that as the cut-off value increases, so does the number of terminated frames. This is likely due to the interaction being missed by the small threshold applied to the iso-surface. The next trend observed is that there is an increase in runtime as the cut-off value increases. This trend is also reflected in the average number of points that are present in the iso-surface. It may seem that Bonder is getting more efficient from the increasing RDG, however this is not the case as it is known that runtime is linear with the number of points present in the iso-surface. This increased "efficiency" is due to the frames that have 0 points not being counted in the average, as this would give a skewed number compared to the amount of points that are in each frame. Bonder spends very little

[2] Per frame with terminated frames excluded

time to determine if a frame is to be dropped. So as the terminated frames increase it falsely appears that the efficiency increases. In Appendix 2 the runtime is plotted against number of points, as the number of points becomes less consistent so do the runtimes. From these initial trends, it appears that 0.2 - 0.3 is the ideal cut-off value for these proteins where few frames are dropped and the runtime is very short in comparison to the larger cut-off values. The cutoff value of 0.2 is suitable to use as there is low runtimes and a decent number of points in the iso-surface to provide numerical stability. However, if the number of points dropped much lower a higher cut-off such as 0.3 would need to be used to provide the stability and ensure that the integrated properties are properly converged. If very accurate results are desired the cut-off could be increased further to 0.4 however there is a trade off with starting to drop more and more frames, and many more runs may need to be completed to acquire the same number of valid frames.

In Figure 3.2.1 the runtimes of each cut-off value are shown. It is clear that as the cut-off increases the runtime increases and also the consistency of the runtime deteriorates. This trend is also seen in Figure 3.2.2 which shows the number of points in the iso-surface for each frame of the simulation. The red and green bands being 0.1 and 0.2 respectively show very consistent results with little deviation. There are small deviations when the cut-off is increased to 0.3 this can be seen where there are blue points that sit above the band. When the cut-off is increased further to 0.4 and 0.5 the band disappears, and the deviations become the majority of the points. This deterioration shows where some iso-surfaces are much larger and have more points in them, ultimately resulting in dropped frames when these become too large. From these results it shows that the higher cut-off values are less consistent and the threshold is around a cut-off value of 0.3.

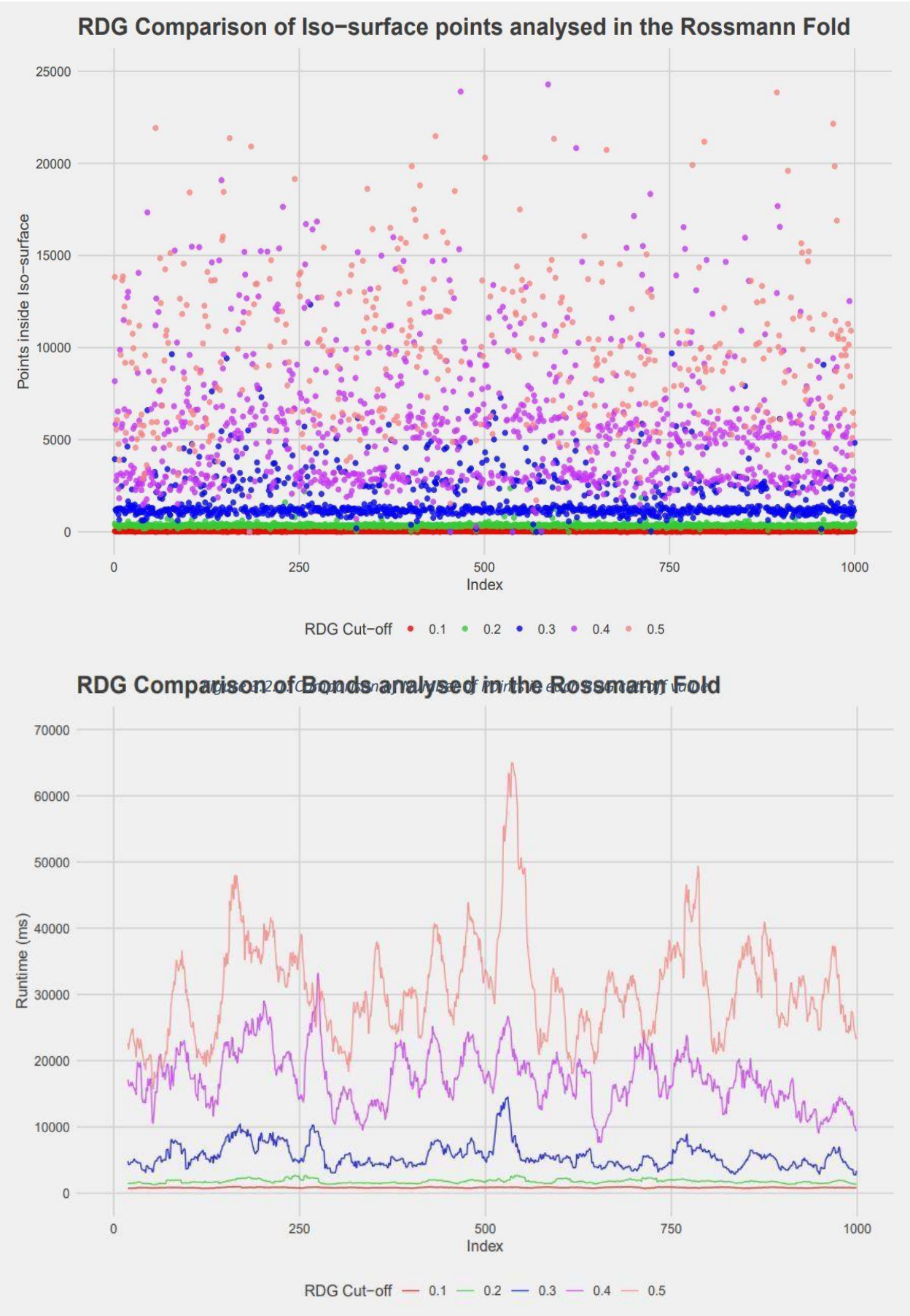


Figure 3.2.2: Comparison of Runtimes using Differing RDG Cut-off Values

3.2.2 Further Testing of RDG Cut-off Values

RDG values of 0.5 and 0.3 were selected for further testing as these have been used frequently to create iso-surfaces in previous works of smaller molecules (42) (43). These values were used on all four interactions in the Rossmann Fold building block, the results are shown in Table 3.2.2. It is clear from these tests that using a 0.5 cut-off to create the isosurface to describe the interaction was too large. This is most likely due to the protein structures being compact and within close proximity of each other. Along with many forces and interactions occurring in the molecule it is not unlikely that another electronegative atom could interfere with the hydrogen bonding interaction being analysed. Proteins are also dynamic in nature, and it is expected that the interaction environment would be changing over the period of the simulation. Comparing the data provided in Table 3.2.2, it is clear that all interactions saw an increase of valid frames when running at 0.3. While Bond 1 was characterised well with a 0.5 cut-off, we needed a general approach that is expected to work for most cases.

Table 3.2.2: RDG Testing of Rossmann Fold Interactions

RDG Value	Interaction	Terminated Frames	Percentage of Valid Frames
0.5	Bond 1	14	98.6%
0.5	Bond 2	562	43.8%
0.5	Bond 3	471	52.9%
0.5	Bond 4	762	23.8%
0.3	Bond 1	0	100%
0.3	Bond 2	9	99.1%
0.3	Bond 3	1	99.9%
0.3	Bond 4	14	98.6%

3.3 Testing of Different Averaging Methods

In general, the data gathered from Hybond for these simulations was noisy and hard to interpret, principally because MD simulations are stochastic in nature. An averaging method was needed to start to compare any potential trends in the data. There are many ways to average this kind of data, but it was decided that a moving average would be suitable as the data sets are tracked over time. Moving averages are commonly used for tracking changes in stock markets and time data series (44). A comparison using five different moving average approaches was conducted. These approaches are outlined below:

- SMA – Simple Moving Average
    - A simple moving average is a rolling mean that has a defined period.
- EMA – Exponential Moving Average
    - Exponential moving averages use an exponentially weighted mean that gives more weight to the most recent observations.
- WMA – Weighted Moving Average
    - Weighted moving averages use a set of pre-determined weights to apply to the average.
- DEMA – Double Exponential Moving Average
    - Double exponential moving average is an extension on EMA and uses and additional volume set to apply to the averaged data.
- ZLEMA – Zero-Lag Exponential Moving Average
    - Zero-lag exponential moving averages use the standard EMA approach but has much less sensitivity overall with a focus on very recent observations.

All five methods were tested on a bond from each protein using the same values for each parameter that the methods required. The area under each average was calculated using Area

Under the Curve (AUC). This parameter was used to determine how similar the methods are and if they do differ, in what ways do they affect the outcome of the results. Table 3.3.1 shows the results of these tests.

Table 3.3.1: Moving Average Tests of an Interaction in each Protein

Method	Rossmann Fold	Alpha-Alpha Barrel	Jelly Roll
SMA	2.694 a.u.	0.926 a.u.	0.709 a.u.
WMA	2.694 a.u.	0.923 a.u.	0.708 a.u.
EMA	2.691 a.u.	0.962 a.u.	0.719 a.u.
DEMA	2.673 a.u.	0.857 a.u.	0.691 a.u.
ZLEMA	2.698 a.u.	0.922 a.u.	0.700 a.u.

From the results there is only one difference which is in the DEMA. This is only a small difference however with the Rossmann fold bond only having 0.77% average difference in the values of the other methods compared to DEMA. The Alpha-Alpha Barrel results were a bit larger with an average difference of 8.1% but fell back down in the Jelly Roll results with a difference of 2.5%. Overall, these differences do not have much impact to the overall conclusion of the data. This can be seen in Figure 3.3.1, where each averaging method overlaps each other except for the ZLEMA (purple) method being more sensitive to spikes in the data. The ZLEMA has a higher sensitivity to changes within the raw data, thus this method didn't show the general trends that the other methods did. Since all the methods were similar in the AUC calculation, and applying Occam's razor, we decided to use the SMA in the rest of this thesis.

Figure 3.3.1: Full Averaging method comparison of a Rossmann Fold Bond (Top) and a zoomed in portion of frames 100-200 of the same Rossmann Fold Bond (Bottom).

3.4 Equivalent Atom Correction

For some hydrogen bonds, multiple effectively equivalent atoms can participate, which will alternate throughout a simulation. For example, if the donor hydrogen is part of an NH_2 group, this can swap between the two hydrogen atoms as the system vibrates. This is particularly problematic for acidic side chains and basic side chains, where under physiological conditions, there are two equal oxygen atoms in a COO^- carboxylate group and three equal hydrogens atoms in an NH_3^+ group. These equivalent atom environments should have been anticipated and were accounted for in Chapter 2.5. After analysing interactions in the Alpha-Alpha Barrel and Jelly Roll proteins there was a distinct pattern observed in the bond angle vs distance relationship relating to these complicated environments. This relationship is shown in Figure 3.4.1 and is highlighted as the blue points and a typical interaction in black. Since the data is spread so much it is hard to tell what is going on. At a

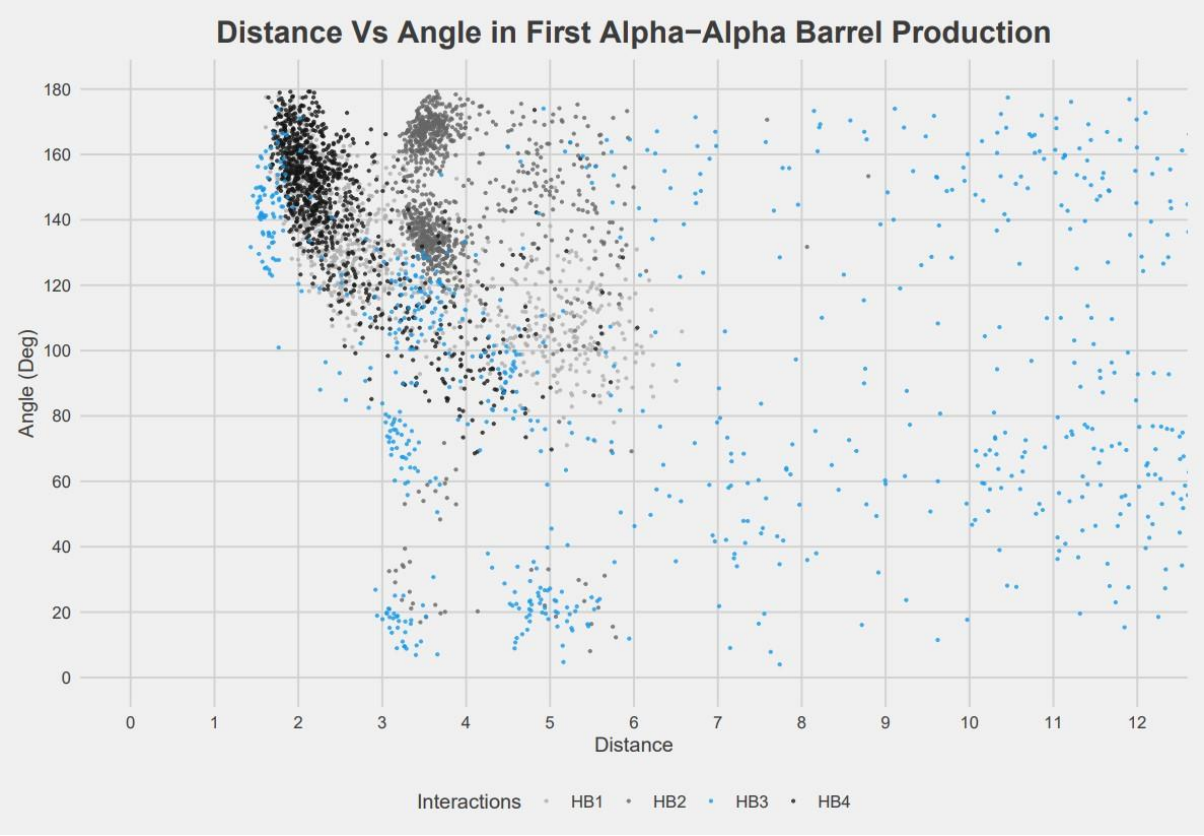


Figure 3.4.1: Distance Vs Angle Relationship Highlighting Anomaly Interaction (Blue)

first glance this looks like a large folding motion of the protein. However, after closer inspection there are multiple clusters of points in this graph, each belonging to a different atom involved in the hydrogen bond. This is due to an NH_3 group present at the terminus of one of the amino acid chains and results in three hydrogen atoms that are in an equivalent environment. Since these hydrogens belong in an equivalent environment to each other they can switch positions relatively freely, as this doesn't change the energy within the neighbouring chains significantly. This is a major problem as Hybond takes an index (position number) from VMD to select the hydrogen involved in the interaction, and if this hydrogen can switch freely the program will be using the wrong hydrogen for characterising the interaction. Therefore, the Hybond output appeared to contain very tight angles at shorter distances because as the hydrogen swaps to a position behind the nitrogen the angle calculation becomes invalid. This situation is shown in Figure 3.4.2, where the correct angle for the interaction is calculated on the left and the incorrect angle is calculated on the right. This obstacle was originally thought to be simple to overcome by just inserting the other hydrogen atom indices with the other parameters and checking for the hydrogen that was the shortest distance to the electronegative atom (Oxygen). While this approach did correct some of these points, any larger structural folding that affects the interaction remained uncharacterised. This correction can be seen to make a major difference when comparing the old distance and angles with the corrected ones (Appendix 3). Comparing the two, there are

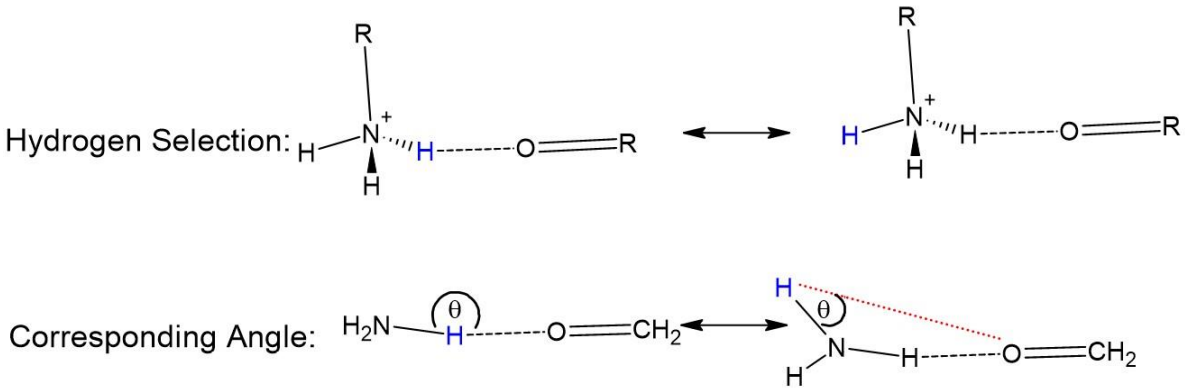


Figure 3.4.2: Angle Calculation Changing Due to Equivalent Hydrogen Environments

now very few points that have an angle of less than 100 degrees and only two clusters can be observed. The difference in the energies is not so impactful, however, with only very slight variations occurring which can be seen in Appendix 4. To understand exactly what is happening in this interaction, the iso-surface of the interaction must be visualised, which is shown in Figure 3.4.3. From observing this iso-surface, it is clear why this bond is troublesome, as both the NH_3^+ donor group and the COO^- group have interchangeable atoms that can form a hydrogen bond. Hence there are 6 possible atom pairs that are effectively equivalent. This issue was fixed by a further modification of the Hybond code to allow for a second oxygen to be defined so that a distance check can be performed across all the potential atom pairs. This ensures that the closest distance between the three hydrogens and the two oxygens present to be selected as the atoms involved in the interaction. The results of this final correction are shown in Figure 3.4.4. This final correction

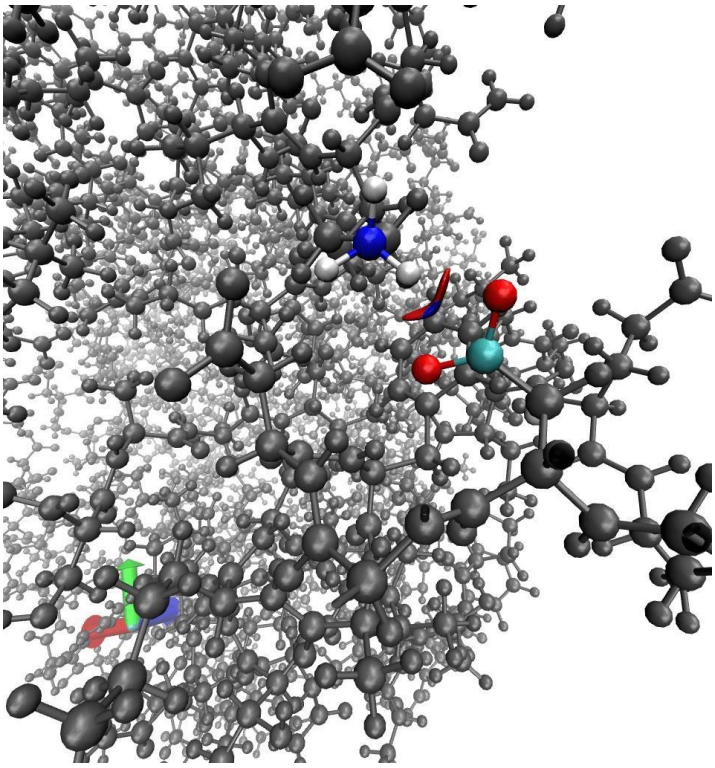


Figure 3.4.3: VMD Visualisation of the Iso-surface (HB3 Alpha-Alpha Barrel) that Bonder Produces

has consolidated all the concentrated clusters of points into one. Having only one cluster of points means that most of the interactions are in roughly the same geometrical environment.

It is worth noting that the remaining cloud of points at larger distances still exists due to the larger protein folding/unfolding motions of the protein chains that separate the atoms involved in the interaction. Little change was observed in the energy graph, this is shown in Appendix 5

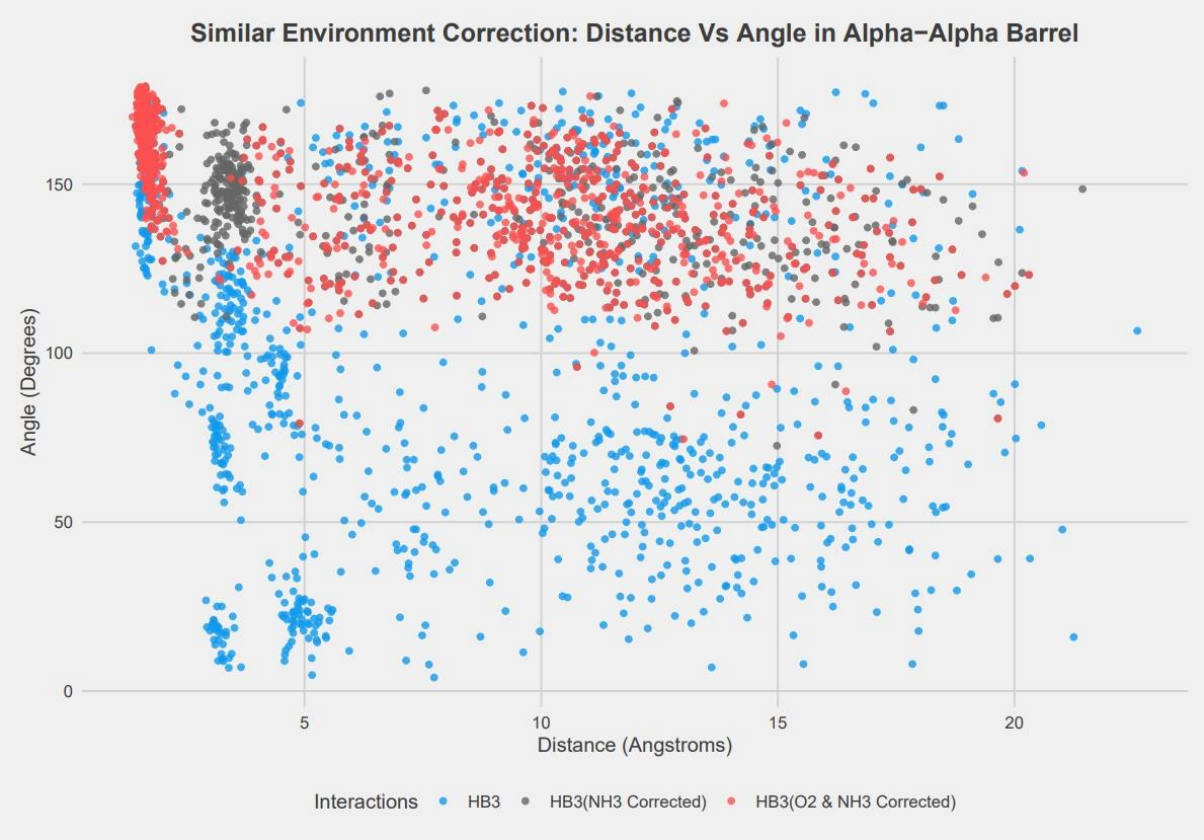


Figure 3.4.4: Comparison of Original (Blue), NH3 Corrected (Grey) and Both NH3 and O2 Corrected (red) Distances and Angles

CHAPTER 4

4.1 Variation in Hydrogen Bond Strength

In the previous chapters, we established the appropriate choices for the RDG cut-off value, averaging approach, and ensuring the right atoms are always selected for analysis with Hybond. This chapter will focus on the aims from the main hypothesis i.e. that hydrogen bonding strength varies appreciably in proteins and over the course of different MD simulations. This aim can be answered by exploring three questions, do the HBs change throughout the simulation? Do the HBs change between different simulations? And, how do different HB compare within the same protein building block? Four hydrogen bonding environments were analysed from each protein building block. These were also done in triplicate using a random seed to determine if there will be variation in the interaction between different, equally valid, simulation runs. There are many different integrated properties that can be considered for each interaction, including kinetic energy density, potential energy density, total energy density, electron localisation function and RHO difference. It has been found that looking at the kinetic energy density is a good descriptor (45) of the strength of a hydrogen bond and will be the focus of analysis throughout this chapter and chapter 5. All simulations of a given protein were started from the same equilibration point. The results from each protein building block will be discussed separately below.

4.1.1 Rossmann Fold Interactions

The Rossmann Fold building block had four interactions analysed through Hybond, which we denote HB1-HB4. Figure 4.1.1 shows the interactions' average kinetic energy density over

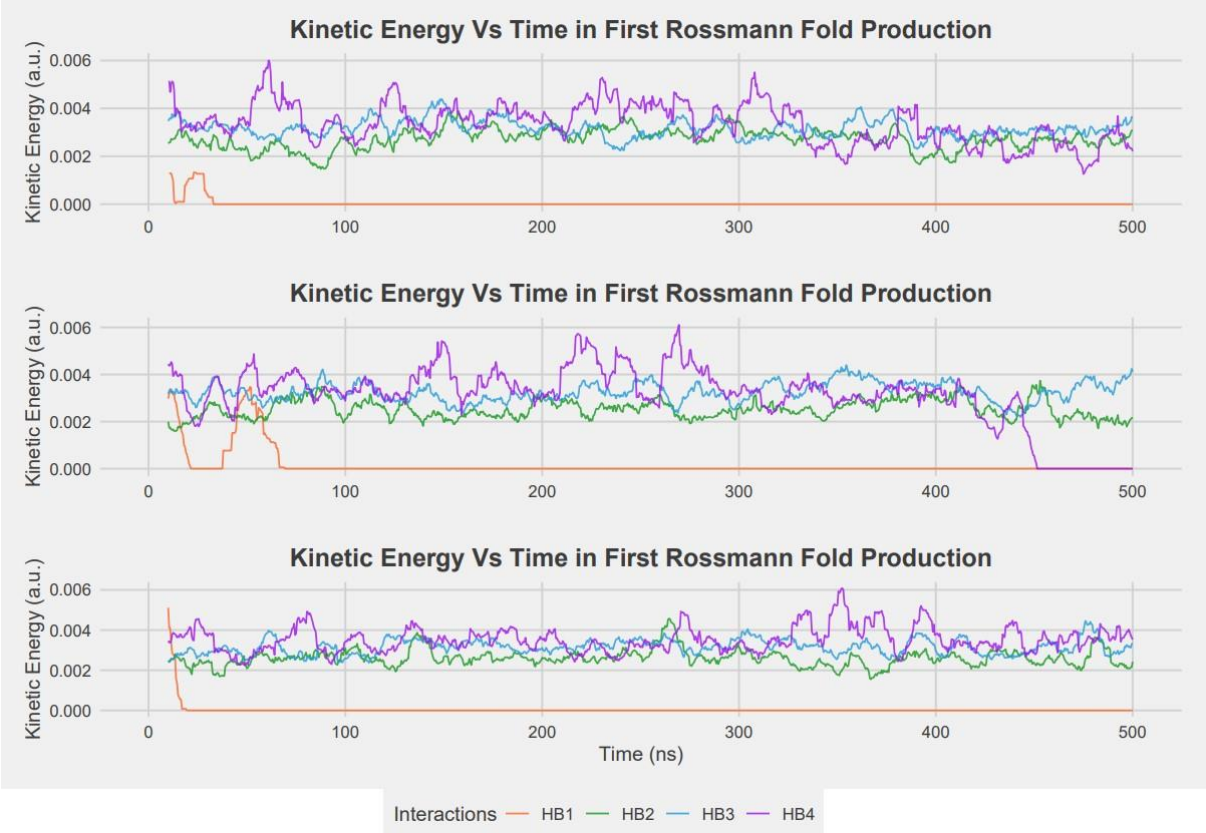


Figure 4.1.1: Simple Moving Average of Kinetic Energy Vs Time in Rossmann Fold Protein

the length of the simulation over all three of the MD simulations. HB1 was not well characterised with the Hybond program, as many points were dropped due to either no interaction being present or the volume of the interaction encompassing more than just the single isolation HB interaction. In each of the three simulation runs, there are slight indications that there is an interaction present for a short amount of time early on. This interaction is vastly different to HB2-HB4 and shows that depending on where an interaction is in the structure, it can lead to very different outcomes. In contrast with HB1, we find that HB2 is a well characterised bond and is present in every run that was conducted. This integrated kinetic energy of this bond fluctuates between 0.002-0.004 a.u., with the main difference between the three simulation runs being where the random peaks and troughs are. HB3 is also very similar to HB2 in that it sits in the same energy range, and again, the main difference is at what time the various peaks and troughs lie in each simulation. The offset of the peaks and troughs could be the same or similar folding motions/vibrations of the

interaction, just occurring either later or earlier in the simulation. This difficult to verify in VMD as the simulation sampling rate is too large to make a sound comparison from frame to frame. In contrast with the other hydrogen bond interactions, HB4 has a much larger range of energies and fluctuates throughout the simulations. From the second production run, the interaction is present up until approx. frame 900 (450ns), after which the interaction is no longer characterised. This is a prime example of why these types of interactions should be characterised over the course of the simulation and not just using one optimised structure. Portions of this interaction are comparable to HB2 and HB3, with the major difference being the regions of higher energy that the bond adopts. These higher regions of energy could arise from a multitude of reasons, including fluctuations in bond energies throughout the protein and could be the reason for these large peaks observed. These peaks could also be due to folding and minor structural changes allowing the atoms to be in a more desirable arrangement, which results in a stronger interaction.

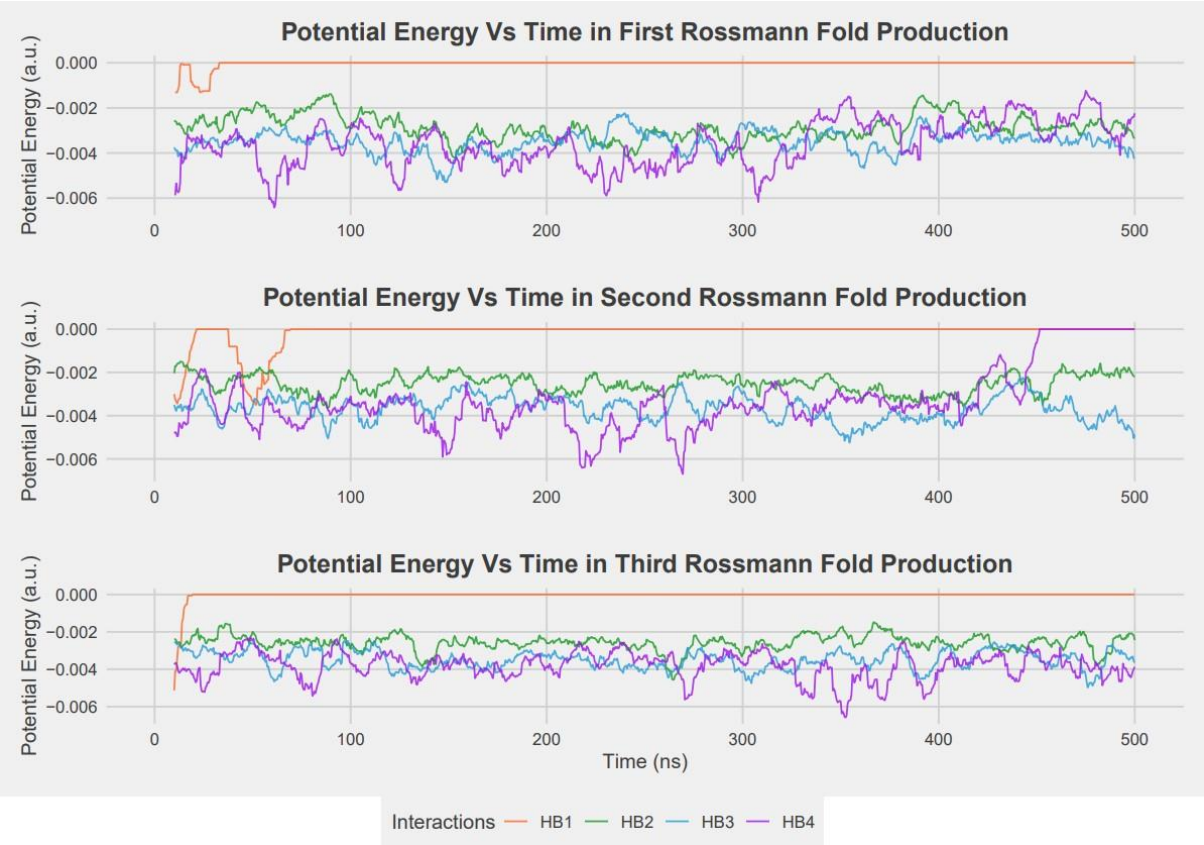


Figure 4.1.2: Simple Moving Average of Potential Energy Vs Time in Rossmann Fold Protein

The integrated potential energy comparison (Figure 4.1.2) for this protein is very similar to the integrated kinetic energy comparison, with the same overall trends albeit with the sign of the values inverted.

The total energy of the interactions, which is the sum of the kinetic and potential energies, holds some more interesting trends. From the graphs shown in Figure 4.1.3, we can start to see the overall nature of some of the interactions. In HB1, we see the similar trend of the interaction only existing in the first part of each simulation. We start to see more interesting variation in HB2. In the first production, the interaction is mostly negative (attractive) with only 5 peaks reaching past into a positive interaction (repulsive). As we move onto the next two simulations there is many more peaks in HB2 that reside in the positive region. HB3 becomes the interaction that is relatively similar between all the simulations, just with varying peaks and troughs. HB4 provides more information as we start to see a trend occur in the data with more regular peaks and troughs. This is most likely due to random sampling of

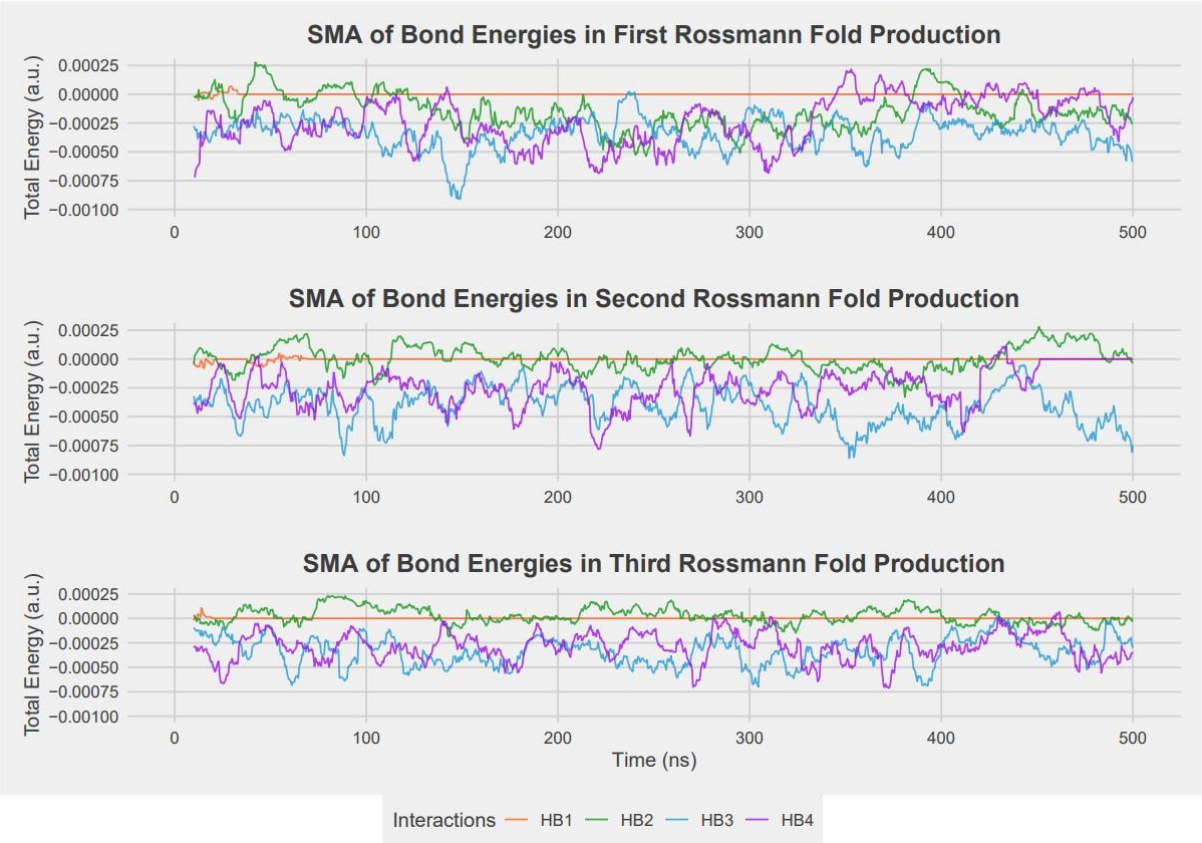


Figure 4.1.3: Simple Moving Average in Rossmann Fold Production Runs

the positions of the atoms and bonds involved in the interaction, which vibrate at a frequency much faster than the sampling rate

Moving on from energies of the interactions, there are other properties that allow us to see how these bonds change over time and throughout each simulation, once again showing the need for temporal analysis. In Figure 4.1.4, we show how the distance and angle of HB1-HB4 vary for each structure in the simulations. Using these geometric parameters, we can start to see some variation in the interactions that were not well characterised by Hybond. HB1 (orange) has quite a different distribution of both bond angle and bond distance, compared with HB2-HB4. For clarity, we define the hydrogen bond distance between the donor hydrogen atom and the acceptor electronegative atom. In the first and second simulation, there is a cluster of points spanning the 7-8 Å range with a spread of points either side ranging from approx. 4 to 10 Å. In the third simulation this spread has dissipated and there is

only a small number of points that are not in the cloud of points that reside around 7-8 Å. The points that

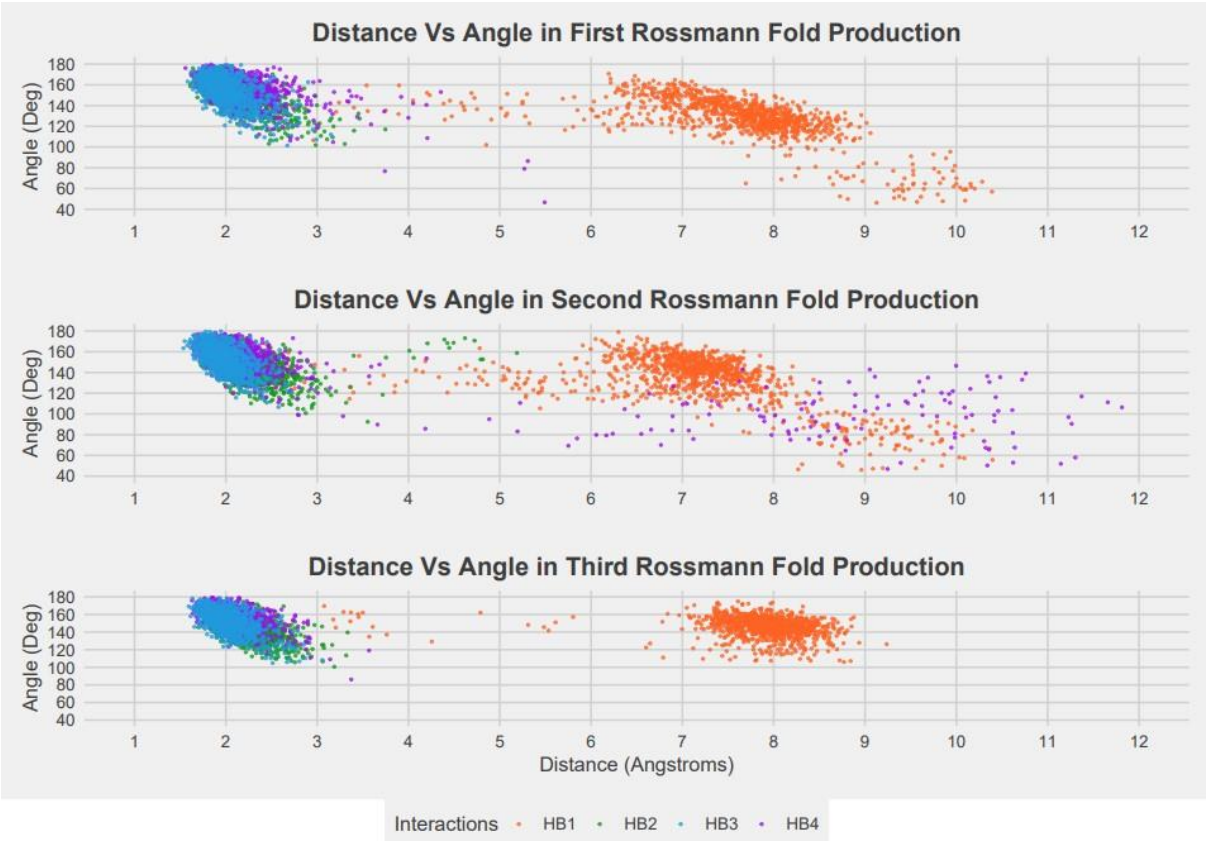


Figure 4.1.4: Distance Vs Angle in Rossmann Fold Simulations

reside outside the average also show trends that give vital information about the interaction. Such as when the bond distance increases the bond angle sits in the lower portion at 40-80 degrees and is consistent with no other points moving out of this range. This could be another slightly favoured environment in terms of energy within the protein, which is why this specific cloud of points appears. The observed variation in the geometric parameters validates why Hybond does not detect a hydrogen bond for HB1 during most of the simulation runs. We see less variation with HB2 and HB3, with most of the bond angles and distances sitting at approx. 120-180 degrees and 1.8 to 2.5 Å. There is also a slight anomaly for HB4, where in the second production run there are points that almost reach a bond distance of 12 Å and some of the smallest angles at approx. 50 degrees. This is due to a large movement in two of the secondary structures in the Rossmann Fold. Visual inspection of the protein in VMD

shows that at the end of the second production simulation the interaction that is acting as a bridge between a beta sheet and a 3-10 helix was split as the structure of the protein changed.

4.1.2 Alpha-Alpha Barrel Interactions

The same analysis was conducted for the Alpha-Alpha Barrel by looking at four different HB interactions in each of the three simulation runs. The kinetic energy graphs for these can be seen in Figure 4.1.5. HB1 fluctuates heavily between each simulation and is present approximately 75% of the time in the first simulation, near to 0% in the second and approximately 50% in the third. This interaction is far from stable throughout each simulation and displays a large variation. As such, a much longer time frame is necessary to be confident as to how frequently the interaction is present. The nature of how often this interaction drops in and out is interesting. In the first simulation, the bond dropped for a short amount of time and was picked back up later in the simulation. On inspection of the trajectory in VMD, we found that this is due to a short folding of the protein chain that then settled back into a position that favoured the interaction. For the second simulation, the interaction drops in and

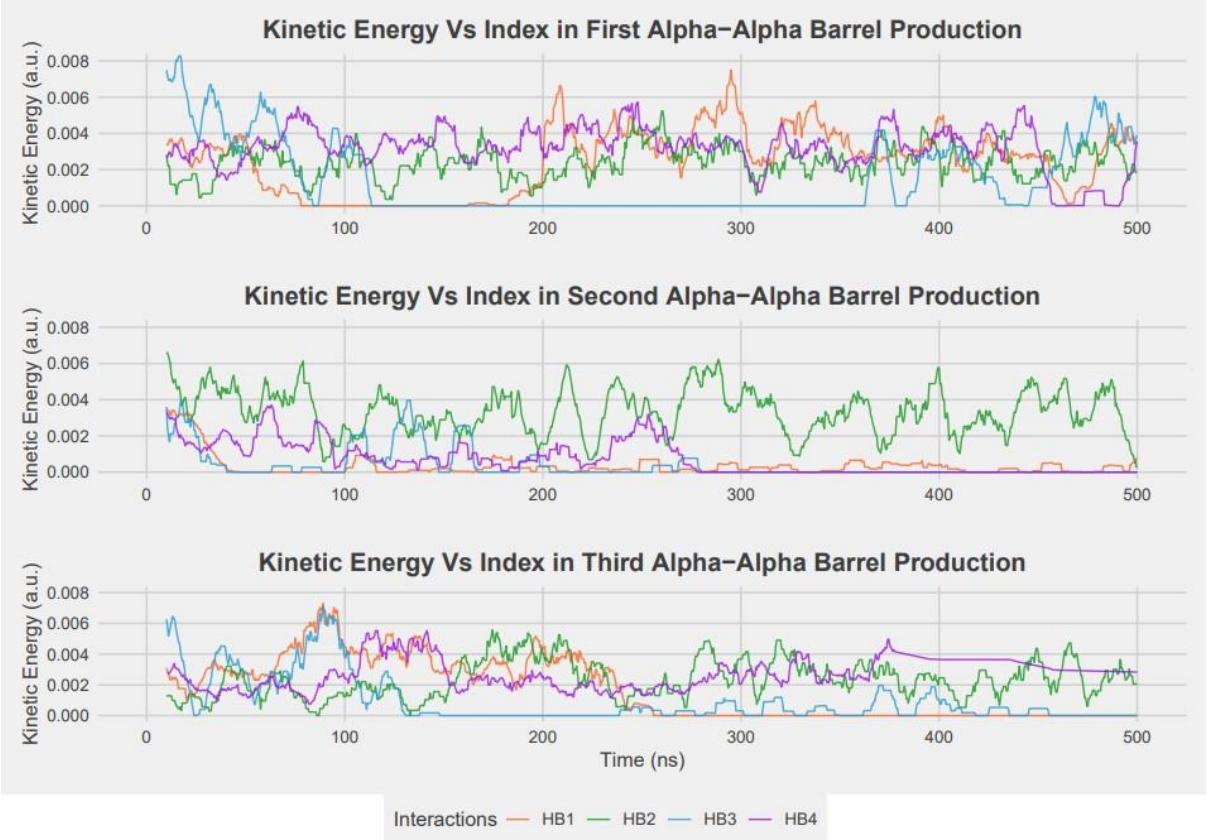


Figure 4.1.5: Kinetic Energy Vs Time in Alpha-Alpha Barrel Simulations

out constantly indicating that the protein chains that hold the atoms involved in the interaction are folding in and out of the viable interaction range at a regular interval and suggests that this simulation is less settled than in the first trajectory. The third simulation provides yet another scenario where the protein chain is conducive to forming a hydrogen bond in the first half of the simulation, but in the second half of the simulation, the interaction disappears completely. This is due to the larger structural change in the protein as the interaction does not appear again in the simulation. HB2 has slight variations between the simulations in terms of the fluctuation in energy. The first simulation has much less variation in interaction energy whereas the second and third simulations show much larger energies, along with fluctuations in these values. This is happening in the second simulation at roughly regular intervals and could be a product of compound vibrations of the atoms around and inside the interaction. HB3 is the problematic interaction described in section 3.4 that has both an NH_3^+ and a COO^- environment. While the results shown for this interaction have had the necessary

corrections applied to it, there is a folding motion that happens through all three simulations which results in the interaction being largely uncharacterizable. In contrast, HB4 is generally well-characterised, appearing throughout simulation run 1 and in the first half of simulation run 2. A quirk can be seen in simulation run 3, where the end of the interaction appears to level out to a non-zero value. However, this is just a result of the averaging method, whereby the last defined value was non-zero and all subsequent frames in the simulation were undefined due to interaction volume exceeding the maximum threshold. This indicates that there is an interaction present, but that it is non-discrete, likely involving several other atoms. As the potential and total energy comparisons for these simulations show roughly the same trends and do not offer any further valuable information on the interaction, these can be found in Appendix 6 and 7 respectively.

Looking at the distance vs angle graph (Figure 4.1.6) of the Alpha-Alpha Barrel gives some insight into the variations seen in the energies of the interactions. HB1 shows some large variation with tight groups of points, which resemble environments where the protein is energetically satisfied and can remain stable. This supports the theory that the energy in the molecule, that resides in these interactions, can be spread around different points of the molecule at any one time. This could either weaken or strengthen certain interactions within the protein. HB2 shows different interaction environments due to the different hydrogens present in an NH_2 group. Both hydrogens are involved in the interaction for the majority of simulation. Throughout the simulation, these hydrogens switch because they are in equivalent environments. HB3 and HB4 show a cloud of points where the geometric properties of the HB interaction are favourable (close to 180 degrees and a distance of approx. 1.5 to 2.5 Å). This is accompanied by large clouds of points that are scattered over the whole range of distances, with HB4 only having a separate cluster in simulation two. This is due to a change in conformer and is responsible for the dropped frames for HB4.

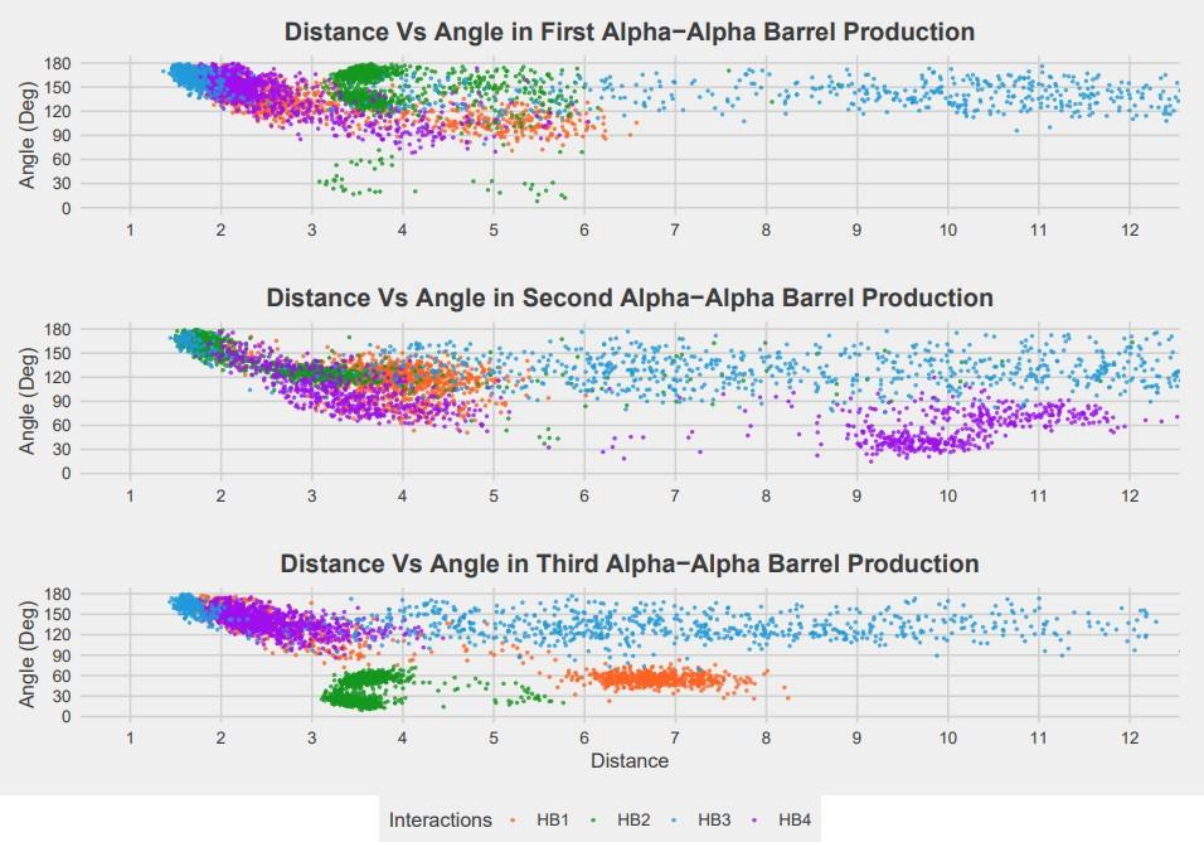


Figure 4.1.6: Distance Vs Angle in Alpha-Alpha Barrel Simulations

4.1.3 Jelly Roll Interactions

Of the 4 interactions that were analysed for the Jelly Roll building block, there were only two interactions that were well characterised. This can be quickly deciphered by looking at the kinetic energy graph shown in Figure 4.1.7, particularly the second and third simulations. Grouping these interactions, HB1 and HB3 share similar interaction characteristics being weakly present in simulation one and present approximately 1% of the time for simulations two and three. These interactions were affected greatly by the structural changes in the simulations which can be seen in VMD as well as angle and distance data. The distance vs angle data of HB1 and HB3 is shown by the orange and blue points respectively in Figure 4.1.8. The weak interactions in simulation one are also backed up by the results shown in the distance vs angle graph. The HB1 interaction has points that are closer to the general trends of “normal” HB interactions. Since these points are on the edge of where HB interactions are

expected we would expect there to be dropped frames along with correctly characterised interactions. This is observed in the kinetic energy of the interaction with many flat peaks which indicate there is no change in the average kinetic energy and these regions are due to dropped frames. HB3 shows reasonable distance values, but the vertical spread of the points indicates the angles of the interaction are too low. This is a limiting factor that could stop the interaction from being characterised properly. When grouping interactions HB2 and HB4, it can be seen that on average, HB2 has double the energy of HB4 although they are both well characterised. The distances, and angles, of these interactions are consistent and mostly sit with the ranges of $110^0 – 180^0$ and 1.7 Å -3.0 Å. This is a good demonstration of how some interactions can exhibit similar structural characteristics while differing in strength.

The other energy comparisons for the Jelly Roll simulations are in Appendix 8 and 9.

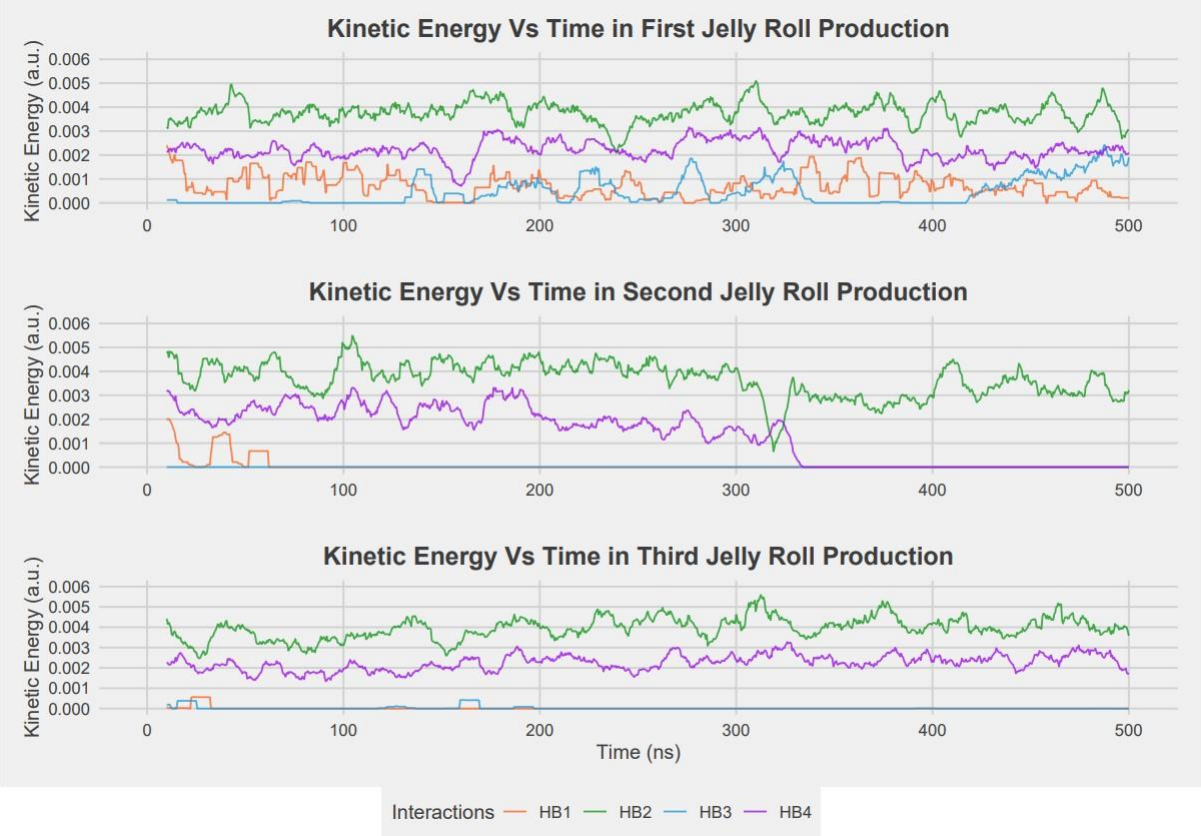


Figure 4.1.7: Kinetic Energy Vs Time in Jelly Roll Simulations

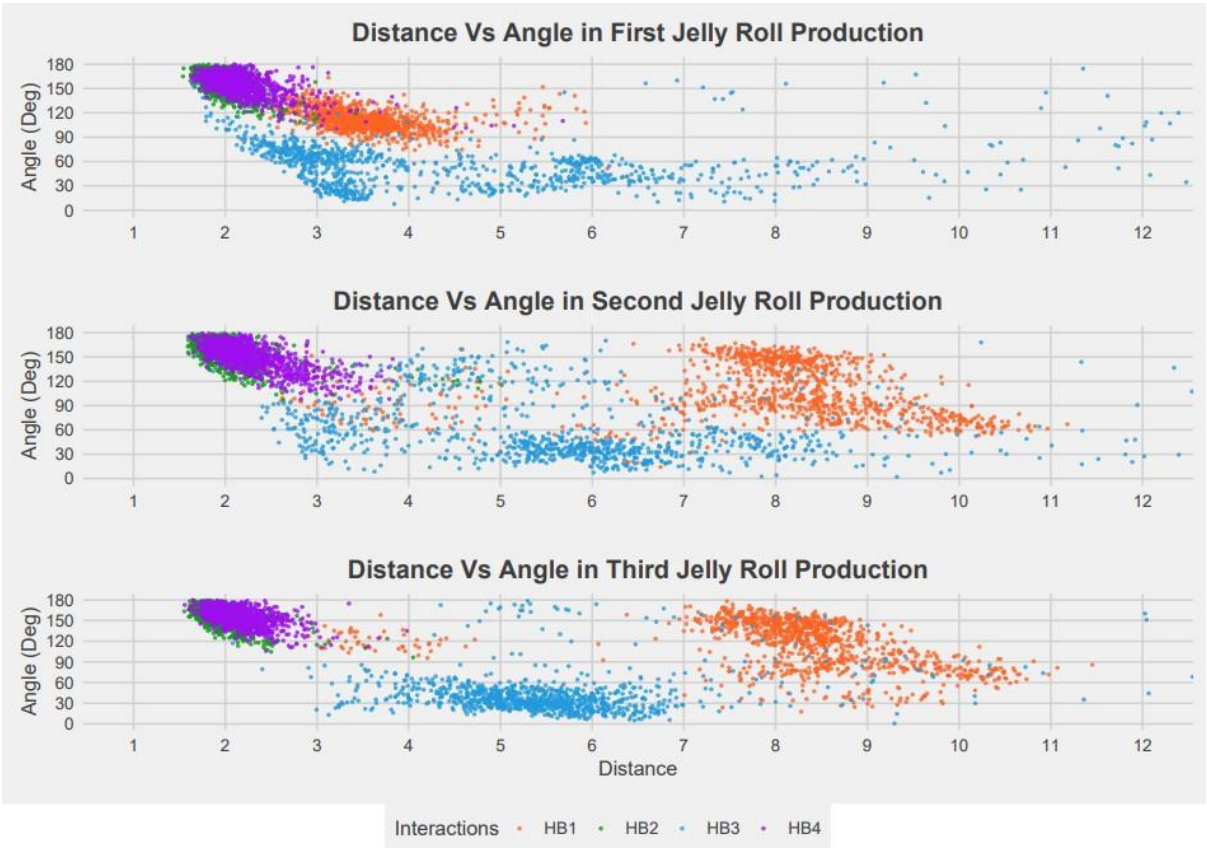


Figure 4.1.8: Distance Vs Angle for the Jelly Roll Simulations

4.2 How Should Bond Strength be Averaged?

This project provided a very large quantity of data that spans 500 ns for each, equally valid, stochastic simulation. For each frame of each simulation, there are three types of output: a non-zero data point from a successful characterisation, a “0” from not finding an interaction and an “N/A” from finding a non-discrete interaction that includes one or more secondary interactions along with it. This raises an interesting problem when deciding the best way to average the data since different approaches yield vastly different outcomes.

The first problem is in the case of no interaction being found for a given frame. If the dropped frames are included in the average, this will drag the average energy of the interaction down and depending on the amount of dropped frames this change could be significant. To include or exclude the dropped frames in the average would depend heavily on what the average was being used to determine. If there was only a need to know how strong the interaction is when it exists, then the dropped frames could easily be forgotten for this averaged value. On the other hand, there is the case where one might want to know the average energy of the interaction area over the entire simulation to determine whether this fluctuation is seen in other places in the protein.

The second problem with averaging the data is the “N/A” values, since these interactions involve a non-discrete HB that includes other secondary interactions as well. In this case the strength of the interaction is not 0 but the exact value is undefined. One option is to average these values based on the frames either side of it, then average those values for the whole interaction. This can either over or underestimate the strength of the interaction. Another option includes ignoring the “N/A” frames. Ignoring a frame from the data simply reduces the amount of data in the average calculation resulting in less numerical stability. From these considerations, four averaging methods were developed to allow the average strength of the

interaction to be calculated, depending on why the interaction was being investigated in the first place:

1. Dropped frames are ignored and N/A data is approximated based on near observations.
2. Dropped frames and N/A values are both ignored leaving just valid points.
3. Dropped frames are included and N/A data is approximated.
4. Dropped frames are included and the N/A values are ignored.

The averaged data obtained with these four methods is shown in Tables 4.2.1, 4.2.2 and 4.2.3.

Table 1.2.1: Integrated Kinetic Energy Density (a.u.) for Attractive Component using 0.3 RDG Cut-off Applying Averaging Methods for Rossmann Fold Building Block

Rossmann Fold					
Simulation	**Interaction**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
1	1	0.003184	0.003319	0.00005413	0.00005316
1	2	0.002738	0.002742	0.00272100	0.00272600
1	3	0.003182	0.003178	0.00317900	0.00317500
1	4	0.003446	0.003427	0.00340100	0.00338200
2	1	0.003913	0.003990	0.00017610	0.00017560
2	2	0.002544	0.002542	0.00251600	0.00251400
2	3	0.003270	0.003270	0.00327000	0.00327000
2	4	0.003487	0.003490	0.00306900	0.00306900
3	1	0.007302	0.007302	0.00010220	0.00010220
3	2	0.002620	0.002611	0.00260700	0.00259800

3	3	0.003151	0.003151	0.00315100	0.00315100
3	4	0.003538	0.003533	0.003535	0.00353

Table 4.2.2: Integrated Kinetic Energy Density (a.u.) for Attractive Component using 0.3 RDG Cut-off Applying Averaging Methods for Alpha-Alpha Barrel Building Block

Alpha-Alpha Barrel

Simulation	Interaction	Method 1	Method 2	Method 3	Method 4
1	1	0.003777	0.003786	0.0024700	0.00246800
1	2	0.007971	0.007971	0.0024150	0.00241500
1	3	0.007077	0.007077	0.0015850	0.00158700
1	4	0.003782	0.003785	0.0031610	0.00316100
2	1	0.002462	0.002464	0.0003669	0.00036360
2	**2**	**0.006399**	**0.006399**	**0.0033850**	**0.00338200**
2	3	0.007931	0.007931	0.0004124	0.00004124
2	**4**	**0.002876**	**0.002876**	**0.0007190**	**0.00071100**
3	1	0.004013	0.003998	0.0016970	0.00167400
3	2	0.007917	0.007928	0.0023430	0.00243100
3	3	0.007494	0.007498	0.0009967	0.00099170
3	**4**	**0.002748**	**0.002808**	**0.0019600**	**0.00250400**

Table 4.2.3: Integrated Kinetic Energy Density (a.u.) for Attractive Component using 0.3 RDG Cut-off Applying Averaging Methods for Jelly Roll Building Block

Jelly Roll Simulation	Interaction	Method 1	Method 2	Method 3	Method 4
1	1	0.003401	0.003432	0.000721	0.0007120
1	2	0.003770	0.003767	0.003766	0.0037630
1	3	0.002473	0.002473	0.000475	0.0004749
1	**4**	**0.002255**	**0.002255**	**0.002224**	**0.0022240**
2	1	0.002914	0.002914	0.000087	0.0000874
2	2	0.003774	0.003765	0.003687	0.0036770
2	3	0	0	0	0
2	**4**	**0.002207**	**0.002209**	**0.00139200**	**0.00139300**
3	1	0.001849	0.001849	0.00001294	0.00001294
3	2	0.003966	0.003967	0.00394200	0.00394300
3	3	0.002023	0.002023	0.00002428	0.00002428
3	**4**	**0.002261**	**0.002263**	**0.00225200**	**0.00225400**

Comparing these interactions with the four different averaging methods shows how either keeping or leaving the dropped frames affects the calculated average strength of the interaction. The general trend is when dropped frames are ignored, the strength of these interactions increases. The well-characterised interactions see very little change in the average value obtained with the different methods, primarily because there are very few 0 or N/A values. The difference between Method 1 and 2 is small in most interactions due to there being minimal frames having N/A values. This is also observed in Methods 3 and 4. The higher RDG cut-off values yielded a larger portion of N/A values which has a larger effect on

the averages. Methods 3 and 4 give the interaction strength over the complete interaction and Methods 1 and 2 give the strength of the interaction when it exists within the protein.

Taking specific interactions[3], we can start to see how much the averaging of the results changes the strength of the interaction over the simulation. Starting with the Alpha-Alpha Barrel simulations (Table 4.2.2), simulation two shows two distinct average kinetic energy densities across the four averaging methods.

Starting with interaction 2, the trace of the kinetic energy on the graph (Figure 4.1.5) has large fluctuations due to dropped frames occurring regularly but not consecutively. The average kinetic energy density observed in Methods 1 and 2 is 0.006399 a.u. and an average of 0.003383 in Methods 3 and 4. The averaged energy, including the 0's of the dropped frames for Methods 3 and 4, is almost 50% of the energy when these frames are ignored for Methods 1 and 2. This is similar to simulation 2, interaction 4 however the trace (Figure 4.1.5) shows that the interaction does not exist for the later portion of the simulation. The energy difference here is also larger as ignoring the dropped frames from the averaged kinetic energy gives a value that is approx. 4 times larger. This difference in energy is quite large and both values can be useful depending on how the energy of the interaction is being considered. This factor demonstrates another hurdle in trying to come up with a general method to describe the strength of an interaction over time.

Simulation 3, interaction 4 for Alpha-Alpha Barrel (Table 4.2.2) shows how the proportion of N/A values influences the average energy value. This influence is the result of the output variation of Hybond as described in Chapter 3.1 and can be seen from ~380 ns of the simulation in Figure 4.1.5. The average energy levels off as the interaction becomes undefinable i.e. Hybond is characterising the interactions as non-discrete. If you only

[3] Interactions in question are bolded in the tables for ease of finding.

consider times when the hydrogen bond is present (Methods 1 and 2), then the average energy value ranges from 0.002748 – 0.002808 a.u. Conversely, considering the hydrogen bond over the entire simulation gives an average energy value range of 0.00196 – 0.002504 a.u. This range is smaller than those of simulation 2, interaction 2 and 4 and demonstrates why multiple averaging methods need to be considered.

For the Jelly Roll protein, interaction 4 is being analysed over the three simulations. The average energy values for interaction 4, simulations 1, 2 and 3 are very similar using Method 1 and 2. The values range from 0.002207 a.u. in simulation 2 to 0.002263 a.u. in simulation 3, with a deviation in interaction 2 for Methods 3 and 4 (observed in Figure 4.1.7). This shows that for well characterised interactions, the averaging methods are similar enough to be confident in saying that when the interaction exists it will be of a certain strength.

CHAPTER 5

5.1 Beyond a Distance Analysis of Hydrogen Bond Strength

Distance has been used to estimate and determine the strength of hydrogen bonding interactions for many years. The third Jelly Roll simulations exhibit two very wellcharacterised interactions, which allows an easy analysis to discover which integrated properties can be used alongside distance to estimate the strength of the interaction.

Correlation lines were fitted to each of the relationships using the R software. The exponential fits were done using the stat_smooth function along with the “nls” method provided in that function. The linear and polynomial fits used the “lm” method. The relationship between kinetic energy density of the interaction and the distance (Figure 5.1.1) is an exponential in the form of equation 3:

$$y \sim a * \exp(b * x) \qquad (3)$$

The exponential fit for HB4 is very accurate with little pulling on the apex of the curve, meaning that the line is centred through the data and follows the points upwards. The fitted line for HB2 is not as good and has a shallower curve compared to how the tendency of the data. Removing these points would shift the line and pull the middle of the curve down, this in turn would line the top of the curve up better with the data. It must be noted that only well characterised bonds can have lines fitted to them as there is large amounts of noise in the data that skew these fitted lines in poorly characterised interactions.

Analysing the kinetic energy vs volume relationship (Figure 5.1.2 and 5.1.3), there is clearly an exponential relationship between these two properties. However, the data for these interactions needed to be cleaned substantially to see these relationships due to how noisy the original data was. The noise in the data is most likely caused by the constant variation in the protein structure and the tight environments that these interactions can be forced into. The form of the exponential relationship is the same as equation 2. This relationship can also be shown to be an exponential relationship by using the natural log of the Y axis and finding a linear relationship. Plotting the natural log of kinetic energy against the volume gives a linear relationship and can be seen in Figure 5.1.3.

The next relationships are observed in the RHO difference and the ELF integrated properties. Starting with the RHO relationship (Figure 5.1.4), we observe a very linear relationship, which is helpful if this integrated property is to be used as a measure of estimating the

hydrogen bond strength. A line was not fitted to this data since it would cover majority of the data points on the graph. The equation for the linear relationship is of the form shown below in Equation 4:

$$y \sim a + b * x \qquad (4)$$

The relationship between the ELF and kinetic energy is of a second order polynomial or quadratic relationship (Figure 5.1.5), where the gradient is largest at y=0 and decreases as the y value increases. This gives a curve that levels off as the values increase. The form of this fit is shown in Equation 5 below:

$$y \sim poly(x, 2) \qquad (5)$$

These well-defined curves all have the potential to be used as estimators for the strength of an interaction. When used in addition to the distance of the interaction, it could be more accurate and impactful than the use of distance alone. Using extra parameters to estimate the strength of interactions in these environments is vital since interactions are unpredictable and influenced by many other factors inside the protein. There are also some parameters that give very weak relationships to the reduced kinetic energy density, such as the hydrogen bond angle, which are of little use.

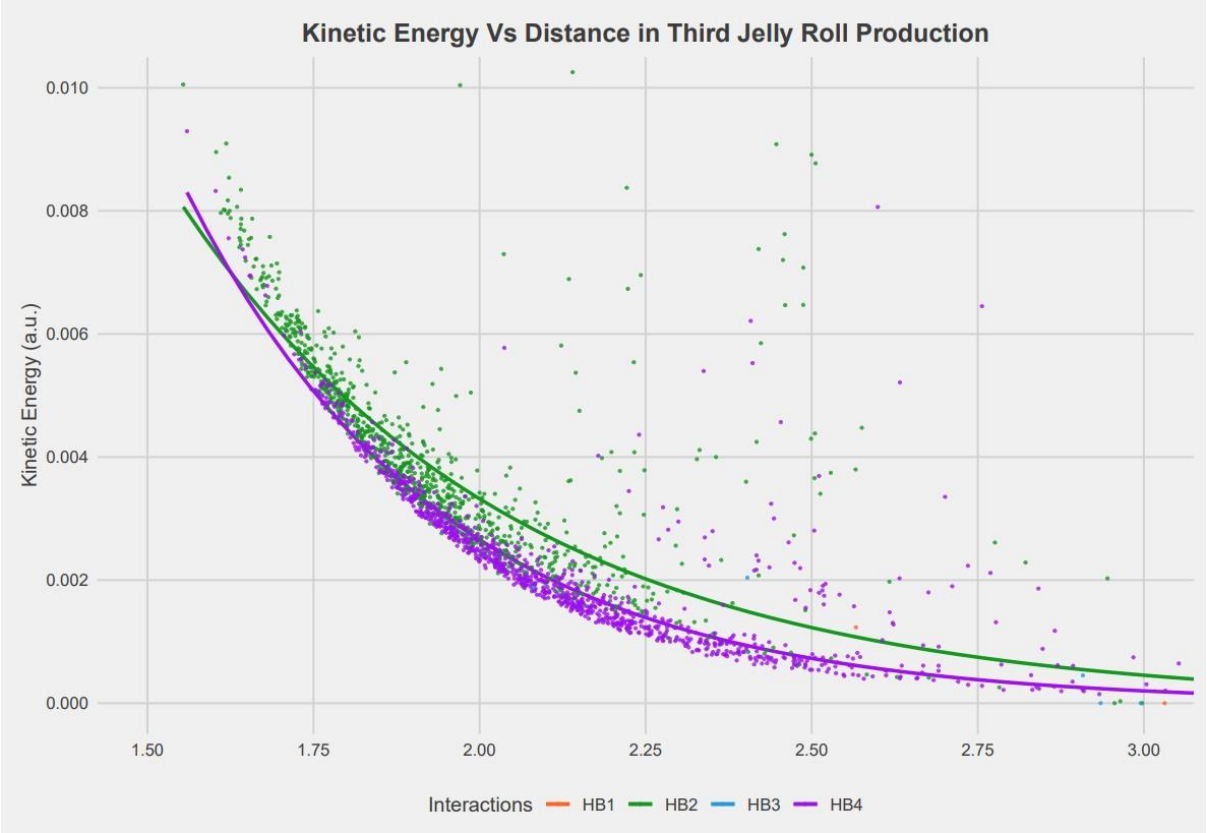


Figure 5.1.1: Exponential Fitted Lines to the Distance vs Kinetic Energy of the Third Jelly Roll Production

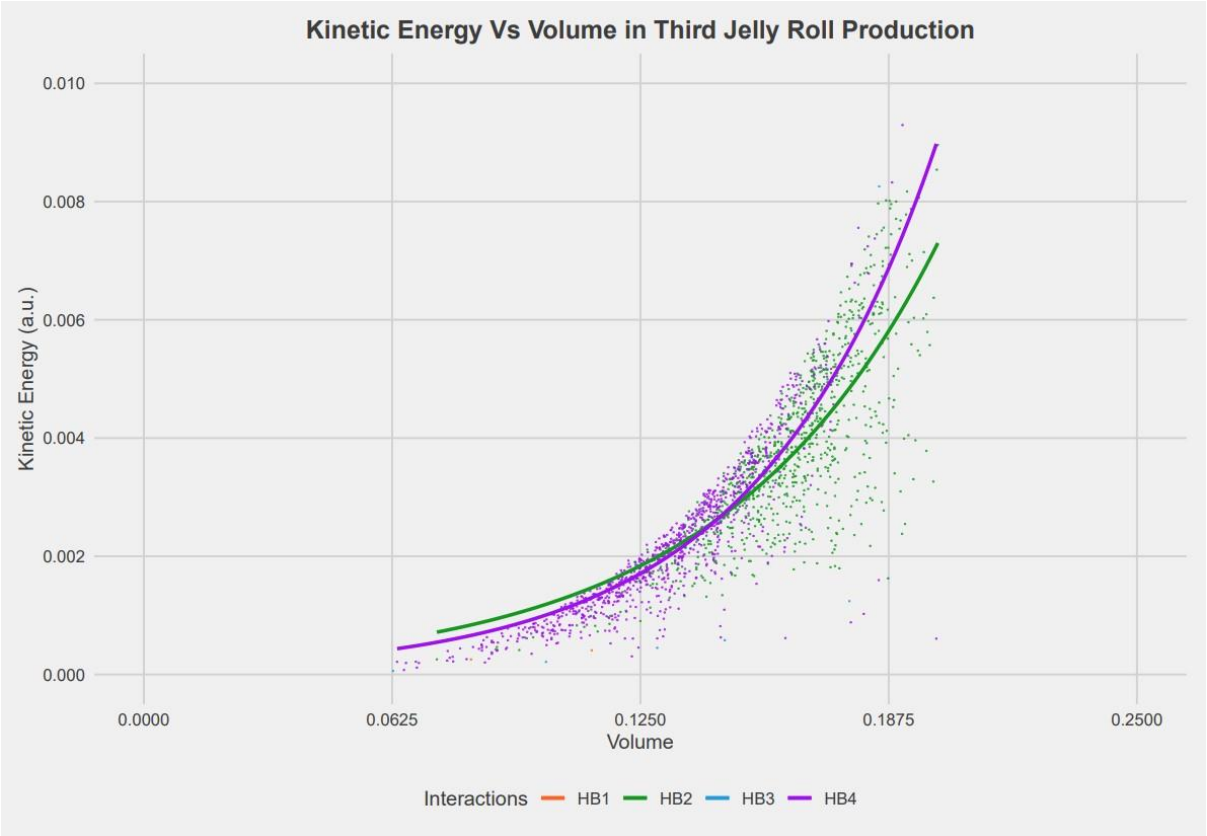


Figure 5.1.2: Exponential Relationship between Kinetic Energy Density and Volume in Third Jelly Roll Production

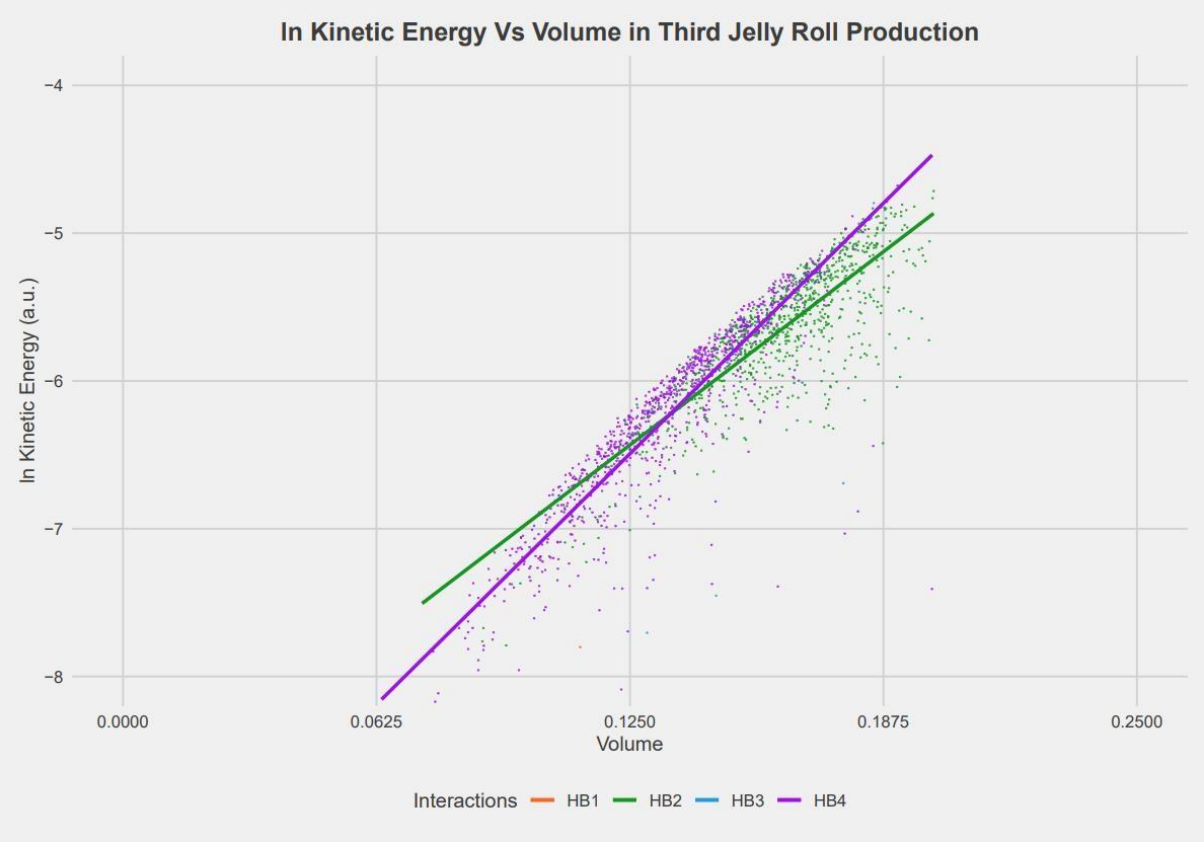


Figure 5.1.3: Linear Volume Vs ln Kinetic Energy Relationship in the Third Jelly Roll Production

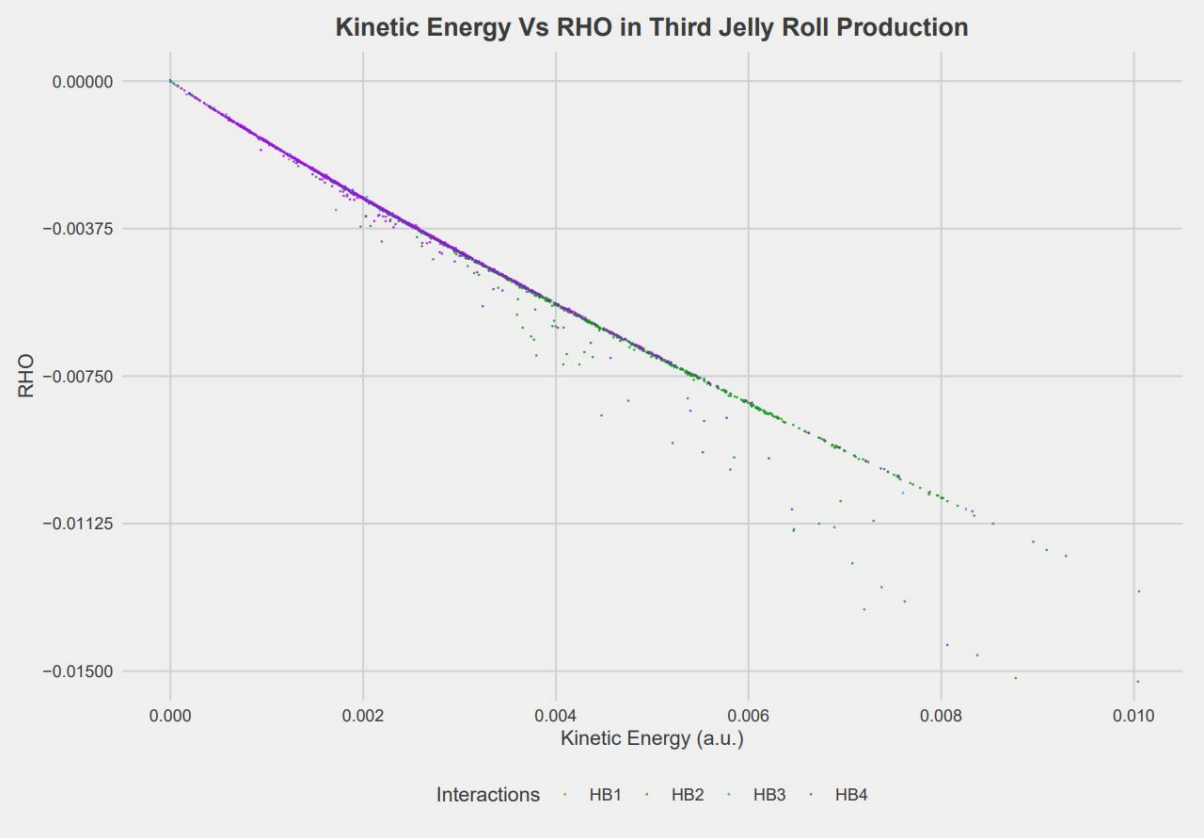


Figure 5.1.4: Linear Relationship between the RHO difference and Kinetic Energy in the Third Jelly Roll Simulation

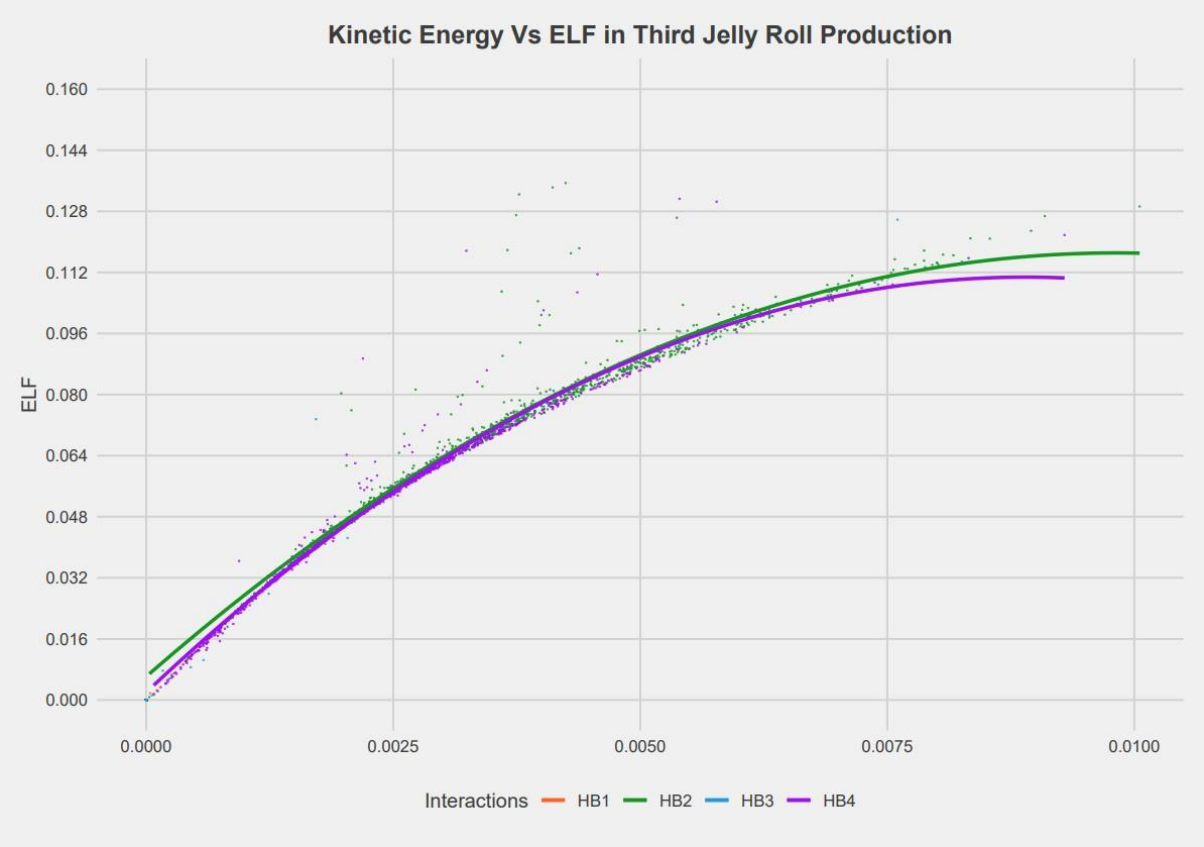


Figure 5.1.5: 2nd Order Polynomial Relationship of Kinetic Energy to ELF in the Third Jelly Roll Production

5.2 Calibration of Hybond

We have shown that Hybond is able to quantitatively describe the dynamic strength of a hydrogen bond using NCI theory. However, to be of most use to researchers, we need to be able to link these results to an absolute strength scale in some physically intuitive energy units i.e., kJ mol^{-1}. To do this, we attempted to construct a simple dimer of two hydrogenbonding fragments (Figure 5.2.1), whereby the absolute energies could also be calculated. This can be done by comparing the curve acquired from calibration to the curve from the raw results and applying the required conversion to transform this raw data line to fit the calibration curve.

A small portion (3 amino acids) of the Rossmann Fold protein was selected (Figure 5.2.3).

This gave the closest representation of a protein environment hydrogen bond without over

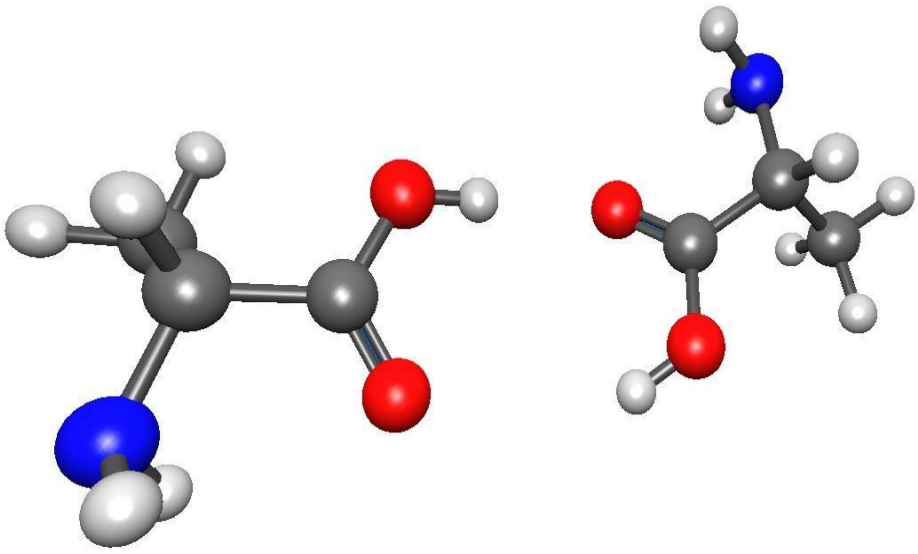


Figure 5.2.1: First Dimer used for Calibration, Alanine Dimer.

complicating the system. By taking the hydrogen bond interaction and separating it by intervals of 0.1 Å (from 1.4Å to 3.3 Å) a set of energies can be obtained for the dimer with the hydrogen bond present. The optimized energy can be obtained for one of the molecules and doubled to get the energy of the dimer without the hydrogen bond. The difference can be calculated from the dimer containing the hydrogen bond and the dimer that was calculated separately. This gives the energy graph of the hydrogen bond energy which starts off large and positive due to the bond being too close and repelling itself strongly. This decreases quickly and starts the energy well where the distance of the hydrogen bond becomes optimal. As the distance increases, the hydrogen bond should get weaker and the energy tends toward 0 with an exponential relationship.

However, the bond energy at the bottom of the well was sitting at approx. -293kJ/mol. Following the well, there is a linear relationship that was tending towards 0 starting from 184kJ/mol but was still present when the hydrogen bond energy should be close to 0kJ/mol. This can be seen in Figure 5.2.2. These energies are unreasonable for the stretching of a single hydrogen bond.

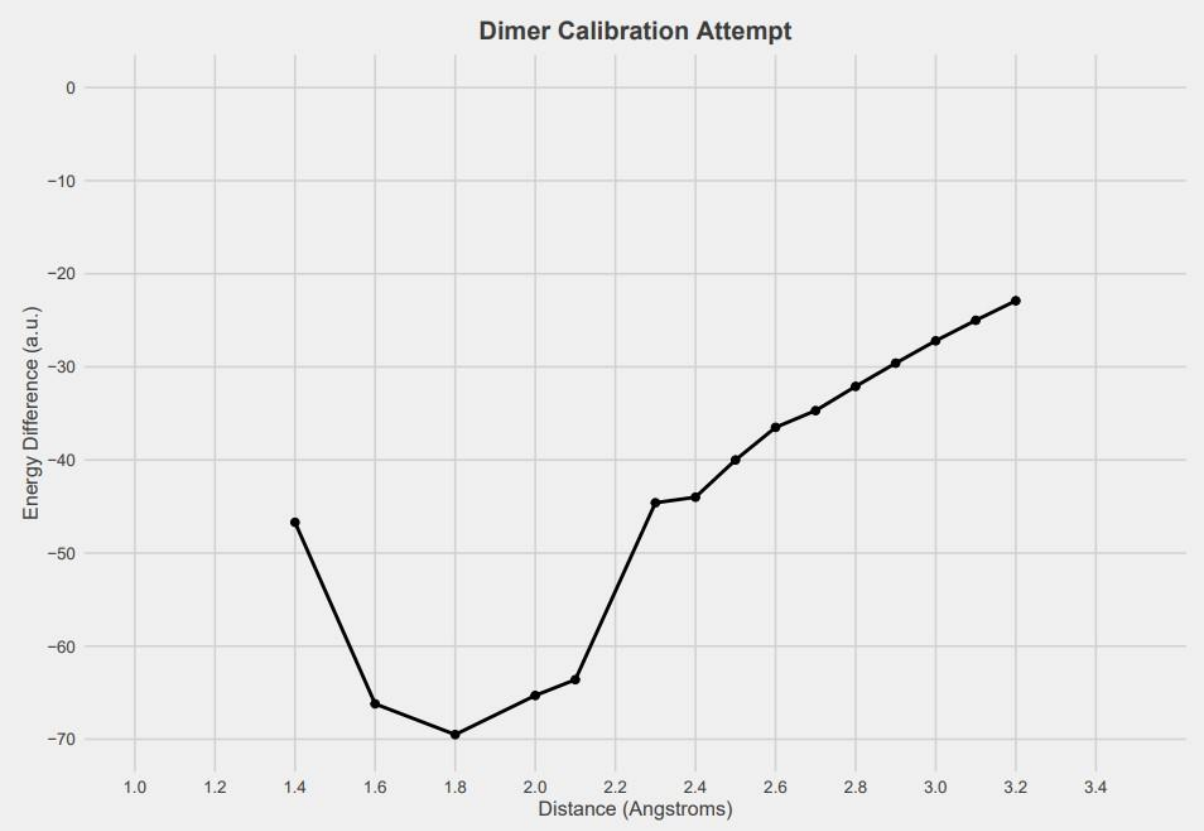


Figure 5.2.2: Calibration Attempt of a Portion of the Rossmann Fold Protein

From visual inspection of the optimised structures there was minimal interaction with other portions of the chain. If there were interactions present, the energy will be larger than expected. Having charged species can also influence the energy heavily. This was shown in a simulation of formaldehyde with hydrogen bonding to NH_3 and NH_4^+ where the energy of the interaction at 2Å with NH_3 was approx. -41 kJ/mol and the NH_4^+ was approx. -1230kJ/mol. This is a much larger difference in energies than is observed in Figure 5.2.2. However, a direct comparison is hard to make due to the difference in the number of atoms in the two systems and the different environments. After discovering these things about the dimer, it was clear that the next step was to find another set of dimers that reduced the effect from the charge and other interaction influences. However, there was simply not enough time to rerun these optimisations and complete the calibration curve. This is now listed as the further research.

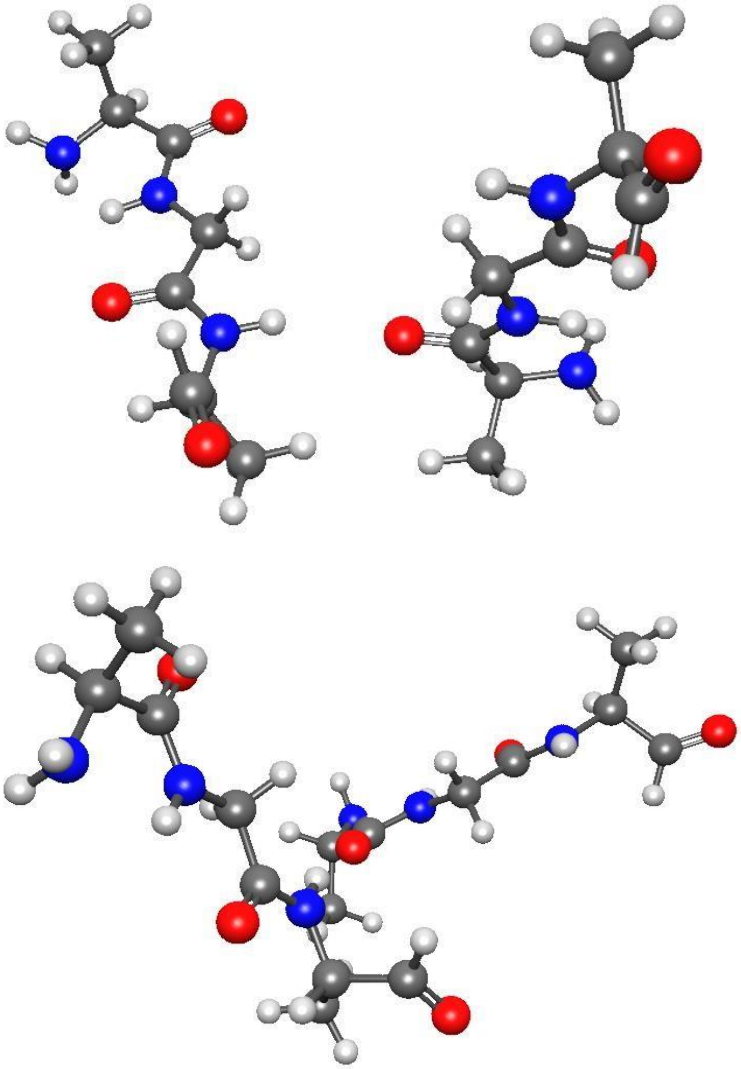


Figure 5.2.3: Portion of the Rossmann Fold Molecule used for calibration, looking across the hydrogen bond between NH and O (Top), looking down the hydrogen bond (Bottom)

CHAPTER 6

6.1 Limitations of this work

There are some trade-offs that must be considered in respect to both the protein environments that were studied and the software that was used. The length of the protein molecular dynamics simulations was a major factor, with 500 ns simulations being too short to capture some of the larger scale folding/unfolding processes that can occur in proteins. However, it is sufficient to observe faster examples of protein folding (100 ns – 1μs) (5) (46). The .dcd output step was set so that 1000 frames were captured over the entire simulation, giving each snapshot a separation of 0.5 ns. Another limitation lies within this .dcd output parameter. Since each step of the simulation being looked at is 0.5 ns, this means that it is not possible to resolve the inherent vibrational frequency ($1*10^{-5}$ ns) of the hydrogen bonds.

The .dcd output step also has another consideration, which is storage space when running the trajectory through the Hybond software. These files can easily reach 100s of gigabytes in total size for simulations with large numbers of frames.

Another limitation of the software being used is that the values that are output from Hybond need some form of calibration curve to relate them to something physically intuitive, like the binding energy of the hydrogen bond. However, developing a calibration curve proved to be more difficult than expected and this is further explained in section 6.3.

6.2 Conclusions

It has been shown in this thesis that hydrogen bonds are dynamic and can vary in strength drastically over a simulation period. This was observed when the energies of each interaction

were compared over the simulations that were conducted in triplicate. For the interactions that were not well-characterised, the difference in the simulations can be observed in the angle and distance data. This data had a mixture of fluctuations and well described interactions in proteins, which is yet another reason why these interactions must be considered in a temporal space. It was also shown that there are interactions that exist at different strengths, at different times. Although more research would be required to find out the reason behind this, it has been theorised that the fluctuations could be due to the protein having a shared total energy from these interactions and this is distributed over the protein so as one interaction gets stronger another in the protein weakens slightly.

From plotting the other integrated parameters of the simulations, it was found that there are many strong correlations with properties such as volume and ELF. RHO shows an excellent linear relationship with the kinetic energy density of the interaction. These properties could be used alongside the distance of the interaction to give better estimates of the overall strength of hydrogen bonding interactions in the future.

Different averaging approaches were also applied. These gave the values that were expected - by dropping frames with uncharacterised interactions, the average bond strength for the interaction increases. There are situations where the bond strength differed significantly between different averaging methods. This contrasts with interaction 4 in the Jelly Roll protein where the kinetic energy density of each method was very similar. The amount of non-discrete and no interaction frames heavily affect the conclusions drawn from these interactions and is why different averaging approaches need to be applied and considered based on the aim of the analysis.

To reiterate, it is important moving forward, that hydrogen bonding interactions are simulated and classed as dynamic interactions. Though more time might have to be invested to characterise these interactions over the simulation time. The results have shown that there is

need for this to be done in systems where there is enough appreciable energy for the proteins to have movement.

6.3 Further Research

This research has proven that there is merit in looking at hydrogen bonding interactions in a dynamic setting. This opens many avenues for further research, such as analysing a protein structure and the interactions occurring in its entirety over the course of a simulation. Learning how interactions inside a protein behave as it folds and vibrates could open discoveries in how enzyme catalysis functions and how reactions occur in proteins. Some sensible next steps to further this research are explained below:

Running Finer and Complete Simulations. This research was done by running simulation of 500 ns and outputting 1000 frames throughout this simulation. This time frame was chosen to be able to characterise faster examples protein folding and stretching. By running finer simulations, the vibrational frequency of the hydrogen bond and the energy fluctuations involved with this could be investigated. This would show how the hydrogen bond changes within the environment that the hydrogen atom exists in. At the other extreme, much longer simulations should be run to show how interactions change for slower folding and stretching movements within a protein. This could start to show how these interactions are affected by reactions and conformer changes within the protein. Also, how these interactions themselves effect parameters such as rates of reactions or enzyme binding energies.

Complete Calibration of Hybond. Hybond creates and uses an iso-surface to characterise the hydrogen bond interaction. However, this doesn't natively give physically intuitive bond strength scale in kJ/mol. A calibration curve of a simple hydrogen bond dimer is necessary to achieve this, which was attempted but led to unreasonable results that could not be resolved within the timeframe permitted. Optimising a molecule that is representative of a protein environment with a hydrogen bond that can be separated at regular intervals with no other

intermolecular interactions is the challenge here. Also charged species need to be far enough away as to not influence the interaction site. This task will allow the values that are output from Hybond to represent the full interaction that is being characterise and make it a tool that is easy to use and can be applied to many different systems.

Develop Automatic Equivalent Environment Detection. There are multiple environments that are present in protein structures such as NH_3^+, NH_2 and $COO^{-\cdot}$ These are the current known environments that allow a hydrogen bonding interaction to switch freely between. There has been a manual correction added to the latest version of Hybond as a part of this research, however automatic detection of these environments would add to the robustness of the program and may allow other environments that are created from the side chains of the proteins to be detected. This is important as there could be many of these interactions present when characterising the entirety of the molecules interactions.

REFERENCES

1. *Inhomogeneous Electron Gas.* **Kohn, P. Hohenberg and W.** 3B, Paris, France : Physical Review, 1964, Vol. 136 .
2. *Insights into Current Limitations of Density Functional Theory.* **A.J. Cohen, P. Mori-Sanchez and W. Yang.** 5890 (792-794), s.l. : Science Mag, August 8th, 2008, Vol. 321.
3. *Metabolic- Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks.* **L. Noda-Garcia, W. Liebermeister, D.S. Tawfik.** 187-216, s.l. : Annual Reviews, Biochem, 2018, Vol. 87.
4. *Eukaryotic DNA polymerases in DNA replication and DNA repair.* **Burgers, P.M.J.** 218-227, s.l. : Springer-Verlag, Chromosoma, 1998, Vol. 107.
5. *Dynamic Personalities of Proteins.* **Kern, K. Henzler-Wildman and D.** s.l. : Nature, December 2007, Vol. 450.
6. *Hydrogen bonding in globular protiens.* **Hubbard, E.N. Baker and R.E.** 97-179, s.l. : Prog. Boiphysical Molecular Biology, 1984, Vol. 44.

7. *Hydrogen Bonds in Proteins: Role and Strength.* **Haider, R.E. Hubbard and M.K.** York, UK : John Wiley & Sons, 2010, Vol. Encyclopedia of Life Sciences.

8. *Definition of the Hydrogen Bond (IUPAC Recommendations 2011).* **E. Arunan, G.R. Desiraju, R.A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D.C. Clary, R.H. Crabtree, J.J. Dannenberg, P. Hobza, H.G. Kjaergaard, A.C. Legon, B. Mennucci and D.J. Nesbitt.** 8 (1637-1641), s.l. : Pure Appl. Chem., 2011, Vol. 83.

9. **McRee, D.E.** Computational Techniques. *Practical Protein Crystallography (Second Edition).* California : Academic Press, 1999.

10. *Short strong hydrogen bonds in proteins: a case study of rhamnogalacturonan acetylesterase.* **A. Langkilde, S.M. Kristensen, L.L. Leggio, A. Molgaard, J.H. Jensen, A.R. Houk, J.N. Poulsen, S. Kauppinen and S. Larsen.** 851-863, s.l. : Acta Crystallographica Section D, Biological Crystallography , 2008, Vol. 64.

11. *3-10 helicies in channels and other membrane proteins.* **R.S. Vieira-Pires, J.H. Morais-Cabral.** 585-592, Porto, Portugal : J Gen Physiol, 2010, Vol. 136.

12. *Occurrence, conformational features and amino acid propensities for the pi-helix.* **Al-Karadaghi, M.N. Fodje and S.** 5 (353-358), May, 2002, Vol. 15.

13. *Short hydrogen bonds in proteins.* **Vishveshwara, S. Rajagopal and S.** 1819-1832, Bangalore, India : FEBS Journal, 2005, Vol. 272.

14. *Hydrogen Bonds: Simple After All?* **Pinney, D. Herschlag and M.M.** California, Department of Biochemistry, Stanford University : Biochemistry, 2018, Vols. 57 (3338-3352).

15. *NCIPLOT: A Progam for Plotting Noncovalent Interaction Regions.* **J. Contreras-Garcia, E.R. Johnson, S. Keinan, R. Chaudret, J-P. Piquemal, D.N. Beratan and W. Yang.** 625-632, s.l. : ACS Publications, J. Chem. Theory Comput, 2011, Vol. 7.

16. *Performance of recently developed kinetic energy density functionals for the calculation of hydrogen binding strengths and hydrogen-bonded structures.* **A.D. Rabuck, G.E. Scuseria.** 439-444, s.l. : Springer-Verlag, Thcor Chem Acc, 2000, Vol. 104.

17. *Kinetic Energy Density as a Predicitor of Hydrogen-Bonded OH-Stretching Frequencies.* **J.R. Lane, A.S. Hansen, K. Mackeprang and H.G. Kjaergaard.** 18 (3452-3460), s.l. : ACS Publications, Journal of Physical Chemistry, 2017, Vol. 121.

18. *Topological Descriptors of the Electron Density and the Electron Localization Function in Hydrogen Bond Dimers at Short Intermonomer Distances.* **Pacios, L.F.** 1177-1188, Madrid, Spain : J. Phys. Chem,, 2004, Vol. 108.

19. *NAMD - a Parallel, Object-Oriented Molecular Dynamics Program.* **M. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. Kale, R.D Skeel, K. Schulten.** Illinois : Theoretical Biophysics Group, 1996.

20. *GROMACS: Fast, Flexible, and Free.* **D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark and H.J.C. Berendsen.** 16 (1701-1718), s.l. : Wiley Online Library, Journal of Computational Chemistry, 2005, Vol. 26.

21. *GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories.* **A.P. Eichenberger, J.R. Allison, J. Dolenc, D.P. Geerke, B.A.C. Horta, K. Meier, C. Oostenbrink, N. Schmid, D. Steiner, D. Wang and W.F. van Gunsteren.** 10 (3379-3390), s.l. : ACS Publications, Journal of Chemical Theory and Computation, 2011, Vol. 7.

22. *CHARMM: The biomolecular simulation program.* **B.R. Brooks, C.L. Brooks III, A.D. Mackerell Jr plus 32 others.** 10 (1545-1614), s.l. : Wiley Online Library, Journal of Computational Chemistry, 2009, Vol. 30.

23. *The Amber biomolecular simulation programs.* **D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A.Onufriev, C. Simmerling, B. Wang, R.J. Woods.** 16 (1668-1688), s.l. : Wiley Online Library, Journal of Computational Chemistry, 2005, Vol. 26.

24. *PyContact: Rapid, Customizable and Visual Analysis of Noncovalent Interactions in MD Simulations.* **M. Scheurer, P. Rodenkirch, M. Siggel, R.C. Bernardi, K. Schulten, E. Tajkhorshid and T. Rudack.** 577-583, s.l. : Biophysical Society , 2018, Feb 6, Vol. 114.

25. *RIP_MD: a tool to study residue interaction networks in protein molecular dynamics.* **S. Contreras-Riqelme, J-A. Garate, T. Perez-Acle and A.J.M. Martin.** Chile : PeerJ, 2018.

26. *MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations.* **R.J. Gowers, M. Linke, J. Barnoud, T.J.E. Reddy, M.N. Melo, S.L. Seyler, J. Domanski, D.L. Dotson, S. Buchoux, I.M. Kenney, O. Beckstein.** s.l. : SCIPY, 2016, Vol. 15th.

27. *MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations.* **N. Michuad-Agrawal, E.J. Denning, T.B. Woolf, O. Beckstein.** 2319-2327, s.l. : Wiley Online Library, 2011, Vol. 32.

28. **Schipper, Daniel.** Bonder Promolecular. *GitHub.* [Online] [Cited: 03 07, 2022.] https://github.com/danielaschipper/Bonder-promolecular.

29. **Arcus, V.** *Meeting on Protein structures.* Hamilton, March 24, 2021.

30. *Proteopedia: Rossmann fold: A Beta-alpha-beta fold at dinucleotide binding sites.* **Hanukoglu, I.** 3 (206-209), s.l. : IUBMB Journals: Biochemistry and Molecular Biology Education, 2015, Vol. 43.

31. *The so far farthest reaches of the double jelly roll capsid protein fold.* **Raaji, C.S. Martin and M.J. van.** 181, s.l. : Virol J, 2018, Vol. 15.

32. **Gullingsrud, J.** CatDCD - Concatenate DCD files. *Theoretical and Compuational Biophysics Group.* [Online] [Cited: 03 11, 2022.] https://www.ks.uiuc.edu/Development/MDTools/catdcd/.

33. *Millisecond-scale molecular dynamics simulations on Anton.* **D.E. Shaw, R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young and M.M. Deneroff + 14 others.** November 2009, s.l. : Association for Computing Machinery, 2009, Vol. Article no.65.

34. *Millisecond dynamics of RNA polymerase II translocation at atomic resolution.* **D.A. Silva, D.R. Weiss, F.P. Avila, L. Da, M. Levitt, D. Wang and X. Huang.** Hong Kong and California : PNAS, 2014.

35. *Routine Access to Millisecond Time Scale Events With Accelerated Molecular Dynamics.* **L.C.T. Pierce, R. Salomon-Ferrer, C.A.F. de Oliveira, J.A. McCammon and R.C. Walker.** California : J. Chem. Theory Comput. , 2012, Vols. 8,9, 2997- 3002.

36. *Picosecond to Millisecond Structural Dynamics in Human Ubiquitin.* **K. Lindorff- Larsen, P. Maragakis, S. Piana and D.E. Shaw.** New York : J. Physs. Chem.B, 2016, Vols. 120, 33, 8313-8320.

37. *Dynamics of hydrogen bonds and vibrational spectral diffusion in liquid methanol from first principles simulations with dispersion corrected density functional.* **Chandra, V.K. Yadav and A.** Kanpur, Department of Chemisty, Indian Institute of Technology : Chemical Physics, 2013, Vols. 415 (1-7).

38. *Enzymes Dynamics During Catalysis.* **E.Z. Eisenmesser, D.A. Bosco, M. Akke, D. Kern.** 22 February, s.l. : ScienceMag, 2002, Vols. 295 (1520 - 1523).

39. *Quantitative Comparison of Flood Fill and Modifed Flood Fill Algorithms.* **Law, G.** 3, Califorina : International Journal of Computer Theory and Engineering, 2013, Vol. 5.

40. **L. Liberti, C. Lavor.** Euclidean Distance Geometry: An Introduction. *Euclidean Distance Geometry: An Introduction.* s.l. : Springer, Springer Undergraduate Texts in Mathematics and Technology, 2017.

41. **ArcCos to Calculate Angle .** ***Math Net.*** **[Online] [Cited: 03 11, 2022.] https://www.math.net/arccos.**

42. ***Pi-Hole spodium bonding in tri-coordinated Hg(II) complexes.*** **R.M. Gomila, A. Bauza, T.J. Moooibroek and A. Frontera. 7545-7553, Baleares, Spain : Royal Society of Chemistry, 2021, Vol. 50.**

43. ***Room Temperature Gibbs Energies of Hydrogen-Bonded Alcohol Dimethylselenide Complexes.*** **A. Kjaersgaard, J.R. Lane and H.G. Kjaergaard. 8427-8434, Copenhagen (Denmark) and Hamilton (New Zealand) : ACS Publications, 2019, Vol. 123.**

44. ***A new approach of moving average method in time series analysis.*** **Hansun, S. 1-4, s.l. : IEEE, 2013 Conference on New Media Studies, 2013.**

45. ***Energetics of hydrogen bonds in peptides.*** **S. Sheu, D. Yang, H.L. Selzle and E.W. Schlag. 22 (12683-12687), Taipei (Taiwan), Garching (Germany) : PNAS, 2003, Vol. 100.**

46. ***Tuning the Attempt Frequency of Protein Folding Dynamics via Transition-State Rigidification: Application to Trp-Cage.*** **R.M Abaskharon, R.M. Culik, G.A. Woolley and F. Gai. 521-526, s.l. : ACS Publications, 2015, Vol. 6.**

47. ***S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures.*** **J. Rezac, K.E. Riley and P. Hobza. 7,8 (2427-2438), Prague, Olomouc : ACS Publication, 2011, Vol. Journal of Chemical Theory and Computation.**

48. ***General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field.*** **Q. Wang, J.A. Rackers, C. He, R. Qi, C. Narth, L. Lagardere, N. Gresh, J.W. Ponder, J-P. Piquemal and P. Ren. 6 (2609-2618), s.l. : ACS Publications, J. Chem. Theory Comput., 2015, Vol. 11.**

49. ***Revealing Non-Covalent Interactions.*** **E.R Johnson, S. Keinan, P. Mori-Sanchez, J. ContrerasGarcia, A.J Cohen, W. Yang. 2010 May 12, J Am Chem Soc., pp. 6498-6506.**

50. ***Cystines and Disulfide Bonds as Structure- Forming Units: Insights From Different Domains of Life and the Potential for Characterization by NMR.*** **C. Wiedemann, A. Kumar, A. Lang and O. Ohlenschlager. Germany : Front Chem, 2020, Vol. 8:280.**

51. ***Protein Folding Guides Disulfide Bond Formation.*** **M. Qin, W. Wang and D. Thirumalai. New York and Nanjing : PNAS, 2105, Vols. 112 (36) 11241-11246.**

52. ***Occurrence, conformational features and amino acid propensities for the Pi-helix.*** **M.N. Fodje, A. Al-Karadaghi. 5 (353-358), May, 2002, Vol. 15.**

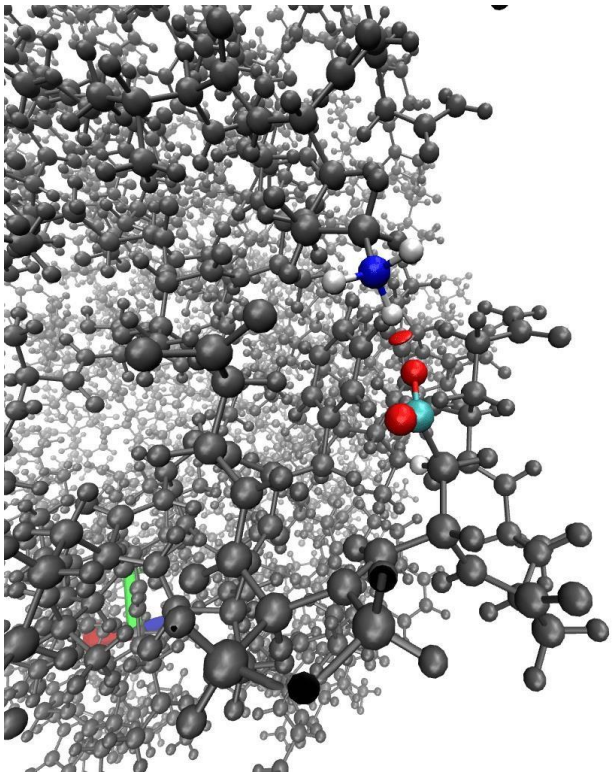
53. ***ELF: The Electron Localization Function.*** **A. Savin, R. Nesper, S. Wengert, T.F. Fassler. 17 (18081832), Germany : WILEY-VCH, 1997, September 17, Vol. 36.**

54. **NeSI Homepage.** ***New Zealand eScience Infrastructure.*** **[Online] [Cited: February 5, 2022.] https://www.nesi.org.nz/.**

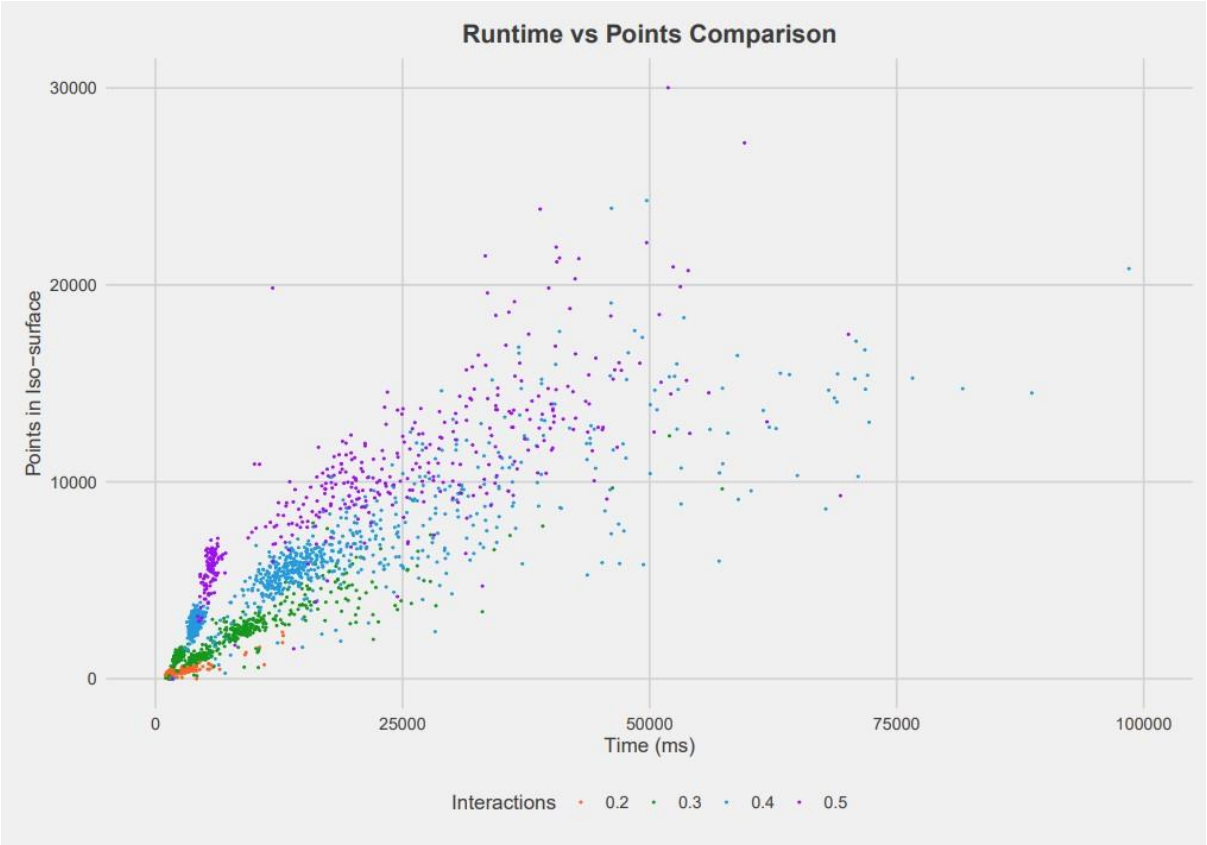
55. **NeSI Homepage.** ***New Zealand eScience Infrastrcutre.*** **[Online] [Cited: February 7, 2022.] https://www.nesi.org.nz/.**

56. ***ProteinNet: A standardized data set for machine learning of a protein structure.*** **AlQuraishi, M. 311, s.l. : BMC Bioinformatics, 2019, Vol. 20.**

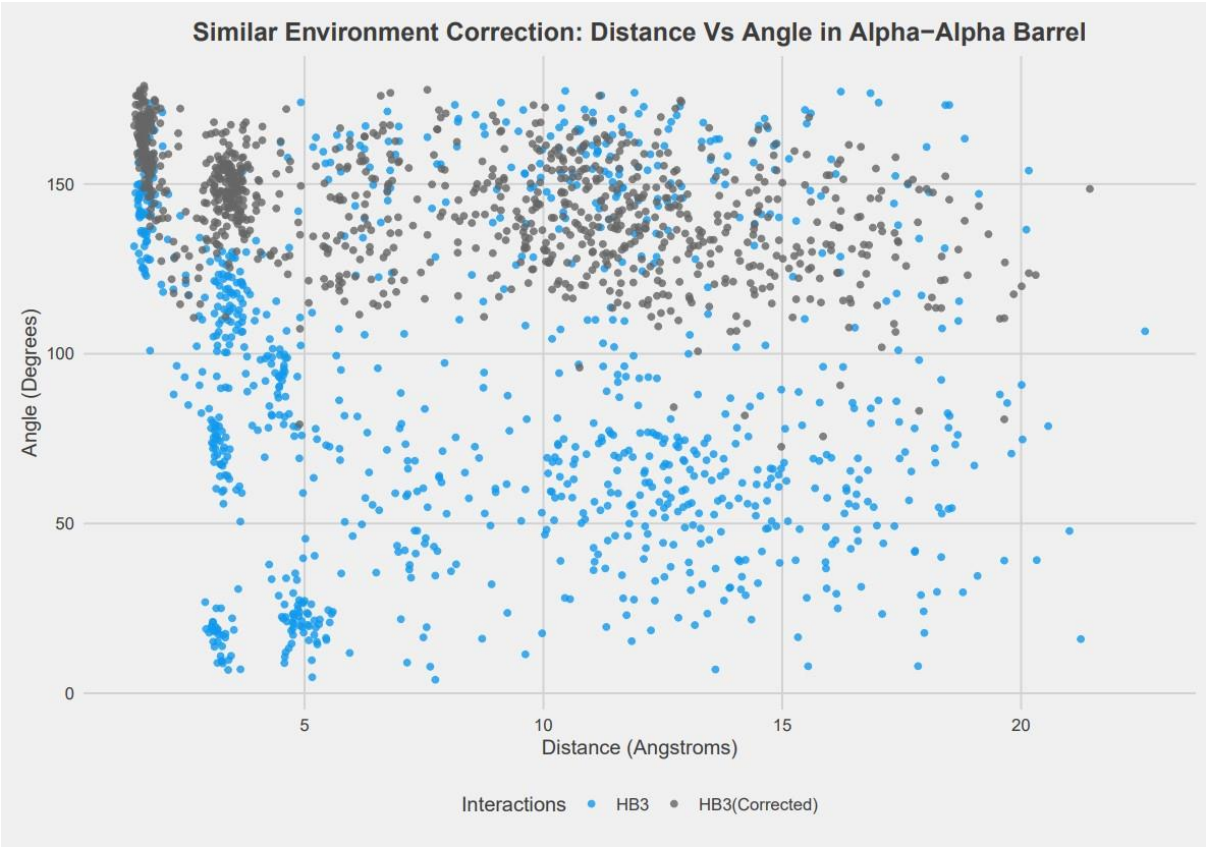
APPENDIX



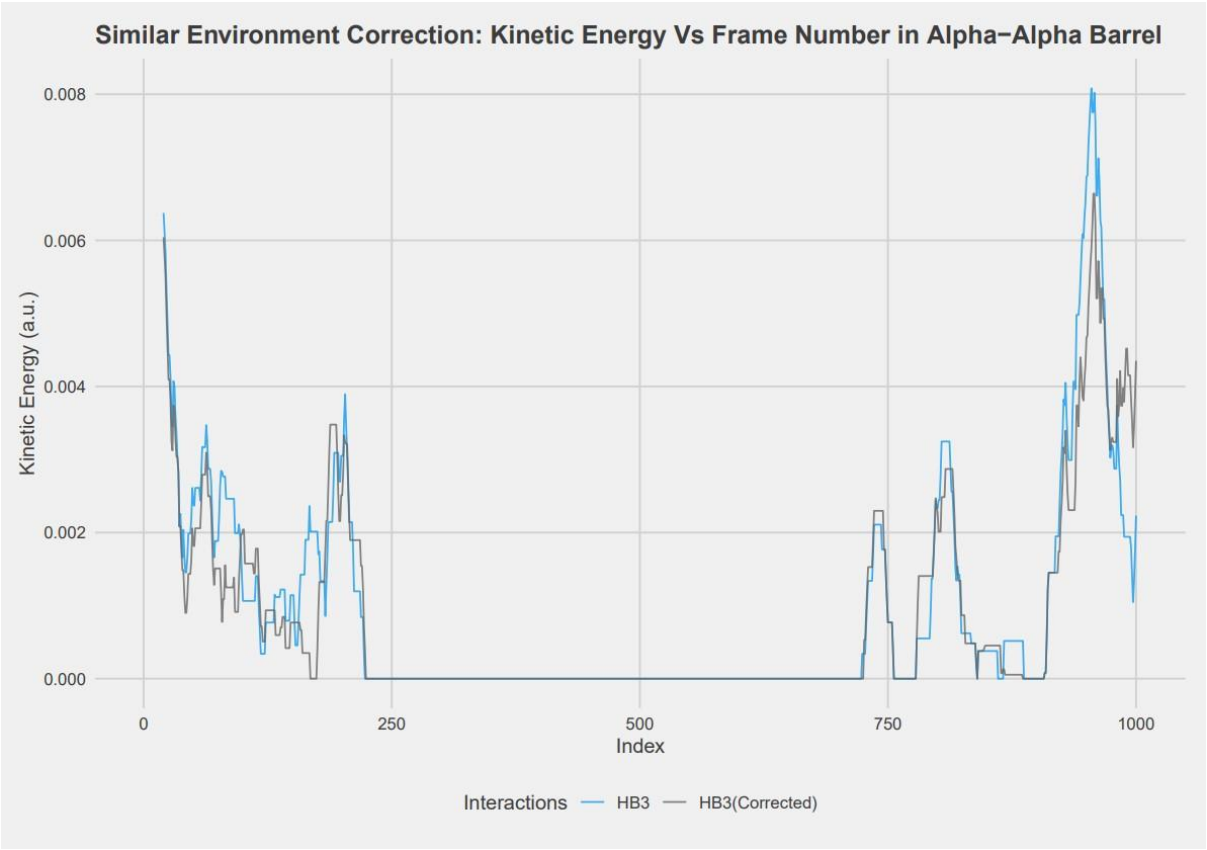
Appendix 1: Second Visualisation of the HB3 Interaction in the Alpha-Alpha Barrel



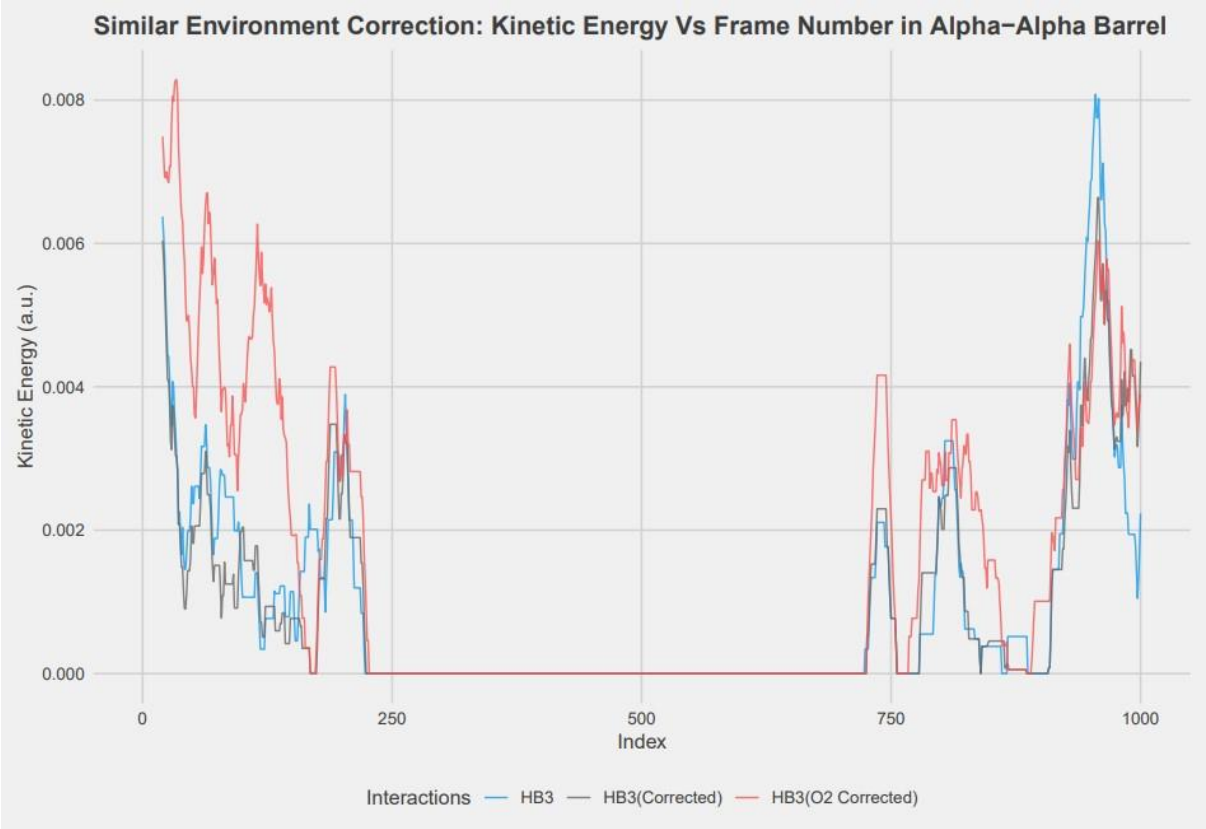
Appendix 2: Runtime vs Number of Points for All RDG values (excluding 0.1)



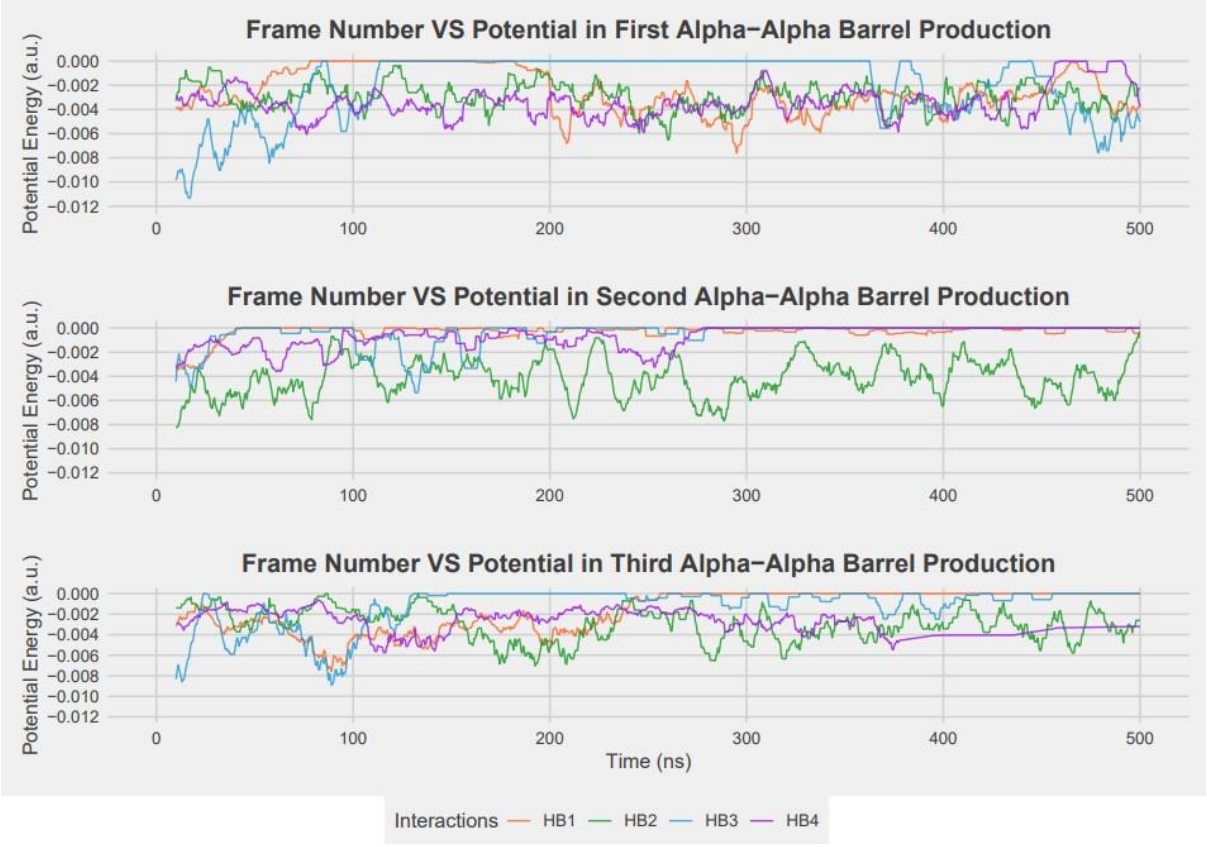
Appendix 3: Comparison of the NH3 Correction on the Bond Distance and Angle



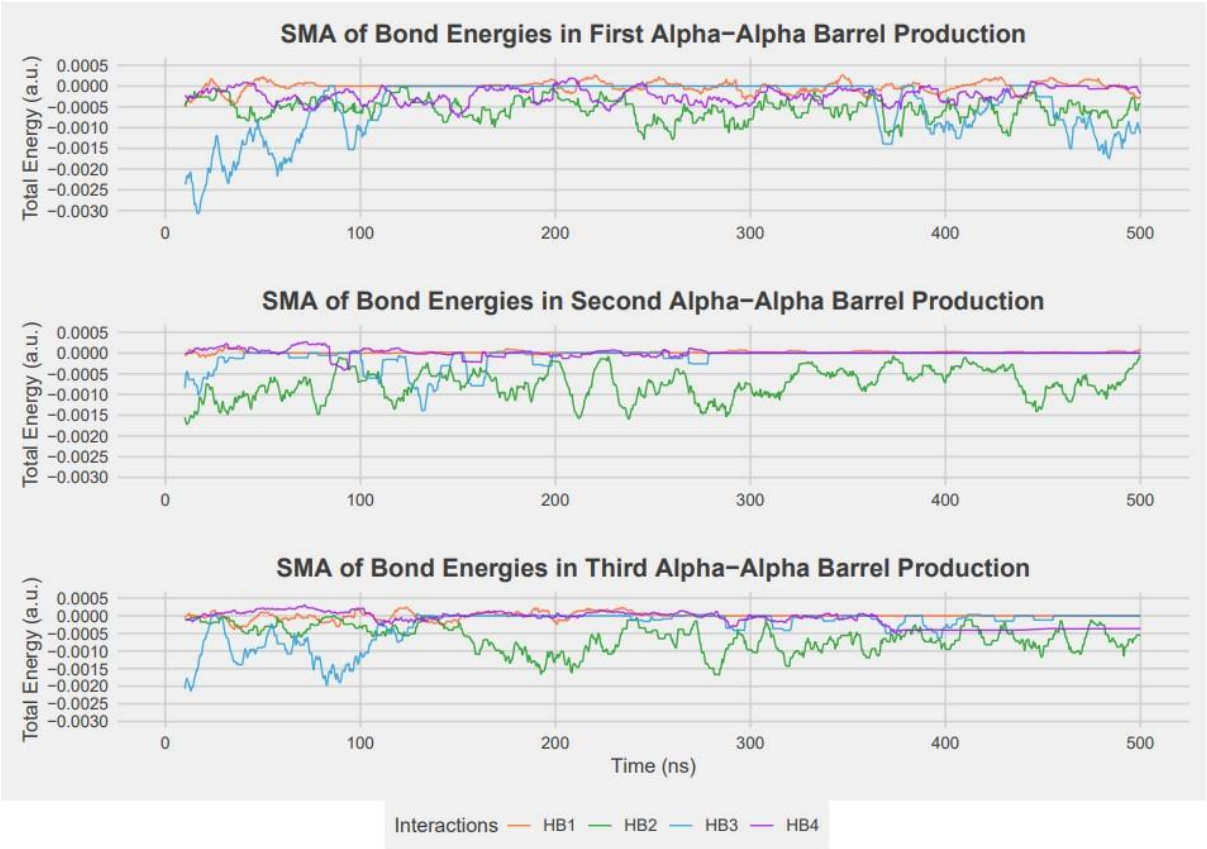
Appendix 4: Comparison of Original (Blue) and Corrected (Grey) Kinetic Energy



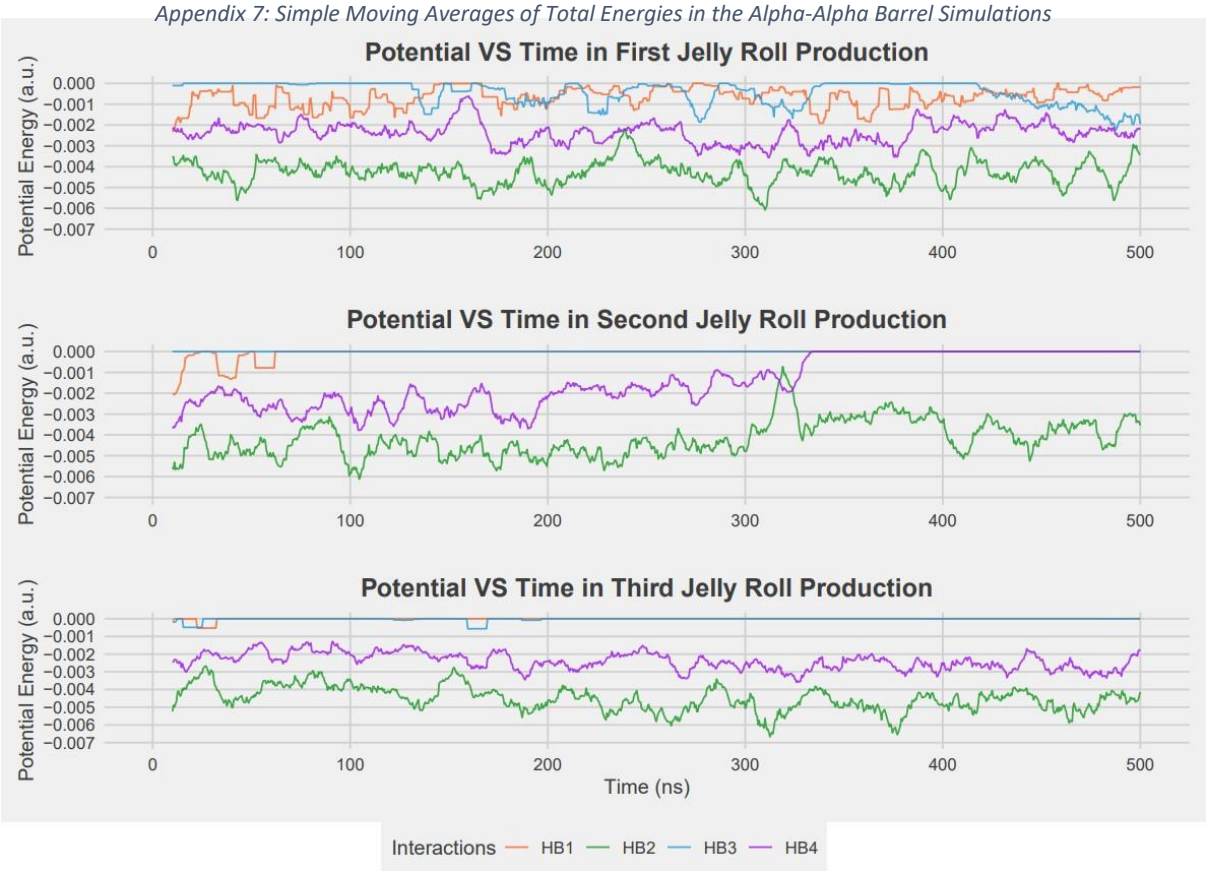
Appendix 5: Comparison of the Energies with each Level of Correction



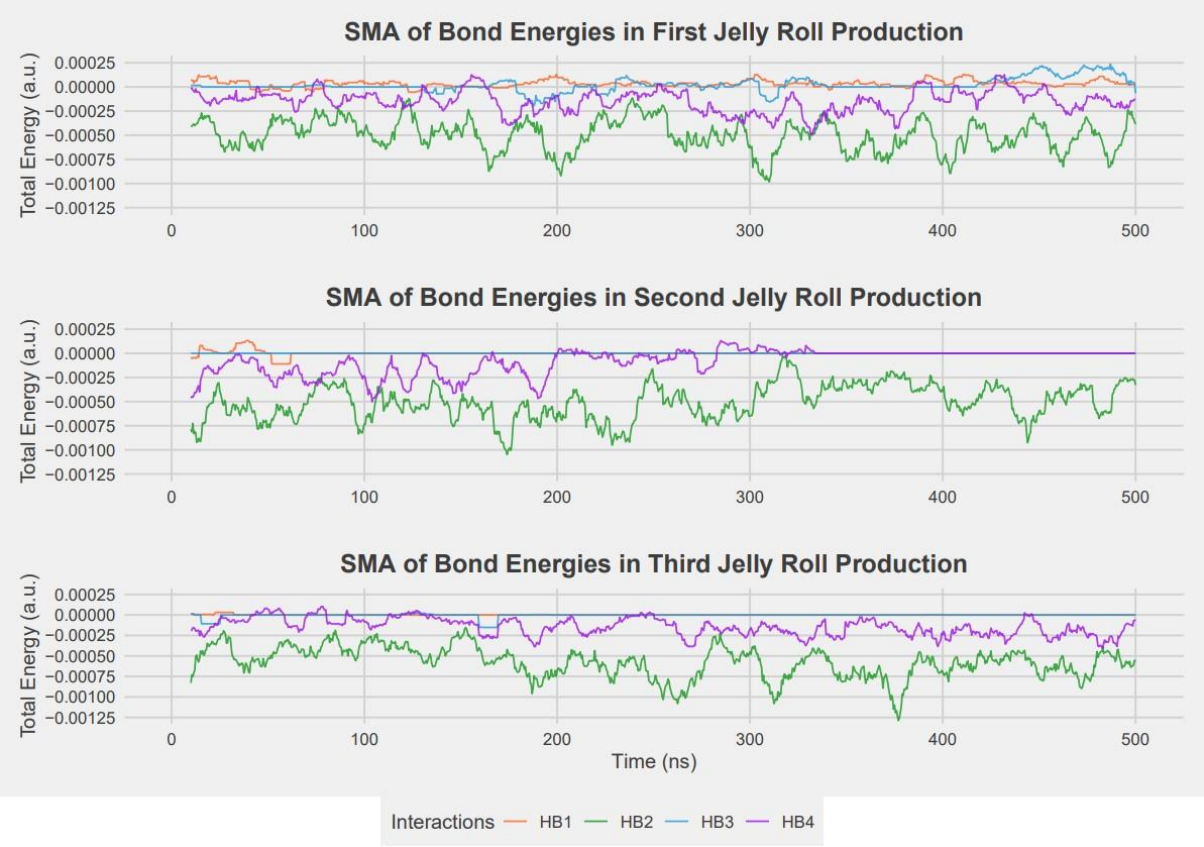
Appendix 6: Potential Energy Vs Time in Alpha-Alpha Barrel Simulations



Appendix 7: Simple Moving Averages of Total Energies in the Alpha-Alpha Barrel Simulations



Appendix 8: Potential Energy vs Time in the Jelly Roll Simulations



Appendix 9: Simple Moving Averages of Jelly Roll Productions