Understanding Nitrogenase: A Computational Model of the MoFe Protein from *Azotobacter vinelandii*

Barði Benediktsson

Faculty of Physical Sciences
University of Iceland
2015

Understanding Nitrogenase: A Computational Model of the MoFe Protein from *Azotobacter Vinelandii.*

Nítrógenasi: Tölvulíkan af MoFe próteini *Azotobacter vinelandii.*

Barði Benediktsson

15 ECTS thesis submitted in partial fulfillment of a Baccalaureus Scientiarum degree in Biochemistry and Molecular Biology

Advisor
Ragnar Björnsson

Co-advisor
Egill Skúlason

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavík, May 2015

Understanding Nitrogenase: A Computational Model of the MoFe protein from *Azotobacter Vinelandii*

15 ECTS thesis submitted in partial fulfillment of a B.Sc. degree in Biochemistry and Molecular Biology

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
VRII, Hjarðarhagi 2-6
107, Reykjavik

Telephone: 525 4000

Bibliographic information:
Barði Benediktsson, 2015, *Understanding Nitrogenase: A Computational Model of the MoFe Protein from Azotobacter vinelandii*, B.Sc. thesis, Faculty of Physical Sciences, University of Iceland.

Printing: Háskólaprent, Fálkagata 2, 107 Reykjavík
Reykjavík, Iceland, May 2015

Útdráttur

Þrátt fyrir að andrúmsloftið samanstandi um það bil 78% af köfnunarefni, þá er einungis lítill hluti lífheimsins sem getur nálgast köfnunarefni beint úr andrúmsloftinu. Það gera niturbindandi lífverur en breyta lífefnafræðilega óvirku nitri í lífefnafræðilega virkt ammóníak sem er lífveran getur notað t.d. í próteinsmíð. Þetta ferli fer fram í einu efnahvarfi hvötuðu af ensíminu nítrógenasa en hvarfgangurinn fyrir þetta hvarf er óþekktur þrátt fyrir margra ára rannsóknir. Vísindamenn hafa þó staðsett líklega staðsetningu hvarfsins í FeMo hjálparþætti í MoFe próteini nítrógenasa. Í þessu verkefni var smíðað sameindafræðilegt líkan af MoFe próteininu út frá þekktri kristalbyggingu með því markmiði að hægt yrði að nota líkanið sem grunn fyrir frekari framhaldsrannsóknir tengdum hvarfgangi nítrógenasa.

Smíði líkansins fór fram í nokkrum skrefum og til að rannsaka stöðugleika þess var það hermt við fast hitastig, rúmmál og atómfjölda. Þegar stöðug 5ns hermun á MoFe próteininu fékkst þá var klippt út úr handahófskenndum tímaramma um það bil 40000 atóma kúlulaga klasamódel með MoFe hjálparþáttinn í miðjunni. Þetta klasamódel er það sem verður notað í frekari QM/MM rannsóknum á hvarfgangi nítrógenasa.

Í einni hermun fékkst áhugaverð hreyfing á amínósýruleifinni Gln432 í A keðju MoFe próteinsins nálægt járn jón. Snerist amíð hliðarhópur glútamínsins í u.þ.b. 180° þannig að vatnssameind slapp fram hjá. Þegar vatnssameindin var farin þá snerist Gln432 aftur um 180°. Gæti þetta bent til mögulegra vatnsganga en þarfnast þó frekari rannsókna.

Abstract

The atmosphere contains approximately 78% nitrogen but only relatively small part of the biosphere can use it. The organisms that can are called diazotrophs and fix the biochemically inactive nitrogen into the biochemically active nitrogen source ammonia. This reaction is catalyzed by the complex metalloenzyme nitrogenase using the metallocofactor FeMo cofactor that resides in the MoFe protein. The mechanism behind this reaction has proven elusive and is not understood even after years of research. In this research project an all-atom molecular mechanics model of solvated MoFe protein was built from the high-resolution crystal structure.

The built model went through a series of preparation steps and rigorously tested by molecular dynamics simulations at constant temperature, volume and number of molecules to determine its stability. When a stable 5ns simulation was achieved a 40000 atom sphere centered on FeMo cofactor was cut-out. This cluster model will be used for further QM/MM studies on the reaction mechanism.

In one of the simulations performed, an interesting movement of Gln432 in chain A near an iron ion of a MoFe protein was observed. The side chain amide group turned approximately 180° which caused a water molecule to move past it. When the water molecule had passed by, it returned back to its resting state. This could point at possible water tunnel but further research must be performed.

Contents

List of Figures

List of Tables

Abbreviations

MD	Molecular dynamics
MM	Molecular mechanics
QM	Quantum mechanics
QM/MM	Quantum and molecular mechanics
MgATP	Magnesium adenosine trisphosphate

Acknowledgements

I would like to thank my advisor Ragnar Björnsson for supporting and helping me throughout this research project. I would also like to thank Bjarni Ásgeirsson and Egill Skúlason for putting aside time to proofread.

I would like to thank my family and my girlfriend Sunna Björnsdóttir for their continued support throughout my education.

1 Introduction

1.1 Nitrogenase

Nitrogen is one of the four most abundant elements that appear in living organisms with the other three being hydrogen, carbon and oxygen. Obviously, this makes nitrogen an extremely important element in the biosphere. In its atmospheric form nitrogen appears as two nitrogen atoms bonded with a triple bond to make a dinitrogen molecule which makes up 78% of the atmosphere. In its dinitrogen form it is very unreactive with a relatively high activation energy barrier for any biochemical reaction and is subsequently unusable by most organisms. Thus, the need for changing atmospheric nitrogen into biological active nitrogen is great and is the process commonly referred to as nitrogen fixation.

Nitrogenases are nature´s solution for atmospheric nitrogen fixation as the enzymes catalyze a reaction that yields two molecules of ammonia for each molecule of dinitrogen. Organisms that are able to fix atmospheric nitrogen are commonly referred to as diazotrophs, which is an umbrella term for prokaryotes that do not need a source of fixed nitrogen. All diazotrophs contain one, two or three types of nitrogenase and are they categorized by their cofactors. These three groups are the molybdenum containing nitrogenase (Mo-nitrogenase), the vanadium containing nitrogenase (V-nitrogenase) and iron-only nitrogenase (Fe-nitrogenase) (Hu & Ribbe, 2015). The cofactor structure is believed to be almost identical between the three groups apart from the identity of a single metal ion in the cofactor (Mo, V or Fe). In terms of catalytic activity, Mo-nitrogenases are the most active with V-nitrogenases coming in second place and the Fe-nitrogenases coming in last place. In an organism that can synthesize all three nitrogenases the metal ion is the only differentiating factor while the protein itself contains the same amino acid residues. The reason why a single organism can synthesize all types of nitrogenase may be an evolutionary response to scarcity of a given metal ion at a particular time point (Bothe H., Newton W. E., & Ferguson S. J., 2007). The nitrogenase that is studied here is the Mo-nitrogenase from *Azotobacter vinelandii*.

The structure of a whole Mo-nitrogenase from *A. vinelandii* is composed of eight subunits. Four of the subunits make up the $\alpha_2\beta_2$ heterotetramer MoFe protein and the other four make up two units of the homodimer Fe protein as can be seen in figure 1. In a single αβ subunit is a FeMo cofactor [7Fe-9S-Mo-C-*R*-homocitrate], P-cluster [8Fe-7S] and a recently discovered iron ion whose role in the MoFe protein is poorly understood but most likely plays a at least part in the stability of the MoFe protein. In a single Fe protein is one 4Fe-4S cluster and two binding sites for MgATP (Seefeldt, Hoffman, & Dean, 2009).

The reaction mechanism how nitrogenase reduces atmospheric dinitrogen to two molecules of ammonia has eluded scientists for decades and is as of today poorly understood. The chemical equation for the reaction is known and looks easy enough but does though contain a mystery. Why a single hydrogen molecule is made as a byproduct and what part it plays in the reaction is completely unknown (Hoffman, Lukoyanov, Yang, Dean, & Seefeldt, 2014).

$$N_2 + 8H^+ + 16MgATP + 8e^- \rightarrow 2NH_3 + H_2 + 16MgADP + 16P_i$$

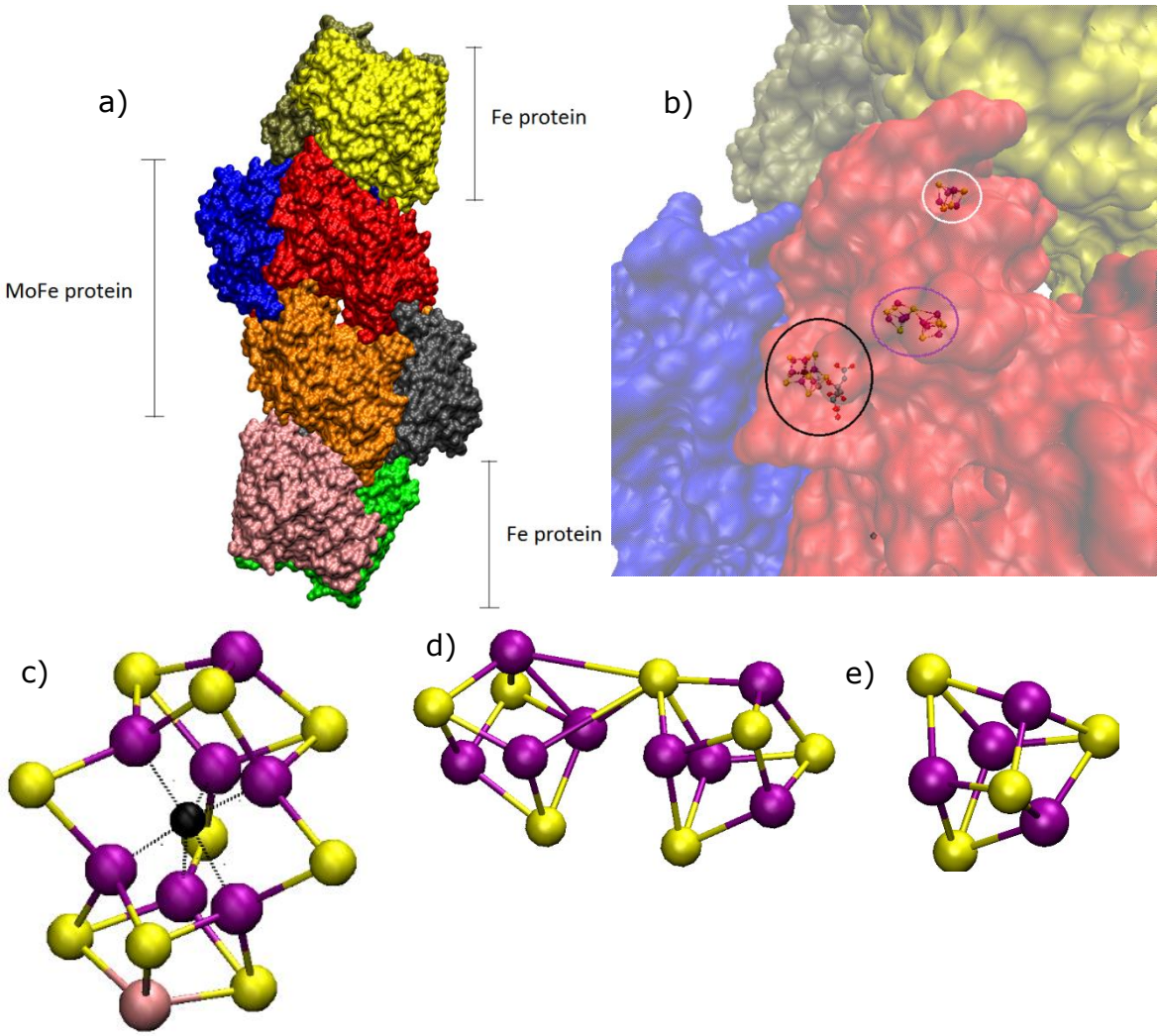

Figure1: a) Nitrogenase and its eight subunits. The red, blue, orange and dark grey area are the MoFe protein while the yellow, light grey, pink and green areas are the Fe protein. b) Close up shot of the upper part of the figure in a) where the Fe protein and MoFe protein meet. The surface area has been made transparent so the position of the cofactors can be seen. The FeMo cofactor and homocitrate is positioned inside the black circle, the P-cluster is positioned inside the light pink circle, the 4Fe-4S cluster inside the white circle and the little dot near the bottom of the image is the iron ion. Figures are created using VMD and the display option surf. Surf simulates the surface as if a probe would scan over the surface (Varshney A., Brooks F. P. Jr., & Wright W. V., 1994). The three cofactors are c) the FeMo cofactor of the MoFe protein without homocitrate, d) the P-cluster of the MoFe protein, e) the 4Fe-4S cofactor of the Fe protein. Purple colored atoms are iron, yellow colored atoms are sulfur, the pink atom is molybdenum and the black atom is carbon. The pdb file of nitrogenase from A. vinelandii with the code 1N2C is used to generate the image (Schindelin, Kisker, Schlessman, Howard, & Rees, 1997).

Although the reaction mechanism is not known the mechanism behind the electron transport chain that supplies electrons for the reduction of dinitrogen is better understood. One theory of electron transport to the FeMo cofactor is the so called deficit spending model (Duval et al., 2013). It assumes that the electron transport takes place in two steps. In the first step a reduced Fe protein (+1 charge) with two MgATP bound to its nucleotide binding sites, associates with a MoFe protein. This association somehow causes electron transport in the MoFe protein from the P-cluster in its resting state to the FeMo cofactor in

its resting state. This results in an oxidized P-cluster and a reduced FeMo cofactor. In the second step, an electron is transferred from the reduced Fe protein to the oxidized P-cluster resulting in the P-cluster returning to its original resting state and the Fe protein to its oxidized state (+2 charge). This electron transfer is coupled to the hydrolysis of two MgATP bound to the Fe protein with two P_i being released and the Fe protein disassociating from the MoFe protein.

1.2 Computational chemistry and molecular mechanics

Calculations and simulations in computational chemistry can be categorized into three groups. The first group is based on molecular mechanics (MM) and models molecules by calculating energy potentials between two or more atoms. The energy potentials can then be used to calculate forces which can then in turn be used to model dynamic behavior of molecules through a molecular dynamics simulation. It is relatively simple to describe a system using MM as the potential energy of a system is calculated between atoms and is the sum of covalent- and non-covalent bonding energies, all described by classical potentials. Because of how relatively simple MM calculations are, it is possible to simulate systems of up to hundreds of thousands of particles in size. However, the biggest drawback is that MM simulations cannot describe chemical reactions and are thus mainly used for research concerning conformational changes. In this research project, MM calculations were used exclusively.

The second group contains quantum mechanical (QM) calculations which utilize either methods based on Wavefunction Theory or the Density Functional Theory to describe a system. Calculations utilizing Wavefunction Theory can be used for calculations involving up to dozens of atoms and are computationally expensive, that is, take long time whilst utilizing powerful computing solutions. Density Functional Theory methods can handle up to few hundred of atoms but is also computationally expensive. As quantum mechanics describe the relation between electrons and nuclei it is possible to predict chemical reactions directly.

The third group is a hybrid method and utilizes both molecular mechanics and quantum mechanics and is commonly referred to as QM/MM. When studying a large system e.g. an enzyme, it is possible to describe a chemical reaction in a reaction center using QM while still accounting for the large enzyme environment using MM. It is the ultimate goal of the nitrogenase research project in Ragnar Björnsson's group to develop a QM/MM model of the MoFe protein.

1.2.1 Molecular Mechanics

As mentioned above, MM calculations are relatively inexpensive computationally when compared to QM calculations. The energy potential of a particle is calculated as the sum of covalent and non-covalent interactions between particles in the system. The covalent energy potential is the sum of bond stretching (described by a harmonic potential between atoms), angle distortion (also described by a harmonic potential) and dihedral strain (usually described as a periodic function). The non-covalent energy potential is described by classical electrostatic forces from Coulomb's law and van der Waals forces are

expressed as Lennard-Jones potential. The total potential energy in a MM system can be expressed with a single equation as a sum of all these energy potentials (Jensen, 2007):

$$E_{tot} = \sum_{Bonds} k_d(d - d_0)^2 + \sum_{Angles} k_\theta(\theta - \theta_0)^2 + \sum_{Dihedrals} k_\phi(1 + \cos(n\phi + \delta)) + \sum_{van\ der\ Waals} 4\varepsilon_{AB}\left(\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right) + \sum_{Electrostatic} \frac{1}{4\pi\varepsilon_0} * \frac{q_A q_B}{r_{AB}} \quad \textbf{(1)}$$

Where k_d, k_θ and k_ϕ are force constants, d_0 and θ_0 are equilibrium constants, q_A and q_B are atom charges and δ is a periodic constant. ε_{AB} and σ_{AB} are Lennard-Jones constants where ε_{AB} represents the depth of a potential well and σ_{AB} is the finite distance where the inter-particle potential is zero. Every single one of these constants has to be fitted to experimental data, estimated or derived from QM calculations. Equation 1 combined with a library of constants for different molecules is called a force field. Evaluation of the nonbonded parameters is usually the most time consuming part of a MM calculation as each particle will interact with all other particles in a large system. Terms that exclude the atoms that are the closest (e.g. connected through a chemical bond) and atoms that are further away than a user-defined distance are usually employed. The MM program used in this research project implements algorithms for fast evaluation of these terms.

Different atom parameters, particularly atom charges, are needed for the element in different functional groups e.g. a nitrogen atom in amide group has different partial charge compared to a nitrogen atom in an amine group and thus needs different parameters to reflect the different chemical behavior of such groups. Because of this, a forcefield for a protein involves a huge number of parameters to describe all the different atom types and interaction present and all need to have been derived some way or another. Even so there exist many different forcefields and there are handful of them especially designed for protein research. The one that will be used in this research project is the CHARMM36 force field (Best et al., 2012).

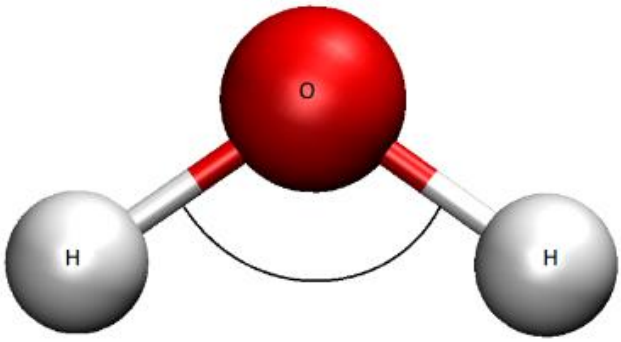

Figure 2: A TIP3P water model. The angle between HOH is 104.52° with bond length of 0.9572 Å. The oxygen carries a partial negative charge of -0.834 (charge of an electron is -1) and each hydrogen carries a positive charge of 0.417 which together makes the TIP3P water molecule carry a neutral charge.

There are different models available to describe water solvent in a system. Considerable time can be saved in calculations by using a simple water model while using complex water model will take longer. Implicit water models express the water molecules as a continuous medium and are time saving in contrast to the more computational expensive explicit water models where each and every water molecule is represented by a

forcefield expression. The explicit water molecule models differ from one to another as they can be 1-site up to 6-site (Skyner, McDonagh, Groom, van Mourik, & Mitchell, 2015). The site number depicts how many Lennard-Jones and electrostatic sites are used on the water the molecule. This means e.g. a 3-site water model includes non-covalent interactions on the hydrogen atoms and the oxygen atom but usually no bonded parameters. As with most solvent forcefields, intramolecular non-covalent interactions are forbidden as they dwarf compared to covalent interactions and would cause longer calculation times (Skyner et al., 2015). In this research project, the Transferable Intermolecular Potential 3P (TIP3P) water model is employed which is a 3-site explicit rigid water model (Jorgensen, Chandrasekhar, Madura, Impey, & Klein, 1983).

1.3 Molecular dynamics

Exploring the dynamical behavior of a chemical system that is describable by a forcefield can be performed with molecular dynamics. A molecular dynamics trajectory is created for the system by solving Newton's equation of motion numerically. Molecular dynamics simulations have a wide array of applications and have been used in various studies. Perhaps one of the biggest part that molecular dynamics has played in biochemistry is in a study from 1976 where molecular dynamics revealed what happens to retinal in a restrictive site during photoisomerization (Warshel, 1976) which was an important step in understanding the importance of protein flexibility.

While the force field describes the interaction between particles, a molecular dynamics algorithm describes the time-dependent behavior of the system. This is where molecular dynamics programs come in. A molecular dynamics program enables movement of atoms that depend on the energy potentials between atoms. This basically means that the program numerically integrates Newton's equation of motion for the particles in the system.

There exist a handful of molecular dynamics algorithms for simulating a moving system. One such popular algorithm is the Velocity Verlet integrator (Verlet, 1967). The forces at the start of the simulation are evaluated by calculating the first derivative of the energy potentials which is usually the most time consuming step. When the forces have been evaluated the acceleration of a given particle with a known mass is easily calculated by using the formula F = m*a. Because there is no way to know the initial velocity of a particle at the start of a simulation, the particle is assigned a random generated velocity from the Maxwell-Boltzmann distribution at simulated temperature. When all these factors have been estimated, the dynamics of the system can be simulated by the Velocity-Verlet integrator which is described by two formulae:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 \qquad \textbf{(2)}$$

$$\vec{v}(t+\ \delta t) = \vec{v}(t) + \frac{1}{2}[\vec{a}(t) + \vec{a}(t+\delta t)]\delta t \qquad \textbf{(3)}$$

where $\vec{r}$(t) is the position vector, $\vec{v}$(t) is the velocity vector and $\vec{a}$(t) is the acceleration vector at time t with the expression t+δt depicting position vector, velocity vector or acceleration vector timestep δt further into the simulation. The equation thus describes how the coordinates and velocities of the system at a timestep later, δt, can be calculated using

the information of the previous coordinates, velocities and acceleration (derived from atomic forces) at time t.

A molecular dynamics simulation should describe a certain thermodynamic ensemble. An ensemble constitutes of all possible microscopic states that a system can have but are identical in macroscopic or thermodynamic state (Nosé, 1984). There are three types of ensembles that are commonly used: Microcanonical (NVE) ensemble, canonical (NVT) ensemble and isothermal-isobaric (NPT) ensemble. In NVE simulations, the system is isolated from changes in number of molecules (N), volume (V) and energy (E), hence NVE. It is a simulation where total energy is conserved while potential- and kinetic energies are being exchanged constantly. This represents an isolated system.

In NVT simulations, the number of molecules, volume (V) and temperature are kept constant while energy is able to enter and escape the system. As in a lab where it is important to keep a heat sensitive reaction at a constant temperature with a thermostat, it is necessary to keep the temperature constant by employing a computational thermostat in NVT simulations. A reason for drifting temperatures in simulated systems may be caused by numeric errors during calculation of energies. This would cause the system to cool down if not for a thermostat. There exist a wide variety of thermostats for NVT simulations with the one used in this research project being the Nosé-Hoover thermostat (Hoover, 1985). In principle, the Nosé-Hoover thermostat functions as particles called chains that can couple with the whole system and keep the temperature stable by calculating the difference between the initial temperature and current temperature and adjusts the velocities of particles correspondingly in the next step.

In NPT simulations, the number of molecules, pressure (P) and Temperature are constant. This ensemble resembles best the environment found in the laboratory as the temperature is either kept constant by the environment or thermostat while the pressure is kept constant by the atmosphere. In NPT simulations, a thermostat is necessary as well as a barostat. An example of a barostat is the Parrinello-Rahman barostat (Parrinello & Rahman, 1982) which functions similarly to the Nosé-Hoover thermostat as an imaginary particle couples with every particle of the system to keep pressure constant.

1.4 Prior MM and QM/MM research

The main role of MM simulations in studies on nitrogenase in recent years has been to identify possible substrate and product channels. One such study identified a possible dinitrogen substrate channel (Smith, Danyal, Raugei, & Seefeldt, 2014) but a later study using xenon gas showed that there are more than one possible substrate channel (Morrison, Hoy, Zhang, Einsle, & Rees, 2015).

A QM/MM study on nitrogenase was performed in 2008 (Xie, Wu, Zhou, & Cao, 2008) when the interstitial atom of the FeMo cofactor had not yet been determined as carbide (C^{-4}) (Lancaster, Hu, Bergmann, Ribbe, & DeBeer, 2013) and the iron ion at the borders of αβ subunits was still wrongly identified as a calcium ion. A new QM/MM model based on the MM model in this project that uses the latest crystal structure from 2011 should be a considerably improved model and will be used for reaction mechanism studies**.**

1.5 What has been done in this research project?

The main goal of this research was the creation of stable and well-built MM model of MoFe protein of nitrogenase. In the course of creating such a model, the protonation states of amino acid residues that can have multiple protonation states have been studied. This is an important aspect of a computational protein model because wrong protonation state could possibly lead to wrong hydrogen bond formation or repulsive forces forming where such forces should not be as they could possibly change the global conformations of the system.

The MM model of the MoFe protein has been built through a series of steps. Firstly, protons were generated as the crystal structure does not contain any information on proton position. Secondly, the MoFe protein model was solvated in a user defined box. Thirdly, the forces between particles were minimized in a two-step process. Fourthly, the MoFe protein was simulated and scoured for artifacts that could have arisen in the previous three steps to determine, at least in part, the quality of the model

A suitable molecular dynamics setup was found for running reliable NVT simulation at equilibrium for extended periods of time. The reliability of the model and simulation was done by monitoring of root-mean-square deviations (RMSD), temperature values and by visual inspection of the model itself as the particles move. From a stable NVT MD trajectory, a spherical cluster model was created consisting of approximately 40000 atoms that will be created will be used for future QM/MM studies on the MoFe protein that will hopefully shed some light on the complex and as yet not understood reaction mechanism. Such study is though out of scope of this research project.

2 Materials and Methods

2.1 Computers and programs

Version 5.0.4 of the molecular mechanics software GROMACS was utilized for MM simulations of the MoFe protein (Pronk et al., 2013). For visualization of results from MM simulations, the version 1.9.2 of VMD was utilized (Humphrey, Dalke, & Schulten, 1996). For long MM simulations the Nordic computer cluster GARDAR and the local computer cluster SOL at the Science Institute, University of Iceland were used. The program PropKa (Rostkowski, Olsson, Sondergaard, & Jensen, 2011) was used to help determine protonation states of amino acid residues that can have variable charge on the R-group.

A modified CHARMM36 forcefield (Best et al., 2012) was used as the force field in GROMACS. Because there are no MM parameters available for the FeMo-cofactor and P-cluster, they were added manually. It was decided that the FeMo cofactor and the P-cluster would be constrained to their crystal structure positions, so only non-bonded parameters were needed. Lennard-Jones parameters for the sulfides in the co-factors were taken from CHARMM36 forcefield where all sulfur atoms connected to iron were set as atom type SM while all Fe and Mo metal ions in the cofactors contained no Lennard-Jones parameters. As for homocitrate, Lennard-Jones parameters and atom charges were modified parameters from a study on citrate (Wright, Rodger, & Walsh, 2013). The atom charges for FeMo cofactor and P-cluster were derived from natural population analysis charges from DFT calculations of the cofactor. These calculations used the BP86 functional and the def2-TZVP basis set using the QM program ORCA (Neese, 2012) and were carried out by Ragnar Björnsson.

2.2 The MoFe protein of nitrogenase and the pdb file 3U7Q

Only the MoFe protein of the enzyme nitrogenase was modelled in this study. The latest crystal structure of the native protein from 2011 was used (Spatzal et al., 2011) for initial preparation of the system, PDB code 3U7Q, that only contains crystallized MoFe protein and no Fe proteins. This crystal structure was used as a base for the model and only some minor changes were made. There were two Ca^{2+} ions on the borders where $\alpha_2\beta_2$ subunit meet, between chains B and D but, as mentioned before, a recent combined X-ray absorption and crystallography study (Zhang et al., 2013) revealed these ions to be ferrous ions (Fe ions) instead. The ions were thus modelled as Fe^{2+} ions instead of Ca^{2+} with 2 bound water molecules (constrained in all simulations) and one Fe^{2+} ion being bound to carboxylic oxygen of residues Asp353D, Asp357D Glu586B and carbonyl oxygen of Arg585B while the other is bound to carboxylic groups of Glu109D, Asp353B, Asp357B and carbonyl group of Arg108D (the oxygen atoms were constrained in all simulations). Missing amino acid residues at the beginning and the end of every chain were not

generated and the MoFe protein is thus modelled with amino acid residues as they appear in the crystal structure.

3 Results and Discussion

3.1 Building the system

3.1.1 Modifications of the 3U7Q pdb file.

In addition to changing the two Ca^{2+} ions at the intersection between chains B and D to Fe^{2+} ions, some residues had to be manually modified to better reflect how the molecular situation is in the MoFe protein. One P-cluster is bound in each αβ dimer through a series of three deprotonated cysteine residues in chain A/C Cys63, Cys89, Cys155 and three in chain B/D Cys71, Cys96, Cys154. These deprotonated cysteine residues have been given an overall negative charge with the sulfur atom being constrained in simulations.

Note that when the naming system A/C and B/D is used, it refers to amino acid residues with the same residue number in different subunits of the protein (subunits A and C are identical with B and D being also identical). For example, something made up of Cys63A and Cys71B in one heterodimer would constitute of Cys63C and Cys71D in the other heterodimer which would be described as two different structures of Cys63A/C and Cys71B/D.

X-ray crystal structures typically do not contain hydrogen atoms. It is not viable to simulate an unphysical deprotonated system and thus all protons have to be added either manually or by the GROMACS.

3.1.2 Protonation state of histidine, glutamate, aspartate and lysine.

GROMACS can guess coordinates for missing hydrogen atoms in the protein structure, including the position of protons of amino acid residues which have titrable R-groups (residues that can have variable protonation state on the R-group). The program needs to be told manually however about the protonation states of these titrable residues. For buried titrable residues, determination of their protonation state is not straightforward as the microenvironment differs often considerably from that of water. To assist with determination of protonation states of residues with titrable R-groups the program PropKa was used. The intuition of a biochemist is also an important tool in deciding protonation states as analyzing possible hydrogen bonds can reveal the protonation state of titrable R-groups.

PropKa calculates theoretical pKa values of the R-groups on the amino acid residues by taking into account the desolvation effect and intraprotein interaction. PropKa uses this information about the molecular environment to estimate pKa values of the titrable R-groups. This renders PropKa very useful for finding amino acid residues within the model that have very abnormal pKa values compared to usual pKa values of the amino acid in aqueous environment. The amino acid residues were manually inspected and protonation state was determined on a case-by-case basis. The amino acid residues that have to be examined carefully are glutamate, aspartate, lysine and histidine. In its default setting, GROMACS automatically assigns negative charge to aspartate and glutamate and

positive charge to lysine. GROMACS decides the protonation state of histidine in comparison to the environment but deciding the protonation state of histidine is complex as it can have four protonation states: No protons, protonated ND1, protonated NE2 and both ND1 and NE2 protonated. For reference for what ND1 and ND2 refers to, see figure 3.

In deciding the protonation state of glutamate in the model, only glutamate with a theoretical pKa value of 7 or more, calculated with PropKa, was given attention to. Of total of 142 glutamate residues, only ten glutamate residues were found to have high enough pKa values that they could possibly be protonated on the carboxylic group. These residues were Glu153A/C, Glu380A/C, Glu440A/C, Glu109B/D, and Glu231B/D. After careful examination it was determined that only Glu153A/C should be protonated as the residues are likely to be hydrogen bond donors to a nitrogen atom on Pro85A/C.

In deciding the protonation state of aspartate, the same methodology as in deciding the protonation state of glutamate was utilized. Only five amino acid residues had theoretical pKa values of over 7 according to PropKa analysis. These amino acid residues were Asp402A/C, Asp357B/D and Asp374B. Careful examination of the hydrogen bond system around these residues showed little to no signs that they should be considered protonated.

In deciding the protonation state of lysine, PropKa analysis showed that possible candidates for deprotonated lysine residues were Lys68A/C, Lys34B/D and Lys365B/D. Careful examination of possible hydrogen bonds resulted in the decision that only Lys365B/D could be deprotonated.

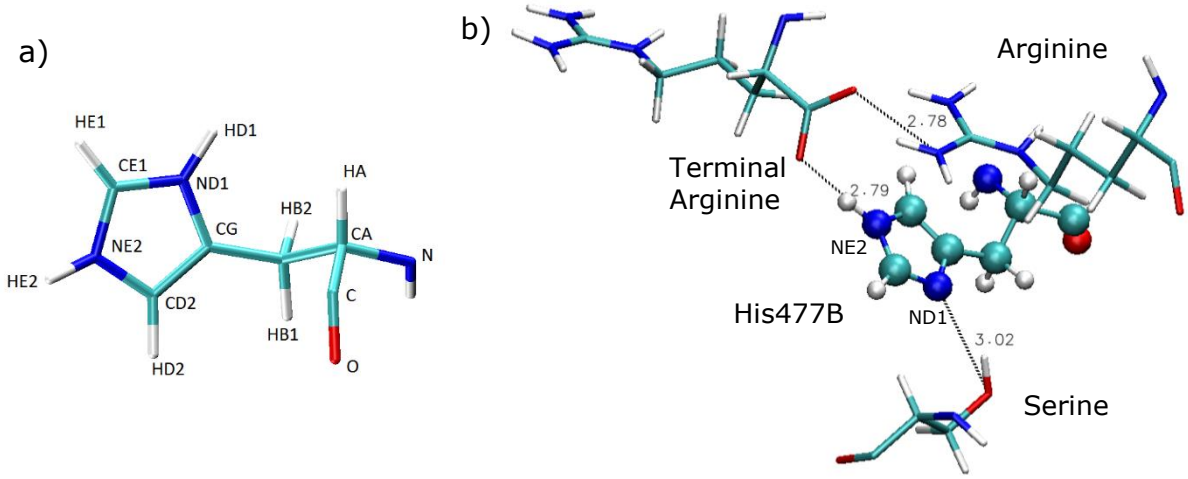

Figure 3: a) CHARMM atom labeling of histidine. b) Example showing how the protonation state of histidine residue His477B/D is decided. The residue is in close contact of two arginine residues and one serine residue. One of the arginine residue is the carboxy terminal of chain D. Here it can be shown that the protonation state of His477B/D must be the one with NE2 protonated as it can be a hydrogen bond donor to the carboxyl group, which according to PropKa is deprotonated. To further support this protonation state, the hydroxyl group on serine can possibly be a hydrogen bond donor for the ND1 nitrogen atom of His477B/D be and hydrogen bond acceptor and deprotonated.

In choosing the protonation state of histidine residues of the MoFe protein, only histidine residues in chain A and B were examined thoroughly. The information on protonation state obtained from chains A and B was directly used to determine protonation states of histidine in chains C and D. The reasoning for this is the fact that the MoFe

protein is a dimer of dimers and the protonation state should be consistent in the two dimers as the microenvironment should be the same.

The histidine residues that were protonated on ND1 were His31A/C, His196A/C, His274A/C, His285A/C, His451A/C, His185B/D, His297B/D, His359B/D, His363B/D and His519B/D.

The histidine residues that were protonated on NE2 were His80A/C, His83A/C, His195A/C, His362A/C, His383A/C, His442A/C, His106B/D, His193B/D, His311B/D, His392B/D, His396B/D, His396B/D, His429B/D, His457B/D, His477B/D, His478B/D, and His480B/D.

Only His91B/D was determined to be doubly protonated and no histidine residue was found likely to be unprotonated and is usually very seldom observed. When all the protonation states of the titrable residues had been safely determined, the system was protonated using GROMACS that is all missing proton coordinates from the X-ray structure were added. Before protonation, the system contained 16295 atoms and after protonation it contained 39566 atoms.

3.1.3 Solvation of the system

The original size of the system in the 3U7Q pdb file containing the MoFe protein was 10.926 nm * 7.704 nm * 12.114 nm. Thus, a 90 nm * 90 nm * 90 nm cubic and continuous system around the existing smaller system was created using GROMACS. By creating a larger system, it was possible to solvate the protein completely as the original system harbors little extra space which newly generated water molecules could fit in.

Crystallized water from the original crystal structure of the MoFe was kept as some crystallized water is important in protein stability e.g. the water molecules around the iron ion. As generated water molecules are not equal to the crystallized water molecules, a list of the crystallized water molecules was created to be able to distinguish between the two series.

The system was solvated in water, using GROMACS to generate water molecules. The resulting system has a density of 1029.76 g/L. Because the charge on the MoFe protein is -23, the charge in the system was neutralized by generating 23 sodium ions which randomly replace existing water molecules. Before solvation and neutralization of charge, the system contained 39566 atoms. After solvation and addition of Na^+ ions, the system contained 39566 atoms.

3.1.4 Energy minimizations.

Before any molecular dynamics simulation can be made, it is important that no force in the system due to atoms in close proximity is exceedingly large. The addition of hydrogen and solvation of the system prepares the system in a state that is far from being in equilibrium. Thus, it is important to do an energy minimization step where the system is relaxed and is this especially true for hydrogen atom coordinates. The steepest descent algorithm as implemented in GROMACS was used here. The algorithm is used to move atoms in close proximity from each other and by doing so the forces between atoms are made much smaller.

The steepest descent algorithm is a simple and fast algorithm that can significantly lower the forces between atoms in a relatively short period of time. However, it has slow convergence properties and cannot minimize forces of a large system perfectly. In this case, the steepest descent algorithm was only used to approximately relax the system

before MD simulations. First, all non-H atoms along with crystal water were constrained to the crystal structure coordinates and the forces on H atoms and generated water molecules were minimized in a 50 step process using the deepest descent algorithm (some protons and generated water molecules had to be manually moved to get rid of large initial forces as the steepest descent algorithm was unable to minimize the forces and was this done six time total with the minimization job being restarted after each modification). The system went from having the unrealistic potential energy of 100 MJ/mol before relaxation to having potential energy of -60.2 kJ/mol.

When the constrained system had been approximately relaxed in regards to protons, the whole system was relaxed constraining only the cofactors and the atoms on residues that connect to the cofactors. This was done by using the same steepest descent algorithm but due to problems only 4 energy minimization steps could be performed. This caused the potential energy to drop even further to -372.7 kJ/mol.

3.2 Determination of optimal run parameters for a NVT simulation.

GROMACS supports a wide variety of integrators, restraints and thermostat algorithms. These algorithms were systematically tested to find a reliable setup for production of long NVT molecular dynamics simulation. Five different simulations were performed with different settings tested as can be seen in table 1. Further information on run parameters can be found in the appendix as a text version of a .mdp file. In these simulations the RMSD deviation of heavy atoms of the MoFe protein was measured along with fluctuation in temperatures. For a stable simulation, the RMSD value (in nanometers) and temperature (in Kelvin) should converge with fluctuation within reasonable limits being allowed.

Table 1: Overview of parameters that were tested.

Simulation	Restrains	Timestep [fs]	Thermal coupling	Integarator	Nosé-Hoover chain
First	All-bonds	1	System	Leapfrog	1
Second	All-bonds	1	Prot-nonprot	md-vv	4
Third	All-bonds	2	System	md-vv	1
Fourth	All-bonds	1	System	md-vv	4
Fifth	All-bonds	1	System	md-vv	1

In the NVT MD simulations, the cofactors and atoms that connect residues to the cofactors were frozen in space. All bonds were constrained with the LINCS algorithm as implemented in GROMACS (Pronk et al., 2013). Non-covalent forces were cut off at 12 Å distance using a force-switch algorithm. The system was heated linearly in 50 ps from 50 K to 300 K with initial velocities being generated by the Boltzmann-Maxwell distribution at 50K. All simulations were set to be done over 1 ns but only the first, second and fourth succeeded due to computer problems with the third simulation ending at 486 ps and fifth simulation at 622 ps.

It was decided to use the same simulation parameters as were used in the fourth simulation for future simulations. This decision was based on two factors. Firstly, this

simulation has a slope in calculated RMSD trend line that is nearest zero of all the 1ns timestep simulations (see figure 4). Secondly, the R^2 value of the trend line is low which suggest very small correlation of the RMSD with the simulation time that also indicates equilibrium and a simulation essentially free of artifacts (see RMSD graphs in appendix B). Regarding temperatures, every simulation gave steady temperatures with averages of each individual simulation being 300.0 K with only the fifth significant number being different between simulations. Fluctuations in temperatures were few degrees which was deemed acceptable (graphs that show fluctuations of temperatures can be found in appendix B).

Every simulation underwent visual analysis and we checked for abnormal movements of amino acid residues using VMD with particular attention paid to the fourth simulation. In the fourth simulation, an unexpected flip of residue Gln432A was observed which will be discussed later.

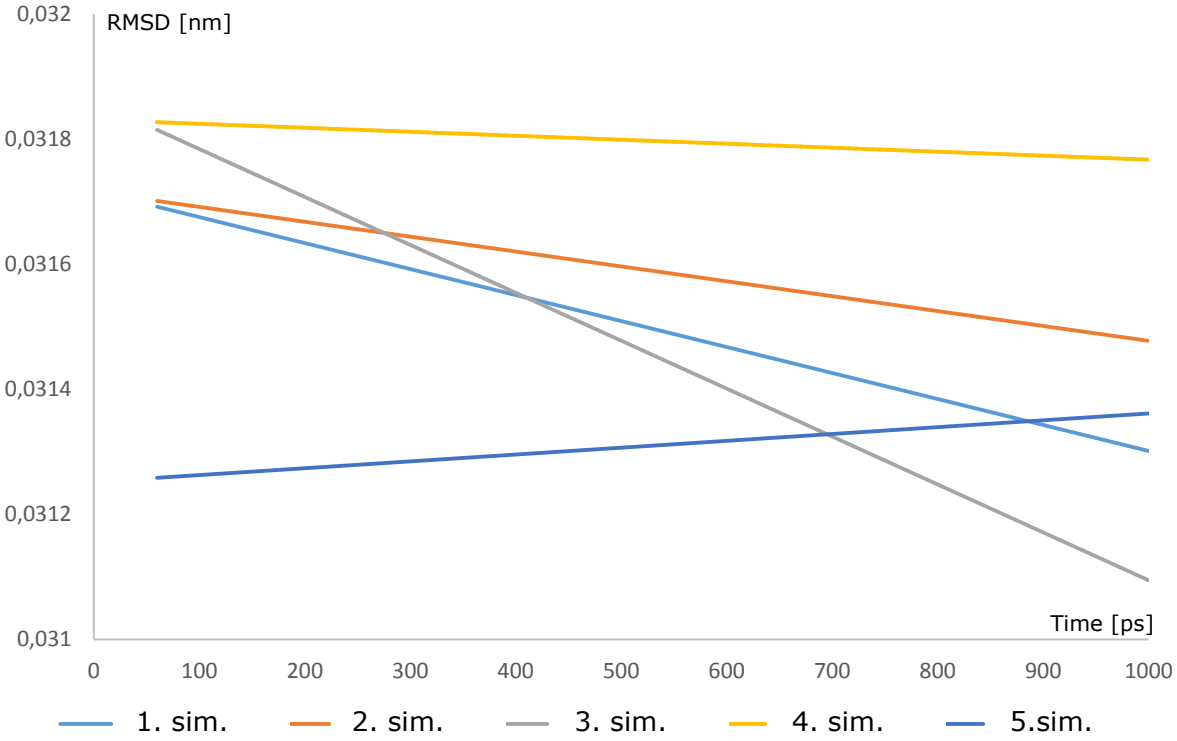

*Figure 4: Best-fit lines (trend lines) for RMSD values for the five different simulations performed. Slope of the first simulation is -4*10^{-7} with R^2= 0.208, second -2*10^{-7} with R^2=0.0834, third -8*10^{-7} with R^2=0.1583, fourth -6*10^{-}8 with R^2=0.0041 and fifth 1*10^{-7} with R^2=0.0068. Data from 60 ps to 1000 ps is used to plot the lines as the first 50 ps are used to warm the system from 50K to 300K.*

3.3 NVT simulation using lysozyme.

The equilibrated model of the MoFe protein has a low RMSD value compared to the original crystal structure but it is important to point out that this is at least partly due to the fact that FeMo cofactor, P-cluster, the iron ions and the atoms on amino acid residues connecting to the protein are artificially frozen in space. This restricts the conformational flexibility of the system. To confirm that the chosen simulation parameters are appropriate, an MD simulation was carried out on lysozyme using the same NVT parameters as for simulation four of the MoFe model. A lysozyme model was created in an analogous way to

the MoFe protein with all protonation states of the protein picked automatically by GROMACS. Lysozyme is a well-known enzyme, has no cofactors that need constraining due to lack of parameters and has been studied in many different MD simulations before (Lerbret et al., 2008), (Wei, Carignano, & Szleifer, 2012) and (Post et al., 1986) and was thus a good candidate for doing a comparative experiment. The lysozyme crystal structure used as a base for the model in this research has the PDB number 1AKI (Artymiuk, Blake, Rice, & Wilson, 1982).

For the NVT simulation performed, timestep was set at 1 fs and the Velocity-Verlet integrator used for a total of 1 ns simulation. A Nosé-Hoover thermostat was utilized with the number of chains being 4 and thermal coupling with the whole system. All bonds were constrained with LINCS and non-covalent forces being cut off at 12 Å distance using a force-switch algorithm. The system was heated linearly for 50 ps from 50 K to 300 K with velocity at the first timestep being generated by Boltzmann-Maxwell distribution at 50 K. A .mdp text file for the MoFe model can be found in the supplementary section but works with minor modifications for the lysozyme model.

Visual analysis of the simulation using VMD showed no obvious anomalies. The RMSD values for lysozyme had an average of 0.0314 nm and standard deviation of 0.00095 nm with a slope of the best fit line of RMSD values being $-1*10^{-7}$ compared to the average value of the fourth MoFe protein simulation being 0.0318 nm with standard deviation of 0.00027 nm and a slope of the best fit line being $-6*10^{-8}$. Looking at the standard deviation, it can be seen that there is in fact more difference in the movement of amino acid residues in the lysozyme model than in the MoFe protein model. As mentioned before, the MoFe protein is bound in place through its constrained cofactors and is this expected behavior.

The mean temperature of the lysozyme simulation was 300.01 K with standard deviation of 1.53 K and the minimum and maximum temperatures being approximately 294 K and 304 K, respectively. Here, a difference was observed as the mean temperature of the fourth MoFe simulation was 300.02 K with standard a deviation of 0.55 K and minimum and maximum temperatures being 298 K and 302 K, respectively. The velocity of a particle is proportional to its kinetic energy and thus temperature. It can be assumed that there are less fluctuations in the speed of particles in the MoFe protein model.

Though there are differences in mobility of the two simulated systems, there were no indications that the parameters used would give an unstable model. The lysozyme protein has more mobility but that was expected as it is not bound down due to constrains to cofactors. The current simulation parameters were deemed usable for the main goal of this research project, creation of stable MoFe protein model for further QM/MM research.

3.4 Molecular Dynamics Studies

3.4.1 Long Molecular Dynamics Simulations.

Two long simulations were performed that differed only in the timestep size, heating time and run time. The first simulation had a timestep of 1 fs, heating time 0-500 ps from 50 K to 300 K and run time of 5 ns. The second simulation had a timestep of 2 fs, heating time 0-1000 ps from 50 K to 300 K and run time of 10 ns. Other parameters were the same as the optimal parameters determined previously (the fourth simulation).

Both simulations proved to be stable with the slope of RMSD values nearing zero and temperature fluctuations normal (graphs in appendix B). Due to simulations with

smaller timestep being more accurate, the 1ns timestep simulation was used to extract snapshots for use in a future QM/MM research.

3.4.2 Effects of excessive heating

To determine better if the effects of cofactor constrains in the MoFe protein model are having effects, it is possible to heat up the MoFe protein system to extreme temperatures and compare it to lysozyme by monitoring RMSD values.

Optimal parameters were used in all simulations with the model system being heated from 50K to goal temperatures in 50 ps, in three different simulations for each model system. The goal temperatures were 1000 K, 1500 K and 2000 K.

As can be seen in figure 5, the RMSD values of the lysozyme model flickers much more relative to the nitrogenase model at extreme temperatures while the nitrogenase system fluctuations stay similar for all three temperatures. This effect is clearly due to the constrained cofactors.

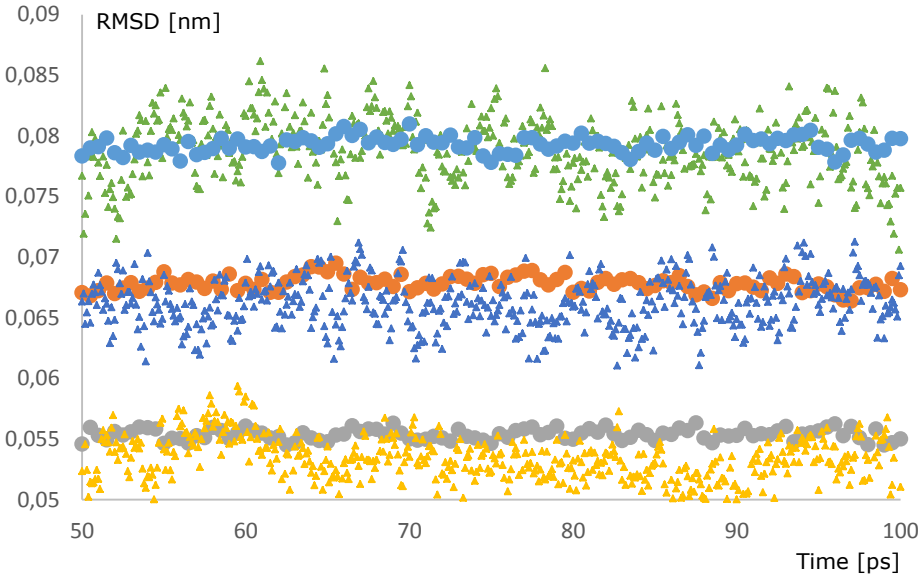

Figure 5: Triangles represent lysozyme while dots represent MoFe protein. Yellow/grey is at 1000 K, orange/deep blue is at 1500 K and light blue/light green is at 2000 K. The RMSD values of lysozyme show much greater variation as opposed those of MoFe protein.

3.4.3 Strange movement of Gln432A

During one of the test simulations, a flip of the amide side chain Gln432A was observed. This residue is close to the recently discovered iron ion in the MoFe protein whose biochemical purpose is not known and only speculations exists as to what role this site plays in nitrogenase. The iron ion site has little in common with other iron containing proteins but resembles mostly the diiron centers of rubrerythrins (Zhang et al., 2013).

The flip happened between 554 ps and 614 ps from the start of the simulation and can be seen in figures 6-8. After thorough examination, no explanation was found as to why this flip occurs but another interesting phenomenon was observed. A generated water molecule (there are two types of water molecules, those who come with the crystallized structure and those generated during solvation) which was close to the iron ion moves away from the iron ion in the process. It is as if Gln432A might be acting as a gatekeeper for water molecules to enter the space near the iron ion (that has 2 bound water molecules).

This result hints at a possible water pathway that leads from the iron site to the surface of the protein which could be controlled by some mechanism connected to the flip of Gln432A. As the purpose of the iron site is not known, the relevance of a possible water pathway in this region is not clear and if this site is related in any way to the nitrogen reduction process. Further simulation studies are required to reproduce this behavior and understand this site.

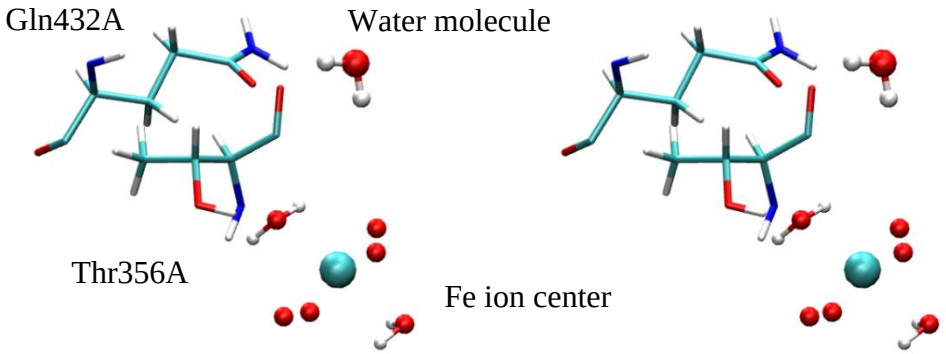

Figure 6: Stereoview – The observed water molecule is in the vicinity of Gln432A with Thr356A being behind. The water is hydrogen bonded to the amide oxygen of Gln432A side chain. The iron ion can been seen in the down right corner with two crystal water molecules. The single red atoms are from the carboxyl groups and carbonyl group that connect to the iron ion (rest of the residues not shown). Figure is taken 554 ps into the simulation.

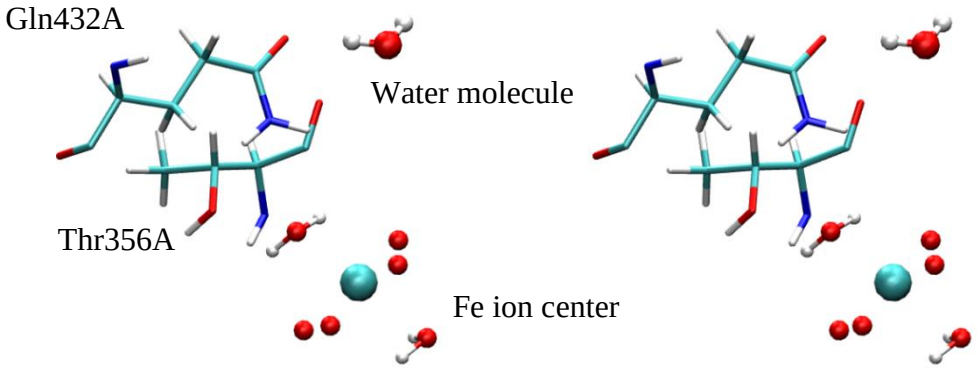

Figure 7: Stereoview – The observed flip with the water molecule still being hydrogen bonded to the oxygen atom of Gln432A amide group. Figure is taken 574 ps into the simulation.

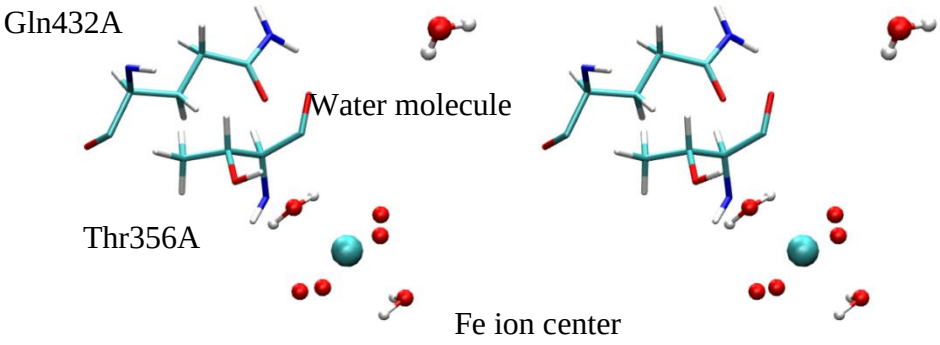

Figure 8: Stereoview – Gln432A has returned to its beginning position with the water molecule driven away.

3.5 Understanding Nitrogenase: Towards a QM/MM model

The main goal of this research project was the creation of an all-atom, stable MM model of the MoFe protein to be used for future QM/MM studies. The QM/MM studies are intended to shed light on the reaction mechanism of nitrogenase, where the FeMo cofactor and its surroundings will be described by quantum mechanics and the rest of the protein solvent environment described by molecular mechanics. A requirement of a QM/MM model is a stable MM model, confirmed by a stable NVT MD simulations from which trajectory snapshots can be extracted and used for geometry optimization in QM/MM.

QM/MM calculations of reaction mechanism in enzymes are performed by doing geometry optimizations of the active site, including the substrate. These type of calculations are usually performed without periodicity and instead spherical clusters are more commonly used. Since it is also unfeasible to perform a reliable QM/MM geometry optimization of a system with over $3*10^6$ atoms (the degrees of freedom being over $9*10^6$) and the QM/MM program Chemshell supporting only system sizes of ~ 40000 atoms, a smaller MM cluster model must be created.

A snapshot after 1862 ps was extracted from the 5 ns MD NVT simulation and cut down. After thorough examination it was decided that the center carbon atom in the FeMo cofactor was to be used as reference point when cutting out a sphere with 42 Å radius. This includes water molecules, sodium ions, the FeMo cofactor, the P-cluster, iron ion complex and amino residues from chain A, B and D. It was decided to exclude chain C in its entirety (being farthest away from the FeMo cofactor). The amino acid residue and other molecules were taken whole and not cut down.

To keep structural integrity of this new system, all amino residues from chain A and B are included. Because there are some atoms from residues in chain D within the newly created 42 Å sphere, some of them are also taken in as a whole for the QM/MM base with these residues being Gln452D, Thr455D, Leu456D, Arg476D, Ser482D, Thr483D and Thr484D. These amino acid from chain D make hydrogen bonds with residues in chain A and B and are potentially important for future QM/MM studies.

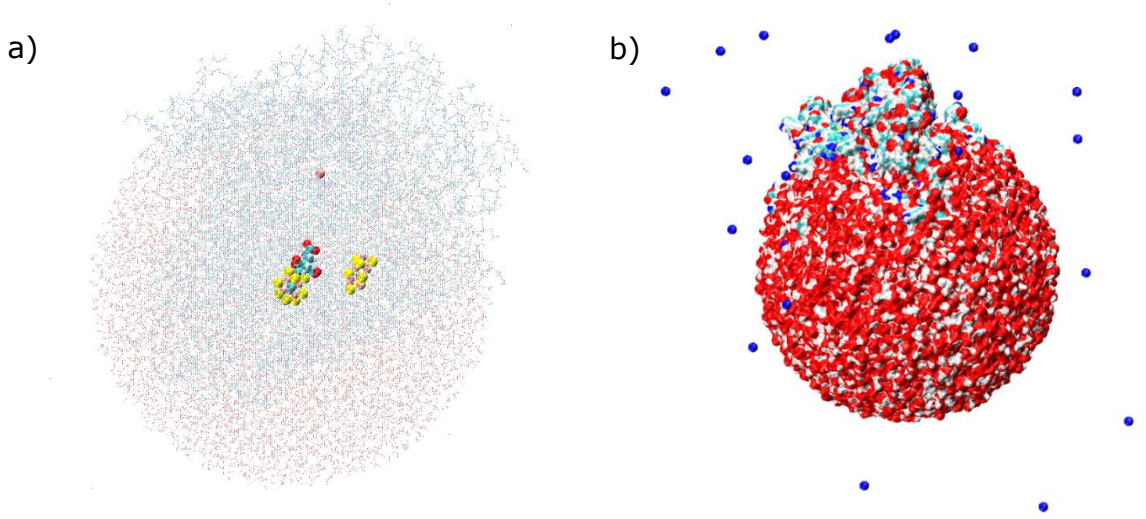

Figure 9: a) Rendered figure of the created system with the cofactors visible solid and residues and water transparent. b) The cut-out system as it would be if it had surfaces and showing Na^+ ions around the sphere.

4 Conclusion

As of 2004, approximately 1% - 2% of humanity's energy consumption was used for creation of ammonia as a fertilizer through the Haber-Bosch process (Smil, 2004). The Haber-Bosch process requires high temperatures and pressures to create ammonia and is not very energy efficient. Nitrogenases on the other hand are able to catalyze ammonia formation at ambient temperature and pressure. By a detailed understanding of nitrogenase, it might be possible to create a catalyst that mimics or improves on the natural process.

The goal of this research project was to create a stable MM model of nitrogenase MoFe protein. Considerable time was spent on creating a reliable model where particular attention was paid to deciding protonation state of residues that can have more than one protonation state. By doing so a very detailed system with correct hydrogen bonds properly in place was made.

Multiple protocols for stable NVT MD simulations were also tested. The results of the MD simulations reveal a stable model over 10 ns, with the mean RMSD being 0.32 Å compared to the 1 Å resolution crystal structure. Due to necessary cofactor constrains, the fluctuations in RMSD are most likely quite a bit lower than they would be without constrains. For further studies, it would be interesting to derive cofactor parameters that would allow limited movement.

Making sure that the system is properly energy minimized was essential. Badly energy minimized system would have too much force between atoms which could in turn cause the model to behave abnormal e.g. seeing chains move excessively. Even though the model of the MoFe protein created here was not completely minimized, due to problems with the steepest descent algorithm, this had no noticeable effect on MD trajectories as seen from stable RMSD and temperature values.

The flip of the residue Gln432A was observed in only a single NVT simulation with the flip taking 60 ps. Whilst the residue flipped, a water molecule moved away from the iron ion possibly hinting at a water tunnel. This needs though further research and simulations.

An approximately 40000 atom spherical cluster model extracted from a snapshot at 1862 ps in a NVT simulation will be used for further QM/MM research in Ragnar's Björnsson group. Hopefully, a fully functional QM/MM simulation can be performed in the future which will increase our understanding of the complex enzyme nitrogenase. By understanding the enzyme and how it performs this reaction in detail, it may be possible to reverse engineer it and create a biocatalyst that could be used as an environmental friendly production of ammonia.

References

Artymiuk, P. J., Blake, C. C. F., Rice, D. W., & Wilson, K. S. (1982). The Structure of the Orthorhombic Form of Hen Egg-White Lysozyme at 1.5 Angstroms Resolution. *ACTA Crystallog Sect B, 38*.

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., & MacKerell, A. D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J Chem Theory Comput, 8*(9), 3257-3273.

Bothe H., Newton W. E., & Ferguson S. J. (2007). *Biology of the Nitrogen Cycle: COST edition* (1 ed.): Elsevier Science.

Duval, S., Danyal, K., Shaw, S., Lytle, A. K., Dean, D. R., Hoffman, B. M., Antony, E., Seefeldt, L. C. (2013). Electron transfer precedes ATP hydrolysis during nitrogenase catalysis. *Proc Natl Acad Sci U S A, 110*(41), 16414-16419.

Hoffman, B. M., Lukoyanov, D., Yang, Z. Y., Dean, D. R., & Seefeldt, L. C. (2014). Mechanism of nitrogen fixation by nitrogenase: the next stage. *Chem Rev, 114*(8), 4041-4062.

Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A, 31*(3), 1695-1697.

Hu, Y. L., & Ribbe, M. W. (2015). Nitrogenase and homologs. *J Biol Inorg Chem, 20*(2), 435-445.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph, 14*(1), 33-38, 27-38.

Jensen, F. (2007). *Introduction to Computational Chemistry* (2 ed.): John Wiley & Sons Ltd.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem phys, 79*(2), 926-935.

Lancaster, K. M., Hu, Y., Bergmann, U., Ribbe, M. W., & DeBeer, S. (2013). X-ray spectroscopic observation of an interstitial carbide in NifEN-bound FeMoco precursor. *J Am Chem Soc, 135*(2), 610-612.

Lerbret, A., Affouard, F., Bordat, P., Wdoux, A., Gulnet, Y., & Descamps, A. (2008). Molecular dynamics simulations of lysozyme in water/sugar solutions. J *Chem Phys, 345*(2-3), 267-274.

Morrison, C. N., Hoy, J. A., Zhang, L. M., Einsle, O., & Rees, D. C. (2015). Substrate Pathways in the Nitrogenase MoFe Protein by Experimental Identification of Small Molecule Binding Sites. *Biochemistry, 54*(11), 2052-2060.

Neese, F. (2012). The ORCA program system. *WIREs Comput Mol Sc, 2*(1), 73-78.

Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys, 52*(2), 255-268.

Parrinello, M., & Rahman, A. (1982). Strain Fluctuations and Elastic Constants. *J Chem Phys.*

Post, C. B., Brooks, B. R., Karplus, M., Dobson, C. M., Artymiuk, P. J., Cheetham, J. C., & Phillips, D. C. (1986). Molecular dynamics simulations of native and substrate-

bound lysozyme. A study of the average structures and atomic fluctuations. *J Mol Biol, 190*(3), 455-479.

Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R.,Shits, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess B., Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics, 29*(7), 845-854.

Rostkowski, M., Olsson, M. H. M., Sondergaard, C. R., & Jensen, J. H. (2011). Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *Bmc Struct Biol, 11*(6).

Schindelin, H., Kisker, C., Schlessman, J. L., Howard, J. B., & Rees, D. C. (1997). Structure of ADP x AIF4(-)-stabilized nitrogenase complex and its implications for signal transduction. *Nat.*

Seefeldt, L. C., Hoffman, B. M., & Dean, D. R. (2009). Mechanism of Mo-Dependent Nitrogenase. *Annu Rev of Biochem, 78*, 701-722.

Skyner, R. E., McDonagh, J. L., Groom, C. R., van Mourik, T., & Mitchell, J. B. (2015). A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys Chem, 17*(9), 6174-6191.

Smil, V. (2004). *Enriching the Earth*: MIT Press.

Smith, D., Danyal, K., Raugei, S., & Seefeldt, L. C. (2014). Substrate channel in nitrogenase revealed by a molecular dynamics approach. *Biochemistry, 53*(14), 2278-2285.

Spatzal, T., Aksoyoglu, M., Zhang, L., Andrade, S. L., Schleicher, E., Weber, S., Rees, D. C., Einsle, O. (2011). Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science, 334*(6058), 940.

Varshney A., Brooks F. P. Jr., & Wright W. V. (1994). Linearly Scalable Computation of Smooth Molecular Surfaces. *IEEE Comput Graph, 14*(5), 19-25.

Verlet, L. (1967). Computer "Experiments" on classical Fluids. I. Thermodynamic Properties of Lennard-Jones Molecules. *Phys Rev, 159*(1), 98-103.

Warshel, A. (1976). Bicycle-pedal model for the first step in the vision process. *Nature, 260*(5553), 679-683.

Wei, T., Carignano, M. A., & Szleifer, I. (2012). Molecular dynamics simulation of lysozyme adsorption/desorption on hydrophobic surfaces. *J Phys Chem B, 116*(34),

Wright, L. B., Rodger, P. M., & Walsh, T. R. (2013). Aqueous citrate: a first-principles and force-field molecular dynamics study. *Rsc Advances, 3*(37), 16399-16409.

Xie, H. J., Wu, R. B., Zhou, Z. H., & Cao, Z. X. (2008). Exploring the interstitial atom in the FeMo cofactor of nitrogenase: Insights from QM and QM/MM calculations. *J Phys Chem B, 112*(36), 11435-11439.

Zhang, L., Kaiser, J. T., Meloni, G., Yang, K. Y., Spatzal, T., Andrade, S. L., . . . Rees, D. C. (2013). The sixteenth iron in the nitrogenase MoFe protein. *Angew Chem Int Ed Engl, 52*(40), 10529-10532.

Appendix A – Commands & parameters

Please note, the following commands were used with GROMACS unless otherwise stated.

A modified version of the original pdb file 3U7Q was used which was created by Ragnar Björnsson. The cysteines connecting to the cofactors were deprotonated, a central carbide ion was added into the FeMo cofactor and iron ions added and calcium ions removed. The file is used to create a protonated system,with the protonation state of lysine, glutamate and histidine being decided manually (hence 0, 1 and 2 as it is the input that the program asks the user to enter). This was done using the following command:

```
echo 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
0 0 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1
1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0
1 0 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 1 1 0 0 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 2 1 0 1 0 1 0 0 1 1 1 1 1 1 1 0 | gmx pdb2gmx -f 3U7Q-form.pdb -lys
-glu -his -o 3U7Q-form-processed.pdb -ff charmm36 -water tip3 >& gmx2pdb.out &
```

It is convenient to have a list available of crystallized water molecules to differentiate between generated water molecules and water molecules from the crystal structure. To create an entry in an index file, the following command was used:

```
gmx make_ndx -f 3U7Q-form-processed.pdb -o index.ndx
#INPUT: 'a OW' og síðan 'q'
```

To create a box to be able to solvate the system, the following command was used:

```
gmx editconf -f 3U7Q-form-processed.pdb -o 3U7Q-form-box.pdb -c -d 1.0 -bt cubic
>& box.out &
```

To solvate the protein inside the defined box, the following command was used:

```
gmx solvate -cp 3U7Q-form-box.pdb -cs spc216.gro -o 3U7Q-form-
solvated.pdb -p topol.top –n index.ndx >& solvated.out &
```

The system had -40 charge and thus it was needed to create positive counter ions. This is a two-step process, first the grompp command (which will create a binary .tpr file from human readable input files) was used with ions.mdp defing the parameters:

```
gmx grompp -f ions.mdp -c 3U7Q-form-solvated.pdb -p topol.top –n
index.ndx –o ions.tpr
```

Where the ions.mdp file containing the run parameters:

```
; ions.mdp - used as input into grompp to generate ions.tpr
```

```
; Parameters describing what to do, when to stop and what to save
integrator     = steep        ; Algorithm (steep = steepest descent minimization)
emtol          = 1000.0       ; Stop minimization when the maximum force < 1000.0 kJ/mol/nm
emstep      = 0.01            ; Energy step size
nsteps         = 50000        ; Maximum number of (minimization) steps to perform

; Parameters describing how to find the neighbors of each atom and how to calculate the
interactions
nstlist            = 1          ; Frequency to update the neighbor list and long range
forces
cutoff-scheme   = Verlet
ns_type            = grid            ; Method to determine neighbor list (simple, grid)
coulombtype        = PME             ; Treatment of long range electrostatic interactions
rcoulomb    = 1.0             ; Short-range electrostatic cut-off
rvdw               = 1.0             ; Short-range Van der Waals cut-off
pbc                    = xyz         ; Periodic Boundary Conditions (yes/no)
```

Secondly, the generated binary file was used to create the charge neutral system with the following command.

```
gmx genion -s ions.tpr -o 3U7Q-make-ionated.pdb -p topol.top -pname
NA -nname CL -np 40
```

To lower the forces between adjacent protons, first grompp is used with the following command:

```
gmx grompp -f minim.mdp -c 3U7Q-make-ionated-MANUAL-H2O-1.pdb -p
topol.top -o hmin.tpr -n index.ndx >& minim.out &
```

Where the minim.mdp contains the following information

```
;minim.mdp - used as input into grompp to generate em.tpr
Integrator             = steep ; Algorithm (steep = steepest descent minimization)
emtol          = 1000.0 ; Stop minimization when maximum energy potential< 1000.0 kJ/mol/nm
emstep         = 0.01 ; Energy step size
nsteps         = 50 ; Maximum number of steps

; Parameters that describe how to find the neighboring forces and energy minimize.
nstlist        = 1 ; Frequency to update the neighbor list and long range forces
cutoff-scheme  = Verlet       ; scheme for cut-off
ns type        = grid ; Determine the neighbor grid
coulombtype    = PME ; Treatment of long range electrostatic interactions
rcoulomb               = 1.0 ; Short-range electrostatic cut-off
rvdw           = 1.0 ; Short-range Van der Waals cut-off
pbc            = xyz ; Periodic Boundary Conditions
freezegrps             = Crystal Protein-H Cofactor ; Defining which groups are restrained
freezedim              = Y Y Y Y Y Y Y Y Y    ; restrained in X Y Z dimensions
nstxout        = 1                          ; information on every step is saved
```

To start the minimization itself after grompp, then the following command is used:

```
gmx mdrun -ntomp 1 -ntmpi 1 -v -c 3U7Q-hmin-sd50.pdb -deffnm hmin >&
hmin.out &
```

To energy minimize the whole system, first grompp was used:

```
gmx grompp -f minim-allminco.mdp -c 3U7Q-hmin-sd50-mod.pdb -p
topol.top -o minim-allminco.tpr -n index.ndx >& minim-allminco-
grompp.out &
```

Where minim-allminco.mdp contains:

```
; minim.mdp - used as input into grompp to generate em.tpr
integrator     = steep         ; Algorithm (steep = steepest descent minimization)
emtol              = 1000.0          ; Stop minimization when energy potentials < 1000.0
kJ/mol/nm
emstep         = 0.01          ; Energy step size
nsteps         = 4             ; Maximum number of (minimization) steps to perform

; Parameters describing how to find the neighbors of each atom and how to calculate the
interactions
nstlist          = 1               ; Frequency to update the neighbor list and long range
forces
cutoff-scheme   = Verlet
ns_type         = grid              ; Determining the neighbor grif
coulombtype     = PME               ; Treatment of long range electrostatic interactions
rcoulomb        = 1.0               ; Short-range electrostatic cut-off
rvdw            = 1.0               ; Short-range Van der Waals cut-off
pbc             = xyz           ; Periodic Boundary Conditions (yes/no)
freezegrps      = Sulfhis FE2P-oxygen Cofactor
freezedim       = Y Y Y  Y Y Y  Y Y Y
nstxout         = 1
```

To do the energy minimization itself, the following command was used:

```
gmx mdrun -ntomp 1 -ntmpi 1 -v -c 3U7Q-minim-allminco.pdb -deffnm
minim-allminco >& minim-allminco.out &
```

There were quite many NVT simulations done, the one explained here refers to the 5 ns NVT simulation which was used for the 40000 atom cut-out cluster model. First the grompp command is used;

```
gmx grompp -f nvt.mdp -c 3U7Q-minim-allminco.pdb -n index -p
topol.top -o nvt.tpr  >& nvt-grompp.out &
```

Where the .mdp file contains:

```
title                = CHARMM-36 Nitrogenase NVT simulation
define               = -DPOSRES     ; position restrain the protein
; This is a NVT simulation
; Run parameters
integrator           = md-vv        ; not a leap-frog integrator
nsteps               = 5000000      ; 1 * 5000000 = 5000 ps =5 ns
dt                   = 0.001        ; 1 fs
; Output control
nstxout        = 10000        ; save coordinates
nstvout        = 10000        ; save velocities
nstenergy      = 10000        ; save energies
nstlog         = 10000        ; update log file
; Bond parameters
continuation            = no          ; first dynamics run
constraint_algorithm    = lincs       ; holonomic constraints
constraints             = all-bonds   ; every bond is constrained
lincs iter              = 1           ; accuracy of LINCS (usually set at 1 but necessary to
be set at 2 for NVE simulations)
lincs_order             = 4           ; also related to accuracy
; Neighborsearching
cutoff-scheme           = Verlet
vdwtype         = Cut-off           ; twin range cut-offs with neighbor list cut-off rlist
and VdW cut-off rvdw, where rvdw ≥ rlist.
vdw-modifier          = Force-switch        ; Smoothly switches the forces to zero between
rvdw-switch and rvdw. This shifts the potential shift over the whole range and switches it
to zero at the cut-off. Note that this is more expensive to calculate than a plain cut-off
and it is not required for energy conservation, since Potential-shift conserves energy just
as well.
ns type            = grid            ; search neighboring grid cells
nstlist            = 10              ; 10 fs, largely irrelevant with Verlet. Remember,
every option relying on nstlist must be a integer of it.
rlist              = 1.2
rcoulomb    = 1.2            ; short-range electrostatic cutoff (in nm)
```

```
rvdw                = 1.2             ; short-range van der Waals cutoff (in nm)
rvdw-switch         = 1.0             ; where to start switching the LJ force and possibly
the potential, only relevant when force or potential switching is used
verlet-buffer-tolerance               = 0.0005       ; Usually at 0.005 but needs to be
lowered for NVE calculations
; Electrostatics
coulombtype             = PME          ; Particle Mesh Ewald for long-range electrostatics
pme order               = 4            ; cubic interpolation
fourierspacing          = 0.16         ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = nose-hoover          ; temperature coupling is switched on in NVT.
tc-grps         = system               ; Temperature is coupled to the whole system
nsttcouple      = -1                   ; we want to let the coupling to be equal to nstlist
nh-chain-length         = 4            ; number of chained Nose-Hoover thermostats.
tau-t           = 1                    ; time constant for coupling
ref-t           = 50                   ; reference temperature for coupling.
;annealing (used for better control of heating in NVT)
annealing               = single       ; The annealing process is set to be linear
annealing-npoints       = 2            ; Let's use 2 points from 50K to 300K.
annealing-time          = 0 500        ; The temperature should rise LINEARILY from 50K to
300K in 50 ps.
annealing-temp          = 50 300       ; We will start at 50K and end in 300K, if everything
works as planned.
; Pressure coupling is off
pcoupl          = no           ; no pressure coupling in NVT
; Periodic boundary conditions
pbc             = xyz          ; 3-D PBC
; Dispersion correction
DispCorr = EnerPres    ; account for cut-off vdW scheme
; Velocity generation
gen vel         = yes          ; assign velocities from Maxwell distribution
gen temp = 50          ; temperature for Maxwell distribution - question to disable this
when running a NVT test
gen_seed = -1          ; generate a random seed
freezegrps              = Sulfhis FE2P-oxygen Cofactor      ; Which groups to freeze
freezedim               = Y Y Y  Y Y Y  Y Y Y              ; Triple Y mean the groups are
frozen in x,y,z coordinates.
```

To start the simulation itself, the following command was used:

```
gmx mdrun -v -deffnm nvt -c 3U7Q-nvt.pdb >& nvt.out &
```

Because most of the NVT simulations were done using the computer cluster GARDAR, the following are commands for doing NVT simulations on the cluster.

To grompp to check if the submit command will work (we will not use the xxx.tpr file, another .tpr file will be generated automatically with the submit command), the following command was used:

```
grompp_mpi -f nvt.mdp -c 3U7Q-minim-allminco.pdb -n index.ndx -p
topol.top -o xxx.tpr
```

If grompp reports everything in order then data was submitted using a submit command created by Ragnar Björnsson:

```
gromacs-submit -m nvt.mdp -coord 3U7Q-minim-allminco.pdb -top
topol.top -index index.ndx -nproc 12
```

To calculate the RMSD value on the computer cluster GARDAR, the generated trajectory file nvt.trr and pdb file from the proton minimization step was used with the following command:

```
gmx_mpi rms -f nvt.trr -s 3U7Q-hmin-sd50-mod.pdb
```

For groups mentioned in the .mdp files under freezedim option, here is a summary of what particle number each group contains. Please note that the particle number used by Gromacs varies from the particle number VMD uses.

```
       Sulfhis group contains: 951   1058   1271   1323   2350   4179   6916   8663   9021
9901  16739   17111  18138  19967  22704  24451  24809 25689
       FE2P-oxygen group contains:
9265   9277   9278   13033  13034  13093  13094  25053  25065  25066  28821  28822  28881
28882  35439  35440  35441  35493  35494  35495  37677  37678  37679  37926  37927  37928
        Cofactor group contain:
31651 31652 31653 31654 31655 31656 31657 31658 31659 31660 31661 31662 31663 31664 31665
31666 31667 31668 31669 31670 31671 31672 31673 31674 31675 31676 31677 31678 31679 31680
31681 31682 31683 31684 31685 31686 31687 31688 31689 31690 31691 31692 31693 31694 31695
31696 31697 31698 31699 31700 31701 31702 31703 31704 31705 31706 31707 31708 31709 31710
31711 31712 31713 31714 31715 31716 31717 31718 31719 31720 31721 31722 31723 31724 31725
31726 31727 31728 31729 31730 31731 31732 31733 31734 31735 31736 31737 31738 31739 31740
31741 31742 31743 31744 31745 31746 31747 31748 31749 31750 31751 31752 31753 31754 31755
31756 31757 31758 31759 31760
        Crystal group contains water molecules in the range 31761-39566
```

To create the cluster model, the following selection was used in VMD

```
index 0 to 15787 or index 30986 to 31575 or index 28631 to 28985 or resname NA or (same residue as within 42 of index 31669) and not index 30921 to 30959 and not index 30449 to 30481 and not index 30396 to 30412 and not index 30109 to 30122
```

Appendix B - Data and Graphs from Simulations

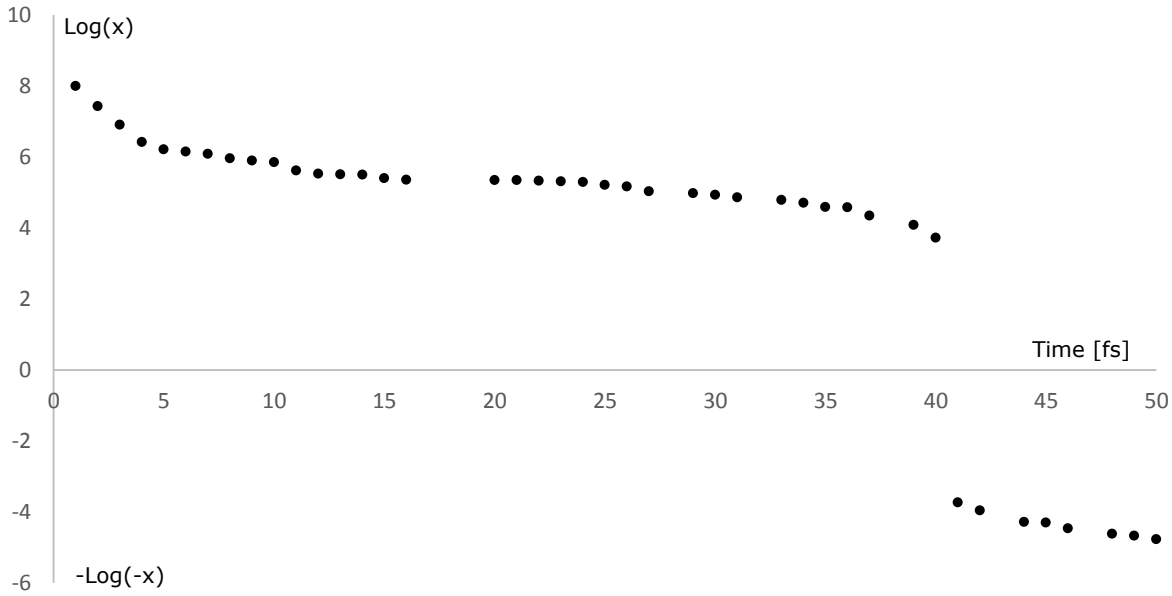

Figure 10: Potential energy during the energy minimization of protons. The y-axis is composed of two scales with the positive half being log([J/mol]) and the negative being –log(-[J/mol]).

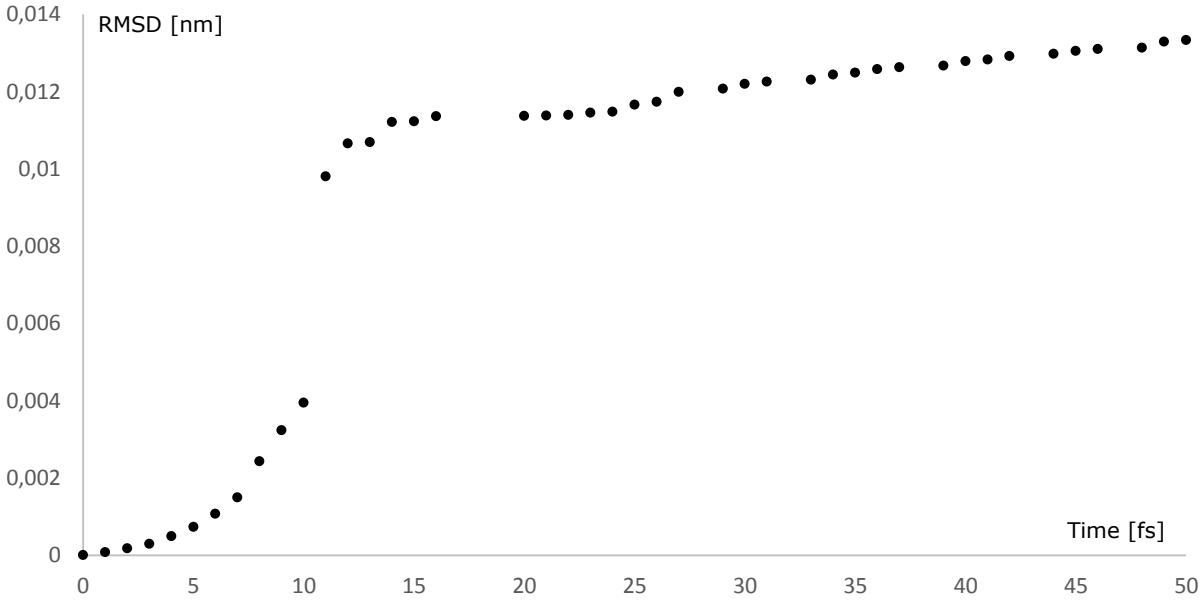

Figure 11: RMSD values of protons as a function of timestep during proton energy minimization.

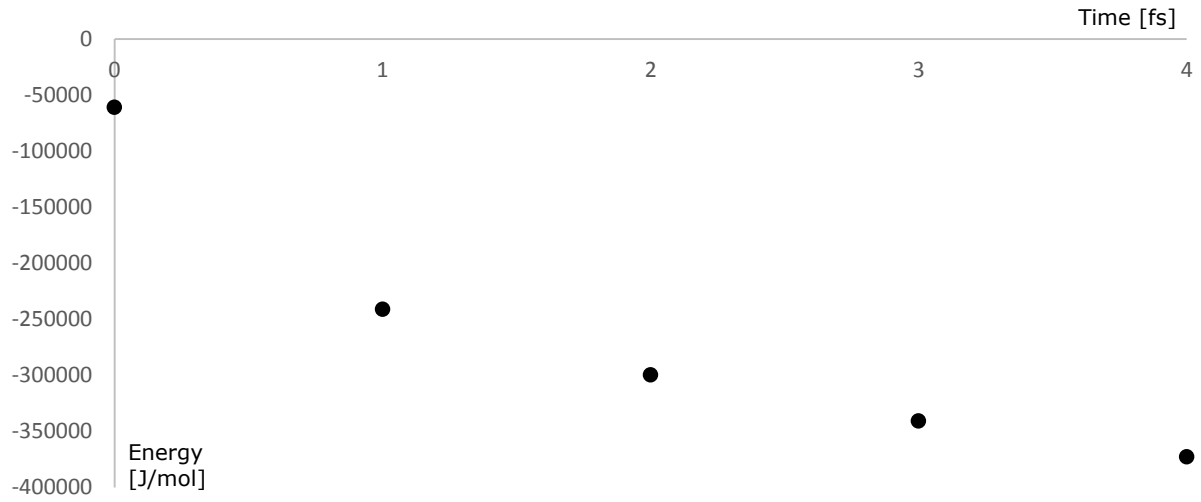

Figure 12: Potential energy as a function of timestep during energy minimization of the whole system.

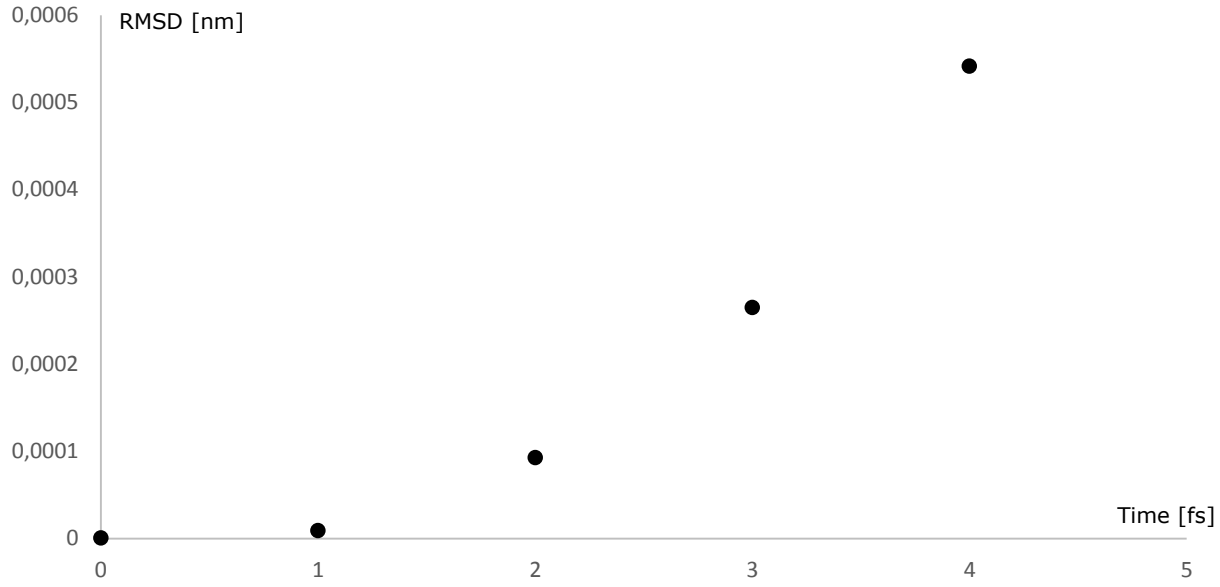

Figure 13: RMSD values of heavy atoms of the protein as a function of a timestep.

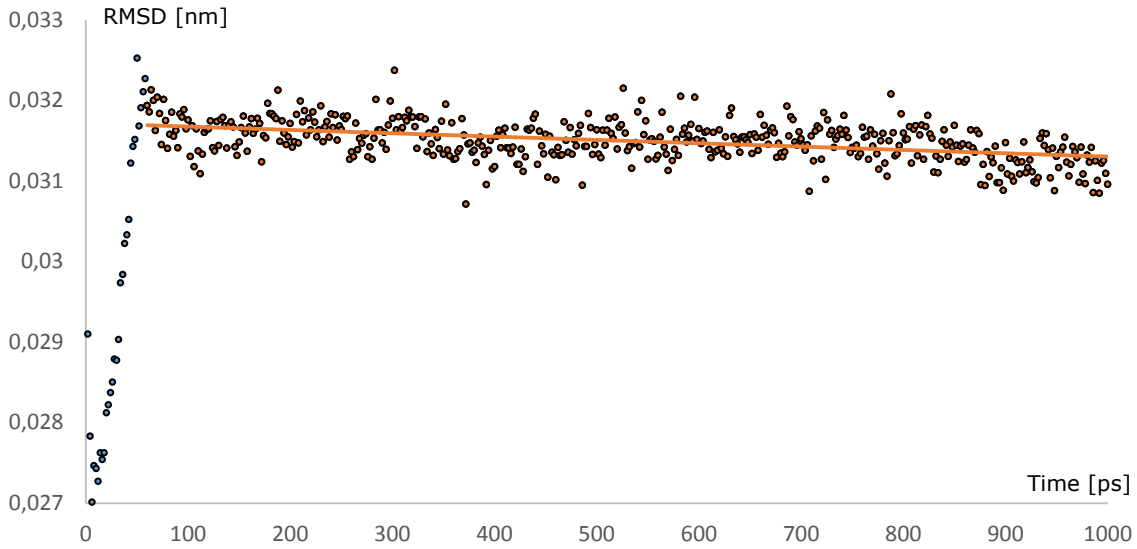

*Figure 14: First simulation (ab-t1000-dt1-leap-sys-nh1). RMSD as function of timestep with the slope of the trendline being -4*10^{-7} and having R^2 = 0.208.*

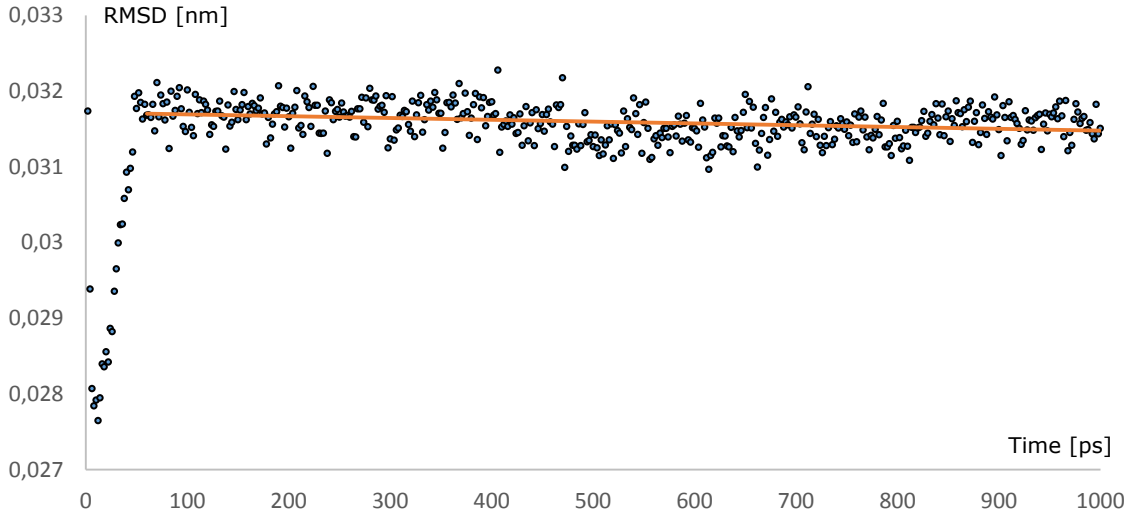

*Figure 15: Second simulation (ab-t1000-dt1-md_vv-prot_nonprot-nh4). RMSD as function of timestep with the slope of the trendline being -2*10^{-7} and having R^2 = 0.084.*

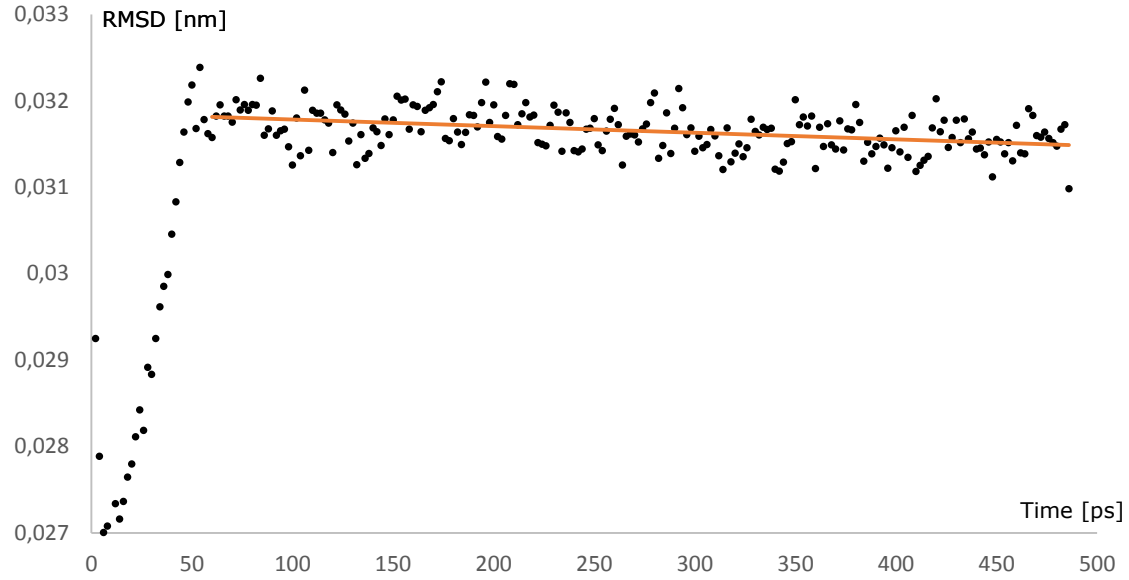

*Figure 17: Third simulation (ab-t1000-dt1-md_vv-sys-nh1). RMSD as function of timestep with the slope of the trendline being -8*10^{-7} and having R^2 = 0.158.*

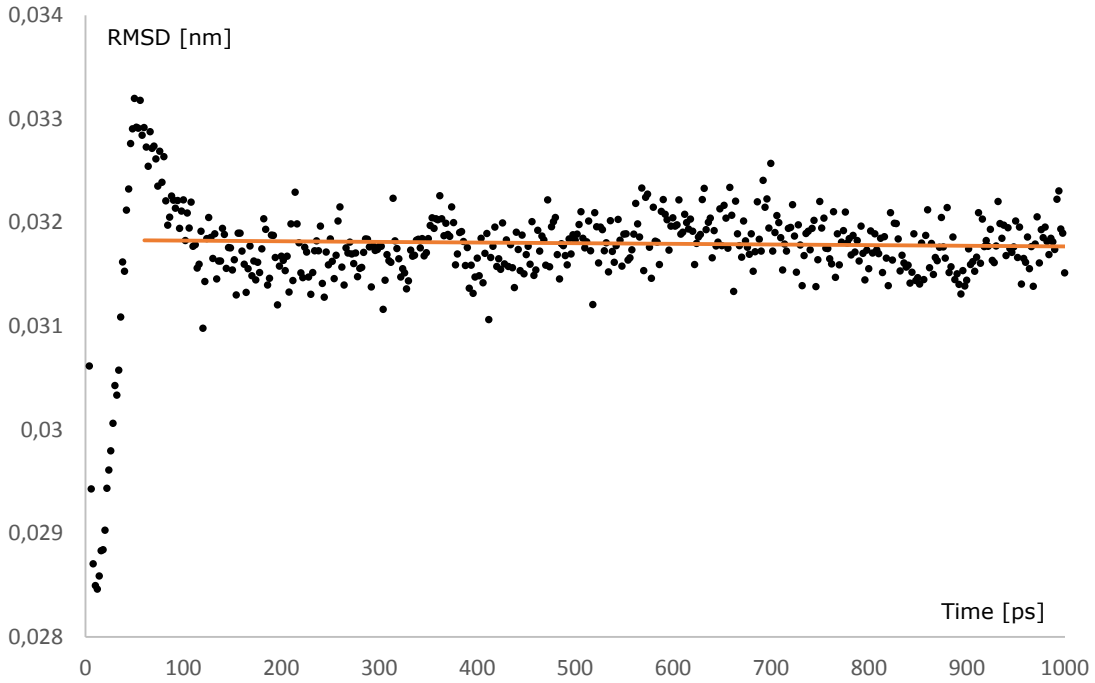

*Figure 16: Fourth simulation (ab-t1000-dt1-md_vv-sys-nh4). RMSD as function of timestep with the slope of the trendline being -6*10^{-8} and having R^2 = 0.004.*

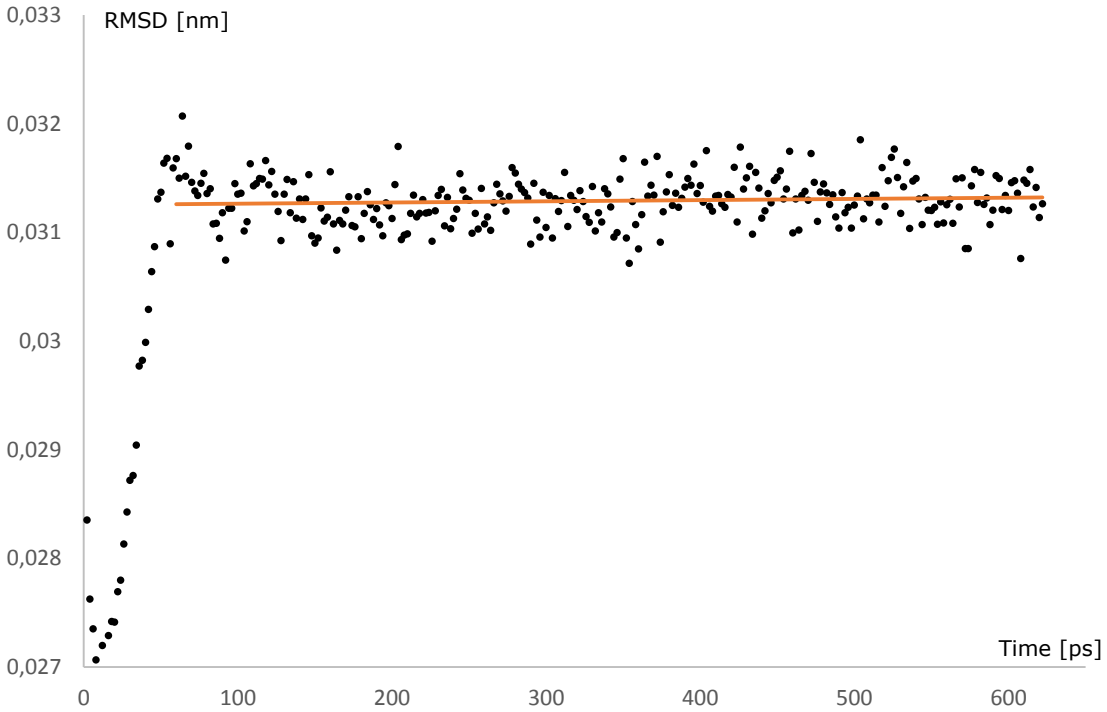

*Figure 18: Fifth simulation (ab-t1000-dt2-md_vv-sys-nh1). RMSD as function of timestep with the slope of the trendline being -1*10^{-7} and having R^2 = 0.007.*

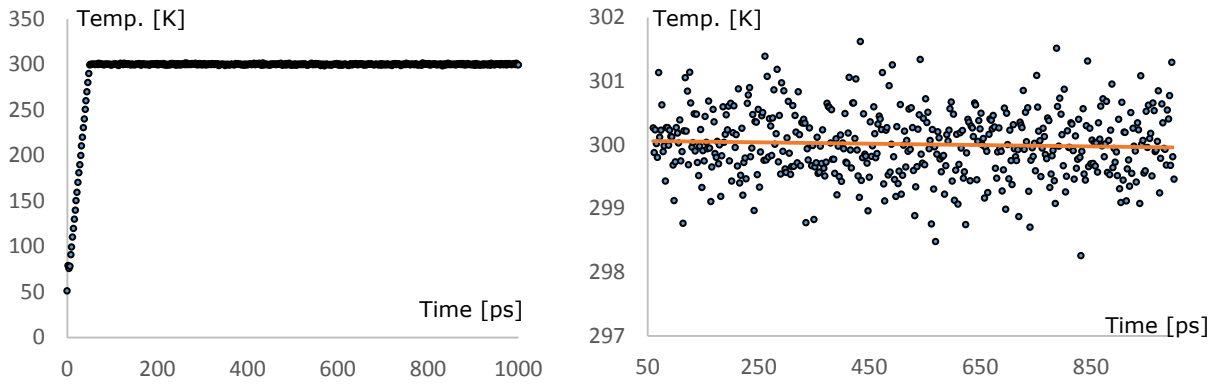

Figure 19: First simulation (ab-t1000-dt1-leap-sys-nh1). The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline -0.0001 and R^2=0.003.

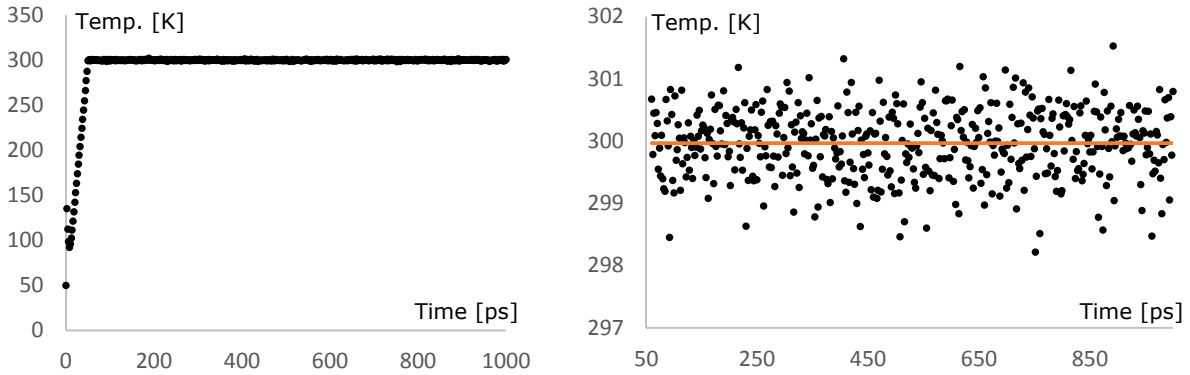

Figure 20: Second simulation (ab-t1000-dt1-md_vv-prot_nonprot-nh4). The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline $-3*10^{-6}$ and $R^2=2*10^{-6}$.

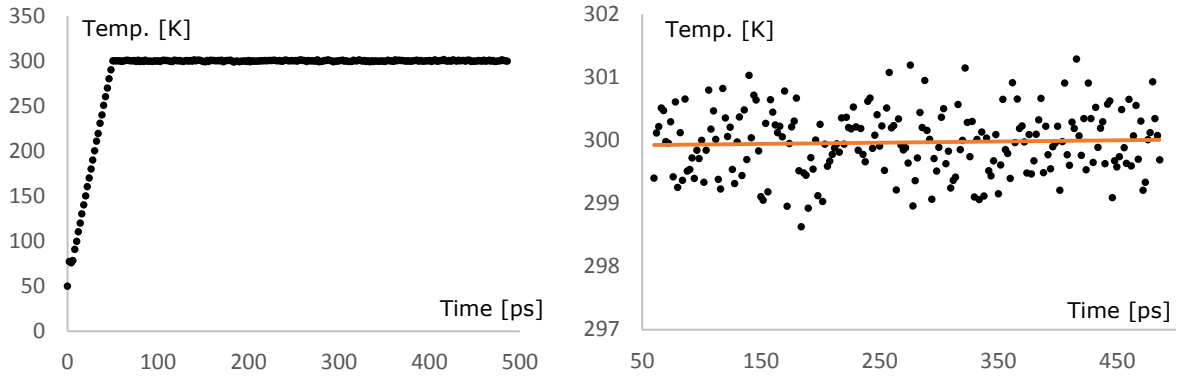

Figure 22: Third simulation (ab-t1000-dt1-md_vv-sys-nh1). The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline 0.0002 and $R^2=0.0021$.

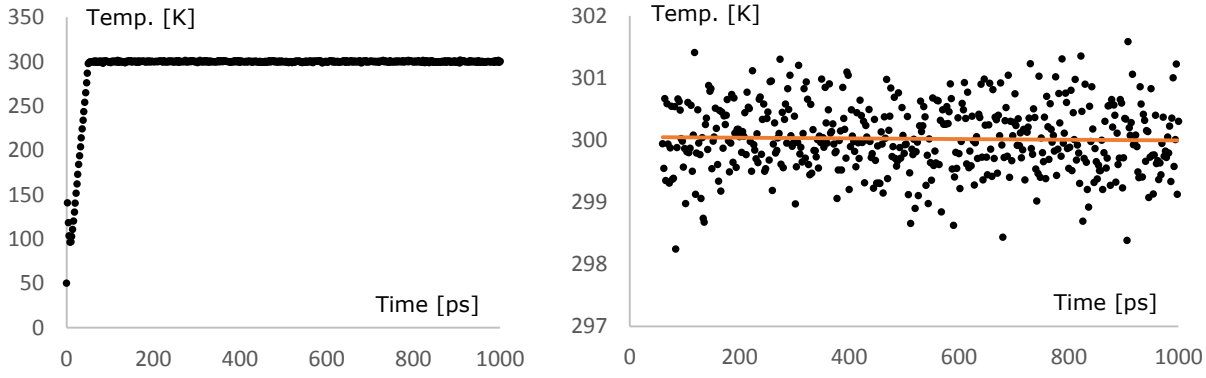

Figure 21: Fourth simulation (ab-t1000-dt1-md_vv-sys-nh4). The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline $-5*10^{-5}$ and $R^2=0.0007$.

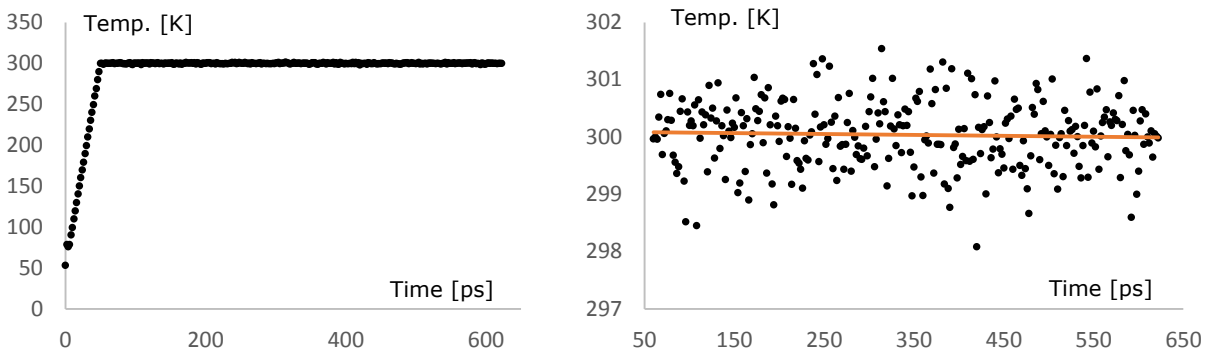

Figure 24: Fifth simulation (ab-t1000-dt2-leap-sys-nh1). The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline -0.0002 and R^2=0.0019.

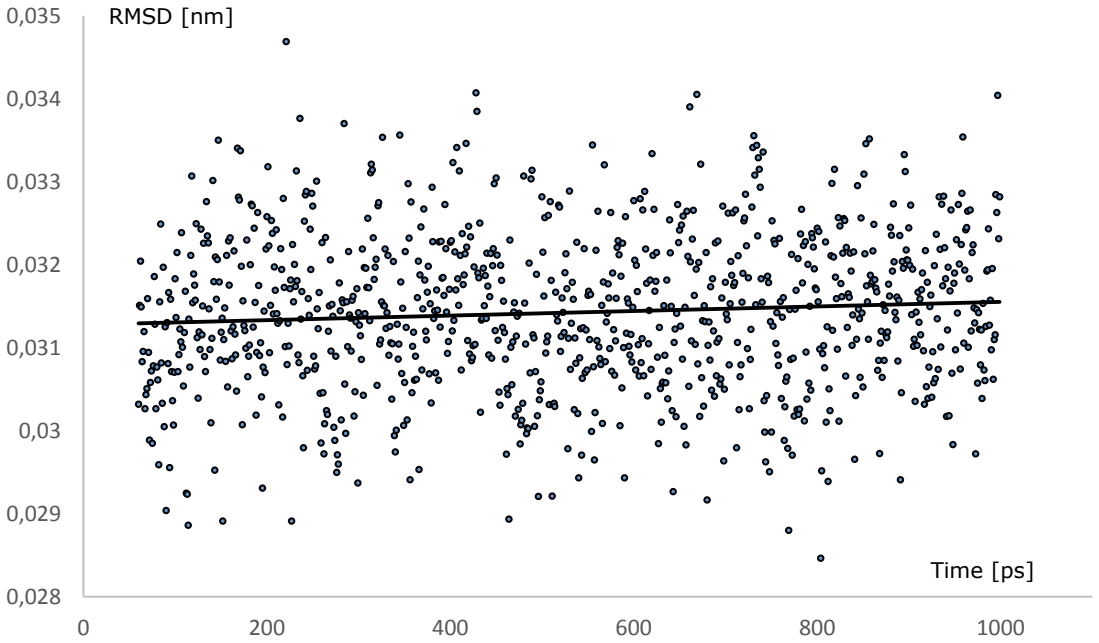

Figure 23: Lysozyme 1 ns simulation using ab-t1000-dt1-md_vv-sys-nh4 parameters. RMSD as function of timestep with the slope of the trendline being $3*10^{-7}$ and having R^2 = 0.006.

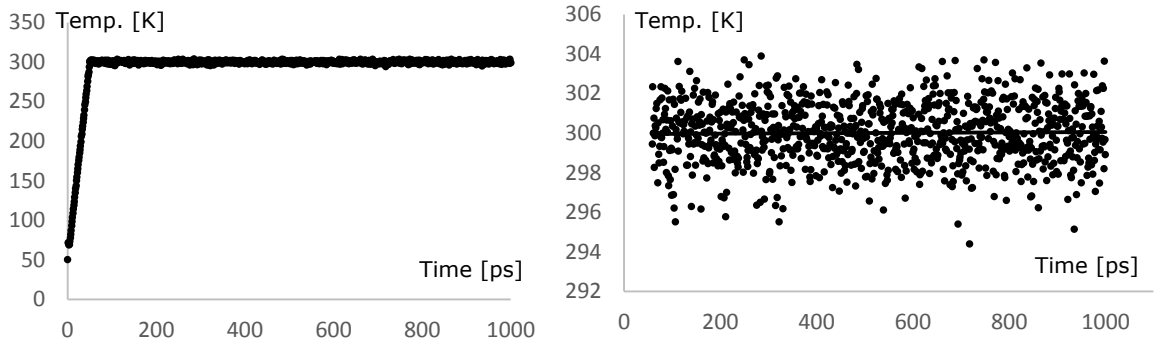

Figure 25: Lysozyme 1 ns simulation using ab-t1000-dt1-md_vv-sys-nh4 parameters. The figure to the left is the whole simulation while the figure to the right is from timestep 50 and shows a close up after the system has been heated. Temperature is a function of timestep with the slope of the trendline 0.0001 with $R^2 = 0.0003$.

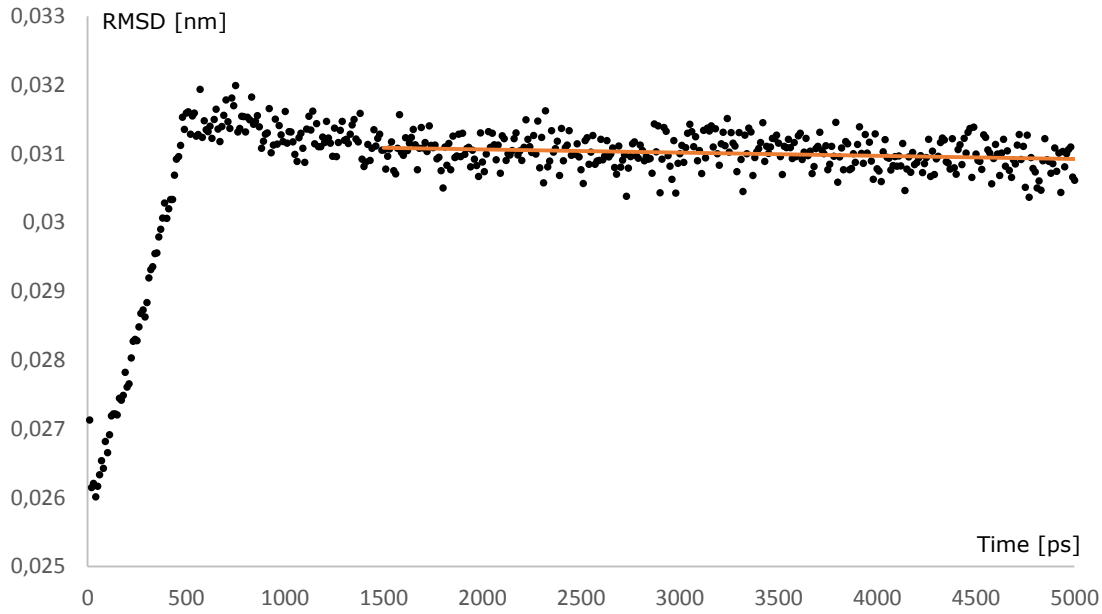

Figure 26: 5 ns simulation using ab-t1000-dt1-md_vv-sys-nh4 parameters. RMSD value is a function of timestep with trendline from 1000 ns to 5000 ns having slope of $-5*10^{-8}$ *and* $R^2=0.043$.

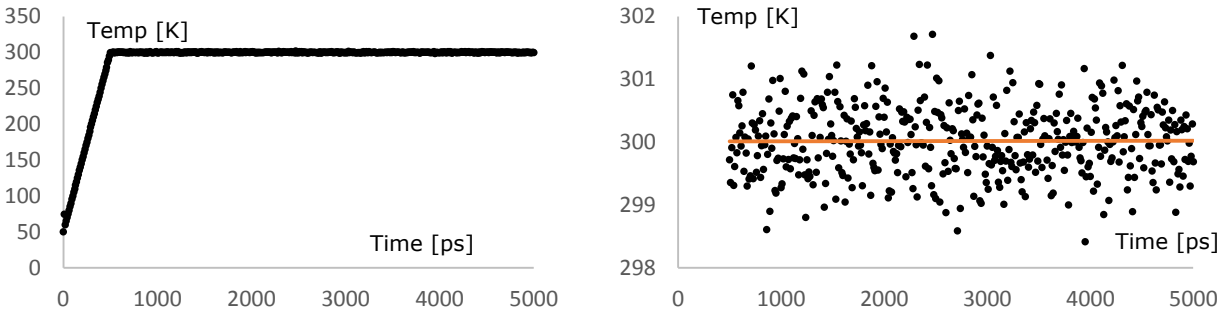

Figure 27: 5 ns simulation using ab-t1000-dt1-md_vv-sys-nh4 parameters. Temperature is a function of timestep. Slope of a trendline from 1000 ns to 5000 ns is $-2*10^{-6}$ with $R^2=2*10^{-5}$.

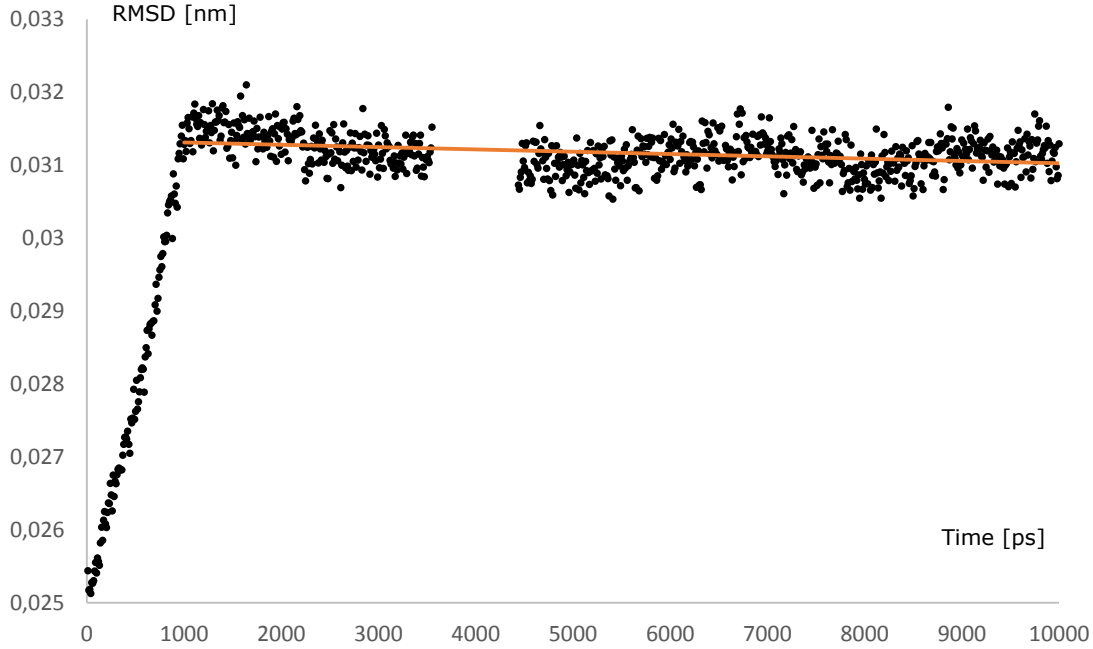

Figure 28: 10 ns simulation using ab-t1000-dt2-md_vv-sys-nh4 parameters. RMSD value is a function of timestep with trendline from 1000ns to 10000ns having slope of $-3*10^{-8}$ and $R^2=0.112$. The absence of data points in the graph are due to corrupted data file.

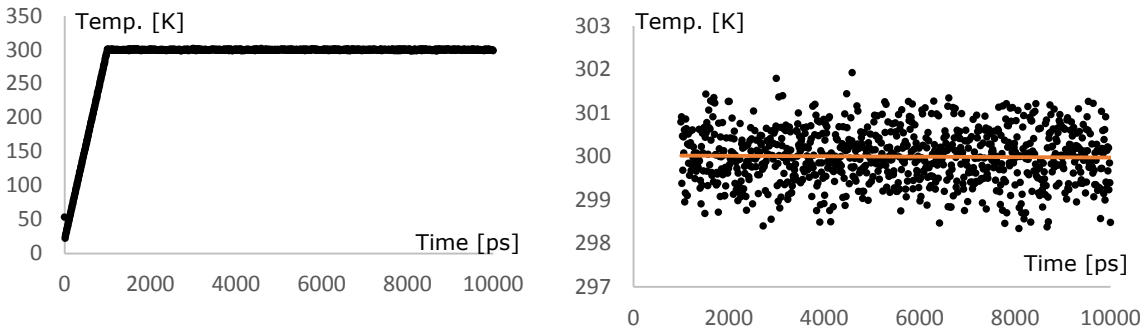

Figure 29: 10 ns simulation using ab-t1000-dt2-md_vv-sys-nh4 parameters. Temperature is a function of timestep. Slope of a trendline from 1000 ns to 10000 ns is $-4*10^{-6}$ with $R^2=0.0004$.