Finding the Good Vibes of Life, Man: Identifying Catalytic Vibrations in Enzyme Catalysis

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Master of Science

at

The University of Waikato

by

Daniel-Albert Schipper

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2020

Abstract

Enzyme catalysis is one of the most vital components of life. As such the elucidation of the exact mechanics of this kind of catalysis is important. The aim of this work is to look into enzyme catalysis though the lens of macromolecular rate theory (MMRT). MMRT is an idea focused around the heat capacity of an enzyme changing over the course of a reaction. Our hypothesis is that this change in heat capacity is strongly tied to the nature of enzyme catalysis. As such, the focus of this thesis is to determine what a large change in heat capacity at the transition state on the reaction pathway means and why it happens.

From previous work I know that there is a difference in heat capacity when an enzyme is bound to a reaction state compared to a transition state analog. One of the components of this project is to investigate the causes of this difference using all-atom molecular dynamics (MD) and calculations based on the MD trajectories.

In the harmonic case, it was found that there was no significant difference between our chosen enzyme in the reaction state and it in a transition state analog; which was possibly either due to insufficient simulation time or higher order anharmonic behavior driving the change in heat capacity. I computed the vibrational frequencies used to test this from the trajectories of a molecular dynamics simulation.

In addition, the patterns of how the change in heat capacity evolves with temperature acts similarly to how changes in heat capacity act during phase transitions. These transitions are characterized by long range fluctuations, which I aimed to detect; however, no change in the distance of fluctuations was detected in the harmonic case in the simulations.

Finally, I looked into computing heat capacities for a large system using the variance in energies. Previous attempts at doing this found reasonable values for the change in heat capacity when looking at an enzyme bound to a transition state analog vs bound to the reactant state. However, the absolute values of the heat capacities were an order of magnitude too high for each state. I investigated why this was the case and concluded that it was almost entirely due to not including quantum effects and the fact that interaction

energies between the system and the solvent are dynamic.

Acknowledgements

I would like to thank Vic Arcus for being my supervisor and driving me though the process. Him being willing to deal with my crazy ideas is what lead to this thesis being what it is. I would like to thank Jo Lane and Bill Rodgers as well for effectively being secondary supervisors for various parts of the project.

I would also like to thank Chris Batterton for explaining the underlying physics and providing the first draft of the mathematics sections of this thesis.

Next I would like to thank the Waikato TSG group for letting me use the undergrad linux computers for the compute requirements of the Thesis.

In addition, I would like to than the editors of this Thesis, namely Chris Batterton, Mackenzie Steele, Adele Williamson and Emily Grout. These people have allowed this thesis to not be a huge mess.

I would also like to thank Mark Van Der Kamp and Adrian Mulholland for letting me do a month long overseas exchange in Bristol.

Finally, I need to thank the C.2.10 lab group along with the rest of my friends and family for keeping me sane over the year that I have been working on this thesis, you have all been a big help.

And of course I need to thank Google for teaching me basically everything I needed to know for this thesis.

Contents

1 Introduction **9**

1.1 Introduction 9

1.1.1 Usefulness of enzymes in green chemistry 10

1.2 Introduction to phase space 11

1.3 Introduction to transition state theory 11

1.4 Hypotheses regarding enzyme catalysis 13

1.4.1 Surface chemistry 13

1.4.1.1 Surface chemistry and enzymes 14

1.4.2 Artificial enzymes and problems with surface catalysis 16

1.4.3 Vibrational catalysis 17

1.4.3.1 Stochastic search 17

1.5 Macromolecular rate theory 18

1.6 Vibrational coupling and MMRT 19

1.7 Statistical thermodynamics 22

1.7.1 Heat capacity 22

1.7.2 Computation 22

1.7.3 Density of states 23

1.8 Molecular Dynamics 23

1.9 MalL 25

2 Computing the Density of States from the Velocity Autocorrelation Function of a Molecular Dynamics Simulation **27**

2.1 Introduction 27

2.2 Theory 28

2.2.1 Density of States 28

2.2.2 Computation of DoS from trajectory data 28

2.3 Derivation for density of states 29

2.3.1 Derivation for heat capacity 34

2.4 Hypothesis 37

2.5 What do we expect DoS to look like 38

2.6 Methods 38

2.7 Results 41

2.7.1 Heat capacity 41
2.7.2 DOS graphs 41
2.8 Discussion 41
2.8.1 Visual inspection of DoS 41
2.8.2 Normal mode analysis 43
2.8.3 Heat capacity 44
2.8.4 DoS shift 44
2.8.5 Interesting features of the DoS 46
2.8.6 Metastable states 47
2.8.7 Anharmonic behavior 47
2.8.8 Detuning 50
2.8.9 Changes in temperature 50
2.9 Conclusion 50

3 Correlation distances as possible evidence for a phase change 52
3.1 Introduction 52
3.2 Hypothesis 52
3.3 Methods 53
3.3.1 MD computations 53
3.3.2 Dot product 53
3.3.2.1 Reason for use 53
3.3.2.2 Implementation 54
3.3.3 Mutual information 54
3.3.3.1 Reason for use 54
3.3.3.2 Implementation 54
3.3.4 Normal modes 55
3.3.4.1 Reason for use 55
3.3.4.2 Implementation 56
3.4 Results and discussion 57
3.5 Conclusion 57

4 Anomalous heat capacity 60
4.1 Introduction 60
4.2 Derived paper 60
4.3 Underlying principles 61
4.3.1 Derivation of heat capacity from variance 61
4.3.2 Hidden assumptions 63
4.4 Quantum issues 64
4.4.1 Quantum derivation 64
4.4.2 Classical derivation 68
4.4.3 A note on anhamonicity 71

4.5 Proposed reasons for error . . . 71
4.5.1 Quantum problems . . . 72
4.5.2 Water coupling . . . 72
4.5.2.1 Charge separation . . . 72
4.5.2.2 Hydrophobic interactions . . . 72
4.5.3 Clustering . . . 72
4.5.4 Non-conservation of energy . . . 73
4.5.5 Force field failure . . . 73
4.5.6 Peptide failure . . . 73
4.5.7 Temperature coupling . . . 73
4.6 Tests for identifying the cause of the anomalous heat capacity 73
4.6.1 MalL . . . 74
4.6.2 Barnase . . . 74
4.6.3 Barnase water . . . 74
4.6.4 Sugars . . . 74
4.6.5 Quad ala/gly . . . 75
4.6.6 Quad ala/gly with GROMOS force field . . . 75
4.6.7 Quad gly with finer time steps . . . 76
4.6.8 Quad gly with tighter buffers . . . 76
4.6.9 ala/gly dimers . . . 76
4.6.10 gly dimer with different sliding windows . . . 76
4.6.11 Gly kinetic energy . . . 76
4.6.12 Other peptides . . . 77
4.6.13 Clustering . . . 77
4.6.14 Changes in interaction energies . . . 78
4.7 Conclusion . . . 78

5 Summary and Conclusions 80
5.1 Summary . . . 80
5.1.1 Density of States . . . 80
5.1.2 Correlation distance and phase change . . . 80
5.1.3 Anomalous heat capacity of proteins . . . 81
5.1.4 Synthesis . . . 81
5.2 Future research . . . 81
5.3 Conclusion and speculation . . . 82

List of Figures

1.1 Energy surface of cyanide isomerization 13

1.2 Enzyme 15

1.3 An example of the types of curves produced under MMRT. . . 19

1.4 Change in $\Delta C_P^{\ddagger}$ for the oxidation of glucose by ssGDH with increasing deuteration for said glucose, taken from "Uncovering the relationship between the change in heat capacity for enzyme catalysis and vibrational frequency through isotope effect studies", used with permission 21

1.5 Dirac combs 24

1.6 MalL structure 26

2.1 Density of states example 38

2.2 Ligands of choice 39

2.3 Calculated density of states for reaction and transition states for MalL 42

2.4 Comparison of density of states between velocity calaculated and energy calculaed 43

2.5 Margin of error for calculations of the differences in density of states between MalL in reaction and transition states 45

2.6 A graph comparing computed density of states of MalL (orange) and KSI (blue) 46

2.7 A graph comparing computed density of states of MalL at 315K(orange) and 280K(blue) 48

2.8 A graph comparing computed density of states of wild type MalL (blue), S536R MalL (red), S536D,E554Q,V556R (green) and just the catalytic domain 49

3.1 A graph of the number of vibrational modes with a given correlation distance in the enzyme MalL comparing a reaction state (blue) to a transition state (orange). 58

4.1 A graph showing how computed heat capacity changes over different sliding windows used to compute heat capacity. 77

Chapter 1

Introduction

1.1 Introduction

Enzymes are incredibly powerful catalysts which can produce reaction rate increases of up to 10^{26} when compared to the corresponding uncatalyzed rate[1]. Enzymes are also incredibly selective in the reactions that they catalyze. A very good example is carbohydrate chemistry. Here $\alpha - 1 - 3$ fucosyltransferase allows for the addition of a six membered fructose ring onto oxygen 3 of another ring and only that specific atom[2]. Compared to most chemical catalysts, enzymes allow for single reactions with incredibly high yield and no need for protecting groups or harsh condition. Compare enzyme based synthesis of complex carbohydrates[2], where all the ingredients are added into a one pot reaction with no need for preparation or protecting groups and chemical synthesis of complex carbohydrates[3], where highly specific protecting groups need to be applied and later removed, for synthesis of specific carbohydrates. As such, engineering enzymes for novel chemical reactions is a very important field of research. The development of engineered enzymes won the 2018 Nobel prize for Frances Arnold[4]. Designed enzymes generally perform very poorly compared to their natural counterparts[5]. This is an indication that we do not fully understand how enzymes work. A recent hypothesis[6] is that the vibrations of the enzyme contributes to catalysis. This hypothesis provides

a reason as to why enzymes are so large, because the size of the enzymes is required to support promoting vibrations. This hypothesis is controversial and warrants detailed investigation[7].

This thesis is focused around identifying catalytic vibrations using a wide variety of computational methods, mostly centered on how vibrations change over the course of an enzyme catalyzed reaction. My hypothesis is that the changes in vibrations along the reaction contribute to catalysis.

1.1.1 Usefulness of enzymes in green chemistry

Enzymes have a major advantage over traditional chemical methods in that they work in water under very mild conditions[8]. This makes enzymes very useful for green chemistry as solvents can be disposed of simply and at low cost[9]. Enzymes are also generally safer than chemical catalysts. For example a purified enzyme is very unlikely to cause harm even if injected, spilled onto skin or in contact with eyes[9]. Compare this to tetrabutyl lithium which is commonly used in the formation of carbon-carbon bonds. Tetrabutyl lithium is incredibly basic, will catch fire on contact with water and air and must be kept at dry ice temperatures for stability[10]. Similar reactions to what is preformed by TBL are done in the body constantly in aqueous solutions at pH 7 and room temperatures[11].

Enzymes follow almost all the rules of green chemistry[12] by using sugar as the only energy source, not requiring any kind of toxic solvent and avoidance of protecting groups. While there are significant byproducts involved with using enzymes, they are easy to dispose of in an industrial setting. Side products left over from a reaction, like AMP and pyruvate, are very biodegradable. Because of the conditions required for enzyme catalysis there is low risk of safety accidents due to chemical spills.

1.2 Introduction to phase space

In many cases a system can be fully described by a list of numbers[13]. As an example our solar system can be described with the location of all the planets. In these cases the overall system can be described as a point in a higher dimensional space. In the case of the solar system there are eight planets, each with an X,Y and Z coordinate, making a total of 24 coordinates. These 24 coordinates describe a 24 dimensional phase space. As such we can reason about the time evolution of the solar system as a point moving though this 24 dimensional space. Phase spaces are useful because they allow for geometric reasoning about complex systems. Some properties, like the energy of a system, can be computed for the entire system, this means that we can assign each point of the phase space with a value. This is usually called a surface as the new value is seen as an additional dimension, creating a surface in n+1 dimensional space. This is very useful for building mathematical models of physical systems as the time evolution of the system acts almost the same as a ball rolling around on a surface.

1.3 Introduction to transition state theory

Transition state theory is a method designed to compute rates of reactions and their temperature dependence[14]. It works by defining three clusters of microstates: the reaction state, the transition state, and the product state. The reaction state is the set of all microstates that can be defined as the reactant; the product state is defined as the set of all microstates that can be defined as the product. The transition state is defined using an enumerable list of microstates that are in between the two sets of reactant and product microstates[15]. Each of these sets of transition state microstates are usually defined by a first order saddle point in positional phase space with respect to energy. At that microstate the gradient of energy with respect to all atomic positions is zero and the Hessian matrix has a single negative eigenvalue, ex-

actly 6 zero eigenvalues and all other eigenvalues being positive. The six zero eignavalues are because the system is transitionally and rotationally invariant. The transition state is defined as the set of microstates in the hyperplane that is orthogonal to the eigenvector that corresponds to the negative eigenvalue. Do note that there can be multiple saddle points and each one will have its own hyperplane, the transition state is the union of all these hyperplane generated sets. The composition of each set of states is entirely determined by the energy landscape of the overall system, as such changing the landscape can alter the nature of the sets of microstates[15]. This can be important when changing solvents, as differences in dielectric constants, lead to changes in the potential energy landscapes. An example of this for a very simple molecule is shown in figure 1.1.

From these sets of microstates Gibbs free energy can be computed for each set. Using Gibbs free energy, as opposed to the set of microstate energies, allows for treating each of the three sets as three macrostates. Transition state theory predicts that only the difference in Gibbs free energy between the reaction macrostate and the transition macrostate determines the rate of reaction, modified by a transmission coefficient (κ)[14]. Because the transmission coefficient cannot be separated from entropy via measurements of reaction rates at different temperatures (which is the easiest way to parameterize the Eyring equation), it is usually taken to be unity, allowing it to be dropped from the formalism. There has been significant research into the nature of the transmission coefficient and how it applies to enzyme catalysis[15]. The transmission coefficient is dependent on the exact dynamics of the system.

Transition state theory formally states that the rate of reaction is the barrierless crossing frequency ($\frac{\kappa k_B T}{h}$) multiplied by the relative population of the reaction state and the transition state ($e^{\frac{-\Delta G^{\ddagger}}{k_B T}}$), forming:

$$k(t) = \frac{\kappa k_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{k_B T}} \tag{1.1}$$

where $k(t)$ is the rate of reaction, k_B is the Boltzmann constant, h is Planck's constant and $\Delta G^{\ddagger}$ is the change of Gibbs free energy between the transition state and the reaction state.

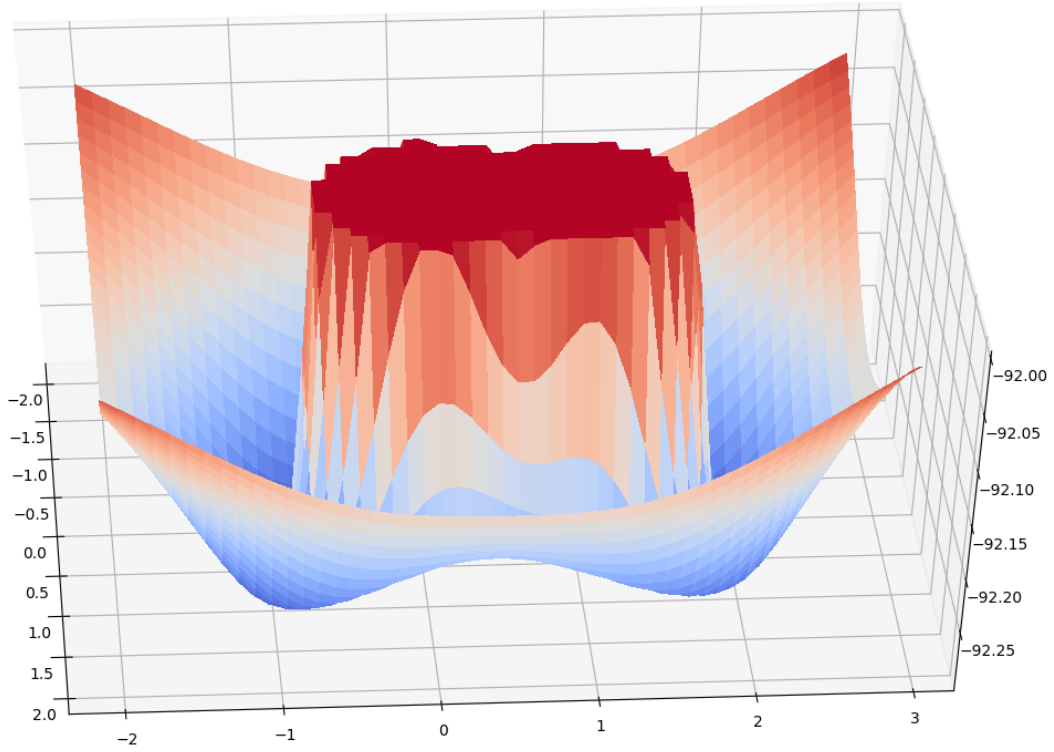

Figure 1.1: An example of the energy phase space for the isomerization of cyanide into isocyanide. Here the locations of the carbon and nitrogen atoms are fixed while the hydrogen can move around. The height of the surface dictates the energy at that configuration. The reaction state is one of the pits and the product state is the other, depending on which direction you want to run the reaction. The transition states are the saddle points between the two states. X and Y axis are in measured in angstroms and describe the location of the hydrogen relative to an arbitrary zero while the Z axis is in Hartrees and describes the energy of that structure. These energies were computed from the gaussian09 software package[16].

1.4 Hypotheses regarding enzyme catalysis

1.4.1 Surface chemistry

Traditional chemical catalysis falls into two categories, small molecule catalysis and surface catalysis. Small molecule catalysts work by creating an entirely new reaction pathway based around the catalyst, as seen by the hydrogenation of alkenes to alkanes[17], where a set of reactions focused on the metal center

drives the conversion. This alternate reaction pathway is focused on additive oxidation and reductive elimination on the metal center[18]. Surface chemistry works by binding the reactants to the surface and then by stabilizing the transition state[19]. This alternate pathway dramatically increases the rate of reaction. A key example of this is catalytic converters in cars. These converters work by adsorbing NO_2 onto its surface and then causing a reaction to break it down into N_2 and O_2[20] . It is currently thought that almost all enzymes work in the same way, by pre-organization of the various reactants to all bind together in the correct conformation and then by stabilizing the transition state[19], part of this thesis is to challenge this idea. This is referred to as the lock and key mechanism[19].

The precise mechanism by which surface type catalysts work is that they alter the energy landscape of phase space in such a way that the reaction state of the reactant-catalyst complex has moderately lowered Gibbs free energy and that the transition state Gibbs free energy is significantly lowered, thus decreasing $\Delta G^{\ddagger}$ [20]. In enzyme kinetics, this is thought to work primarily by altering the electrostatic environment around the reaction and transition state as opposed to directly bonding to the substrate like a metal catalyst does[19].

1.4.1.1 Surface chemistry and enzymes

Enzymes are large macromolecules composed of one or more chains of amino acids linked by peptide bonds[21], see figure 1.2. These chains are folded into very complex three-dimensional structures. A small subset of the enzyme's surface is an active site where the substrate can bind and where the reaction occurs.

Historically, the mechanism for enzyme catalysis was thought to be effectively surface catalysis similar to a catalytic converter[19]. The traditional view is that enzymes are capable of creating the highly specific electrostatic environment around the active site for catalysis to be effective due to the exact way the enzyme is folded[22]. Everything around the active site is there

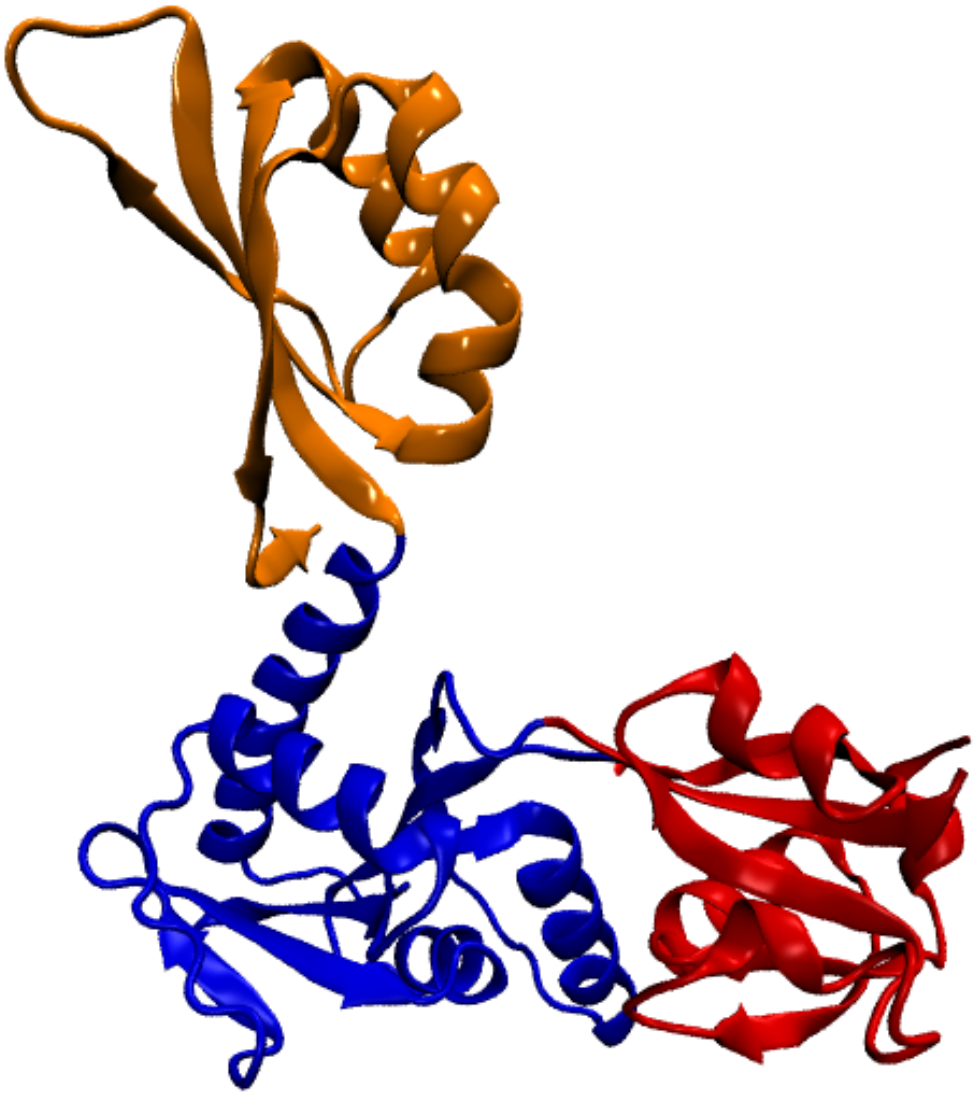

Figure 1.2: An example of an enzyme (pdb code 4YB5) colored by the different domains, red and blue are both catalytic domains while yellow is a more flexible allosteric domain.

to ensure that the active site is folded in the exact way required, thus acting as a scaffold[19].

Note that there are sometimes secondary binding sites for allosteric ligands which regulate enzyme activity[21]. These allosteric sites are for ligands that can alter catalysis despite being far from the active site[21]. This is enigmatic because in a static surface model, it is unclear how a ligand binding far from the active site can influence the rate of reaction[23].

This static surface model was then updated by including conformational change following substrate binding in which the addition of the substrate to

the enzyme causes a shift in the structure of the enzyme. This new structure allows for precise surface chemistry to occur leading to catalysis[7].

1.4.2 Artificial enzymes and problems with surface catalysis

The construction of artificial enzymes has long been a goal for biochemists. One approach has been to utilize antibodies which are adept at binding to an arbitrary predefined ligand[24]. Antibodies have been engineered to bind to transition state analogs with the aim of stabilizing the transition state for a reaction, the central hypothesis for enzyme catalysis[25]. In general, catalytic antibodies have not been very successful as a route to enzyme design[6]. This suggests that static transition state binding is not sufficient for catalysis as effective as native enzymes.

Alternatively, modern computational design can also be used create a structure that strongly binds to a transition state analog[26]. Directed evolution can also be used for constructing enzymes; this insight won the 2018 Nobel prize[4]. In all of these cases, the enzyme is designed to bind to the transition state as tightly as possible. However, these enzymes tend to be much smaller than natural enzymes, and they are far worse catalysts. Most natural enzymes tend to have a catalytic improvement of between 10^{12} to 10^{26}[1] while artificial enzymes show improvements on the order of 10^{6}[6], which is a million to trillion fold slower than naturally occurring enzymes. This discrepancy indicates that current methods to develop enzymes are insufficient. Given that evolution has led to the formation of large enzymes, when the production of enzymes takes up a large portion of the cell's resources[21], it is fair to say that the difference in catalysis may be driven by the difference in enzyme size. If enzymes could be made smaller with the removal of amino acids then there would be far smaller enzymes in nature due to the very strong evolutionary pressure.

1.4.3 Vibrational catalysis

One influential paper proposing the concept of vibrationally driven catalysis is "Good vibrations in enzyme catalysis" by Hay et al. [27]. In brief, vibrational coupling states that, in the transition state, the velocity phase space is restricted in addition to the positional phase space. These ideas are further explained and expanded in section 1.6. In the Schwartz and Schramm[27] paper the authors describe their use of the kinetic isotope effect to show how the catalysis of purine nucleoside phosphorylases (PNP) could not be done simply using only surface chemistry. This was firstly shown by comparing two PNP enzymes from humans and cows, these enzymes have the same active site (where surface chemistry happens) but different transition states. The change in transition states was determined via temperature dependent kinetic isotope effects, while the change to the active site was determined by binding of a transition state analog. A modification of the human enzyme was performed by swapping out two amino acids from the human enzyme with the bovine, $25\mathring{A}$ away from the active site. From the kinetic isotope effect it was determined that this chimeric enzyme had the transition state from the bovine enzyme. This cannot be easily explained via just surface chemistry as the surface of the active site does not change. Using quantum mechanical/molecular mechanical molecular dynamics simulations (QM/MM), the authors determined that this shift in amino acids led to changes in the vibrational modes of the enzyme, and further, that the vibrational modes were implicated in catalysis.

1.4.3.1 Stochastic search

The hypothesis proposed in [27] as to what is happening in the PNP reactions is that there is a hammer and anvil effect. If specific vibrations are at specific amplitudes and specific phases, then the width of the transition state is compressed. Normally this would not lead to catalysis as the height of the transition state barrier determines rate of reaction. But since only one proton is being transferred, quantum tunneling becomes available. This com-

pression of the distance required to traverse across the transition state means that tunneling from one state to another becomes more probable. The precise set of conformations required for this to happen is quite rare. Thus, in order for catalysis to happen, the enzyme substrate complex undergoes a stochastic search through the phase space of conformations and motions. This explains why turnover is on the order of microseconds while the actual reaction occurs over the order of femtoseconds.

Vibrational coupling in enzyme catalysis is extremely controversial and other authors have argued that this does not occur[28]. Mulholland et al have shown by simulation and experiment that the contribution of vibrational coupling is very small, at least in the case of dihydrofolate reductase [29].

1.5 Macromolecular rate theory

This thesis is focused on macromolecular rate theory (MMRT), which states that the well-known optimum temperature of enzyme catalysis is due to a change in heat capacity along the reaction coordinate[30]. Heat capacity is defined as the first derivative of the expectation value of the energy function for the system ensemble with respect to temperature $\partial_T\langle\phi\rangle$ and is directly related to the variance of the energy function in the system. Because enzyme systems are usually treated as canonical ensembles, holding temperature, pressure and number of moles constant, C_P is used. $C_P \equiv \partial_T\langle H\rangle = T\partial_T S$. MMRT is an extension of the Eyring equation (equation 1.2, 1.3), by allowing heat capacity to change between the reaction state and the transition state. This perturbs the values for entropy and enthalpy with respect to temperature leading to the MMRT equation (see equation 1.6). See figure 1.3 for an example of how MMRT describes a rate equation. This perturbation has been shown to fit with how enzymes respond to temperature across numerous enzymes and up to an ecosystem level [30][31]. The question of how heat capacity changes with reaction coordinate is the focus of this thesis.

$$k = \frac{k_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{k_B T}} \tag{1.2}$$

$$\ln(k) = \left(\frac{k_B T}{h}\right) - \frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B} \tag{1.3}$$

$$\Delta H^{\ddagger} = \Delta H^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P (T - T_0) \tag{1.4}$$

$$\Delta S^{\ddagger} = \Delta S^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P \ln(\frac{T}{T_0}) \tag{1.5}$$

$$\ln(k) = \left(\frac{k_B T}{h}\right) - \frac{\Delta H^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P (T - T_0)}{k_B T} + \frac{\Delta S^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P \ln(\frac{T}{T_0})}{k_B} \tag{1.6}$$

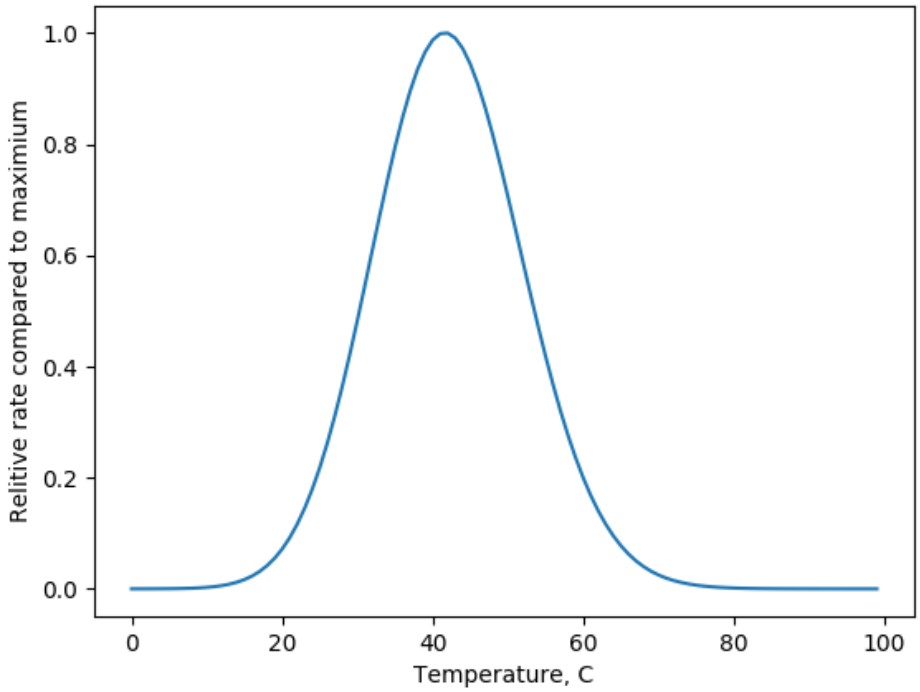

Figure 1.3: An example of the types of curves produced under MMRT.

1.6 Vibrational coupling and MMRT

In a traditional model of reaction dynamics, it is assumed that there is an abstract variable, known as the reaction coordinate, which defines the progress along the reaction[15]. The normal way to define a reaction coordinate[20] is to first define a reaction path, this can be seen as the shortest path between

the reaction and the product state that has its highest energy at the transition state. The transition state is a saddle point. A saddle point is formally defined in this context as a point in positional phase space for the atomic system where all first derivatives in terms of spatial coordinates are zero and the Hessian matrix only has a single negative eigenvalue, with all others being positive and at most six being zero (those vibrations being translation and rotation)[15]. When computing thermodynamic properties for the transition state, the potential energy surface of a hyperplane that is orthogonal to the single negative eigenvalue is considered[14].

Now because the reaction coordinate has a fuzzy definition, being defined by its properties, as opposed to a more fundamental definition, questions can be asked about whether the reaction coordinate is only a function of positions of the molecular system or if there are other factors. In a quantum system ψ when moving a system from state $|A\rangle$ to state $|B\rangle$ the reaction coordinate can be defined as λ such that $\psi = \lambda|A\rangle + \sqrt{1-\lambda^2}|B\rangle$ where $\lambda \in [0, 1]$. There is currently significant debate about whether the location of the system in velocity phase space is coupled to the reaction coordinate[6].

Previous research indicates that $\Delta C_P^\ddagger$ is strongly influenced by vibrational coupling along the reaction coordinate[32]. This means that the changes to the system that cause $\Delta C_P^\ddagger$ are intimately linked to the nature of the reaction. For example, the changes in $\Delta C_P^\ddagger$ for a proton-transfer reaction in glucose dehydrogenase and deuterium-transfer reaction for the same reaction is far higher than can be explained using traditional methods of computing heat capacity[33]. If the transition state can be fully described by a hyperplane in phase space then the structure of said hyperplane is conserved under isotopic substitution, as per the Born-Oppenheimer approximation[34]. There will be some slight differences in thermodynamic properties due to changes in masses leading to changes in zero point energy[20], however minor changes in zero point energy due to a single isotopic substitution cannot lead to a significant change in heat capacity for either the reaction state or the transition state[35][33]. There are

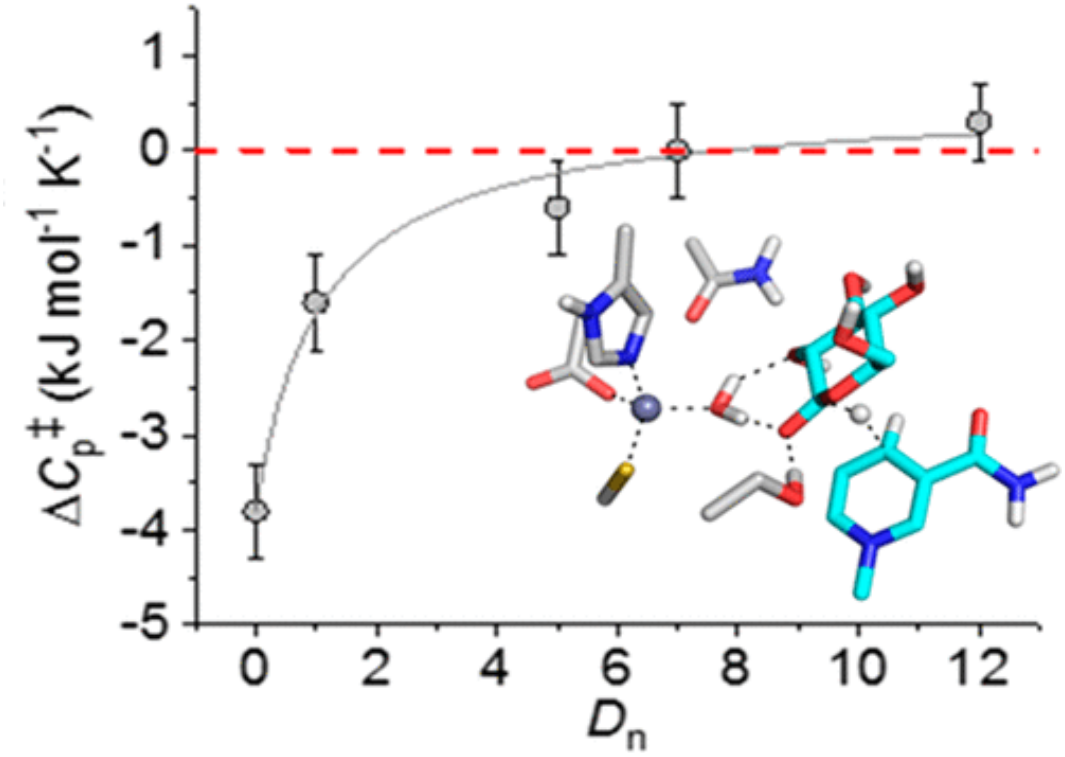

Figure 1.4: Change in $\Delta C_P^{\ddagger}$ for the oxidation of glucose by ssGDH with increasing deuteration for said glucose, taken from "Uncovering the relationship between the change in heat capacity for enzyme catalysis and vibrational frequency through isotope effect studies", used with permission

two options that could lead to this significant shift in $\Delta C_P^{\ddagger}$. One option is significant vibrational coupling (where location of the system in velocity phase space is as important as the location of the system in position phase space for defining a transition state). The other option is a change in the nature of the transition state. If the Born-Oppenheimer approximation is assumed, the potential energy surface for the entire system is conserved under isotropic substitution. In the classical case this would mean that the transition state will remain the same. In the quantum case, quantum tunneling can cause a difference between the two transition states, as deuterium is significantly less able to quantum tunnel than a proton is.

1.7 Statistical thermodynamics

1.7.1 Heat capacity

The core of MMRT is the change in heat capacity between the reaction state and the transition state, labeled $\Delta C_P^{\ddagger}$. Heat capacity can be defined as the derivative of the expectation value of system energy $\langle \phi \rangle$ with respect to temperature. Heat capacity can change if different constraints are put on a system. C_P, heat capacity at constant pressure, will be different from C_V, heat capacity at constant volume. In biochemical systems in an aqueous environment, it is assumed that pressure remains constant and therefore C_P is the most appropriate parameter; however, because water is imcompressible $C_P \approx C_V$[36]. As a transition state is a very small subset of the overall phase space, heat capacity for a transition state involves adding energy while remaining a transition state. This, combined with vibrational coupling restricting the velocity phase space of the transition state, is what I hypothesize causes the change in heat capacity on the reaction coordinate for enzyme catalyzed reactions.

1.7.2 Computation

At the time of the experimentation done for this thesis there were no known tools (to the author) for identifying vibrationally coupled transition states, so all computations are required to be at equilibrium. Computing statistical properties at equilibrium is significantly easier as a dynamic system without complex constraints can be used. While a true transition state cannot be analyzed this way, as it is not in equilibrium, an analog can be used instead. Indeed, transition state theory can be regarded as a quasi-equilibrium between reactant state and transition state[14].

Using classical statistical thermodynamics, the heat capacity can be computed from the partition function, the partition function can be computed from vibrational information of a single enzyme in isolation. If a harmonic assumption is taken, then each vibrational mode can be viewed independently

and their partition functions can be computed independently and then multiplied to give the partition function of the entire enzyme. Normally, rotational and translational effects have to be included, however their contributions are likely very small in this context as they are highly damped by the solvent and because there are only 6 modes of this form, compared to the tens of thousands of vibrational modes for an enzyme system.

1.7.3 Density of states

In order to compute the partition function, a list of vibrational frequencies is required. The equations for computing statistical properties can be converted from a sum to an integral via a continuum approximation. Here we replace the list of vibrational frequencies with a function called density of states (DoS) $G : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$[20]. If a list of vibrational frequencies can be enumerated then G is a Dirac comb, see following equation and figure 1.5.

$$G(\omega) = \sum_{i=1}^{N} \delta(\omega - \omega_i) \tag{1.7}$$

However this version of density of states allows for a less exact description, which is useful for empirical measurements, especially given that exact frequencies cannot be measured in a finite time. Due to the time frequency trade off of Fourier transforms[37], any approach based on the motions of atoms or wave-functions cannot produce truly accurate frequencies, as such the density of states is a useful way to allow for that uncertainty.

1.8 Molecular Dynamics

Bio-molecular systems have a key issue in that they are so large that, while full potential energy landscapes can be constructed, they cannot be usefully analyzed. Given that a key hypothesis of this thesis is that vibrations are required to properly understand the phase space of reaction and transition

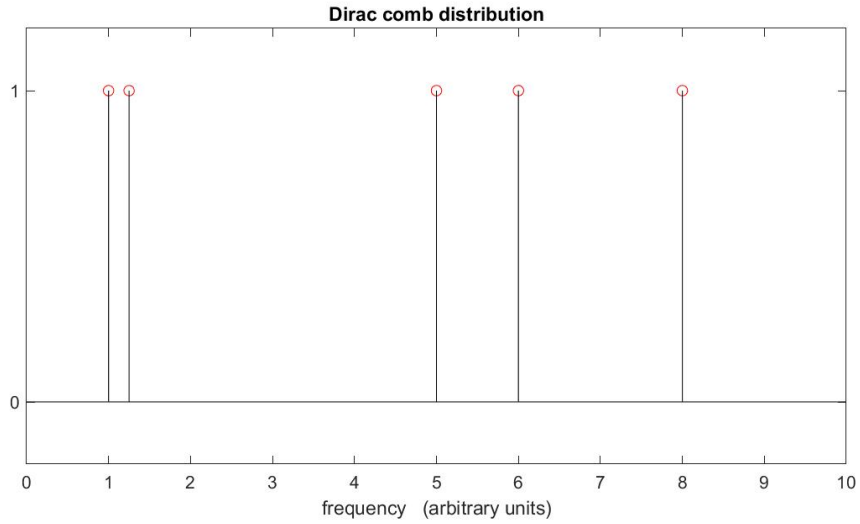

Figure 1.5: An example of a Dirac comb based on randomly generated delta functions.

states, normal Monte Carlo sampling is insufficient, as that only uses positional data[38]. As such, to sample large sections of phase space, molecular dynamics is required. Molecular dynamics is a way to explore the phase space of how position and vibration interact by taking a structure of a protein and allowing it to evolve in time. In molecular dynamics atoms are treated as classical point particles and time-evolve as per classical Hamiltonian mechanics[39]. The Born-Oppenheimer approximation is taken, assuming that the potential energy of the system is purely defined in terms of the location of nuclei. To compute the potential energy surface required for Hamiltonian physics, there are two approaches. The first is to use a quantum calculation. Here the wavefunction for the electrons, given the locations of the nuclei, is computed. From this wavefunction, the energy of the system is calculated. This is very computationally expensive and does not scale well to enzyme sizes. The other approach, which is a computationally cheaper option, is molecular mechanics[39]. With molecular mechanics the potential energy of a system is defined from a relatively small closed form solution. The following is a suitable example of a force field for molecular dynamics as applied to an enzyme[39].

$$V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2 + \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

In this equation the potential energy of the system (V) is, given a list of coordinates (r^N), the sum of the energy in the bonds between atoms (modeled as a harmonic oscillator in positional space), the energy of the angles (modeled as a harmonic oscillator in angular space), the energy of the dihedrals (modeled as a finite Fourier expansion of an arbitrary function), the energy of Van Der Waals forces (modeled as a Lennard-Jones potential) and electrostatic interactions (modeled with Coulomb's law)[39]. This force field has a major limitation, in the investigation of catalysis, as it cannot simulate bonds forming and breaking[39].

There are additional factors placed on top of Hamiltonian mechanics to convert the chaotic system from an NVE ensemble (holding number of atoms, volume and energy constant) to other thermodynamic ensembles. NPT is the most common, holding number of atoms, pressure and temperature constant[39].

1.9 MalL

The model enzyme for this work is MalL, an oligo-1,6-alpha-glucosidase[30]. MalL cleaves alpha-glycosidic bonds on the non-reducing end of starch chains[30]. This enzyme was chosen because it has a high $\Delta C_P^{\ddagger}$ of -10kj/mol K[30]. In addition, MalL was selected as it is the enzyme of choice for the Arcus lab experimental work allowing any hypotheses emerging from this theoretical work to be experimentally tested. MalL is the enzyme of choice for experimental research because it has a very simple and inexpensive assay, expresses and pu-

rifies very well and is highly stable. MalL also has multiple crystal structures, which means that its conformational phase space is well-defined. See figure 1.6 for the structure of MalL.

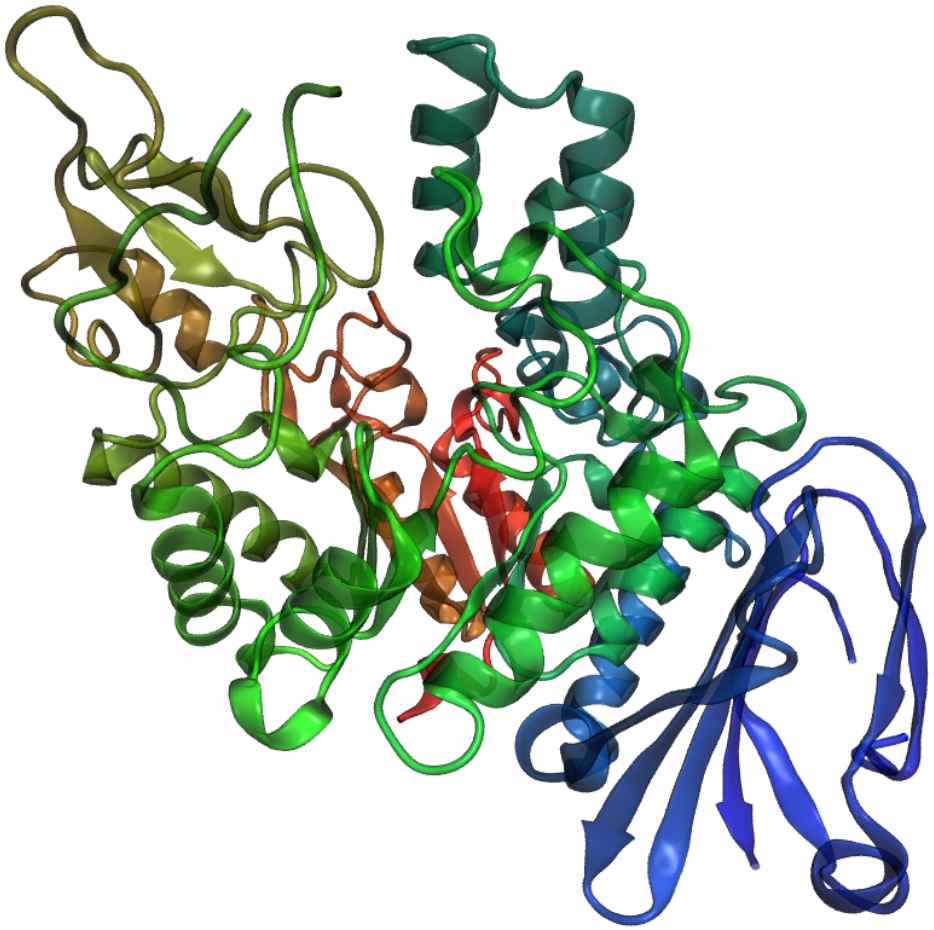

Figure 1.6: MalL structure

The structure of the enzyme MalL, coloring is dependent on position along the the chain, with red being N-terminal and blue being C terminal.

Chapter 2

Computing the Density of States from the Velocity Autocorrelation Function of a Molecular Dynamics Simulation

2.1 Introduction

The current major focus of theoretical MMRT is explaining the detailed statistical thermodymanics origins of $\Delta C_P^{\ddagger}$. If $|\Delta C_P^{\ddagger}| < R$ then it could be explained as the loss of a vibrational mode by virtue of being in a transition state as opposed to a reaction state (as the transition state is defined as a hyperplane), given that a single degree of freedom usually has a heat capacity of R[36]. Our hypothesis is that there is some form of vibrational coupling in the reaction that leads to differences in overall vibrational modes when comparing the enzyme-substrate complex with the enzyme-transition state complex. In the simplest case where vibrational frequencies change between these two states, this could be measured in differences in the density of states. As such, the aim of this section is to compute the density of states for both the enzyme-substrate complex and a pseudo enzyme-transition state complex from molecular dynamics

simulations.

2.2 Theory

2.2.1 Density of States

The density of states (DoS) is a useful approximation for describing systems with a very large number of vibrational modes. The definite integral between two frequencies of the DoS shows how many vibrational modes there are in that region. The non-integer component indicates the probability of an additional mode. Theoretically, in the harmonic case, the DoS should be a Dirac comb, with each delta function corresponding to a single vibrational mode. However, the methods we use are not accurate enough to isolate those deltas, causing them to be smeared out and significantly overlap.

Molecular vibrations are quantized[40]. As such, each vibrational mode has an integer number of quantum steps above the ground state that it may occupy. For a particular vibrational mode this is defined as the phonon occupation number. These follow a Bose-Einstein distribution in the gas phase at thermal equilibrium[36] under the rigid rotor model. However, in aqueous phase the vibrations of the water can resonate with some frequencies, because of a lack of weak coupling between the system and the surroundings, creating spikes along with other interesting behaviors.

2.2.2 Computation of DoS from trajectory data

The density of states can be computed from the trajectory information of a molecular dynamics simulation. The Fourier transform of the mass-weighted velocity autocorrelation (VAC) function gives the spectral energy distribution of the system in the harmonic case. This can be converted into the DoS by dividing by the expected energy of an independent mode at a given temperature (k_BT). This will lead to errors under a resonance situation or if the system is anharmonic. Because the system is anharmonic and rotating, the VAC will

tend to zero over time, thus only a finite time slice can be used. Due to the time-frequency trade off, this does mean that the DoS will be fuzzy and not an exact Dirac comb. In the case of sufficiently large molecules like enzymes, the states are so concentrated that the overall graph is very smooth and states cannot be separated out.

2.3 Derivation for density of states

For all derivations in this thesis the Braket notation is used $|v\rangle$ describes a vector v in some vector space, usually implied by context. $\langle a|b\rangle$ is defined as the inner product of $|a\rangle$ and $|b\rangle$, while $|a\rangle\langle b|$ is the outer product.

As has been previously stated, the density of states should contain all the relevant information to derive thermodynamic properties. We wish to derive a density of states from first principles, and in such a way as to take molecular dynamics data as an input, this data is restricted to the position, velocities, forces and energies of a system.

The classical mass-weighted velocity (energy) autocorrelation function for time lengths (durations) between zero and t can be expressed as

$$S(t) \;=\; \langle V(t)\,|\,M\,|\,V(0)\rangle$$

where V is the vector of velocites as indexed by atom, and M is the diagonal masses of the associated atom. Note that the correlations are time symmetric, $S(t) = S(-t)$. If we define the system time translation (evolution) operator to be based on a simple harmonic potential, then we have

$$|V(t)\rangle \;=\; U(t)\,|V(0)\rangle$$

where, time translation is unitary, $U^\dagger = U$. By assuming a simple harmonic

temporally evolving the atomic velocity, we have

$$S(t) \;=\; \langle V(0)\,|\,U(t)\,M\,|\,V(0)\rangle$$

Because the time evolution is harmonic, its eigendecomposition is given from

$$\begin{aligned}\partial_t^2 X \;&=\; -\,AX\\ \Rightarrow\; A \;&=\; \sum_k \omega_k^2\,|k\rangle\langle k|\end{aligned}$$

where X is the positional vector, centered on the zero vector.

Using the result that $U(t) \;=\; \exp(t\sqrt{-A})$, we have

$$U(t) \;=\; \sum_k e^{\pm i\omega_k t}\,|k\rangle\langle k|$$

where the summation is over eigenvectors k, and the eigenfrequencies and eigenbases are ω_k and $|k\rangle\langle k|$, respectively. Thus, we have an orthonormal basis set ($\langle i|j\rangle \;=\; \delta_{ij}$) in atomic numbering which allows us to change the velocity and momentum basis to an eigenbasis

$$\begin{aligned}|V(0)\rangle \;&=\; \sum_k a_k|k\rangle\\ M|V(0)\rangle \;&=\; \sum_k m_k a_k|k\rangle\end{aligned}$$

where $a_k = a_k^\dagger$ are the velocity magnitudes for each atom (collectively, the k^{th} eigenstate). We are now ready to evaluate the autocorrelation function in this

system

$$\begin{aligned} S(t) &= \langle V_0 \,|\, U(t)\, M \,|\, V_0 \rangle \\ &= \Big(\sum_i a_i^\dagger \, \langle i| \Big) \Big(\sum_k e^{\pm i\omega_k t} \, |k\rangle\langle k| \Big) \Big(\sum_j a_j m_j \, |j\rangle \Big) \\ &= \sum_{i,j,k} m_j a_i^\dagger a_j \, e^{\pm i\omega_k t} \, \langle i|k\rangle\langle k|j\rangle \\ &= \sum_{i,j,k} m_j a_i^\dagger a_j \, e^{\pm i\omega_k t} \, \delta_{ik} \, \delta_{kj} \end{aligned}$$

The summation can be collapsed around $i = j = k$ via the δ-functions, resulting in

$$S(t) \;=\; \sum_k m_k |a_k|^2 \, e^{\pm i\omega_k t} \;=\; \sum_k E \, e^{\pm i\omega_k t}$$

where the total energy of the k^{th} atom, E, is given by $m_k a_k^2$. This is the symmetric autocorrelation function, to elucidate the density of states, we rotate this result by $\pi/2$ in the complex plane (Fourier transform) to frequency basis

$$\tilde{S}(\omega) \;=\; \mathscr{F}[S(t)](\omega)$$

This can be processed numerically using standard software packages, or we can proceed with an analytic derivation.

$$\begin{aligned} \tilde{S}(\omega) &= \mathscr{F}[S(t)](\omega) \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt \; e^{-i\omega t} \; S(t) \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt \; e^{-i\omega t} \; \sum_k E \, e^{\pm i\omega_k t} \end{aligned}$$

The order of operations can be swapped such that each normal mode is Fourier transformed independently

$$\tilde{S}(\omega) \;=\; \frac{1}{2\pi} \; \sum_k E \int_{-\infty}^{+\infty} dt \; e^{i(-\omega \pm \omega_k)t}$$

The time-integrals can be evaluated into their Cauchy principle components

via

$$\int_0^\infty d\tau\ e^{\pm i\Omega\tau}\ =\ \pi\,\delta(\Omega)\ \pm\ i\mathscr{P}\frac{1}{\Omega}$$

where $\mathscr{P}$ is the Cauchy principal component.

To use this result, we use the symmetry of $S(t)$ to change the integration region to $t \in [0, +\infty)$, and double the size of the resultant integral

$$\begin{aligned}\tilde{S}(\omega)\ &=\ \frac{1}{\pi}\ \sum_k E\ \int_0^\infty dt\ e^{-i(\omega\pm\omega_k)t}\\ &=\ \frac{1}{\pi}\ \sum_k E\left(\pi\,\delta(\omega\pm\omega_k)\ -\ i\mathscr{P}\frac{1}{\omega\pm\omega_k}\right)\end{aligned}$$

Separate real and imaginary components

$$\begin{aligned}\tilde{S}(\omega)\ &=\ \frac{1}{\pi}\ \sum_k E\ \int_0^\infty dt\ e^{-i(\omega\pm\omega_k)t}\\ &=\ \sum_k E\,\delta(\omega\pm\omega_k)\ -\ \frac{i}{\pi}\sum_k E\ \ \mathscr{P}\frac{1}{\omega\pm\omega_k}\end{aligned}$$

The Dirac δ-function is undefined under summation if its argument is a continuous variable, we therefore take the continuum limit over the k-mode frequencies ($\omega_k \in [0, \infty)$) and in doing so introduce the density of states

$$\tilde{S}(\omega)\ =\ \int_0^\infty d\omega_k\ g(\omega_k)\ E\,\delta(\omega\pm\omega_k)\ -\ \frac{i}{\pi}\int_0^\infty d\omega_k\mathscr{P}\frac{g(\omega_k)\ E}{\omega\pm\omega_k}$$

By observation of the frequency-matching δ (resonance) term under the integral, we see that it allows a non-zero value if and only if $\omega \pm \omega_k = 0$. The resonance term can in general be true for $\omega - \omega_k = 0$, however, $\omega = \omega_k = 0$ can only be true for $\omega_k = 0$. In the specific case of $\omega_k = 0$, $\omega + \omega_k = \omega - \omega_k$. So the case of $\omega - \omega_k = 0$ can be taken in this case without loss of generality.

$$\tilde{S}(\omega)\ =\ \int_0^\infty d\omega_k\ g(\omega_k)\ E\,\delta(\omega-\omega_k)\ -\ i\int_0^\infty d\omega_k\mathscr{P}\frac{g(\omega_k)\ E}{\omega-\omega_k}$$

Note the removal of the $\pm$ symbol.

The integral over ω_k is now readily solvable, and note that the imaginary component represents a frequency (Lamb) shift to the system, $\Delta\omega$

$$\tilde{S}(\omega) \;=\; g(\omega)\, E(\omega) \;-\; i\Delta\omega$$

If we assume the Lamb shift is sufficiently small, we can omit it, and we thus arrive at an analytic result for the density of states

$$g(\omega) \;=\; \frac{\tilde{S}(\omega)}{E(\omega)}$$

If we use the classical, harmonic approximation for energy, $E = k_B T$ (energy is purely due to thermal contributions), and reinsert the discontinuous $\tilde{S}$ with summation over atoms, we have

$$g(\omega) \;=\; \sum_k \frac{E}{k_B T}\delta(\omega - \omega_k)$$

This is non-dimensionalised by the ratio of oscillatory to thermal energy. If each atom has energy $E_k = k_B T$, then the ratio is unity and $g(\omega)$ is simply a Dirac comb distribution of unit weight

$$g(\omega) \;=\; \sum_k \delta(\omega - \omega_k)$$

Thus, we have analytically derived the density of states for a classical, harmonic system from an initial state that is knowable via molecular dynamics.

2.3.1 Derivation for heat capacity

The ensemble partition function, Q, for an ideal gas composed of N atoms is the collection of partition functions q of each individual molecule according to

$$\ln Q \;=\; N\,\ln q \;-\; \ln N \tag{2.1}$$

where the microstates q_i are given by the ensemble

$$\begin{aligned} q \;&=\; \prod_{i=0}^{n} q_i \\ q_i \;&=\; \sum_{n=0}^{\infty} e^{-n\beta\hbar\omega_i} \;=\; \frac{1}{1-e^{-\beta\hbar\omega_i}} \end{aligned}$$

by the geometric progression. Taking the natural log of q for use in (2.1)

$$\begin{aligned} \ln q \;&=\; \ln\left(\prod_{i=0}^{n} q_i\right) \;=\; \sum_{i=0}^{n} \ln q_i \\ &=\; \sum_{i=0}^{n} \ln\left(\frac{1}{1-e^{-\beta\hbar\omega_i}}\right) \end{aligned}$$

If we let $n \to \infty$, and then take the continuum limit for state spacings, an integral density of states can replace the sum

$$\ln q \;=\; \int_0^{\infty} d\omega_i\, g(\omega_i)\, \ln\left(\frac{1}{1-e^{-\beta\hbar\omega_i}}\right)$$

The heat capacity of a system (for a generalized energy function ϕ), C, is defined via the partition function

$$q \;\equiv\; \sum_{n} e^{-\beta\phi_n}$$

From this, the expectation value of energy in this system is

$$\begin{aligned}
\langle \phi \rangle &= -\frac{\partial \ln q}{\partial \beta} \\
&= -\frac{1}{q}\frac{\partial q}{\partial \beta} \\
&= -\frac{1}{q}\frac{\partial}{\partial \beta}\sum_n e^{-\beta\phi_n} \\
&= -\frac{1}{q}\sum_n (-\phi_n) e^{-\beta\phi_n} \\
&= \sum_n \phi_n \frac{e^{-\beta\phi_n}}{q} \\
&= \sum_n \phi_n P_n
\end{aligned}$$

where the thermodynamic β is the inverse of energy due to temperature, $\beta \equiv 1/k_B T$. The heat capacity is defined as the change of expectation values of energy for the ensemble over temperature

$$C \equiv \frac{\partial \langle \Phi \rangle}{\partial T}$$

The use of Φ for energy is here to denote that the generalized energy function allows for variance in volume and chemical potential and across all ensembles, so that we are not restricted to microcanonical ensembles (isolated systems).

Thus, the heat capacity is derived from the energy expectation value in terms of the partition function

$$\begin{aligned}
C &\equiv \frac{\partial \langle \phi \rangle}{\partial T} \\
&= \frac{\partial}{\partial T}\Big(-\frac{\partial \ln Q}{\partial \beta}\Big) \\
&= -\frac{\partial \beta}{\partial T}\frac{\partial^2}{\partial \beta^2}\ln Q \\
&= \frac{1}{k_B T^2}\frac{\partial^2}{\partial \beta^2}\ln Q
\end{aligned}$$

Substitute in (2.1)

$$C \;=\; \frac{1}{k_B T^2}\,\frac{\partial^2}{\partial\beta^2}\Big(N\,\ln q \;-\; \ln N\Big)$$

If we assume that the number of atoms does not change with temperature, as with a canonical ensemble, then

$$C \;=\; \frac{N}{k_B T^2}\,\frac{\partial^2}{\partial\beta^2}\ln q$$

Now substitute the integral (continuum) representation of $q(\beta)$ and differentiate with $\beta = 1/k_B T$

$$C \;=\; \frac{N}{k_B T^2}\,\frac{\partial^2}{\partial\beta^2}\int_0^\infty d\omega_i\, g(\omega_i)\,\ln\Big(\frac{1}{1-e^{-\beta\hbar\omega_i}}\Big)$$

If we assume that the density of states does not change with temperature, which is true in the harmonic case, then the derivative can move into the frequency integral and over the microstates

$$\begin{aligned} C \;&=\; \frac{N}{k_B T^2}\int_0^\infty d\omega\, g(\omega)\,\frac{\partial^2}{\partial\beta^2}\ln\Big(\frac{1}{1-e^{-\beta\hbar\omega}}\Big) \\ &=\; \frac{N}{k_B T^2}\int_0^\infty d\omega\, g(\omega)\,\frac{(\hbar\omega)^2\, e^{\beta\hbar\omega}}{(1-e^{-\beta\hbar\omega})^2} \end{aligned}$$

This is a solution in terms of the β-derivative of the Planck thermal occupation number distribution

$$\begin{aligned} \bar{n}(\omega,\beta) \;&\equiv\; \frac{1}{e^{\beta\hbar\omega}-1} \\ \partial_\beta\,\bar{n} \;&=\; \frac{\hbar\omega\, e^{\beta\hbar\omega}}{(e^{\beta\hbar\omega}-1)^2} \end{aligned}$$

Thus the heat capacity for an continuum ensemble of harmonic oscillators is

$$C \;=\; \frac{N}{k_B T^2}\int_0^\infty d\omega\,\hbar\omega\;\, g(\omega)\;\,\partial_\beta\, n(\omega)$$

If the density of states is known, the heat capacity is fully determined (up to a scalar N) and readily computed numerically for arbitrary $g(\omega)$

$$\begin{aligned} C &= \frac{N\hbar^2}{k_B T^2} \int_0^\infty \omega^2 \, d\omega \; g(\omega) \; \frac{e^{\beta\hbar\omega}}{(1-e^{-\beta\hbar\omega})^2} \\ &= k_B N \int_0^\infty \left(\frac{\hbar\omega}{k_B T}\right)^2 d\omega \; g(\omega) \; \frac{e^{\beta\hbar\omega}}{(1-e^{-\beta\hbar\omega})^2} \\ &= k_B N \int_0^\infty \left(\beta\hbar\omega\right)^2 d\omega \; g(\omega) \; \frac{e^{\beta\hbar\omega}}{(1-e^{-\beta\hbar\omega})^2} \\ &= \frac{N}{k_B T^2} \int_0^\infty d\omega \, \hbar\omega \; g(\omega) \; \partial_\beta \, n(\omega) \\ &= \frac{N}{T} \int_0^\infty d\omega \; g(\omega) \; \beta\hbar\omega \; e^{\beta\hbar\omega} \; n^2(\omega) \end{aligned}$$

We could change the integration variable to the ratio of oscillation-to-thermal energy: $dx = d(\beta\hbar\omega) = \beta\hbar \; d\omega$

$$\begin{aligned} C = \; & \frac{N}{T} \int_0^\infty \frac{dx}{\beta\hbar} \; g(x/\beta\hbar) \; \frac{x \; e^x}{(e^x+1)^2} \\ = \; & \frac{N k_B}{\hbar} \int_0^\infty dx \; g(x) \; \frac{x \; e^x}{(e^x+1)^2} \end{aligned}$$

2.4 Hypothesis

I predicted that the calculated heat capacities for MalL in the reaction state and with a transition state analog bound, as calculated from the velocity autocorrelation function derived density of states, should be approximately 10 kJ/mol K apart as has been calculated from experiment and simulation previously. The prior simulation computed the heat capacity from molecular dynamics trajectories using the variance of the internal energy[41]. On average there should be fewer states in the medium frequencies for the enzyme bound to the transition state analog.

2.5 What do we expect DoS to look like

In figure 2.1 the key features of an enzyme's DoS can be seen. There is a smooth and wide peak from 0 to 500 cm^{-1} and a sharp peak at 1500 cm^{-1}. There is also a peak at 3000 cm^{-1} but this is beyond the range the DoS VAC method can compute based on the chosen sampling rate. In a quantum system these very high frequencies do not contribute significantly to the heat capacity at temperatures compatible with life and thus can be ignored.

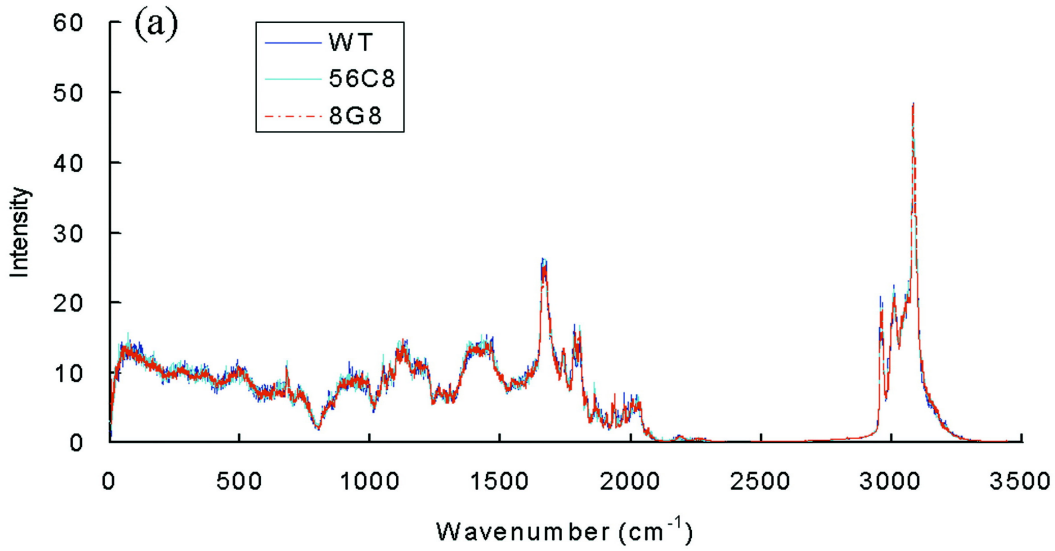

Figure 2.1: A graph of the density of states of an enzyme taken from "Protein dynamics in a family of laboratory evolved thermophilic enzymes"[42], as calculated using neutron diffraction. The three lines are a wild type para-nitrobenzyl esterase and two mutants. Used with permission

2.6 Methods

Six 160-ns MD simulation trajectories were computed. All computational work was done using GROMACS for molecular dynamics, computation of DoS information from trajectory data and calculation of C_P from DoS data. Four of these simulations were done with the substrate bound to the active site of the enzyme and the other two were with a transition state analog bound to the active site of the enzyme. The reaction state of choice is isomaltose and the transition state is a variant of it, see figure 2.2 . The starting structures and

topologies were taken from the paper: "Dynamical origins of heat capacity changes in enzyme-catalysed reactions"[41]. In addition, in accordance with the paper an additional interaction was inserted to stop the reactant from leaving the active site. The parametrization was also taken from this paper[41]. The GROMACS computing software was used to preform the simulation, with the following parameters: temperature of 315 K, pressure of 1 bar, step size of 2 fs and the V-rescale temperature coupling with a time scale of 0.1 ps as a half-life.

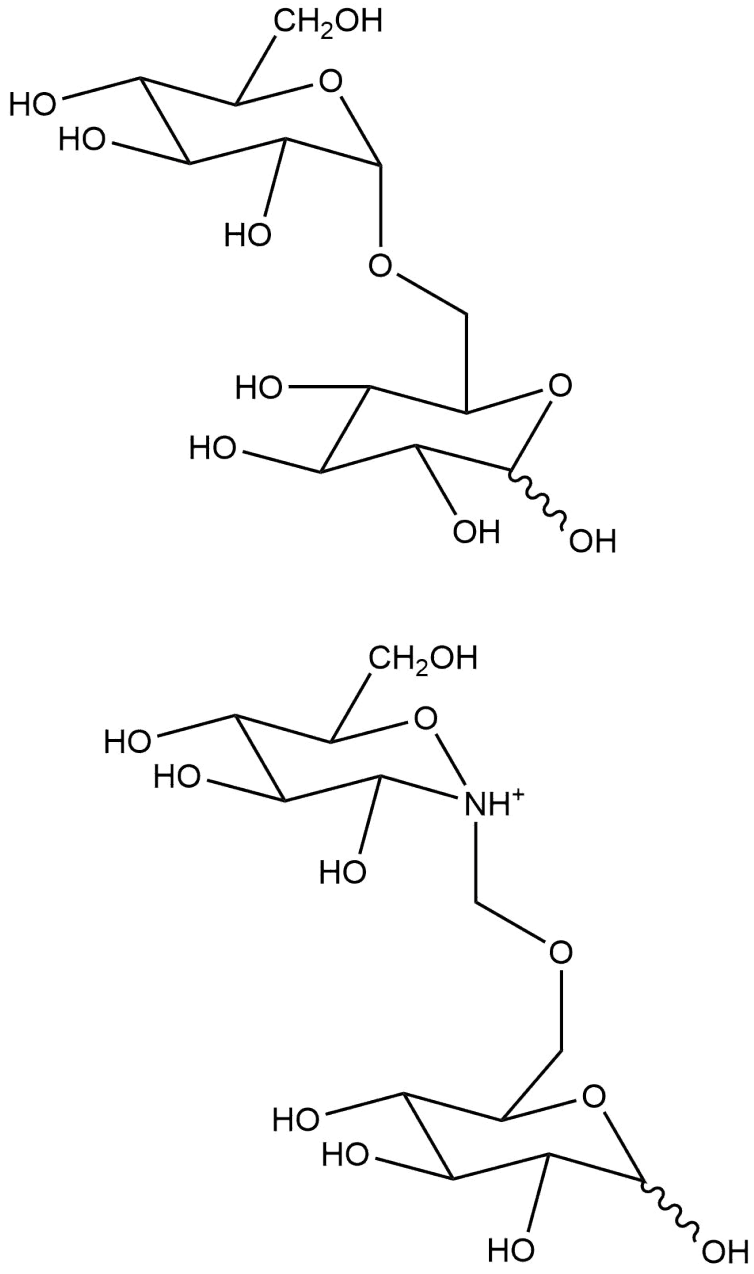

Figure 2.2: The ligands of choice for simulations of MalL, first is isomaltose, second is our transition state analog

A snapshot of the position and velocity data was taken every 500 ps. The first 40 ns were ignored as that time was required for the enzyme to equilibrate from its crystal state to its solvated state, as per [41]. Then a series of 100 ps simulations were run, one for each snapshot, taking that snapshot as the starting conditions. These simulations recorded the velocity of each atom in the simulation every 2 fs. From this trajectory data the VAC was calculated for each atom, the VAC was then transformed into the DoS. All thermodynamic properties can be computed from the DoS, so the heat capacity of each enzyme was computed. Because enzymes have multiple conformations the average of each DoS and heat capacity had to be considered. The DoS was normalized so that there were a total of $3N - 6$ vibrational modes. The analysis generated a constant value for DoS for frequencies greater than 2000 cm^{-1} and so this region was ignored. The heat capacity for each structure was computed as the average of the heat capacities of each individual 100 ps run.

Additionally, normal mode analysis was conducted to test if the DoS calculation method produced reasonable results. Normal mode analysis was conducted starting from the end result one of the 160ns simulations. The structure was then energy minimized using steepest decent, followed by conjugated gradients. Topology information remains the same as with the molecular dynamics simulations. Normal mode analysis was then conducted using standard GROMACS tools in double precision.

Secondary DoS calculations were performed on mutations of MalL. Namely, a single mutant (S536R), a triple mutant (S536D,E554Q,V556R) and the catalytic domain of MalL alone (i.e. removing the entire C-terminal auxiliary domain of MalL). These mutants were constructed by Carlin Hamill in our lab and had been characterized experimentally. This was only done with a single run in the reaction state to look for evidence of detuning of the enzyme. Here parametrization was generated from the amber99sb-ildn force field for the protein and the GAFF force field for the ligand. Additional simulations were done with Ketosteriod isomerase and MalL again at $270K$, holding all

other parameters the same to see how temperature drives the DoS.

2.7 Results

2.7.1 Heat capacity

The average heat capacity, as calculated using the equations generated in section 2.3.1, of the reaction state complex is $52,209 \pm 21$ J/ mol K, while the transition state analog complex is $52,237 \pm 30$ J/ mol K. This value is anomalously low, see section 2.3.1.

2.7.2 DOS graphs

The vibrations extracted from the many simulations can be seen in the following figures. Figure 2.3 shows the DoS comparison between the reaction state and transition state, as computed per a Fourier transform of the velocity autocorrelation function, the key takeaway from this is the lack of difference between the two. Figure 2.4 shows the comparison of DoS between normal mode analysis and the VAC method, this shows that the structure of the two versions are almost the same, but there are significant normalization issues. Figure 2.7 shows how the DoS changes in response to temperature, between 280K and 315K, which seems to be a coherent shift of the frequencies. Figure 2.8 shows how small mutations in MalL's auxiliary domain affect the DoS, namely that there are not major differences.

2.8 Discussion

2.8.1 Visual inspection of DoS

The DoS generated from the VAC, visually matches up with the previously published spectra based on normal mode analysis, see figure 2.1[42], showing the same sorts of patterns that are found in the literature: with a smooth

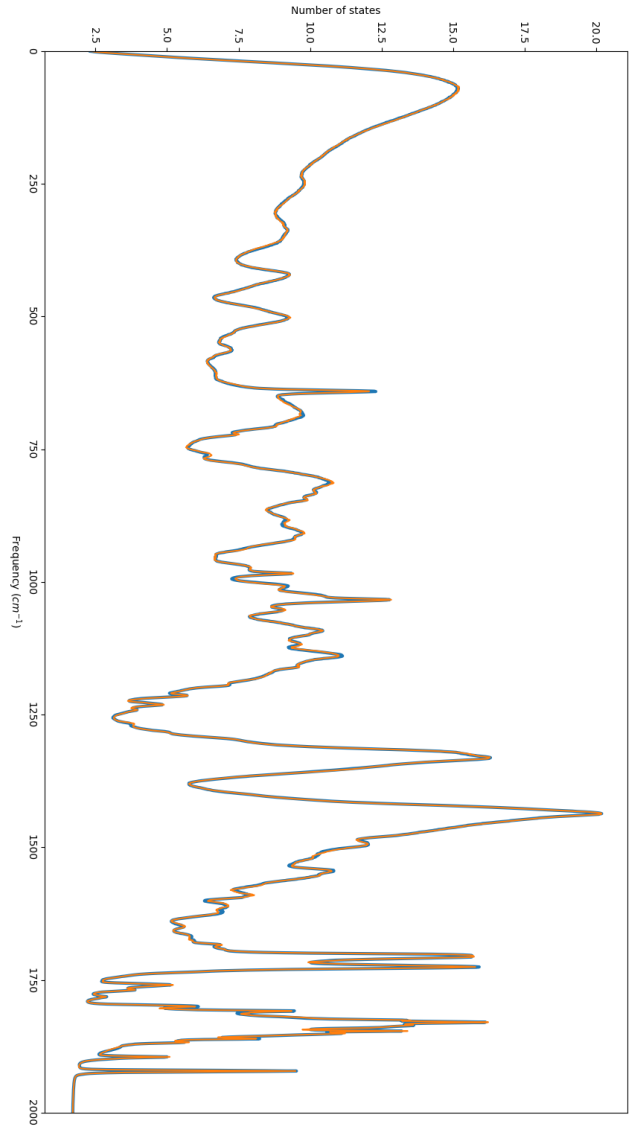

Figure 2.3: A graph of the computed density of states of MalL with both the substrate and a transition state analog bound to the reaction site. The reaction state is colored blue and the transition state analog is in orange. The reaction state has a wider line to better show the overlap of the two graphs. The most notable feature of this comparison is the near-equivalence of the two DoS traces. See figure 2.5 for the difference between the two density of states

region at $0-200$ cm^{-1}, and three amide peaks at $1300, 1400$ and 1700 cm^{-1}[43]. Furthermore, a clear peak separation at 1700 cm^{-1} can be seen, due to MalL being a mixture of alpha helices and beta sheets[43].

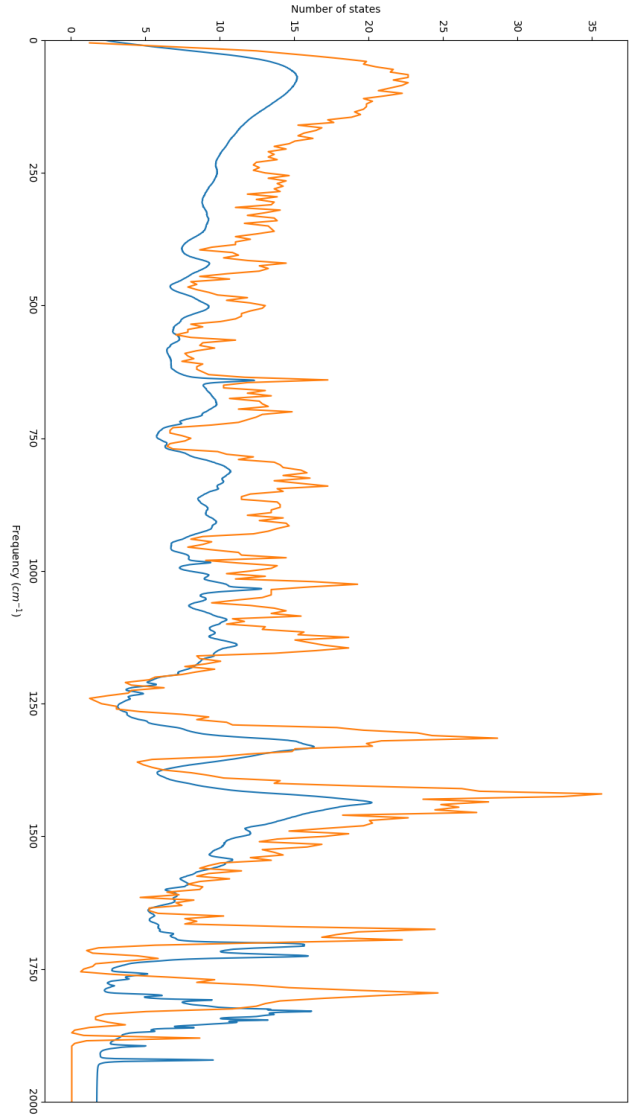

Figure 2.4: A graph of the calculated density of states for the reaction state as calculated from velocity data (blue) and from a normal mode analysis of the same system (orange). The difference in heights is due to issues of normalization, see section 2.8.2 for an extended discussion on this.

2.8.2 Normal mode analysis

The DoS generated from the VAC is similar to the one generated by normal mode analysis (NMA), however normal mode analysis is scaled up significantly and is slightly shifted to the left, see figure 2.4. The left shifting is likely

due to anharmonicity as discussed in section 2.8.9 since NMA is effectively computed at 0K. Normal mode analysis only considers the nature of dynamics directly around a local energy minima, like a system approaching $0K$ would do, while VAC methods look at a wider range of phase space and tries to fit those dynamics to harmonic oscillation. The differences in height is likely due to normalization working incorrectly and scaling down the DoS for the VAC calculations significantly. Given that there is a constant value for the DoS after 2000 cm^{-1}, the overall integral for the DoS was likely far higher than expected, this lead to significant over correction for normalization by multiplication due to a constant factor, for future experiments normalization should not be attempted. This would also mean that the heat capacities were also improperly scaled. Relative heat capacities will also be incorrectly scaled in the same way, but since both the reaction state and transition state analog are scaled to the same degree the difference may still be useful.

2.8.3 Heat capacity

On average the enzyme with the bound transition state analog had a heat capacity 28 J/ mol K higher than the enzyme with the bound reactant state, with a margin of error of $37J$ J/ mol K. This is very far away from the theoretical value of $10K$ J/ mol K and not statistically significant. The absolute values are somewhat lower than what is expected at (150 KJ/mol K) and this is likely due to incorrect normalization as seen by the normal mode analysis.

2.8.4 DoS shift

Plotting the difference of the DoS, along with a confidence interval, shows a statistically significant reduction, at 95% confidence, in the DoS between the reaction state and the transition state between 500-600 cm^{-1} and 800-1000 cm^{-1}, using a 100 point moving average, see figure 2.5. However, this absolute value of the difference is very small, being a difference of less than a twentieth of a state per cm^{-1}. As such, it is more likely that this difference

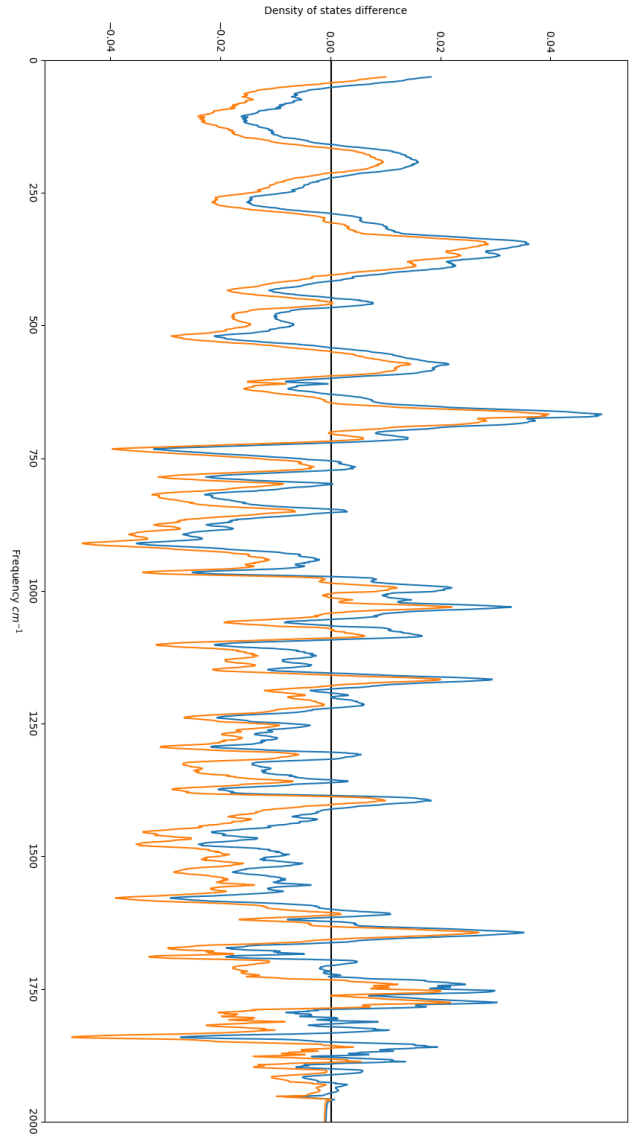

Figure 2.5: A graph of the difference of computed density of states of MalL between the substrate and a transition state analog bound to the active site, while showing the upper and lower bounds for the 95% confidence interval.

is entirely due to the slightly different vibrations of the substrate versus the transition state analog with the vibrations of the overall enzyme not being changed in any way. At the very least such a small shift has no explanatory power for the differences between the two states.

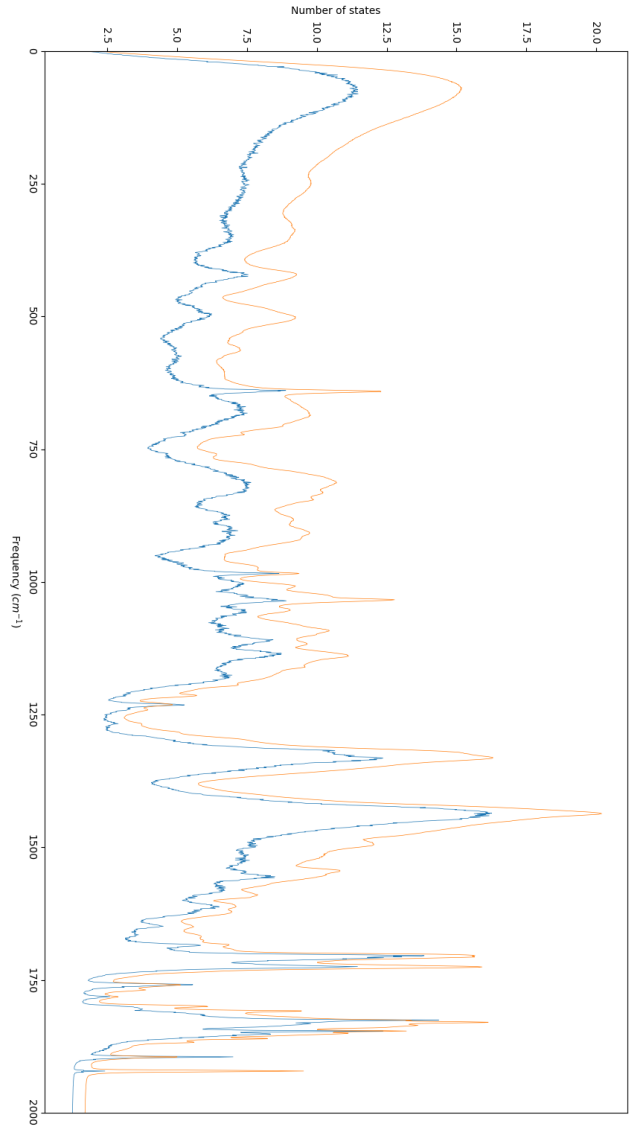

Figure 2.6: A graph comparing computed density of states of MalL (orange) and KSI (blue)

2.8.5 Interesting features of the DoS

There is one major feature found in the Mal-L DoS plot not found in the reference systems (see figure 2.1)[42]: two thin spikes in the DoS plot at 600 and 1000 cm^{-1}. These could be signs of resonance in the enzyme, and they do match up with bond breaking modes in the maltose system. These modes were identified by visually inspecting vibrational modes computed from a frequency

analysis using the Gaussian software package[16]. However, when a ketosteroid isomorase enzyme was tested under the same methodology it also showed spikes at the same frequency. Because these spikes are at the same frequency, they are more likely an artifact due to excess symmetry in the underling Hamiltonian as opposed to anything real. In the chosen molecular dynamics Hamiltonian all bonds of the same type are treated as identical, leading to symmetries that are not present in the real system.

2.8.6 Metastable states

There are two hypotheses for why no difference in heat capacity could be seen from the density of states approach despite being seen in simulations that calculate $\Delta C_P^{\ddagger}$ using variance[41] and being measured by experiment[44]. The first hypothesis is that anharmonic behavior drives $\Delta C_P^{\ddagger}$ and that this is not captured in our approach which assumes harmonic behavior for the system. Alternatively, the simulations did not include all states. Given that MalL has two metastable states, the chosen simulation length is likely to only sample one of these states and the two states have different populations at equilibrium. It could be that the two metastable states[45] both have the same heat capacity for the reaction and the transition state, but the metastable states have different heat capacities, as compared to the other metastable state. Do note that while the published results of "dynamical origins of heat capacity in enzyme catalysis" showed six conformations, there are only two metastable states[45]. Thus $\Delta C_P^{\ddagger}$ is driven by the relative populations of the two metastable states. This could be tested by using a longer simulation, 1μs , to allow for more metastable states to be included in the time ensemble that makes up the density of states.

2.8.7 Anharmonic behavior

The other option is that the vibrations of an enzyme show significant anharmonic shifts when they move from reaction state to transition state. Given

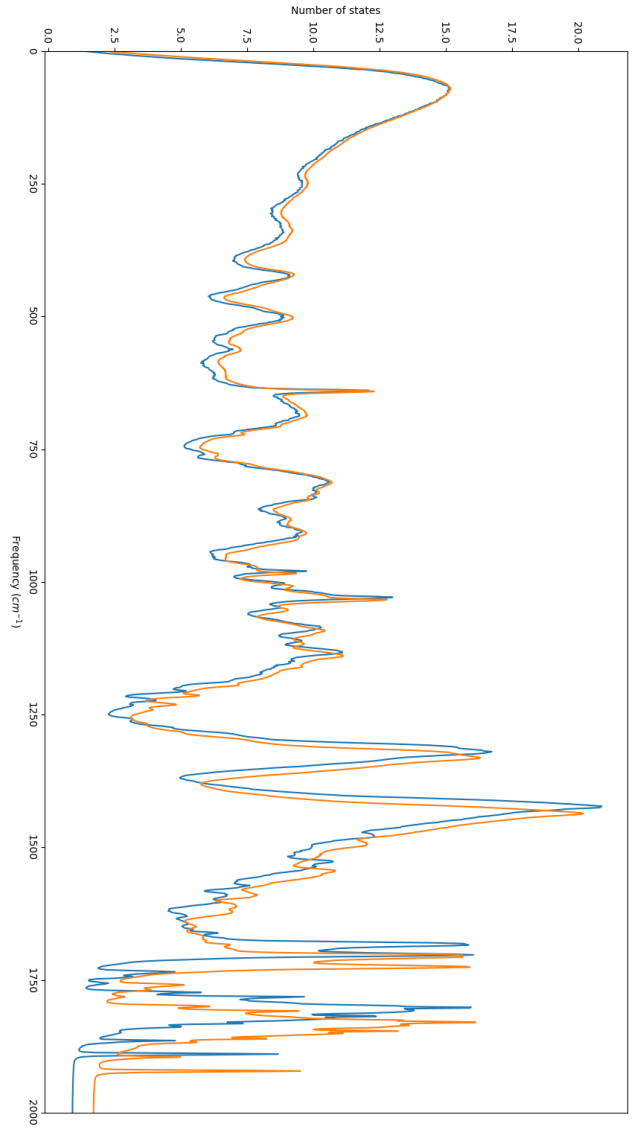

Figure 2.7: A graph comparing computed density of states of MalL at 315K(orange) and 280K(blue)

that the correct heat capacity is shown in the reaction state (after correcting for erroneous normalization), we can assume that either there is no significant anharmonic components or they all cancel out. However, the transition state may have large anharmonic components that cannot be measured using the methods in this chapter. There is not currently a known method to test this. Recent results from our collaborators at the University of Bristol

show dramatic increases in correlated motions at the transition state which may have anharmonic components (personal communication), which acts as further evidence for the anharmonic hypothesis.

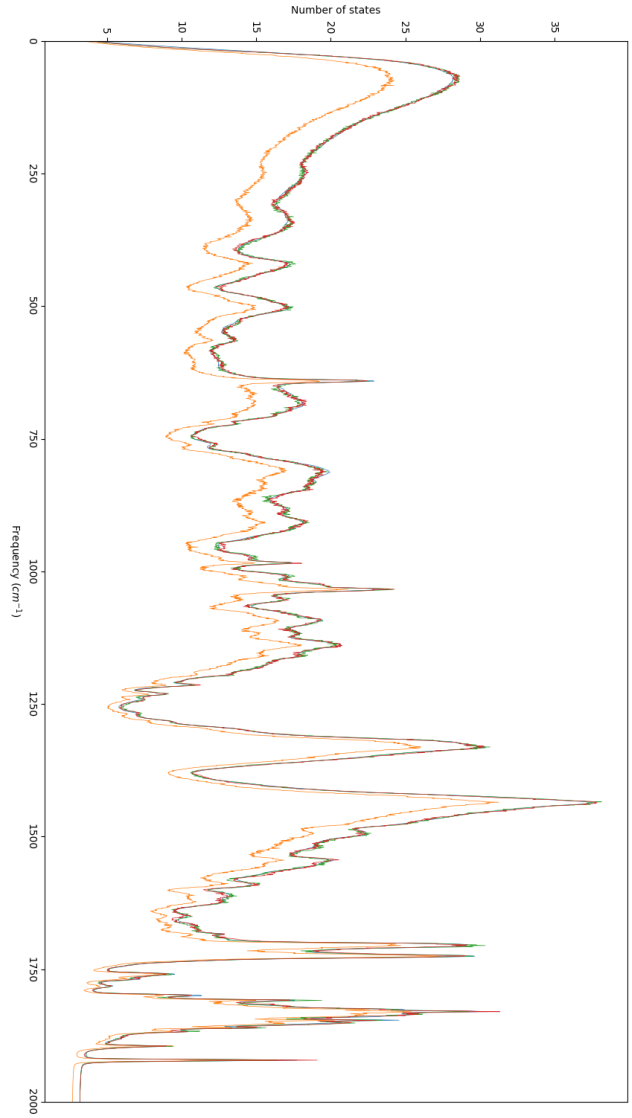

Figure 2.8: A graph comparing computed density of states of wild type MalL (blue), S536R MalL (red), S536D,E554Q,V556R (green) and just the catalytic domain

2.8.8 Detuning

From Figure 2.8, we can see that despite the mutations having a change in $\Delta C_P^{\ddagger}$ experimentally there is no real change in the DoS for a point mutation and a triple mutant, this is further evidence of the insufficiency of DoS based modeling. Apart from the total removal of the auxiliary domain, there seems to be no shift in frequencies beyond the margin of error. In the case of the removal of the auxiliary domain, all amplitudes are reduced due to the smaller size of the enzyme, but the patterns don't seem to change.

2.8.9 Changes in temperature

Figure 2.7 shows that there is a shift to higher frequencies at higher temperatures. If the normal mode analysis is seen as computing the frequencies of the system at 0K then it will fit this pattern. All features of the graph remain consistent but shifted. The degree of shifting seems to be related to the frequency in a non-linear but increasing fashion. There are some differences in the intensity of the system, but this is likely due to insufficient simulation time for the cold simulation, as seen by the fact that the cold simulation is much rougher. This is a strong indicator of anharmonic behavior at high frequencies. Interestingly this also indicates that the second anharmonic term is very strongly correlated to the frequency in all vibrational modes and this is strong evidence for a Lamb shift.

A Lamb shift is a shift in the frequency of a vibrational mode due to interactions with a solvent bath, these shifts are temperature dependent[46]. Again this is strong evidence of a major anharmonic component to enzyme dynamics.

2.9 Conclusion

To conclude, the DoS produced by this method does seem to be accurate for the case of MalL in the reaction state, however there is no evidence for the heat

capacity shift at the transition state as shown previously[41]. I suspect that this is caused by either very anharmonic behavior or an insufficient simulation duration. There is strong evidence for some form of anharmonic behavior shown by the temperature shifts.

Chapter 3

Correlation distances as possible evidence for a phase change

3.1 Introduction

A phase transition is defined as a discontinuity in free energy over some thermodynamic variable[47]. A classic example is of water boiling. Here energy is put into the system causing the temperature to rise until it hits 100°C, at this point adding more energy will not increase the temperature until the water has completely boiled. From the perspective of energy as a function of temperature this produces a discontinuity. We hypothesize that there is a discontinuity in the second derivative of free energy with respect to the reaction coordinate during enzyme catalysis. If the enzyme-substrate complex is relatively mobile and the enzyme transition-state complex is very rigid, then this may manifest as a phase change along the reaction coordinate. A key sign of a phase transition is long ranged energy correlations[47]. As such, it would be useful to test this by attempting to detect these correlations.

3.2 Hypothesis

My hypothesis is that there will be some method for discerning how correlation in the velocities of particles falls off with distance, and that enzymes with a

transition state analog bound in their active site will have a slower fall-off in correlation than enzymes in the reaction state conformation.

3.3 Methods

3.3.1 MD computations

For the trajectory analysis used in sections 3.3.2 and 3.3.3, 100-ps MD simulations were computed. These computations were done using the same setup as with the density of states calculations. Starting structures were taken from the end of the 160-ns simulations that were used for the DoS calculations. If they were found to be usable, more simulations would be done with starting structures based on other time slices in the simulations. Normal mode analysis was performed using the standard GROMACS method, using a structure from the end of a 160 ns simulation from the DoS chapter which was then optimized using a mixture of gradient descent and conjugated gradients. A Hessian was generated and the eigen decomposition performed, giving eigenfrequencies and eigenvectors. The values that define the topology of the system were randomly changed by up to 5%, to prevent there being too much symmetry in the system. This would lead to high frequency vibrations being larger than they should, as highly symmetric enzymatic systems have wider vibrations[23]. All code was written in C++.

3.3.2 Dot product

3.3.2.1 Reason for use

A dot product method was first chosen because it is a simple way to measure correlation. By treating velocity over time as a single vector, the divergence can be measured as the angle between the two vectors. This method was inspired by the autocorrelation methods used in calculating density of states. There could be an issue in that this method is invariant to time rearrangement.

In addition, two coupled oscillators that are 90 degrees out of phase would not contribute to the measure of correlation and thus could not be detected.

3.3.2.2 Implementation

Define Z_{ij} to be a measure of correlation between atom i and atom j. Defining T to be the total time of the simulation, $\mathbf{V_i}(t)$ to be the velocity of atom i at time t.

$$Z_{ij} = \frac{1}{T} \frac{\int_0^T \mathbf{V_i}(t) \cdot \mathbf{V_j}(t) dt}{\mathbf{V_i}(\mathbf{0}) \cdot \mathbf{V_j}(\mathbf{0})} \tag{3.1}$$

Running this on a 100-ps simulation of MalL produced a correlation matrix. The expected result is that the correlation value of two atoms would decline exponentially with the distance between them.

$$Z_{ij} \propto e^{-kd_{ij}} \tag{3.2}$$

However this relationship was not observed in the data ($r^2 = 10^{-6}$). This indicates that the inner product is a poor measure of the correlation that we seek.

3.3.3 Mutual information

3.3.3.1 Reason for use

Given that a trivial vector dot product was found not to be useful, a more general approach was used. What is measured with this value is: given the knowledge of one particle's velocity, how much can be known about another particle's velocity. While this method is time invariant, it does solve the phase problem of the previous section.

3.3.3.2 Implementation

Mutual information measures how much information is gained about a random variable from observation of another random variable. Here we define $P_i(X)$

to be the probability that atom i has velocity magnitude of X. Due to RAM limitations only the magnitude of the velocity is used and not its direction, in addition only alpha carbons were tested for the same reason. We define $P_{ij}(X, Y)$ to be the probability that atom i has a velocity of X and atom j has a velocity of Y. Because continuous probability distributions cannot be generated some degree of binning is required. As such, $P(X)$ is split into 100 bins ranging from the lowest to the highest recorded values in the trajectory data. Defining Z_{ij} to be our measure of correlation between atom i and atom j

$$Z_{ij} = \int_0^\infty \int_0^\infty P_{ij}(x, y) \log_2 \left(\frac{P_{ij}(x, y)}{P_i(x) P_j(y)} \right) dx \; dy \tag{3.3}$$

However, this also failed for the same reason as the dot product metric. With a 100 ps simulation of MalL, Z_{ij} was completely uncoupled from d_{ij} with $r^2 = 10^{-5}$. I suspect that this is the case because there are too many interacting vibrations, causing any signal to be lost in the noise.

3.3.4 Normal modes

3.3.4.1 Reason for use

Given that the previous two methods failed, it was suspected this was due to too many interacting vibrations destroying any ability to measure correlations. As such, normal mode analysis was used to elucidate the vibrational modes. In a non-homogeneous medium each vibrational mode has a different extent of vibrations (which will be referred to as the size of the vibration as this is not the amplitude). From this, the extent of how correlated distant motions are can be seen. Instead of trying to fit a distance decay parameter, a plot of sizes of fluctuations is created. This will only look at harmonic vibrations, but it was believed that this may be sufficient.

3.3.4.2 Implementation

Assuming that all vibrations in the enzyme are harmonic, the equations of motion can be reduced down to a simple differential equation. Let X be a vector of atom locations, X is also translated such that $X = 0$ defines the energy minimum, A is defined as the mass weighted energy hessian matrix or dynamical matrix for the harmonic system. $A = M^{-1/2}HM^{-1/2}$. If there are N atoms, there are $3N$ dimensions inside of X.

$$\frac{\partial^2 \mathbf{X}}{\partial t^2} = -A\mathbf{X} \tag{3.4}$$

Solving this equation gives:

$$\mathbf{X}(t) = \sum_{i=1}^{3N} a_i \psi_{\mathbf{i}} \cos(\omega_i t + \varphi_i) \tag{3.5}$$

where a_i and φ_i are the amplitude and the phase of each normal mode, these are determined by the starting conditions. $\psi_{\mathbf{i}}$ and ω_i are the $i'th$ eigenvector and the square root of the $i'th$ eigenvalue of A, these are not dependent on the initial conditions of the system.

So my next aim was to determine the extent of the normal modes. This can be done by taking a single mode and seeing how concentrated the vibrations are. If all the vibrations are clustered in one location, then atoms far away from each other are uncorrelated, as these modes are independent. A simple approach of just looking for when the displacement becomes zero would not work as all modes affect all atoms to some degree. As such a notion of the center of mass has been developed for these modes based on the mass weighted displacement values. Let $\bar{\mathbf{X}}_{\mathbf{i}}$ be this center of mass for mode i, let m_j be the mass of atom j, let $\mathbf{X_j}$ be the position of atom i at the energy minima, let N be the number of degrees of freedom.

$$\bar{\mathbf{X}}_{\mathbf{i}} = \frac{\sum_{j=1}^{N} m_j \psi_{ij} \mathbf{X_j}}{\sum_{j=j}^{N} m_j \psi_{ij}} \tag{3.6}$$

Note that the precise norm of ψ_i is irrelevant. A concept of the mean correlation distance, σ_i, can be developed using a weighted standard deviation.

$$\sigma_i^2 = \frac{\sum_{j=1}^{N} m_j \psi_{ij} (\mathbf{X_j} - \bar{\mathbf{X}}_\mathbf{j}) \cdot (\mathbf{X_j} - \bar{\mathbf{X}}_\mathbf{j})}{\sum_{j=1}^{N} m_j \psi_{ij}} \tag{3.7}$$

Originally the process produced very large values of σ_i. This was due to the model of the system being overly symmetric in the force field, requiring some fuzzing of the force field data. The data originally showed a large difference between the reaction state and the transition state, however this was due to the transition state being over the edge of the simulation box. When that problem was fixed there was no difference between the reaction state and the transition state. This did produce useful results which will be discussed below.

3.4 Results and discussion

From Figure 3.1, it is clear that there is no major difference in correlation distance between the reaction state and the transition state analog. This is a strong indicator that in the harmonic case there is no correlation in energy fluctuations required for a phase change. However, in this experiment only harmonic vibrations were looked at. Given that MalL has two metastable states[45], each with different population statistics for the reaction and transition states. As such it may be that either anharmonic behavior drives phase transitions or that only one metastable state was considered. If the second scenario is the case then more structures need to be considered, this was not done due to time limitations.

3.5 Conclusion

To conclude, there was no difference in chosen measures of correlation between the chosen enzyme in the reaction state and the chosen enzyme with a bound transition state analog. Due to the limitations of this experiment it does not

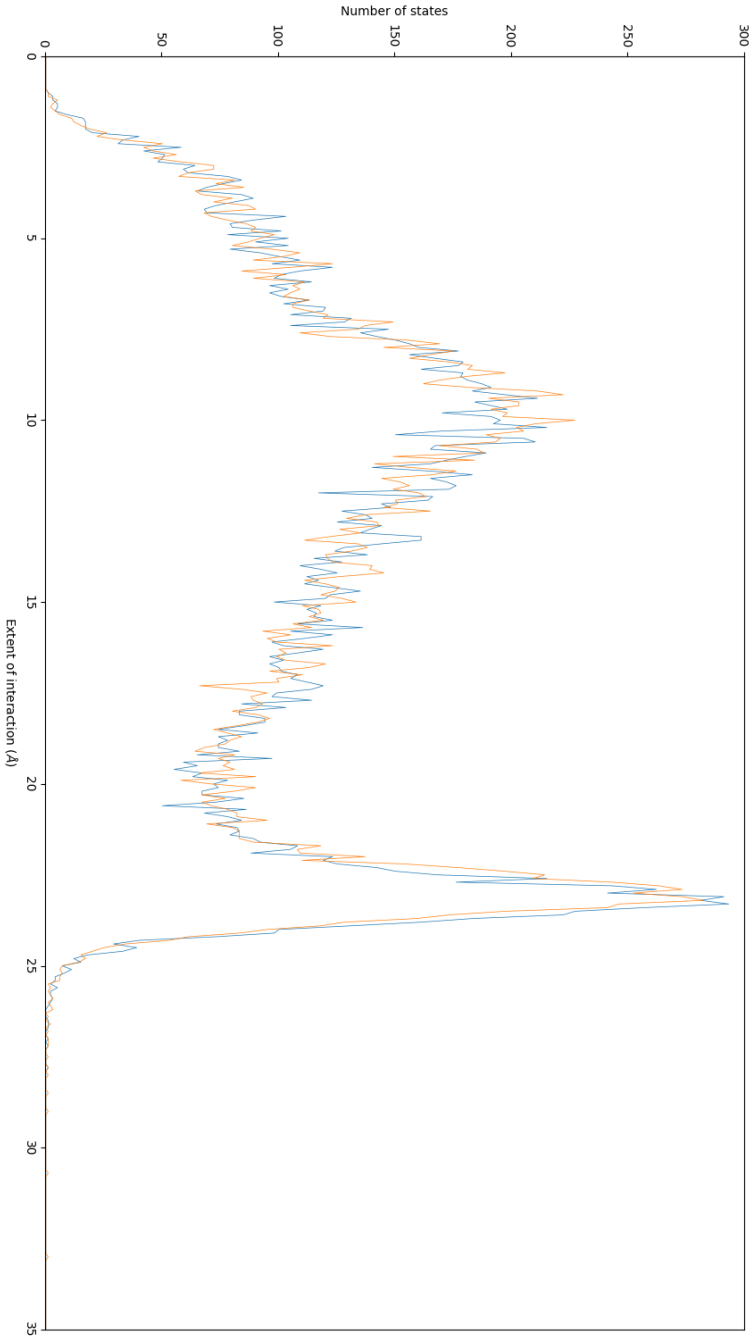

Figure 3.1: A graph of the number of vibrational modes with a given correlation distance in the enzyme MalL comparing a reaction state (blue) to a transition state (orange).

fully explore the phase change hypothesis, other than to suggest that a more in-depth approach is required. It is likely that one of the main contributions to the change in heat capacity along the reaction coordinate is the large fluctuations in the reactant state that traverse multiple conformations and that these are anharmonic and not being sampled in the MD trajectories here. The solution to this is to use much longer MD trajectories or to use more sophisticated normal mode analysis. Alternatively, if anharmonic contributions dominate the changes at the transition state, then entirely new approaches are required.

Chapter 4

Anomalous heat capacity

4.1 Introduction

Previous work on computing the heat capacities of enzymes, from MD trajectories, has focused on using the variance of system energy to compute heat capacity[41]. While this has produced values of $\Delta C_P^{\ddagger}$ that match up with experimental values calculated from kinetics data, the absolute values of C_P produced are not accurate[41].This is expected as calculating the absolute values of thermodynamic properties in liquids is notoriously difficult. As such, this chapter is focused on investigating the possible origins of the anomalous absolute value of the heat capacity from the variance in MD trajectories. The other major goal was to find a method that does compute reasonable absolute values of heat capacity from calculations of energy variance, as this could be useful for general computational chemistry.

4.2 Derived paper

The core of this work was revisiting the paper "Dynamical origins of heat capacity changes in enzyme-catalysed reactions"[41]. This paper (Dynamical origins for short) took MalL and computed the heat capacity in both the reaction state and a transition state analog. From the MD data, the enthalpy at each time step was computed, a sliding window of variance was then taken

and the heat capacity was calculated for each window. This paper reproduces the experimental differences in heat capacity between the reactant state and the transition state for two model enzymes. However, the absolute values for the heat capacity for each species is approximately 10 times larger than what we expect from experimental values calculated for model proteins. A normal protein has a heat capacity of 1.33 J/gK, while the measured heat capacity of MalL was on the order of 10 J/gK. This is an intriguing discrepancy that was worthy of investigation despite the inherent difficulties in calculating absolute values for thermodynamic systems of such complexity.

4.3 Underlying principles

Heat capacity is defined by how much energy is required to add one extra kelvin of temperature to a system. As such, it is equal to $\frac{\partial\langle\phi\rangle}{\partial T}$. By computing a full derivation of heat capacity from the partition function it can be shown that in all cases under the assumptions of classical statistical thermodynamics, heat capacity is directly proportional to energy variance.

4.3.1 Derivation of heat capacity from variance

Define $\beta \equiv \frac{1}{k_B T}$ to be thermodynamic beta, ϕ to be the ensemble's thermodynamic energy function and a sum over Ω to be a sum over all states in the ensemble.

Start with the normal definition of the partition function Z over a list of states

$$Z = \sum_{\Omega} e^{\phi\beta}$$

$$P(\Omega_\phi) = \frac{e^{\phi\beta}}{Z}$$

Taking the derivative of the log of the partition function with respect to ther-

modynamic beta

$$\begin{aligned}\frac{\partial \ln(Z)}{\partial \beta} &= \frac{1}{Z}\sum_{\Omega} \phi e^{\phi\beta} \\ &= \sum_{\Omega} \phi \frac{e^{\phi\beta}}{Z} \\ &= \sum_{\Omega} \phi P(\Omega_\phi) \\ &= \langle \phi \rangle\end{aligned}$$

The expectation value of energy is computed.

Take the definition of heat capacity to be

$$\begin{aligned}C \equiv \frac{\partial \langle \phi \rangle}{\partial T} &= \frac{\partial}{\partial T}\frac{\partial \ln(Z)}{\partial \beta} \\ &= \frac{1}{k_B T^2}\frac{\partial^2 \ln(Z)}{\partial \beta^2} \\ \frac{\partial T}{\partial \beta} &= \frac{1}{k_B T^2}\end{aligned}$$

And then compute the derivative

$$\begin{aligned}
\frac{\partial^2 \ln(Z)}{\partial \beta^2} &= \frac{\partial}{\partial \beta} \frac{\sum_\Omega \phi e^{\phi\beta}}{Z} \\
&= \frac{(\partial_\beta \sum_\Omega \phi e^{\phi\beta}) Z - (\sum_\Omega \phi e^{\phi\beta}) \partial_\beta Z}{Z^2} \\
&= \frac{Z(\sum_\Omega \phi^2 e^{\phi\beta}) - (\sum_\Omega \phi e^{\phi\beta})^2}{Z^2} \\
&= \frac{\sum_\Omega \phi^2 e^{\phi\beta}}{Z} - \left(\frac{\sum_\Omega \phi e^{\phi\beta}}{Z} \right)^2 \\
&= \sum_\Omega \phi^2 \frac{e^{\phi\beta}}{Z} - \left(\sum_\Omega \phi \frac{e^{\phi\beta}}{Z} \right)^2 \\
&= \sum_\Omega \phi^2 P(\Omega_\phi) - \left(\sum_\Omega \phi P(\Omega_\phi) \right)^2 \\
&= \langle \phi^2 \rangle - \langle \phi \rangle^2 \\
&= var(\phi) \\
C &= \frac{\partial \langle \phi \rangle}{\partial T} = \frac{var(\phi)}{k_B T^2}
\end{aligned}$$

This gives the relation between heat capacity and energy variance.

4.3.2 Hidden assumptions

There are five key hidden assumptions in this derivation. The first and the most important is that whatever energy function used is accurate and conserved. This is trivial but needs to be stated[48]. Second is that the partition function equation is based on interaction energies between the system and the surroundings being constant[48]. Third is that there is no positional/vibrational coupling between the system and the surroundings that is not included in the energy term (ϕ). That is, knowing the exact state of the system will only tell you the energy of the surroundings. In a classical system such as under an MD simulation[39], this assumes that vibrations are isolated to the system and the surroundings but not spread out over both, in a quantum system this means that the wavefunction can be factorized into a system wavefunction and a surrounding's wavefunction. Fourth, to properly use the

variation over time as an ensemble, it has to sample a representative set of points. Finally, to simulate an MD enzyme we assume that enzymes are ideal gasses when considering how to compute the global partition function from a single molecule partition function[48]. A system can be considered an ideal gas assuming that each instance of a system has independent and identically distributed microstates. This assumption is true assuming that there are no long range interactions between instances and that concentration is sufficiently low. The fact that the enzymes are solvated in water and not in a vacuum does not change this. Given that the typical concentration of an enzyme in a typical assay is 1ppm by mass[8], this is acceptable.

Note that from this derivation there is no requirement for the system to have normally distributed energy and takes into account multiple clusters, as the partition function assumes an arbitrary density of states function, so the sliding window used in dynamical origins of heat capacity[41] does not seem appropriate. Instead, the entire simulation run time should be used to act as a proper ensemble as the moving window gives the heat capacities of the sub-states and not of the overall system.

4.4 Quantum issues

One other major issue with using MD to compute energy is that MD is a classical system without quantization, this means that the energy function (ϕ) is an estimate. A good example of the differences between classical and quantum treatments for heat capacity is section 4.4.1 and section 4.4.2.

4.4.1 Quantum derivation

Here we define n_P as the number of particles in a single instance of the system and n_F as the number of degrees of freedom, $n_F = 3n_P$. N is the number of molecular instances in the ensemble, if looking at specific heat capacity it is

the number of instances in a gram or a mole.

$$C \equiv \frac{\partial \langle \phi \rangle}{\partial T} = \frac{1}{k_B T^2} \frac{\partial^2}{\partial \beta^2} \ln Q \tag{4.1}$$

with ensemble partition function given as

$$\ln Q = N \ln q - \ln N!$$

if the chemical potential is zero, $\mu = \partial_\beta N = 0$ (ie, $N \neq N(\beta)$), then

$$\frac{\partial}{\partial \beta} \ln Q = \frac{\partial}{\partial \beta} \Big(N \ln q - \ln N! \Big) = N \frac{\partial}{\partial \beta} \ln q$$

thus, for systems where N is independent of temperature, the ensemble partition function is a homogeneous scaling of the individual microstates, looking at q for a harmonic oscillator

$$\begin{aligned} q &= \prod_{i=0}^{n_F} q_i \\ q_i &= \sum_{k=0}^{\infty} e^{-k\beta\hbar\omega_i} = \frac{1}{e^{-\beta\hbar\omega_i} - 1} \end{aligned}$$

where we have used the geometric progression. Taking the natural log of q

$$\begin{aligned} \ln q &= \ln \prod_{i=1}^{n_F} q_i \\ &= \sum_{i=0}^{n_F} \ln q_i \\ &= \sum_{i=0}^{n_F} \ln \left(\frac{1}{1 - e^{-\beta\hbar\omega_i}} \right) \end{aligned}$$

If we allow $n \to \infty$, we can take the continuum limit for microstates by

introducing the density of states, $g(\omega)$, for the system

$$\ln q \;=\; \int_0^\infty d\omega\, g(\omega)\, \ln\left(\frac{1}{1-e^{-\beta\hbar\omega_1}}\right)$$

We can now use this to evaluate C_p

$$\begin{aligned} C &= \frac{1}{k_B T^2}\frac{\partial^2}{\partial\beta^2}\ln Q \\ &= \frac{N}{k_B T^2}\frac{\partial^2}{\partial\beta^2}\ln q \\ &= \frac{N}{k_B T^2}\frac{\partial^2}{\partial\beta^2}\int_0^\infty d\omega\, g(\omega)\, \ln\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right) \end{aligned}$$

If we take the density of states to be constant with temperature, then the β derivative can be moved into the integral

$$C \;=\; \frac{N}{k_B T^2}\int_0^\infty d\omega\, g(\omega)\, \frac{\partial^2}{\partial\beta^2}\ln\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right)$$

Looking at the derivative in isolation

$$\begin{aligned} \frac{\partial^2}{\partial\beta^2}\ln\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right) &= \frac{\partial}{\partial\beta}\left[\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right)^{-1}\frac{\partial}{\partial\beta}\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right)\right] \\ &= \frac{\partial}{\partial\beta}\left[(1-e^{-\beta\hbar\omega})\left(\frac{-1}{(1-e^{-\beta\hbar\omega})^2}\right)\frac{\partial}{\partial\beta}(1-e^{-\beta\hbar\omega})\right] \\ &= \frac{\partial}{\partial\beta}\left[\left(\frac{-1}{1-e^{-\beta\hbar\omega}}\right)(-\frac{\partial}{\partial\beta}e^{-\beta\hbar\omega})\right] \\ &= \frac{\partial}{\partial\beta}\left[\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right)\left(\hbar\omega\, e^{-\beta\hbar\omega}\right)\right] \\ &= \frac{\partial}{\partial\beta}\left[\frac{\hbar\omega\, e^{-\beta\hbar\omega}}{1-e^{-\beta\hbar\omega}}\right] \\ &= \hbar\omega\,\frac{\partial}{\partial\beta}\left[\frac{1}{e^{\beta\hbar\omega}-1}\right] \\ &\equiv \hbar\omega\,\frac{\partial\,\bar{n}}{\partial\beta} \end{aligned}$$

Note that the term in the square braces is the Planck thermal distribution, $\bar{n}(\beta,\omega)$, for occupational number. Therefore, we can write C in terms of this

well-known form

$$C \;=\; \frac{N\hbar}{k_B T^2}\int_0^\infty d\omega\,\omega\,g(\omega)\,\frac{\partial\bar{n}}{\partial\beta} \tag{4.2}$$

In this form, heat capacity can be interpretead as the frequency mean-average (first moment integral in ω) for the thermal β derivative of the Planck distribution scaled by $N\,\hbar/k_B T^2$, at the frequencies given by the density of states for the system.

The derivative of $\bar{n}$ with β is

$$\begin{aligned}
\frac{\partial\,\bar{n}}{\partial\beta} &= \frac{-1}{(e^{\beta\hbar\omega}-1)^2}\,\frac{\partial}{\partial\beta}(e^{\beta\hbar\omega}-1)\\
&= \frac{-\hbar\omega\,^{\beta\hbar\omega}}{(e^{\beta\hbar\omega}-1)^2}\\
&= \hbar\omega\,\frac{e^{\beta\hbar\omega}}{(e^{\beta\hbar\omega}-1)(-e^{\beta\hbar\omega}+1)}\\
&= \frac{\hbar\omega}{(e^{+\beta\hbar\omega}-1)(e^{-\beta\hbar\omega}-1)}
\end{aligned}$$

which can be expressed in terms of the hyperbolic cosine

$$\frac{\partial\,\bar{n}}{\partial\beta} \;=\; \frac{\hbar\omega}{2}\,\frac{1}{1-\cosh(\beta\hbar\omega)} \tag{4.3}$$

Using this for the form due to q, we have

$$\frac{\partial^2}{\partial\beta^2}\ln\left(\frac{1}{1-e^{-\beta\hbar\omega}}\right) \;=\; \hbar\omega\,\frac{\partial\,\bar{n}}{\partial\beta} \tag{4.4}$$

$$= \frac{(\hbar\omega)^2}{(e^{+\beta\hbar\omega}-1)(e^{-\beta\hbar\omega}-1)} \tag{4.5}$$

$$= \frac{(\hbar\omega)^2}{2}\,\frac{1}{1-\cosh(\beta\hbar\omega)} \tag{4.6}$$

This is a symmetric function about $\beta\hbar\omega = 0$ (since temperature and frequency are both taken to be real and positive numbers) and, in terms of β, rises from negative infinity at zero, $\bar{n}'(0) = -\infty$, smoothly and monotonically almost through the point $(+1, -1)$ then asymptotes to zero from underneath,

$\bar{n}'(\infty) = 0$.

Connecting equation 4.6 with equation 4.3

$$C = \; Nk_B \int_0^\infty d\omega \, g(\omega) \, \frac{(\beta\hbar\omega)^2}{2\cosh(\beta\hbar\omega) \; - \; 2} \tag{4.7}$$

4.4.2 Classical derivation

Taking the same definitions are the previous section.

$$C \; \equiv \; \frac{\partial \langle\phi\rangle}{\partial T} \; = \; \frac{1}{k_B T^2} \, \frac{\partial^2}{\partial \beta^2} \ln Q \tag{4.8}$$

$$\ln Q \; = \; N \, \ln q \; - \; \ln N!$$

if the chemical potential is zero, $\mu = \partial_\beta N = 0$ (ie, $N \neq N(\beta)$), then

$$\frac{\partial}{\partial \beta} \ln Q \; = \; \frac{\partial}{\partial \beta} \Big(N \, \ln q \; - \; \ln N! \Big) \; = \; N \, \frac{\partial}{\partial \beta} \ln q$$

thus, for systems where N is independent of temperature, the ensemble partition function is a homogeneous scaling of the individual microstates

$$q \; = \; \prod_{i=0}^{n_F} q_i$$

The Hamiltonian (energy) for a simple harmonic oscillator with mass m and frequency ω is given in (normalized) position-momentum (x, p) space as

$$\mathcal{H} \; = \; \frac{p^2}{2m} \; + \; \frac{1}{2} m\omega^2 x^2 \tag{4.9}$$

The classical partition function can be generally represented as the phase space ξ integral over the region $\mathcal{R}$ of the system, with dynamics due to the Hamiltonian reduced by the thermodynamic beta $\beta = 1/k_B T$

$$q_i \;=\; \int_{\mathcal{R}} d\xi \;\exp\Big[-\beta\, \mathcal{H}_i(\xi)\Big]$$

for the i^{th} oscillator defined above

$$q_i \;=\; \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{dx\; dp}{h} \;\exp\Big[-\beta\Big(\frac{p^2}{2m} \;+\; \frac{1}{2}m\omega^2 x^2\Big)\Big]$$

Under these approximations, the position and momenta can be disentangled/separated (note this is not true in the quantum correction since position and momentum do not commute) and each integral can be performed independently

$$q_i \;=\; \frac{1}{h}\int_{-\infty}^{+\infty} dp \;\exp\Big[-\beta\frac{p^2}{2m}\Big] \int_{-\infty}^{+\infty} dx \;\exp\Big[-\beta\frac{1}{2}m\omega_i^2 x^2\Big]$$

These integrals are Gaussian integrals which can be immediately solved

$$\begin{aligned} q_i \;&=\; \frac{1}{h}\,\sqrt{\frac{\pi}{(\beta/2m)}}\,\sqrt{\frac{\pi}{(\beta m\omega_i^2/2)}} \\ &=\; \frac{1}{h}\,\frac{2\pi}{\beta\omega_i} \end{aligned}$$

Using the reduced Planck constant we obtain the partition function for the classical harmonic oscillator

$$q_i \;=\; \frac{1}{\beta\hbar\omega_i} \tag{4.10}$$

which is a dimensionless ratio of thermal to oscillator energy, and is commonly encountered in statistical and quantum thermodynamics. To use this result to

determine this system's heat capacity, take the natural log

$$\begin{aligned}
\ln q &= \ln \prod_{i=0}^{n_F} q_i \\
&= \sum_{i=0}^{n_F} \ln q_i \\
&= \sum_{i=0}^{n_F} \ln \Big(\frac{1}{\beta \hbar \omega_i} \Big) \\
&= -\sum_{i=0}^{n_F} \ln \big(\beta \hbar \omega_i \big)
\end{aligned}$$

Separating thermal and oscillatory components since here it is assumes that the temperature homogeneous across all microstates (thermal equilibrium),

$$\begin{aligned}
\ln q &= -\sum_{i=0}^{n_F} \Big(\ln \beta \, + \, \ln(\hbar \omega_i) \Big) \\
&= -\ln \beta \sum_{i=0}^{n_F} 1 \; - \; \sum_{i=0}^{n_F} \ln \big(\hbar \omega_i \big) \\
&= -n_F \, \ln \beta \; - \; \sum_{i=0}^{n_F} \ln \big(\hbar \omega_i \big)
\end{aligned}$$

Note the microstate summation over the temperature dependent aspects of the system amounts to simply scaling a constant by the number of states. We can now evaluate the heat capacity of this system

$$\begin{aligned}
C_p &= \frac{N}{k_B T^2} \frac{\partial^2}{\partial \beta^2} \ln q \\
&= \frac{N}{k_B T^2} \frac{\partial^2}{\partial \beta^2} \Big(-n_F \, \ln \beta \; - \; \sum_{i=0}^{n_F} \ln \big(\hbar \omega_i \big) \Big)
\end{aligned}$$

Note that in this system the oscillator do not contribute to heat capacity (because $\omega_i \neq \omega_i(T)$), and the expression vastly simplifies into

$$\begin{aligned}
C_p &= n_F \frac{N}{k_B T^2} \frac{\partial^2}{\partial \beta^2} \Big(-\ln \beta \Big) \\
&= n_F \frac{N}{k_B T^2} \frac{\partial}{\partial \beta} \Big(-\frac{1}{\beta} \Big) \\
&= n_F \frac{N}{k_B T^2} \frac{1}{\beta^2}
\end{aligned}$$

Returning the definition of β we arrive at a very simple result for heat capacity

$$C_p = n_F\, N\, \frac{(k_B T)^2}{k_B T^2} \tag{4.11}$$

$$= 3 n_P\, N\, k_B \tag{4.12}$$

$$= 3 n_P\, R \tag{4.13}$$

where we have used the 'ideal gas constant' $R = k_B N$ in the last result. We can non-dimensionalise this expression by moving k_B, and thus obtain the form as it appears in the exponentiated rate equations

$$\frac{C_p}{k_B} = 3 n_P\, N \tag{4.14}$$

$$\frac{C_p}{R} = 3 n_P \tag{4.15}$$

Do note that quantum heat capacity converges to classical heat capacity as $T \to \infty$.

4.4.3 A note on anhamonicity

In a classical system there can be a heat capacity that is not $3n_P R$ if and only if there is deviation from harmonic behavior (see section 4.4.2). The extent of anharmonic behavior will drive the deviation from the trivial heat capacity. As such, anharmonicty could be the origin of ΔC_P in the dynamical origins paper[41]. Because the calculation of ΔC_P was based on an MD simulation, the change cannot be of quantum origins. The variance method used by dynamical origins is able to pick up of changes of heat capacity because it does not assume harmonicity unlike my analysis of heat capacity based on density of states.

4.5 Proposed reasons for error

There were many reasons that could be hypothesized as the cause of the anomalous heat capacity.

4.5.1 Quantum problems

One potential error is that the MD simulation do not include quantum effects, leading to discrepancy in heat capacity of unknown magnitude. However, I do not think that the effect is strong enough here to justify that, especially given that C_P is significantly greater than $3R$ per atom.

4.5.2 Water coupling

Another option is that there is energy coupling between the enzyme and the surrounding water bath. Effectively $\mathrm{Var}(total) << \mathrm{Var}(enzyme)+\mathrm{Var}(water)$. This is really interesting if it is true because it indicates complex interactions with the solvent bath.

4.5.2.1 Charge separation

The strong charges found in charged residues and at the termini may cause the anomalous heat capacity due to water interactions behaving oddly.

4.5.2.2 Hydrophobic interactions

One idea that may cause the water to behave erroneously is not forming proper hydrophobic cages around hydrophobic regions. Water is well know to behave oddly in simulations[49] so this could be doing something to the energy flows. Indeed, the change in heat capacity associated with protein folding is made very complex due to the interaction between water and exposed hydrophobic residues in the unfolded state of the protein[50].

4.5.3 Clustering

One option is that the proper way to compute heat capacities is to look at the heat capacities of each conformation. This can be done quite simply using machine learning.

4.5.4 Non-conservation of energy

One minor problem with molecular dynamics is by numerically integrating the equations of motion, energy is only approximately conserved. One distinct possibility is this non-conservation of energy is creating a larger variance. There are two possible ways to reduce this error, firstly use finer time slices, secondly raise the cutoff distance for electrostatic interactions.

4.5.5 Force field failure

The most interesting possibility is that this anomalous heat capacity is indicative of a failure in the force field itself, and this could be tested by switching to a different force field.

4.5.6 Peptide failure

Some experiments indicated that the peptide bond seems to be the common cause in the anomalous heat capacity. If this is the case then altering the parameters of the dihedral energy function could lead to better results.

4.5.7 Temperature coupling

It is well known that the Berendsen thermostat, as used by the simulation, does dampen vibrations due to a linear temperature response[51]. This should lead to vibrations being dampened, which should reduce overall variance. One possibility is that the separation of temperature coupling, as the system and surroundings are coupled separately, is leading to more variance than it should.

4.6 Tests for identifying the cause of the anomalous heat capacity

Unless stated otherwise all simulations were run for 400 ns. All proteins were parameterized via AMBER, other structures were parameterized via GAFF.

Step size was 2 fs and energy is reported every 500 ps. Everything else was set to the same parameters as seen in the DoS calculations. C_V was chosen as the measure heat capacity as system energy was faster to extract and more consistent than system enthalpy.

4.6.1 MalL

The first test was of the MalL enzyme replicating the "Origins of heat capacity"[41] paper on my own MD trajectories, as collected in chapter 2. This was trajectory was run for 160 ns. The anomalously high absolute value for the heat capacity was replicated, at approximately 10 times the accepted experimental value, indicating that there was not a mathematical error in the original calculations.

4.6.2 Barnase

To ensure that this problem can be replicated the Barnase enzyme was tested. The anomalously high heat capacity was also reproduced for barnase which is significantly smaller than MalL. The calculated value for the heat capacity from the single MD trajectory using the variance method was 10 J/gK as opposed to the experimental value of 1.33 J/gK[52].

4.6.3 Barnase water

The water surrounding the barnase was also tested and found to have a heat capacity of 6 J/gK, as opposed to the experimental value of 4 J/gK. This provides a strong indication of the water being involved in the anomalously high heat capacity.

4.6.4 Sugars

To see if this anomalous heat capacity occurred with different substances, isomaltose and a four-unit starch oligomer was tested. Both showed heat capac-

ities that were approximately accurate, ignoring the quantum effects. These heat capacities were still somewhat high, but were more reasonable. This indicates that the variance method does have some value for calculating absolute C_V. The dimer had a C_V of 950 J/ mol K as opposed to the experimental value of 436 J/ mol K [53]. The tetramer had a C_V of 1892 J/ mol K compared to the experimental value of 839 J/ mol K. These values are approximately two times greater than the experimental values and indicates that the error may scale with system size.

4.6.5 Quad ala/gly

To see if peptides also showed anomalous absolute heat capacity values an alanine and a glycine tetramer were tested. Both also had anomalously high heat capacities. The ala tetramer had a C_V of 1920 J/ mol K as opposed to the experimental value of 733.4 J/ mol K[52]. The gly tetramer showed a C_V of 6280 J/ mol K as opposed to the experimental value of 283 J/ mol K. The difference for the gly tetramer seems to be unusually high and may provide clues as to the origin of the error. This also shows that system size has no effect on the error for heat capacity. Glycine is particularly flexible when compared to other amino acids[21] and so conformational changes may not be well treated in simulation. Because of these results, gly oligomers was chosen for subsequent experiments as they are small and thus very fast to simulate.

4.6.6 Quad ala/gly with GROMOS force field

An alternate force field was tested (GROMOS) to see if AMBER was the primary cause. The GROMOS force field produced heat capacities of approximately the same magnitude. With a C_V of 2960 and 3637 J/mol K for ala and gly respectively. The AMBER force field was used in subsequent experiments because of this.

4.6.7 Quad gly with finer time steps

To test if a lack of conservation of energy due to numerical integration is the cause of anomalous heat capacity, a timestep of 1 fs was tested, with a glycine tetramer, showing no real change, with a heat capacity of 4828 J/mol K.

4.6.8 Quad gly with tighter buffers

For other issues that could drive non-conservation of energy, the PME buffer parameter was tightened by a factor of four. Again no real change, with a calculated heat capacity of 6180 J/mol K.

4.6.9 ala/gly dimers

More testing was done on the even simpler system of ala and gly dimers. Again it was found to have an anomalously high heat capacity. This was found to have a heat capacity of 707 and 602J/mol K as opposed to experimental values of 303 and 99 J/mol K for ala and gly respectively.

4.6.10 gly dimer with different sliding windows

Different sliding window sizes were tested to see if they were the cause of the heat capacity differences. There was a decline in heat capacity over the sliding window, but it did not stabilize (See figure 4.1). A window size of 100 ps did produce a value equivalent to the experimental value, although this point is very arbitrary and cannot be used predictively.

4.6.11 Gly kinetic energy

In a classical system, the heat capacity of the subsystem entirely composed of the kinetic energy of the particles should be $1.5Rn$ or 12 J/mol K per atom[51]. So this was tested with the gly dimer. Here only the variance of the kinetic energy was looked at. This was found to have a kinetic energy heat capacity of 9.8 J/mol K per atom. This is lower than expected, however the Berendsen

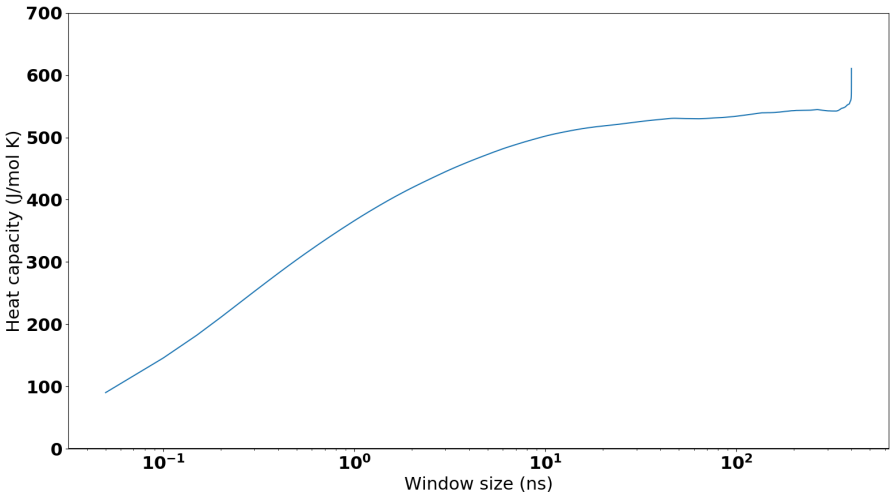

Figure 4.1: A graph showing how computed heat capacity changes over different sliding windows used to compute heat capacity.

thermostat does reduce energy fluctuations more than it should, so this is to be expected[51]. This indicates that there is a problem in the computation of potential energy, as kinetic energy can be eliminated from consideration. The high level dynamics of the system are acting as they should, which implies that the anomalously high heat capacity is due to how we are analyzing the system.

4.6.12 Other peptides

Other peptides were tested, namely a kevlar dimer and a 3-3 nylon dimer. Both were tested to see if it was just amino acids that had the anomalous issue with heat capacity. The aromatic system had a C_V of 2.67 J/gK and the nylon system had a C_V of 1.49 J/gK. The aromatic system seems too high but the nylon system seems within expected bounds.

4.6.13 Clustering

One idea is that the variance of each conformation should be investigated separately. A simulation of the glycine dimer was taken. All dihedral angles were recorded and clustered using expectation minimization/Gaussian mixture

clustering[54]. Each cluster had its energy variance measured and averaged out. This would underestimate heat capacity because it does not include energy held in higher energy states, meaning that more energy would need to be put into the system than expected by this method. This still showed an abnormally high heat capacity of 543 J/ mol K.

4.6.14 Changes in interaction energies

Given that kinetic energies were approximately correct, the Hamiltonian mechanics of the system were likely also correct. Thus it is likely that we are looking at the energy of the system incorrectly. It could be that the interaction energies of the system and the surroundings were strongly negatively coupled to the energy of the system. This would mean that $\mathrm{Var}(E_{system}) << \mathrm{Var}(E_{system} + E_{interaction})$. The new value for the systems energy was $E_{new} = E_{system} + E_{interaction}/2$. Because we can't measure the interaction energy directly, we compute it by $E_{interaction} = E_{total} - (E_{system} + E_{suroundings})$. This produced a heat capacity of 1.58 J/gK for barnase. This is close enough to the experimental value of 1.33 J/gK to say the rest of the difference may be due to the omission of quantum effects.

4.7 Conclusion

To conclude, the anomalous heat capacity was determined to be due to the system boundary not being as weakly coupled as is required for predicting heat capacity from variance data. Using some form of implicit solvation method was found to be useful for computing heat capacity. However, I was not able to fully reconcile experimental heat capacities with the theoretical values. But on the other hand, I was able to reconcile them down to a level where the fundamental sources or error of using MD energy calculations became apparent. The only way to get a better value would be to use some form of quantum calculation.

Given that the anomalous absolute heat capacity seems to be due to in-

teractions between the system and the surroundings it is likely that relative heat capacity between reaction and transition state analog is not affected and produces accurate values. However, since we do not have a useful model of what changes drive $\Delta C_P^{\ddagger}$ we cannot say this for certain.

Chapter 5

Summary and Conclusions

5.1 Summary

5.1.1 Density of States

In the analysis of density of states performed using the velocity autocorrelation function, no difference between the reaction state and the transition state analog could be found. This method was found to be a good way to calculate the DoS. The DoS was found to be strongly temperature dependent. There is evidence that $\Delta C_P^{\ddagger}$ is driven more by anharmonicity than by a shift in harmonic frequencies. The computations of DoS seemed to be accurate, but had major normalization problems.

5.1.2 Correlation distance and phase change

In the analysis of correlation distances, no difference in correlation distances could be measured between the reaction state and the transition state in the harmonic case. Trajectory based analysis was not found to be a good way to identify correlation distances, I think this is due to all the interacting vibrational modes. If there is a phase transition, it is more likely caused by rapid fluctuations between states than in the structure of an individual state.

5.1.3 Anomalous heat capacity of proteins

The anomalous absolute heat capacity of enzymes was determined to be due to energy fluctuations between the system and the system boundary. As such, including the energy of the system boundary into the energy of the system significantly reduces the calculated heat capacity down to reasonable levels. However, $\Delta C_P^{\ddagger}$ remains accurate, meaning that it is not being driven by iterations with the solvent. This method was not found to be a good method for computing absolute heat capacities of system that are solvated, having at best an error on the order of a factor of 2.

5.1.4 Synthesis

Harmonic approaches were found to be of limited use in analyzing the differences between the reaction state and the transition state analog for enzyme catalysis. While a harmonic analysis produced useful results when it came to computing DoS it failed to produce a useful result when comparing the DoS between a reaction and transition state. A harmonic analysis provided useful information of correlation distance but then failed to generate useful differences in correlation distances. As seen by my analysis of heat capacities there are differences in dynamics between reaction states and transition states, meaning that these differences are anharmonically driven, which has significant implication for further computational analysis of enzyme catalysis.

5.2 Future research

The focus of future research for this work is to extend the simulations to see if the lack of positive results is due to not fully covering the possible phase space of MalL[41]. Due to a significant lack of computational resources available to me I was only able to sample 1 μs of computational time while it seems that at least 20 μs are required[41], with additional resources I could extend the simulations. Additionally, metadynamics[55] could be used to see if a phase

change discontinuity occurs in free energy over the course of a reaction. Again this would require more computational resources than presently available. Finally, reanalysis of the MalL trajectories using implicit solvation[56] should be performed to see if that changes values for $\Delta C_P^{\ddagger}$. Additionally, this method has some potential for computing heat capacities of small molecules.

5.3 Conclusion and speculation

To sum up all the work done in this thesis, all solutions that are based in harmonic approximations fail to create any notable difference between reaction states and transition states. However, approaches that do not require harmonic approximations do produce significant differences in heat capacity[41]. This provides a strong indicator that anharmonic vibrations are what drive these differences in heat capacity. This has some interesting thermodynamic implications in that most systems of coupled anharmonic oscillators can act as continuous quantum engines [57]. A continuous quantum engine is a quantum equivalent of a classical heat engine that operates continuously and without discrete strokes[57], although work in a quantum engine is defined as putting energy into a heat bath of infinite temperature. As such, I think that the enzyme is acting in a similar process, pumping energy into a "work bath" composed of the bond breaking mode. As a passive system, an enzyme cannot change expectation values of energy in the bond breaking mode[36], I think that the enzyme acts as what I would like to call a "quantum probability pump", reducing the variance in energy inside itself to increase variance in energy in the substrate;s bond breaking mode. This would allow the system to more effectively go over the transition state barrier.

References

[1] David R Edwards, Danielle C Lohman, and Richard Wolfenden. Catalytic proficiency: the extreme case of s–o cleaving sulfatases. *Journal of the American Chemical Society*, 134(1):525–531, 2012.

[2] Pamela Sears and Chi-Huey Wong. Intervention of carbohydrate recognition by proteins and nucleic acids. *Proceedings of the National Academy of Sciences*, 93(22):12086–12093, 1996.

[3] Kathryn M Koeller and Chi-Huey Wong. Synthesis of complex carbohydrates and glycoconjugates: enzyme-based and programmable one-pot strategies. *Chemical Reviews*, 100(12):4465–4494, 2000.

[4] The Royal Swedish Academy of Sciences. Press release: The nobel prize in chemistry 2018, 2018.

[5] Olga Khersonsky, Daniela Röthlisberger, Orly Dym, Shira Albeck, Colin J Jackson, David Baker, and Dan S Tawfik. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the ke07 series. *Journal of molecular biology*, 396(4):1025–1042, 2010.

[6] Sam Hay and Nigel S Scrutton. Good vibrations in enzyme-catalysed reactions. *Nature chemistry*, 4(3):161, 2012.

[7] David R Glowacki, Jeremy N Harvey, and Adrian J Mulholland. Taking ockham's razor to enzyme dynamics and catalysis. *Nature chemistry*, 4(3):169, 2012.

[8] John E Coligan and Virginia Benson Chanda. Current protocols in protein science. 2000.

[9] New Zealand Ministry for Primary Industries. Transitional facilities for biological products. `https://www.mpi.govt.nz/dmsdocument/1640/direct`, 2019.

[10] Merck. tert-butyllithium sds. `https://www.merckmillipore.com/NZ/en/product/msds/MDA_CHEM-814147`, 2019.

[11] Henry M Miziorko. Enzymes of the mevalonate pathway of isoprenoid biosynthesis. *Archives of biochemistry and biophysics*, 505(2):131–143, 2011.

[12] VK Ahluwalia and M Kidwai. Basic principles of green chemistry. In *New Trends in Green Chemistry*, pages 5–14. Springer, 2004.

[13] Arthur Cayley. Chapters in the analytical geometry of (n) dimensions. *Cambridge Mathematical Journal*, 4:119–127, 1843.

[14] Henry Eyring. The activated complex and the absolute rate of chemical reactions. *Chemical Reviews*, 17(1):65–77, 1935.

[15] Donald G Truhlar. Transition state theory for enzyme kinetics. *Archives of biochemistry and biophysics*, 582:10–17, 2015.

[16] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[17] Kathrin Helen Hopmann and Annette Bayer. Enantioselective imine hydrogenation with iridium-catalysts: reactions, mechanisms and stereocontrol. *Coordination Chemistry Reviews*, 268:59–82, 2014.

[18] Michael J Hostetler, Matthew D Butts, and Robert G Bergman. Scope and mechanism of alkene hydrogenation/isomerization catalyzed by complexes of the type r2e (ch2) 2m (co)(l)(r= cp, me, ph; e= phosphorus, tantalum; m= rhodium, iridium; l= co, pph3). *Journal of the American Chemical Society*, 115(7):2743–2752, 1993.

[19] Joseph Kraut. How do enzymes work? *Science*, 242(4878):533–540, 1988.

[20] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2012.

[21] Robert K Murray, Daryl K Granner, Peter A Mayes, and Victor W Rodwell. *Harpers illustrated biochemistry*. Mcgraw-hill, 2014.

[22] William R Cannon and Stephen J Benkovic. Solvation, reorganization energy, and biological catalysis. *Journal of Biological Chemistry*, 273(41):26257–26260, 1998.

[23] Tom CB McLeish, Martin J Cann, and Thomas L Rodgers. Dynamic transmission of protein allostery without structural change: spatial pathways or global modes? *Biophysical journal*, 109(6):1240–1250, 2015.

[24] Alfonso Tramontano, Adrian A Ammann, and Richard A Lerner. Antibody catalysis approaching the activity of enzymes. *Journal of the American Chemical Society*, 110(7):2282–2286, 1988.

[25] Yang Xu, Noboru Yamamoto, and Kim D Janda. Catalytic antibodies: hapten design strategies and screening methods. *Bioorganic & medicinal chemistry*, 12(20):5247–5268, 2004.

[26] Anastassia N Alexandrova, Daniela Rothlisberger, David Baker, and William L Jorgensen. Catalytic mechanism and performance of computationally designed enzymes for kemp elimination. *Journal of the American Chemical Society*, 130(47):15907–15915, 2008.

[27] Steven D Schwartz and Vern L Schramm. Enzymatic transition states and dynamic motion in barrier crossing. *Nature chemical biology*, 5(8):551, 2009.

[28] Shina CL Kamerlin and Arieh Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Structure, Function, and Bioinformatics*, 78(6):1339–1375, 2010.

[29] Louis YP Luk, J Javier Ruiz-Pernía, William M Dawson, Maite Roca, E Joel Loveridge, David R Glowacki, Jeremy N Harvey, Adrian J Mulholland, Iñaki Tuñón, Vicent Moliner, et al. Unraveling the role of protein dynamics in dihydrofolate reductase catalysis. *Proceedings of the National Academy of Sciences*, 110(41):16344–16349, 2013.

[30] Joanne K Hobbs, Wanting Jiao, Ashley D Easter, Emily J Parker, Louis A Schipper, and Vickery L Arcus. Change in heat capacity for enzyme

catalysis determines temperature dependence of enzyme catalyzed rates. *ACS Chemical Biology*, 8(11):2388–2393, 2013.

[31] Louis A Schipper, Joanne K Hobbs, Susanna Rutledge, and Vickery L Arcus. Thermodynamic theory explains the temperature optima of soil microbial processes and high q10 values at low temperatures. *Global change biology*, 20(11):3578–3586, 2014.

[32] Vickery L Arcus and Christopher R Pudney. Change in heat capacity accurately predicts vibrational coupling in enzyme catalyzed reactions. *FEBS letters*, 589(17):2200–2206, 2015.

[33] Hannah BL Jones, Rory M Crean, Christopher Matthews, Anna B Troya, Michael J Danson, Steven D Bull, Vickery L Arcus, Marc W Van Der Kamp, and Christopher R Pudney. Uncovering the relationship between the change in heat capacity for enzyme catalysis and vibrational frequency through isotope effect studies. *ACS Catalysis*, 8(6):5340–5349, 2018.

[34] Eduard August Grüneisen and Max Planck. *Annalen Der Physik*, volume 44. JA Barth, 1914.

[35] RM Ibberson, WIF David, O Yamamuro, Y Miyoshi, T Matsuo, and H Suga. Calorimetric, dielectric, and neutron diffraction studies on phase transitions in ordinary and deuterated acetone crystals. *The Journal of Physical Chemistry*, 99(38):14167–14173, 1995.

[36] Stephen J Blundell and Katherine M Blundell. *Concepts in thermal physics*. OUP Oxford, 2009.

[37] Victor Havin and Burglind Jöricke. *The uncertainty principle in harmonic analysis*, volume 28. Springer Science & Business Media, 2012.

[38] Jie Hu, Ao Ma, and Aaron R Dinner. Monte carlo simulations of biomolecules: The mc module in charmm. *Journal of computational chemistry*, 27(2):203–216, 2006.

[39] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[40] Albert Einstein. Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt. *Annalen der physik*, 322(6):132–148, 1905.

[41] Marc W van der Kamp, Erica J Prentice, Kirsty L Kraakman, Michael Connolly, Adrian J Mulholland, and Vickery L Arcus. Dynamical origins of heat capacity changes in enzyme-catalysed reactions. *Nature communications*, 9(1):1177, 2018.

[42] Patrick L Wintrode, Deqiang Zhang, Nagarajan Vaidehi, Frances H Arnold, and William A Goddard III. Protein dynamics in a family of laboratory evolved thermophilic enzymes. *Journal of molecular biology*, 327(3):745–757, 2003.

[43] Hyuntae Na, Guang Song, and Daniel Ben-Avraham. Universality of vibrational spectra of globular proteins. *Physical biology*, 13(1):016008, 2016.

[44] Ross S Firestone, Scott A Cameron, Jerome M Karp, Vickery L Arcus, and Vern L Schramm. Heat capacity changes for transition-state analogue binding and catalysis with human 5-methylthioadenosine phosphorylase. *ACS chemical biology*, 12(2):464–473, 2016.

[45] Christian Hounslow. Explaining heat capacity changes in the enzyme mall.

[46] Daniel F Walls and Gerard J Milburn. *Quantum optics*. Springer Science & Business Media, 2007.

[47] Nigel Goldenfeld. *Lectures on Phase Transitions and the Renormalization Group*. Westveiw press, 1992.

[48] Terrell L Hill. *An introduction to statistical thermodynamics*. Courier Corporation, 1986.

[49] Pak K Yuet and Daniel Blankschtein. Molecular dynamics simulation study of water surfaces: comparison of flexible water models. *The Journal of Physical Chemistry B*, 114(43):13786–13795, 2010.

[50] Mikael Oliveberg, Yee-Joo Tan, and Alan R Fersht. Negative activation enthalpies in the kinetics of protein folding. *Proceedings of the National Academy of Sciences*, 92(19):8926–8929, 1995.

[51] Stephen C Harvey, Robert K-Z Tan, and Thomas E Cheatham III. The flying ice cube: velocity rescaling in molecular dynamics leads to violation of energy equipartition. *Journal of Computational Chemistry*, 19(7):726–740, 1998.

[52] George I Makhatadze. Heat capacities of amino acids, peptides and proteins. *Biophysical chemistry*, 71(2-3):133–156, 1998.

[53] Lars-Erik Briggner and Ingemar Wadsö. Heat capacities of maltose, maltotriose, maltotetrose and α-, β-, and γ-cyclodextrin in the solid state and in dilute aqueous solution. *The Journal of Chemical Thermodynamics*, 22(11):1067–1074, 1990.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[55] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 2011.

[56] Alexey Onufriev. Implicit solvent models in molecular dynamics simulations: A brief overview. In *Annual Reports in Computational Chemistry*, volume 4, pages 125–137. Elsevier, 2008.

[57] Robert Alicki and Ronnie Kosloff. *Introduction to quantum thermodynamics: History and prospects*. Springer Cham, 2019.