Title
Biomolecular Dynamics and Function: A Study on Amino Acids and Enzymes

Author
Belsare, Saurabh

Publication Date
2017

Peer reviewed|Thesis/dissertation

Biomolecular Dynamics and Function: A Study on Amino Acids and Enzymes

by

Saurabh Belsare

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Teresa Head-Gordon, Chair
Professor Patricia Babbitt
Professor Mohammad Mofrad
Professor Judith Klinman

Spring 2017

Biomolecular Dynamics and Function: A Study on Amino Acids and Enzymes

Abstract

Biomolecular Dynamics and Function: A Study on Amino Acids and Enzymes

by

Saurabh Belsare

Joint Doctor of Philosophy
with University of California, San Francisco in Bioengineering

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

Proteins are biomolecules involved in cellular structure as well as function. These molecules are long chain polymers consisting of amino acids, which are organic compounds containing many different functional groups such as amine ($-NH_2$) and carboxylic acids ($-COOH$). Actin proteins form part of the cellular structure, membrane proteins act as channels for transfer of ions, and enzymes catalyze critical cellular reactions. While structure is a well-appreciated determinant of function, the role of dynamics of proteins and solvent are less well studied. In my thesis work, I have studied the statistical fluctuations and dynamics of the basic amino acids in water and up to the full complexity of enzymes. The combination of experimental and computational techniques is a powerful combination for obtaining insight into dynamical events. I have used force-field based classical molecular dynamics simulations using an advanced polarizable force field to study the behaviour of these biomolecules in solution and have simulated experimental observables to understand conformational motions.

In the first part of my thesis, I have characterized the dynamical modes of a basic protein unit - a single zwitterionic amino acid in solution - to make quantitative comparisons to the low frequency Terahertz (THz) absorption spectra. An analysis protocol for decomposing the THz absorption spectrum has been previously developed for analyzing zwitterion simulations performed using *ab-initio* molecular dynamics (AIMD). In this work we extend the analysis method to simulations performed by force field molecular dynamics, which are computationally far less intensive, and setting the stage for decomposing the THz spectra for larger proteins that are not affordable by AIMD. We also show that the main impact of the solvation on the dynamical modes of zwitterions comes from the first solvation shell around the zwitterion only, and presence of waters further out does not affect the dynamics of these molecules significantly.

In the second part of my thesis work, I have explored the role of statistical fluctuations of solvation for artificial enzymes - which have poor activity - and have evaluated how the entropic features change upon mutation through laboratory directed evolution in which the enzymes show much greater activity. I have used two Kemp Eliminases (KE07 and KE70)

and show that the active sites of these two enzymes have starkly contrasting interactions with solvent. KE07 incorporates the water into the active site to enhance the catalysis of the Kemp Elimination reaction while KE70 creates a strong hydrophobic pocket leading to the catalysis being driven entirely by the protein residues at the active site. Different entropic species of waters based on their vibrational dynamics are identified, and we observe varying behaviour of waters between mutants, as well as with the presence of the ligand.

In the final part of my thesis, I have looked at the dynamical correlations between residues in KE07 and have evaluated how the dynamical correlations change upon mutation through laboratory directed evolution. In particular, I have characterized the residue-residue interactions where we find that there is correlated motion between surface loops in KE07, which potentially could modulate access of the ligand to the active site of the enzyme; we observe that the binding of the ligand increases the correlation between the residues of the protein in the higher performing variants of the enzyme.

To my father Milind, mother Swati, and sister Sayali.

Contents

List of Figures

List of Tables

Acknowledgments

For this work, and my entire PhD journey, I owe a deep gratitude to my research advisor Prof. Teresa Head-Gordon, who has guided me through this wonderful, but often mysterious landscape of research. I have learned a lot about how one should critically examine a problem, and how to formulate an approach to tackling it, from her. I would also like to thank MSc. Alexander Esser, MSc. Viren Pattni, Prof. Dominik Marx, and Prof. Matthias Heyden for work on various collaboration projects. It was a wonderful learning experience to work with people with different expertise, and to be able to contribute together towards new research.

I would like to express my sincere gratitude to my Graduate Advisor Prof. Ian Holmes, our former program administrator Rebecca Pauling and last, but certainly not least, our current program administrator Kristin Olson, for guidance and support at various points of uncertainty and self-doubt during the course of this journey. I would like to thank Arezou Razavi, our lab administrative assistant, for help with all the bureaucracy, and our cluster systems administrator Kelley McDonald for his help with our computational resources.

All of the people in the Head-Gordon lab, current and former, have been great friends as well as co-workers, and in particular, I would like to thank Dr. K. Aurelia Ball, Dr. Asmit Bhowmick, Alex Albaugh, Sukanya Sasmal, Dr. Omar Demerdash, Dr. Sudhir Sharma, and Eugene Yedvabny for a multitude of both, academic and non academic discussions over the years. Many teachers all the way from school through college to graduate school have been instrumental in my journey thus far, and I would like to thank all of them too. I am still fortunate to be in touch with friends who I have known for ten, fifteen, or in some cases even twenty-five years (you know who you are), and regular conversations with them, even if over a phone call, video conferencing, or just over text, have often been a great way to relax on many weekends during this journey.

Above all, I would like to thank my parents and my sister, who, despite being on the other side of the continent or even other side of the planet, have always only been just a phone call away. They have been the greatest supports through this journey, and they deserve a significant part of the credit of this process.

Chapter 1

Introduction

Naturally occurring enzymes are very strong catalysts, and have been observed to increase the rates of biological reactions over 11 orders of magnitude compared to the uncatalyzed reactions[1] ($\frac{K_{cat}}{K_M}/k_{uncat}) > 10^{11}$. There has been much discussion over the mechanism through which these enzymes can catalyze reactions at this extraordinary rate. The pre-organized electrostatic effects resulting from the structural fold of the enzyme - including the residues at the active site as well as further distal residues - have long been thought to be the dominant factor controlling catalysis[2]. But recent research has also suggested that flexibility in the active site and the scaffold, as well as the internal dynamics of the protein, also affect catalytic activity[3–14].

There is often not a clear definition of what is meant by "dynamics". One definition has considered dynamics as "non-equilibrium barrier crossing effects occurring during catalysis[8]". An alternative definition of dynamics, however, considers equilibrium fluctuations, like side chain motions, as well as "time dependent changes in atomic coordinates" which have been shown to contribute to catalytic activity[7]. Multiple studies have shown these effects in proteins such as di-hydrofolate reductase(DHFR) and liver alcohol dehydrogenase (LADH)[4–6, 9, 10]. In addition to the enzyme structure and dynamics, solvent fluctuations have been suggested to control protein motions and functions[15]. The protein folding pathway has been shown to be slaved to solvent motions, i.e. the activation enthalpy of folding is dominated by the solvent[16]. Certain experiments have shown the existence of a hydration layer extending out to $\sim 20\mathring{A}$ from the surface of the protein[17]. The waters in this hydration layer are different from the bulk in terms of their diffusion and low frequency intermolecular vibrations. In addition, the dynamical signatures of waters near the surfaces of biomolecules have been shown to depend on the chemical nature of the biomolecular species, in addition to the topography of the molecular surface[18]. Hence, the interactions of proteins and solvent with each other are inextricably linked, and solvation in the active site would play an important role in determining the catalytic activity of the enzyme.

1.1 Solvation in Small Biomolecules

Chapter 1 discusses the effect of interactions of small biomolecules, namely single zwitterionic amino acids, with water. These amino acids are the building blocks of proteins, and understanding their dynamic landscape would be the first step towards understanding dynamics in larger biomolecules. In particular, one of the experimental techniques used to study molecular dynamics is TeraHertz (THz) spectroscopy[19]. THz spectroscopy explores the infrared absorption spectrum between the region of 0.1 to 30 THz frequency (3 to 1000 cm^{-1}). The signals in this spectrum for pure water have been well studied and understood in terms of component motions, including hydrogen bond network vibration with a peak at $200cm^{-1}$ and librational motions of waters in their hydrogen bonded environment with a peak at $650cm^{-1}$. The IR spectrum of water shows further intra molecular peaks at $\sim$ 1600 cm^{-1} for the H-O-H angle bending mode and around $\sim$ 3700 cm^{-1} for the O-H bond stretch motions. However, the decomposition of the equivalent spectra for biomolecules in water into the component motions isn't known. A method has recently been developed [20] to decompose the simulated THz spectra for these systems into component motions using AIMD ($Ab-initio$ molecular dynamics) simulations. However, AIMD simulations are computationally intensive, in comparison to force field based molecular dynamics simulations. In our work, we extend this method to perform the dynamic mode decomposition based on force field molecular dynamics (FFMD) simulations using an advanced polarizable force field, AMOEBA. We have found excellent agreement with the AIMD, setting the stage for using FFMD for larger systems that are inaccessible to AIMD.

1.2 Solvation in Designed Enzmyes

Chapter 2 discusses the impact of solvation in the catalytic activity of artificial enzymes, in which we have calculated the vibrational density of states for waters using AMOEBA with molecular dynamics simulations in order to calculate the solvent entropies using a spatially decomposed 2-state thermodynamic theory. In this study we have characterized the solvation entropy signatures in the apo, ligand bound (EL), and transition complex ($EL^{\dagger}$) for designed Kemp Eliminase enzymes modeled and synthesized by the Baker Group at the University of Washington[21]. One of the enzymes, KE07, was further mutated by subjecting it to 8 rounds of directed evolution[22], leading to a final improvement of $\sim$ 200 fold in the $\frac{k_{cat}}{K_M}$. Another designed Kemp Eliminase, KE70, was improved through 9 rounds of directed evolution[23], resulting in a $\sim$ 400 fold improvement in $\frac{k_{cat}}{K_M}$. Looking at the number of waters as well as the solvent entropies for waters in KE07 and KE70, we observe that these two enzymes show very distinct solvation signatures. The best performing KE07 variant allows access to larger number of waters in both the apo and ligand bound states, compared to the original designed enzyme. In KE70, on the other hand the active site is much more compact, and does not allow water in the presence of the ligand. This study shows how laboratory directed evolution can adopt very distinct paths to catalyze the same reaction.

1.3 Residue Correlation

The final part of my thesis work considers true dynamics in the form of distance-distance time correlations between residues in a particular Kemp Eliminase enzyme, namely KE07. In particular we analyze the dynamic correlation spectra, which are the Fourier Transforms of the atomic position autocorrelation functions for the various residues in the proteins. Since some of the residues mutated were in the active site, while a lot of mutations were distal, we hypothesized that the effect of these distal mutations is to alter the overall dynamics of the enzyme, leading to an improvement in the catalysis. We show that as with the scaffold TIM barrel protein from which this enzyme was designed, the motions of the loops are strongly correlated with the catalytic residues in the high performing variants as opposed to the original design, showing that incorporating residue dynamic information can further be used to enhance enzyme design.

1.4 Contributions

Since some of the work includes collaborations with other labs, I am outlining my contributions within each of these chapters.

Chapter 2: Small Molecule THz Spectra - I performed the new parameterization of the Zwitterionic Amino Acids, I performed all the simulations, Alexander Esser and I jointly came up with the method for adapting the mode decomposition to work with the AMOEBA force field. Alexander Esser performed the mode decomposition analysis and we jointly discussed the interpretations of the results. The paper was written together.

Chapter 3: Solvent Contributions in Kemp Eliminases - I performed the parameterization of the ligand, I performed all the simulations for all enzymes, I performed all the analysis except for the analysis included in the section "Hydration Water Species Analysis", which was performed by Viren Pattni. i.e. I performed the analysis for the Catalytic Site Geometry, Active Site Waters, the Solvation Layer, and Entropic Contributions to Catalysis. I have also written this entire chapter.

Chatper 4: Residue Correlation - This chapter is entirely my work, not part of a collaboration.

Chapter 2

Small Molecule THz Spectra

Note: This chapter is reproduced from *A. Esser*, S. Belsare*, D. Marx, and T. Head-Gordon (2017). Mode specific THz spectra of solvated amino acids using the AMOEBA polarizable force field. Phys. Chem. Chem. Phys. 19, 5579-5590* with permission.

2.1 Introduction

Understanding the molecular motions that arise from solute-solvent interactions is one of the key problems in Solvation Science. Terahertz (THz) and related spectroscopies have proven to be a sensitive tool in order to probe solvation shell dynamics around a variety of solutes, from simple single monovalent ions to complex biological systems like proteins and enzymes [19, 25–30]. But in order to understand the experimentally observed spectra in molecular detail, theoretical methods are required. For small molecular species and simple ions, *ab initio* molecular dynamics (AIMD) simulations [31] have proven to give a reasonably faithful description of the THz experimental observable, and hence can be relied upon to decompose the motions of solvation shell dynamics [20, 32–35]; see Ref. 36 for a review of the techniques underlying the AIMD approach to theoretical infrared (and thus THz) spectroscopy. In particular for zwitterionic glycine in aqueous solution, AIMD interpreted the THz observable to have three major modes of motion, including rigid body translational motions of the whole molecule at low frequencies (~ 80 cm^{-1}), intermolecular cross correlation modes due to the interaction of the zwitterion with the solvent at ~ 200 cm^{-1}, after which purely intramolecular angle bending modes are present[20].

However, the computational cost is a limiting factor for extending AIMD to larger systems or to faithfully probe solvation dynamics beyond the second solvation shell. In principle, force field simulations should easily allow extension to larger systems due to their more tractable cost even when using polarizable versions [37–50], however there is a need for validation against AIMD to ensure that the THz spectra and mode decomposition are consistent using the simpler model to describe the interatomic interactions. In the THz regime, electronic polarization and/or charge transfer effects, which are included in AIMD simulations since

they rely on solving self-consistently the electronic structure problem on-the-fly[31], are of particular importance[33]. In turn the more tractable force fields can probe any potential problems with finite system size effects, as well as cross-validate the AIMD protocols for simulating the THz spectra and assumptions for interpreting the low frequency modes.

In this study the AMOEBA polarizable model[46, 51] is tested for its ability to reproduce the results given by AIMD on the solvent induced intramolecular and intermolecular motions of the zwitterioinic form of single glycine and valine molecules in water. We have chosen AMOEBA since validation studies on bulk water have demonstrated that the THz observable is qualitatively reproduced (Fig.2.1). It is noteworthy that the signature of the intermolecular vibrations of the water network in the $\approx 200\ cm^{-1}$ (or ≈ 6 THz) region is captured by the direct polarization iAMOEBA[52] and full mutual polarization AMOEBA models[53], whereas if we turn-off the many-body polarization component, this feature is lost from the simulated THz spectrum (Fig.2.1). This suggests that more standard fixed partial charge models would be insufficient for representing intermolecular interactions probed by the THz experiment [33, 54], hence we require at least many-bodied polarization[49] as a minimum level of physics for the force field that might replace the AIMD simulations for larger systems. However there are other important quantum mechanical features that are not currently accounted for in AMOEBA, but are clearly present in the AIMD simulations, including charge penetration and charge transfer. Although active work for incorporating these important short-ranged interactions are under active development within the force field community [45, 55–61], and are starting to be introduced in more standard molecular mechanics models[62–70], they are not present in the current version of AMOEBA.

As will be detailed based on comprehensive analyses for glycine and valine in water, we find that the AMOEBA model performs well in comparison to AIMD in terms of capturing the intramolecular modes and the hindered translation (cage rattling) and hindered rotation (libration) modes of the zwitterions, as well as the intermolecular cross correlation modes of the zwitterion with water. It is noted in passing that the AMOEBA parameters for the two single zwitterionic amino acids in AMOEBA14 water had to be developed based on a systematic protocol. What is remarkable is the level of agreement between the polarizable force field and electronic structure based treatments given the differences in how the molecular dipole moments are calculated and the assumptions that go into the mode decomposition that uses a charge weighted velocity cross-correlation matrix. An additional benefit to the AMOEBA investigation here is to examine the potential influence of finite size effects on the calculated THz observables in the AIMD study, for which we find no issues, except for the simple loss of information for outer water shell dynamics beyond the first solvation shell in the present case.

The remainder of this paper is outlined as follows. In section 2.2 we describe the theoretical models and methods. The resulting data and analysis for AMOEBA simulations of the THz spectra and mode decompositions of the two amino acids are compared against the AIMD results and discussed in Sec. 2.4. The insights gained from these benchmark calculations of AMOEBA are discussed in Sec. 2.5 and plans for future studies are discussed.

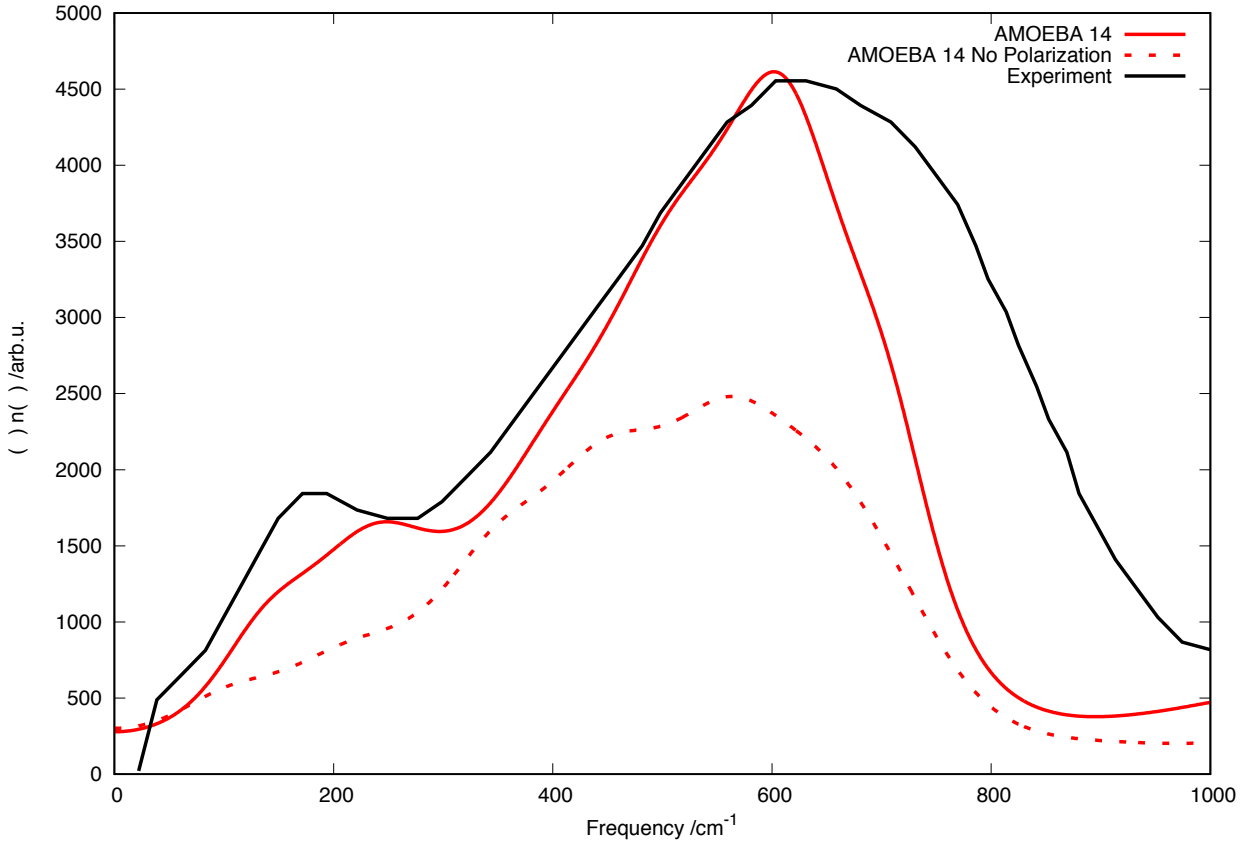


Figure 2.1: THz spectra, $\alpha(\omega)n(\omega)$, of pure bulk water for AMOEBA water with (solid red line) and without (dashed red line) the induced dipole contributions (see text) compared to the corresponding experimental data (adapted from the SI of Ref. 52). Note: AMOEBA 14 is only a water model, and that water model had been consistently used throughout all the simulations in this paper, for both the AMOEBA original and AMOEBA new simulations outlined in the methods and results sections.

2.2 Theoretical Models and Simulation Methods

AMOEBA Model

In particular, the AMOEBA potential energy is formulated as

$$U = U_b + U_\theta + U_{tors} + U_{b\theta} + U_{oop} + U_{vdW} + U_{ele}^{perm} + U_{ele}^{pol} \tag{2.1}$$

where U_b and U_θ correspond to harmonic bond and angle deformations, U_{tors} is a truncated Fourier series to describe rotations around bonds, $U_{b\theta}$ is a Urey-Bradley coupling term, and U_{oop} comprises the out-of-plane bending energy, while the last three terms embody the non-bonded interactions. Given that non-bonded terms are the most important aspect of solute-solvent interactions, we describe them in more detail.

The first non-bonded term is the permanent electrostatics (U_{ele}^{perm}) based on an atom-centered point multipole on each atomic site i, comprising monopole (q_i), dipole ($\boldsymbol{\mu}_i$), and quadrupole ($\mathbf{Q}_i$) moments:

$$\mathbf{M}_i = [q_i, \mu_{ix}, \mu_{iy}, \mu_{iz}, Q_{ixx}, Q_{ixy}, Q_{ixz}, Q_{iyy}, Q_{iyz}] \tag{2.2}$$

The total permanent electrostatics contribution is then evaluated as the pairwise sum of interactions between different atomic sites:

$$E_{ele}^{perm} = \sum_{i<j} \mathbf{M}_i \mathbf{T}_{ij} \mathbf{M}_j \tag{2.3}$$

where $\mathbf{T}_{ij}$ is the "composite" multipole interaction tensor between sites i and j, whose exact form can be found in Ref. 71. Accounting for the fixed charges through a higher order moment distribution(dipoles, quadrupoles) leads to better accuracy at short distances ($< 5\mathring{A}$).

The polarization effect in AMOEBA is modeled by induced dipoles placed on each atomic site, whose magnitude is determined by the site-specific isotropic polarizability and the total external electric field exerted:

$$\boldsymbol{\mu}_i^{ind} = \alpha_i (\mathbf{E}_i + \mathbf{E}_i') \tag{2.4}$$

where $\mathbf{E}_i$ is the electric field owing to the permanent multipoles on other fragments, and $\mathbf{E}_i'$ is the field generated by the induced dipoles on all the other atomic sites:

$$\mathbf{E}_i = \sum_j \mathbf{T}_{ij}^d \mathbf{M}_j^{perm} \tag{2.5}$$

$$\mathbf{E}_i' = \sum_{j \neq i} \mathbf{T}_{ij}^{d-d} \boldsymbol{\mu}_j^{ind} \tag{2.6}$$

Since the RHS of Eq. 6 relies on the induced dipoles, $\boldsymbol{\mu}_i^{ind}$'s are solved self-consistently to capture many-body polarization effects. With converged $\{\boldsymbol{\mu}_i^{ind}\}$, the associated energy lowering (the contribution of induced electrostatics) is determined by

$$E_{ele}^{ind} = -\frac{1}{2}\sum_i \boldsymbol{\mu}_i^{ind} \cdot \mathbf{E}_i \tag{2.7}$$

One artifact of a distributed interactive induced electrostatics model is the so-called "polarization catastrophe" [72], i.e., the electric field generated by point multipoles can severely overpolarize at short range and even lead to divergence. To ensure the finite nature of intermolecular induction effect, a Thole-style damping scheme is employed by AMOEBA, which is equivalent to replacing a point multipole with a smeared charge distribution. The damping function forms for all multipoles are reported in Ref.41. In practice, the damping functions are built in the formation of multipole interaction tensors in Eq. 2.5 and 2.6. Atomic polarizabilities are obtained as derived by Thole [72] and used with a modified damping factor (0.39) from Thole's original (0.567) as outlined in Ref.51.

For the van-der-Waals interactions, AMOEBA adopts a pairwise additive buffered 14-7 potential as formulated by Halgren [73].

$$E_{vdW} = \sum_{i<j} \epsilon_{ij} \left(\frac{1+\delta}{\rho_{ij}+\delta}\right)^7 \left(\frac{1+\gamma}{\rho_{ij}^7+\gamma} - 2\right) \tag{2.8}$$

where ϵ_{ij} is the depth of the potential well, ρ_{ij} is the dimensionless distance between sites i and j: $\rho_{ij} = R_{ij}/R_{ij}^0$, where R_{ij}^0 is the equilibrium distance. γ and δ are two constants whose values are set to 0.12 and 0.07, respectively. The combination rules for heterogeneous atom pairs that determine ϵ_{ij} and R_{ij}^0 are:

$$R_{ij}^0 = \frac{(R_{ii}^0)^3 + (R_{jj}^0)^3}{(R_{ii}^0)^2 + (R_{jj}^0)^2}, \quad \epsilon_{ij} = \frac{4\epsilon_{ii}\epsilon_{jj}}{(\epsilon_{ii}^{1/2} + \epsilon_{jj}^{1/2})^2} \tag{2.9}$$

Parameterization of Zwitterionic Amino Acids

The parameters in the AMOEBA force field have been developed with applications to large proteins in mind, and thus no parameters exist for single amino acid side chains in their zwitterionic form[74]. Furthermore given the centrality of water to the solvation study and the need for accuracy, we opted to work with the new AMOEBA14 water model[53] which provides a robust description of bulk water properties, but which requires reparameterization to work with other solutes. Thus a new set of non-bonded parameters of the glycine and valine solutes were required for compatibility with the AMOEBA14 water model. The standard AMOEBA parameterization protocol [51] was followed, with the exception for deriving the multipoles, since the first step of performing geometry optimization in vacuum converts the zwitterions into neutral molecules as expected. Instead, 100 structures spaced 2 ps apart from a 200 ps AIMD trajectory served as input structures for the following parameterization calculations.

While most of the valence parameters in Eq. 2.1 are defined and thus transferable from the existing AMOEBA13 parameter set, the parameters for the $N - C_\alpha - C$ bond angle

parameter specific to a zwitterion do not exist, and the vdW parameters on the carboxyl oxygens and amino hydrogens required optimization to account for modified interactions with the AMOEBA14 water model. The ForceBalance software[75] was used to derive the van der Waals parameters and bond angle force constant using static quantum chemical *ab initio* calculations as reference data. We generated the quantum mechanical energy and force calculation fitting data based on the MP2 method together with the 6-311G(1d,1p) basis, to maintain consistency with the charge and multipole parameterization protocol below, using *Q-chem*[76]. All of the newly derived zwitterionic non-electrostatic parameters (vdW and bond angle) parameters obtained for glycine were transferred to valine without re-optimization, demonstrating transferability.

Using the five lowest energy of the 100 available structures, we obtained the permanent atomic multipoles from the distributed multipole analysis *via* Stone's *DMA* program[77] based on single point MP2/6-311G(1d,1p) calculations using *Gaussian*[78]. The *TINKER poledit* utility was used to rotate the atomic multipoles obtained from *DMA* to *TINKER* defined local frames. This also defines Thole intramolecular polarization, for which polarization groups are defined based on Ref. 74: methyl, carbonyl and amine groups. This gives us an initial estimate of the multipole values. These values are further refined by performing a single point MP2/aug-cc-pVTZ calculation in *Gaussian*[78], which is used to derive the electron density and subsequently construct the electrostatic potential on a grid of points outside the vdW envelope using *Cubegen*[78]. The *TINKER potential* program then refines the atomic multipoles based on the quantum mechanical electrostatic potential. The DMA monopole values are not modified from the initial values in the refinement step.

Validation: AMOEBA *versus* AIMD

Figure 2.2 shows a comparison of the glycine-water radial distribution functions (RDFs) computed using data obtained from simulations from the original and modified AMOEBA parameters, and compared with the same RDFs from the AIMD calculations and from experiment [79]. It is observed that the first solvation shells of the two charged groups of the zwitterion as obtained from the AIMD simulations agree convincingly with the experimental data within rather small differences in peak positions, whereas the second shells feature increased deviations. The original AMOEBA model, in stark contrast, does not reliably capture even the first solvation shell structure, both in terms of peak positions and peak heights, and thus does not accurately represent the hydrogen bonding pattern of these important hydrophilic functional groups (note, however, that AMOEBA was never parameterized to study individual zwitterionic amino acids in water!). After reparameterization without any reference or fitting to our AIMD data, the RDFs of the resulting modified AMOEBA model shift much closer to the AIMD results and thus to experiment. Similar agreement between AIMD and AMOEBA is found for valine as shown in the supplementary information (SI). Further structural comparisons in terms of intramolecular angle and dihedral distributions are also shown in the SI, in which the modified AMOEBA and AIMD distributions match

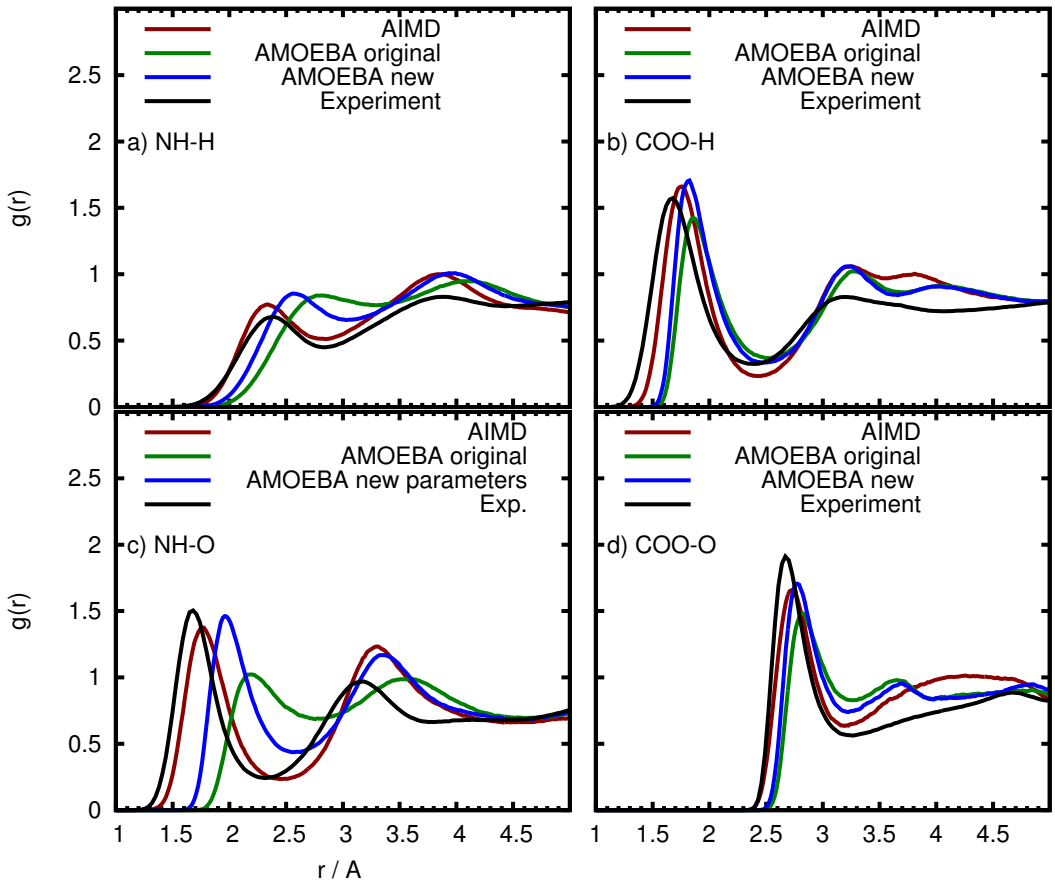


Figure 2.2: Radial distribution functions for a single zwitterionic glycine molecule in water: (a) amide-H and water-H, (b) carboxyl-O and water-H, (c) amide-H and water-O, (d) carboxyl-O and water-O sites according to the legend; see text for the corresponding methods and references.

similarly well. Hence, its is clear that the principal structural information of the amino acids is retrieved, and comparable to the results obtained from the AIMD calculations.

Within the AIMD approach to theoretical infrared spectroscopy[36], the corresponding linear absorption cross sections, $\alpha(\omega)$, are obtained from auto-correlation functions of the dipole moments that are obtained from concurrent electronic structure calculations. Thus, in addition to the solvation shell structure around the solute molecule, also its dipole moment in solution is of key importance. Figure 2.3 compares the dipole distributions of glycine and valine of the original and reparameterized AMOEBA force field to AIMD. As expected, the water dipole distribution shows good agreement when compared with the AIMD distribution whereas for the amino acids, the original parameters does not agree with the AIMD distribution well enough for the present purpose. However, after reparameterization the dipole

distributions align with the AIMD to a reasonable extent.

Simulation Protocols: AIMD and AMOEBA

For the AIMD calculations[31], the glycine and valine aqueous solutions were simulated using the PBE functional with pseudo potentials and a plane wave cut-off energy of 400 Ry and a TZV2P basis set using the *CP2k* program package [80]; see the SI of Ref. 20 for comprehensive computational details. A cubic simulation cell is used with a side length of 9.85 Å for glycine and 12.49 Å for valine. Each cell contains one amino acid molecule and 30 water molecules in the glycine case and 60 water molecules in the valine case. After equilibration long AIMD simulations have been carried out in the NVT ensemble using Nosé-Hoover chain thermostats[31] at a rescaled temperature of 400 K to approximately counterbalance its systematic underestimation by about 20–30 % thus following the approach introduced some time back for pure water and aqueous solutions[20, 32, 33, 81]. This *ad hoc* method not only provides agreement of radial distribution functions and the diffusion coefficient of water, but also accounts for the proper THz intensities compared to experiment[32]. From the NVT trajectory, 80 independent starting structures (and corresponding velocities) for glycine and 60 for valine are sampled at equidistant points as starting structures for NVE trajectories. Each NVE trajectory is simulated for 20 ps with an integration time step of 0.5 fs, and the maximally localized Wannier functions (MLWFs)[31] are computed every 2 fs.

For the AMOEBA force field simulations, glycine is solvated with 30 water molecules and valine with 60 water molecules, resulting in a cubic boxes of side lengths of 10.20 Å for glycine and 13.01 Å for valine. After initial equilibration in the NPT ensemble, we generated a 200 ps long NVT trajectory at 300 K using a Bussi thermostat, and sampled conformations and velocities every 2 ps. These served as starting structures and velocities for 100 independent NVE trajectories that were also simulated for 20 ps. The time step of integration is 0.5 fs and the configurations and induced dipoles are written out every 2 fs. Due to much faster computations using the AMOEBA model, the systems can be easily increased in the number of water molecules in order to probe finite size effects. Towards that goal, larger systems of both glycine and valine with 253 and 256 water molecules, respectively, were also studied. These systems had box lengths of 19.73 Å for glycine and 20.17 Å for valine. All the AMOEBA simulations were performed in the *TINKER* molecular dynamics package.

2.3 Theoretical THz Spectroscopy

Computing THz Spectra

In the limit of classical nuclear motion [82], the total linear infrared absorption cross section is given by [36]

$$\alpha(\omega) = \frac{1}{n(\omega)} \frac{1}{6\epsilon_0 V c} \frac{1}{k_{\mathrm{B}} T} I(\omega) \ , \tag{2.10}$$

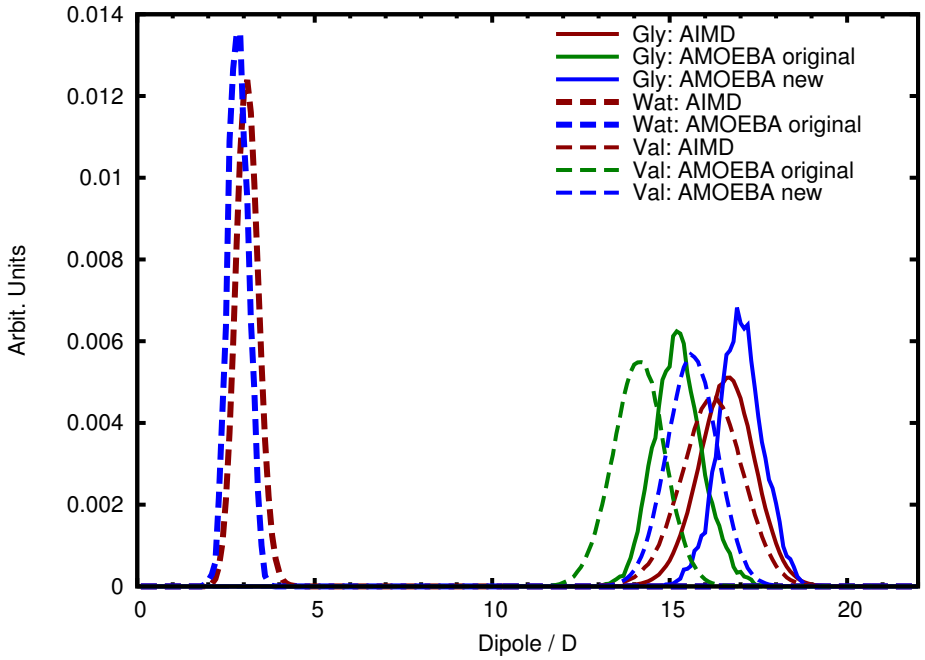


Figure 2.3: Dipole moment distribution functions for a single zwitterionic glycine and valine molecule in water split into solute and solvent contributions according to the legend; see text for the corresponding methods.

where V is the volume of the solution that is simulated at temperature T, ω is the frequency, $n(\omega)$ is the refractive index, and ϵ_0 and c are the vacuum permittivity and speed of light, respectively. It is noted in passing that the prefactor in that expression includes the frequency dependence of what is sometimes called the "harmonic quantum correction factor" [82] if using $I(\omega)$ expressed in terms of charge current time-correlations as specified next (thus taking into account the resulting extra ω^{-2} factor in front of $I(\omega)$ in Eq. 2.10). Here, the lineshape function $I(\omega)$ is given *via* the Fourier transform of the dipole velocity auto-correlation as follows

$$I(\omega) = \int_{-\infty}^{+\infty} dt \, \langle \dot{\mathbf{M}}(0) \, \dot{\mathbf{M}}(t) \rangle e^{-i\omega t} \quad , \tag{2.11}$$

where the total dipole moment $\mathbf{M}(t)$ can be defined as the vector sum of the (effective) molecular dipole moments $\boldsymbol{\mu}_J(t)$ in solution,

$$\mathbf{M}(t) = \sum_{J=1}^{N_M} \boldsymbol{\mu}_J \quad , \tag{2.12}$$

where N_M is the total number of molecules in the entire system. The time-derivative of the total dipole moment vector (being the total charge current) of the simulation box, $\dot{\mathbf{M}}(t)$, is computed as a finite difference quantity from consecutive configuration frames. Restricted

summation within Eq. 2.12 allows one to compute spectral contributions stemming from specified subsystems, for instance "solute-only" spectra if only the (effective) dipole moment (velocity) for instance of glycine, $\boldsymbol{\mu}_{J=\mathrm{Gly}}$, is considered. Since $I(\omega)$ is easily accessible to molecular dynamics trajectories, it is the product of the absorption cross section and the refractive index, $\alpha(\omega)n(\omega)$, that is straightforwardly obtained and thus mostly reported in the literature. However, $n(\omega)$ can be computed by applying the Kramers-Kronig relation as described in the SI of Ref. 20 so that the absorption coefficient $\alpha(\omega)$ itself is obtained which is indeed the experimental observable.

Computation of the molecular dipole moments $\boldsymbol{\mu}_J$ is fundamentally different between AIMD and AMOEBA simulations. For AIMD the molecular dipole moment is simply the sum of the product of charges and their cartesian positions of all charge centers $i = 1, \ldots, N_J$ in the molecule J that is considered[31],

$$\boldsymbol{\mu}_{J,\mathrm{aimd}} = \sum_{i=1}^{N_J} q_i \mathbf{r}_i \tag{2.13}$$

where i labels the charge centers irrespective of their nature. In AIMD, each nucleus is a charge center that contributes its positive nuclear core charge $q_i = +Z^{\mathrm{core}}|e|$ (thus taking into account the reduction of the bare nuclear charge Z whenever pseudo potentials are used to replace core electrons), whereas each Wannier charge center position carries a negative charge of $q_i = -2|e|$ in case of the mostly used doubly-occupied closed-shell representation of the valence electronic structure in terms of (maximally localized Wannier valence) molecular orbitals; note that this charge would be $-|e|$ in open-shell spin-polarized calculations where singly-occupied spin orbitals are used.

AMOEBA, being a polarizable point multipole based force field, uses both permanent and induced dipoles centered at each atom I, in addition to monopoles, all of which contribute to the dipole moment of molecule J,

$$\boldsymbol{\mu}_{J,\mathrm{amoeba}} = \sum_{I=1}^{N_J} \left(\boldsymbol{\mu}_I^{\mathrm{perm}} + \boldsymbol{\mu}_I^{\mathrm{ind}}\right) \quad , \tag{2.14}$$

where $\boldsymbol{\mu}_I^{\mathrm{perm}}$ is the contribution to the dipole from the permanent electrostatics; thus no additional pseudo interaction sites carrying only charges and/or multipoles are introduced in AMOEBA. In this fashion, the total dipole moment for molecule J is calculated by summing over all multipolar contributions of all atoms $I = 1, \ldots, N_J$ in that molecule. The effective molecular and thus also the total dipole moments obtained this way from AIMD and AMOEBA for solvent and solute species are comparable as already demonstrated in the validation section (cf. Fig. 2.3).

Decomposing THz Spectra

In order to understand the signals in the total THz spectra $\alpha(\omega)$ at the molecular level, decomposition of the THz observable in terms of atomic motions is necessary. Our mode

decomposition as developed in Refs. 36, 83, 84, and 85 to analyze infrared spectra of floppy molecules in the gas phase and extended to dissect THz spectra of aqueous solutions in Ref. 20 leads to mode specific lineshape functions and thus absorption cross sections $\alpha_k(\omega)$ that allow for an understanding of the spectrum in terms of explicit molecular displacements. In the following, we provide only a concise exposition of the key ideas and refer the interested reader to a review [36] and to the SI of Ref. 20 for comprehensive theoretical background including the treatment of nuclear quantum effects in the realm of theoretical infrared spectroscopy and technical details of the present approach, respectively.

Our particular computational approach [20] to decompose total infrared spectra $\alpha(\omega)$ into dynamical modes k and associated lineshape functions $\alpha_k(\omega)$ has been formulated specifically for analyzing AIMD trajectories including the electronic structure based on the charge current cross-correlation matrix that involves all charge centers i in the system,

$$C_{\zeta,\xi}(\omega) = \int_{-\infty}^{+\infty} dt \, \langle \, \dot{\mu}_{i,\zeta}(0) \, \cdot \, \dot{\mu}_{j,\xi}(t) \, \rangle e^{-i\omega t} \tag{2.15}$$

$$= \int_{-\infty}^{+\infty} dt \, \langle \, q_i v_{i,\zeta}(0) \, \cdot \, q_j v_{j,\xi}(t) \, \rangle e^{-i\omega t} \ , \tag{2.16}$$

as expressed in terms of charge-weighted velocities, where charges q_i weight the velocity of charge center i at position $\mathbf{r}_i(t)$; here cartesian velocity components ζ and ξ are used that have been rotated into a molecular frame of reference. Summation over all charge centers in the system thus not only includes the contributions due to the motion of the (positive) nuclei, but also the electron dynamics as represented by the motion of the Wannier centers being the (negative) electronic charge centers that are obtained from the maximally localized Wannier valence molecular orbitals. This charge-weighting obviously provides the required dipolar (cross-) correlations in terms of the time-derivative of the dipole moment vectors of all charge centers, $\{q_i \dot{\boldsymbol{r}}_i(t)\}$, thus including also the full electronic contribution to the charge current. Importantly, the corresponding sum over all nuclear and electronic charge centers within a specific molecule J at time t leads to its total charge current $\dot{\boldsymbol{\mu}}_{J,\mathrm{aimd}}(t)$ corresponding to Eq. 2.13, which is finally required according to Eq. 2.12 in order to compute the total infrared spectrum $\alpha(\omega)$ from Eq. 2.10 *via* Eq. 2.11 where the resulting total charge current $\dot{\mathbf{M}}(t)$ enters.

We note in passing that a decomposition approach which neglects the explicit electronic contributions alltogether and cross-correlates the $\sqrt{\mathrm{mass}}$–weighted atom velocities instead of the dipole velocities introduces a generalization of the vibrational density of states (VDOS) [32] and thus provides access to its decomposition in terms of modes (something that is accessible experimentally via inelastic neutron scattering). This procedure obviously does not provide infrared intensities, and thus no access to THz spectra [33], yet the same mode decomposition as performed here for the dipole correlations yields very similar mode displacement patterns of the atoms in real space as explicitly demonstrated for the present example in the SI.

It is key to observe that the cross-correlation matrix as defined *via* Eq. 2.15 not only includes the particle velocities with the associated core charges (and thus the contribution of

the molecular skeleton like in non-polarizable force field simulations), but in particular also the velocities of the Wannier centers which represent the electron dynamics in AIMD simulations within the Born-Oppenheimer approximation [31]. By taking into account the Wannier orbital dynamics in that sense, the purely electronic contributions to infrared absorption spectra, such as polarization and charge transfer effects, are included in the computation of $\alpha(\omega)$ based on AIMD trajectories.

At this stage, the mode-specific absorption cross sections (or mode spectra) $\alpha_k(\omega)$ are obtained after diagonalization of $\mathbf{C}(\omega)$, where the off-diagonal rest term α_{cross} is a measure of the remaining cross-correlations. The dipole displacement vectors corresponding to the kth mode can be determined from the transformation matrix that approximately diagonalizes the cross-correlation matrix (as explained in the SI of Ref. 20), which is close in spirit to the atomic displacement vectors that are obtained in traditional normal mode analysis. Finally, the total absorption cross section,

$$\alpha(\omega) = \sum_k \alpha_k(\omega) + \alpha_{\text{cross}}(\omega) \ , \tag{2.17}$$

can be recovered by summing over all decoupled modes k after adding the remaining cross terms. Most importantly, this provides one with a systematic tool to probe, mode by mode, how the lineshape of the total THz spectrum is generated by considering selected subsets of modes.

On the other hand, AMOEBA uses a completely different approach for calculating the molecular dipoles according to Eq. 2.14. As a result, the aforementioned computational approach and in particular Eq. 2.15 cannot be applied directly to the AMOEBA data, since the polarization contributions to the modes is located at the atom centers and thus are coupled directly into the molecular motion itself, whereas they are represented explicitly by the Wannier center dynamics in AIMD. Thus, in order to decouple the polarization modes (arising mostly from solute-water interactions) from the intramolecular atomic displacements solely for the purpose of spectra calculations, we need to introduce charged pseudo-sites in order to capture the polarization contributions separately as explained in the following.

For a water molecule J in aqueous solution, we can do this in a straightforward way by computing one effective charge center or pseudo-site, $\tilde{\mathbf{r}}_J$, which exclusively carries a charge $\tilde{q}_J$ in order to approximately capture the polarization contributions *via*

$$\tilde{q}_J \tilde{\mathbf{r}}_J(t) = \boldsymbol{\mu}_{J,\text{amoeba}}(t) - \sum_{I=1}^{N_J} q_I \mathbf{r}_I(t) \ , \tag{2.18}$$

where N_J is the number of atoms in molecule J, q_I is the charge of atom I at position $\mathbf{r}_I(t)$ at time t and $\boldsymbol{\mu}_{J,\text{amoeba}}(t)$ is the total dipole moment of water molecule J in solution as given by the full AMOEBA force field including the instantaneous polarization effects at time t according to Eq. 2.14. In order to conform as closely as possible with the AIMD approach to spectral decomposition, the atom charges q_I have been chosen to be identical to the nuclear core charges which underly the pseudo potential representation of the electronic structure in

AIMD (i.e. $q_{\mathrm{O}} = +6|e|$ and $q_{\mathrm{H}} = +1|e|$ in case of oxygen and hydrogen atoms). Thus, the effective charge $\tilde{q}_J$ attached to the pseudo-site at position $\tilde{\mathbf{r}}_J$ is identical to the full valence charge of the molecule as given by the sum of all its Wannier charges, which is $-8|e|$ in case of a neutral water molecule.

Following this procedure, the position of the pseudo charge site, $\tilde{\mathbf{r}}_J(t)$, can be uniquely computed from Eq. 2.18 once the total dipole moment of water molecule J at time t is determined in the AMOEBA simulation of the solution for the corresponding molecular configuration $\{\mathbf{r}_I(t)\}$ of that (polarized) water molecule. Based on this approach, the total dipole moment of water molecule J in solution is represented by the following expression in AMOEBA,

$$\boldsymbol{\mu}_{J,\mathrm{amoeba}}(t) = \sum_{I=1}^{N_J} q_I \mathbf{r}_I(t) + \tilde{q}_J \tilde{\mathbf{r}}_J(t) \ , \tag{2.19}$$

which is solely used in that form when evaluating the dipolar cross-correlations according to Eq. 2.15 with the aim to dissect the total infrared spectrum of the solution, $\alpha(\omega)$, in terms of modes k and the corresponding mode-specific absorption cross sections $\alpha_k(\omega)$. Thus, a single water molecule in AIMD is composed of three nuclear and four electronic (Wannier) charge centers, the latter representing the eight paired valence electrons, whereas it is approximated in AMOEBA by the same three nuclear charge centers at the atom positions together with one effective electronic pseudo charge center that carries the full valence charge.

Since a single pseudo charge center would be a rather crude approximation for molecules much larger than water, we adopted a fully additive "divide and conquer" approach by introducing one such pseudo-site $\tilde{q}_J$ for each functional group. In case of the two amino acids, the following such fragments have been defined: the protonated amino NH_3^+ and deprotonated carboxyl COO^- groups, the side chain groups for glycine (H) and valine ($CH(CH_3)_2$), as well as the $C_\alpha H_\alpha$ group. Each effective center is computed exactly in the way described above for a single water molecule while only taking into account all those atoms (including the hydrogens) that belong to the respective functional group. Upon representing a molecule additively by a sum of atoms requires to cut covalent bonds that connect these functional groups, which is done here in the crudest way by dividing up the respective electron pairs democratically between the two functional groups. This results also in the correct net charge in case of charged groups, for instance NH_3^+ consists of one N and three H atom sites (thus providing a total nuclear charge of $+8|e|$ in view of the nuclear core charges $q_{\mathrm{N}} = +5|e|$ and $q_{\mathrm{H}} = +1|e|$) and seven electrons (three electron pairs $-2|e|$ from the N–H bonds and one electron $-1|e|$ from the cut C–N bond) and thus a pseudo charge of $\tilde{q}_{NH_3^+} = -7|e|$, which leads to a net charge of $+1|e|$ as required. At this stage, the same spectral analysis machinery as developed for AIMD can be carried over to analyze the AMOEBA trajectories, which also carries over to much larger systems such a peptides by virtue of the additive fragmentation approach in terms of well-defined functional groups.

Finally, the THz modes, in particular those that couple solute and hydrogen-bonded solvent molecules, are obtained by employing the supermolecular solvation complex (SSC)

approach as introduced in Ref. 20. The SSC is composed of the solute molecule, in this case glycine or valine, and either three water molecules at the amino group (which is denoted as SSC(+)) or one water molecule at the carboxylate group (SSC(-)) as assessed in the SI of Ref. 20. Employing the SSC(+/-) analysis enables us to compute modes that take into account the coupled motions of the solute with the solvation water molecules at the hydrophilic sites.

2.4 Results

THz Spectra: AMOEBA *versus* AIMD

Overall the new set of AMOEBA parameters provide properties of aqueous solutions of amino acids that are in reasonable agreement with AIMD, and undoubtedly are an improvement over the original set of parameters. Therefore we can now proceed to compare the theoretical THz spectra resulting from AIMD and AMOEBA simulations for glycine and valine, and the cross-correlation spectra of each amino acid with their respective water environment. We will then compare the mode-specific absorptions computed *via* the mode decomposition scheme described above. Since the zwitterionic amino acid modes are found to be very similar in most cases, glycine will be discussed in detail and any differences with valine will be shown when applicable.

The total spectrum of glycine computed from the AMOEBA simulations shows qualitative agreement with the one obtained from AIMD simulations as seen in Fig. 2.4. An absorption peak is seen at 80 $cm^{-1}$where the low frequency rigid body motions are located, as well as the most characteristic absorption, the NCCO open/close mode at 300 cm^{-1}, which is much sharper due to the more harmonic nature of the corresponding intramolecular motion according to the AMOEBA force field. The intermolecular absorption signal, originating from the interaction with the solvating water molecules due to hydrogen-bonded stretching is present at 200 cm^{-1}as shown in Fig. 2.5. It originates from polarization effects since fixed charge force fields do not exhibit this cross-correlation signal. From this comparison, however, it is clear that not all of the cross-correlation is necessarily captured in view of some missing intensity. This could arise due to the need to conform to using charged-weighted velocities to represent the dipoles *via* Eq. 2.15 and the sum rule that recovers the total spectrum from Eq. 2.17. This computational approach has been shown to work well for AIMD but is a less natural definition for the AMOEBA force field where is requires the introduction of charged pseudo-sites *via* Eq. 2.18 to approximately capture polarization contributions. Another possibility is that the solute-solvent mode is comprised of more than just pure polarization, such that the remaining missing intensity might be attributed to lack of charge transfer in the AMOEBA model while it is captured by AIMD, since simulated infrared spectra are known to be sensitive to this molecular interaction. Valine shows very similar behavior to that observed for glycine (see SI material). Nonetheless, while the AMOEBA intensities are smaller and frequencies are slightly shifted, the principal lineshapes follow the trends observed in AIMD.

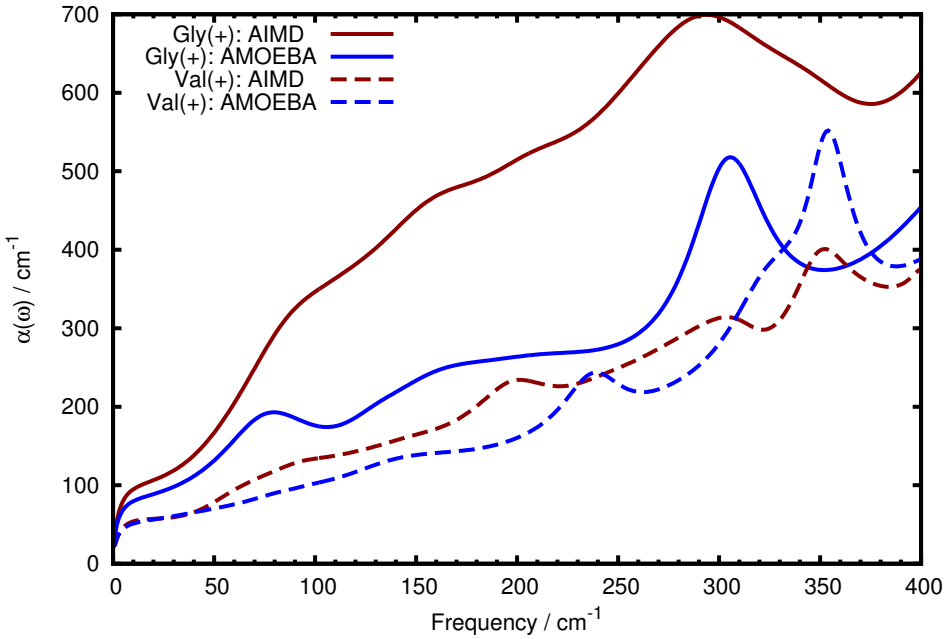


Figure 2.4: Total THz spectrum $\alpha(\omega)$ of glycine and valine based on AMOEBA (blue) and AIMD (red) simulations obtained within the SSC(+) approach. The three signals can be assigned to intramolecular bending modes (at ∼300 cm^{-1}), rigid body motions (∼80 cm^{-1}) and intermolecular solute-solvent coupling modes (∼200 cm^{-1}).

THz Modes and Spectral Decomposition

From our previous AIMD analysis[20] we have found intramolecular motions of the amino acid itself (e.g. opening and closing of NCCO, twist around the CC bond), quasi rigid body motions that describe the hindered translations of the molecule within the water environment (rattling) as well as hindered water rotations (librations) and water stretching and bending motions that describe intermolecular interactions of water with the amino acid directly. In Fig. 2.6 representative examples of the glycine modes are visualized in terms of the displacement vectors, with a similar set found for valine which also includes additional rotomeric motions of the aliphatic side chain. Therefore, in order to further compare the AMOEBA and AIMD calculations, we decompose the AMOEBA spectrum by assigning each band a molecular displacement using the SSC(+/-) approach.

Comparing the resulting modes obtained from the AMOEBA simulation and from AIMD, we see very good agreement for the glycine modes shown in Fig. 2.7 and for the valine modes in Fig. 2.8. From visual inspection it is evident that the intramolecular modes at the high frequency end of the THz spectrum, Fig. 2.6(a-c), are very similar and show basically identical displacement patterns between AMOEBA and AIMD. Glycine and valine both show the characteristic NCCO open/close mode above 300 cm^{-1}(305 $cm^{-1}$for glycine and

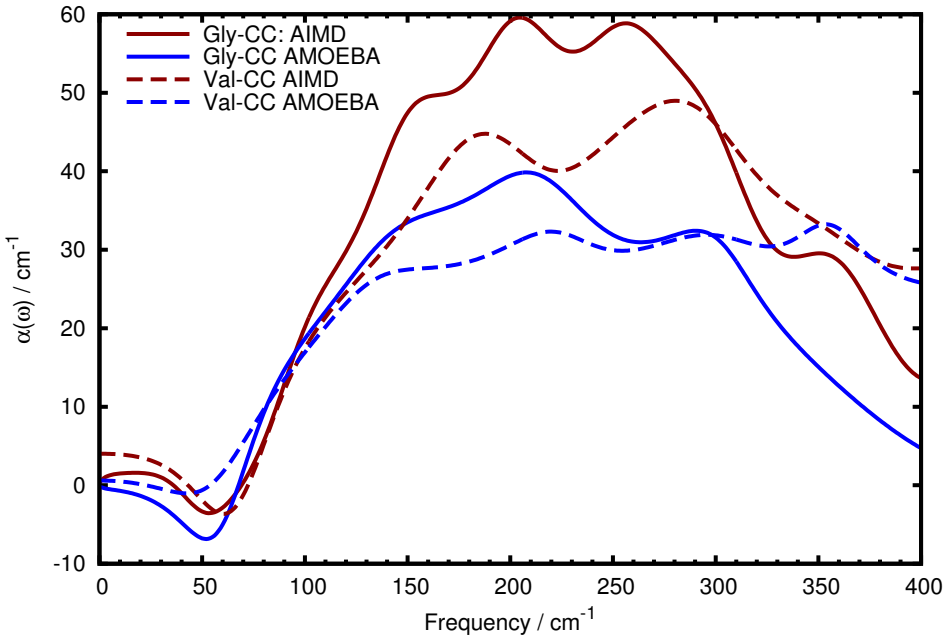


Figure 2.5: Cross-correlation spectrum of glycine with surrounding water molecules based on AMOEBA (blue) and AIMD (red) simulations.

352 $cm^{-1}$for valine); while this mode is broad in the AIMD case, it is a sharper mode in glycine due to the harmonic nature of the corresponding force field term in AMOEBA. Valine shows additional intramolecular modes (at 335 $cm^{-1}$and 321 cm^{-1}) involving the side chain rotomers, although they are slightly red-shifted compared to AIMD (317 $cm^{-1}$and 281 cm^{-1}, respectively).

The AIMD study revealed that the CC-twisting mode shows a strong coupling to the water hydrogen-bonded network in the first solvation shell, as shown by the coupled motion of the twisting atoms together with the hydrogen-bond stretching of the water molecules. Furthermore, this mode dominates the 200 cm^{-1}signal that is associated with the solute-solvent coupling. Both of these key observations are also true for the same mode obtained from the AMOEBA simulation. In direct comparison to the AIMD mode (Fig. 2.7) the intensity of the mode is reduced by roughly half, while the displacement is very similar (Fig. 2.6b). This is in agreement with the overall lower cross-correlation signal of glycine with water (Fig. 2.5) with reduced intensity that could stem from the pseudo charges that are adopted from the AIMD result. The modes derived by AMOEBA for valine show very similar trends, but additional intramolecular modes are obtained due to the side chain, which are in good accord between the AIMD results and the AMOEBA force field as well.

The rattling modes due to hindered translations as observed in the AIMD simulation show a concerted uni-directional motion of all amino acid atoms and water molecules (within the SSC(+/-) approaches), whereas the AMOEBA modes show a more disorganized motion for

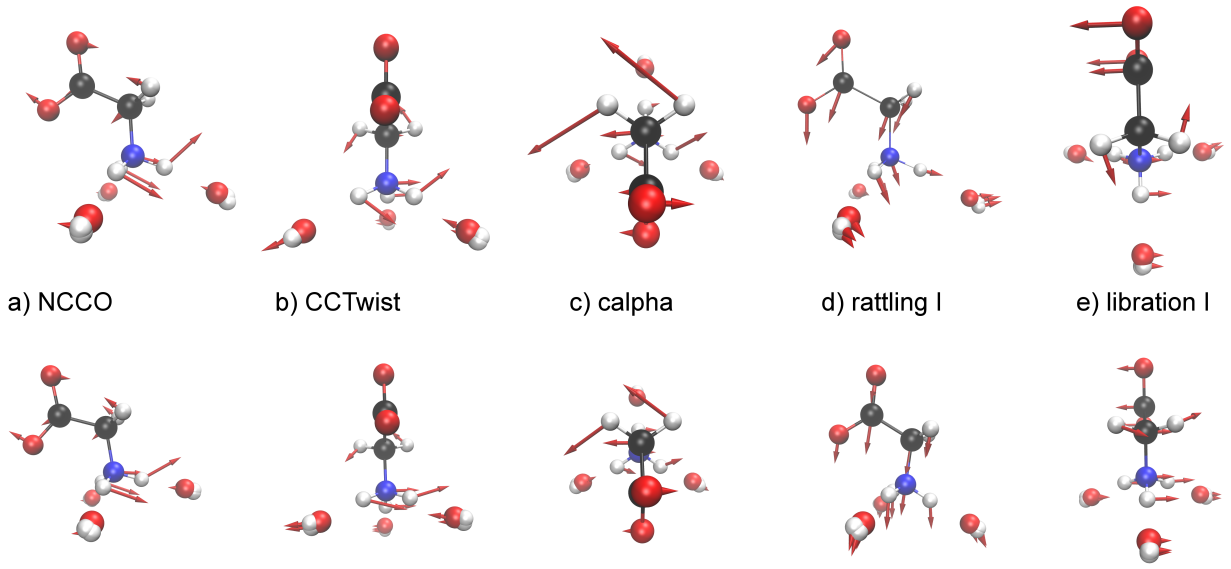


Figure 2.6: Glycine mode displacement vectors for AMOEBA(above) and AIMD(below) obtained with the SSC(+) approach: NCCO open/close, C-C twist coupled to hydrogen-bond stretch, C_α out-of-plane, quasi rigid body cage rattling I, and quasi rigid body cage libration I. The corresponding mode-specific THz spectra are shown in Fig. 2.7.

both glycine and valine. It is a systematic problem that could arise from either the method of introducing the additional centers to separate out the polarization modes, the charges taken from AIMD to define the pseudo charge center, or the result of the lack of charge transfer in the AMOEBA force field itself that manifests mostly in the solute-solvent interactions.

Assessing Finite Size Effects

Simulations with the polarizable AMOEBA model allow for much larger system sizes than the AIMD simulations. Since the spectrum and the modes obtained from AMOEBA agree well in general to AIMD data, increased system sizes were investigated with AMOEBA in order to probe finite size effects. Tables 2.1 and 2.2 show the peak frequencies of the modes obtained from the small and large simulation boxes. This analysis is based on the SSC(+) approach thus including the first solvation shell of the protonated amino group. Only minor

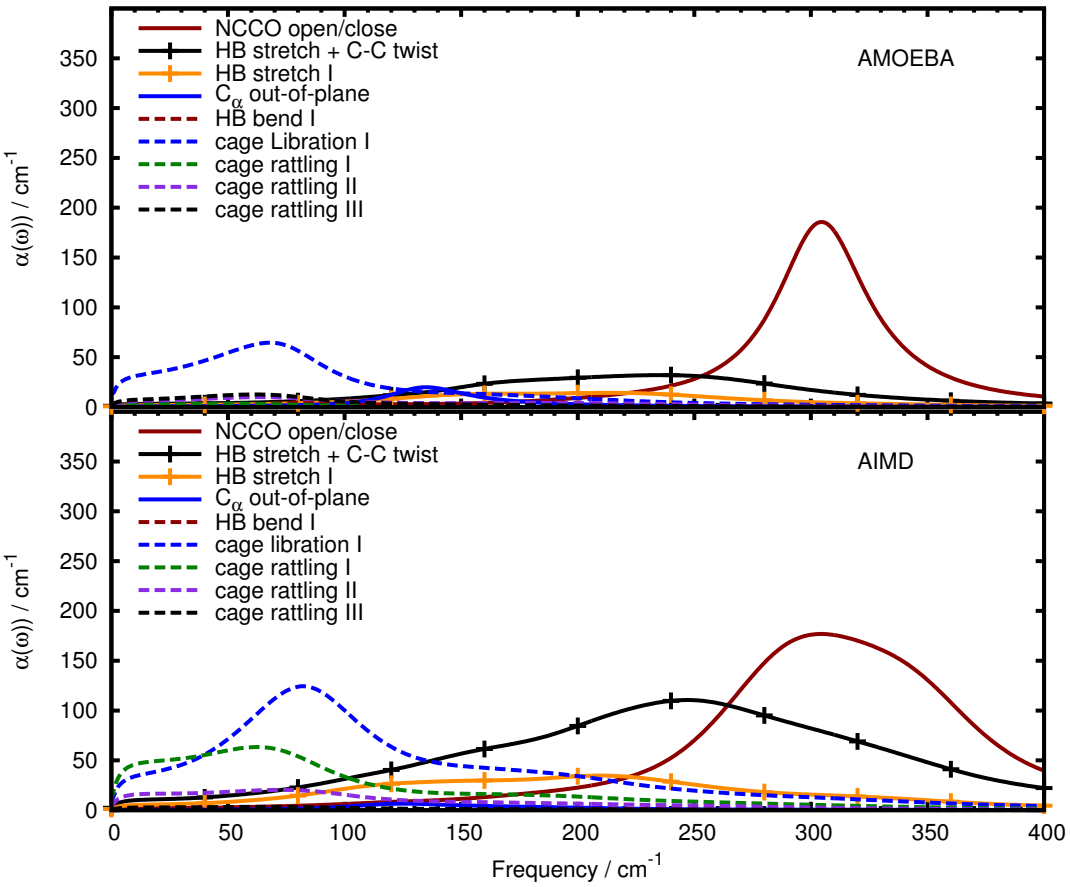


Figure 2.7: Mode-specific THz absorption spectra $\alpha_k(\omega)$ of glycine based on AIMD and AMOEBA simulation data obtained within the SSC(+) approach. Only the THz modes with intensity greater than one wavenumber are shown.

frequency shifts are observed which are attributed to the broad overall peak shape since only the maximum is reported. After inspection of the mode displacement assignments it is clear that all modes of the small and large simulation boxes agree with each other, see SI for visual inspection. We conclude that although the simulation box sizes used in AIMD appear to be small, the THz spectra and their interpretation in terms of the mode displacement vectors according to the SSC(+/-) approaches do not suffer from finite size effects at the required level of accuracy.

2.5 Discussion and Conclusions

Overall we conclude that our reparameterized AMOEBA polarizable model provides qualitative, and in some instances quantitative agreement with the AIMD reference results for

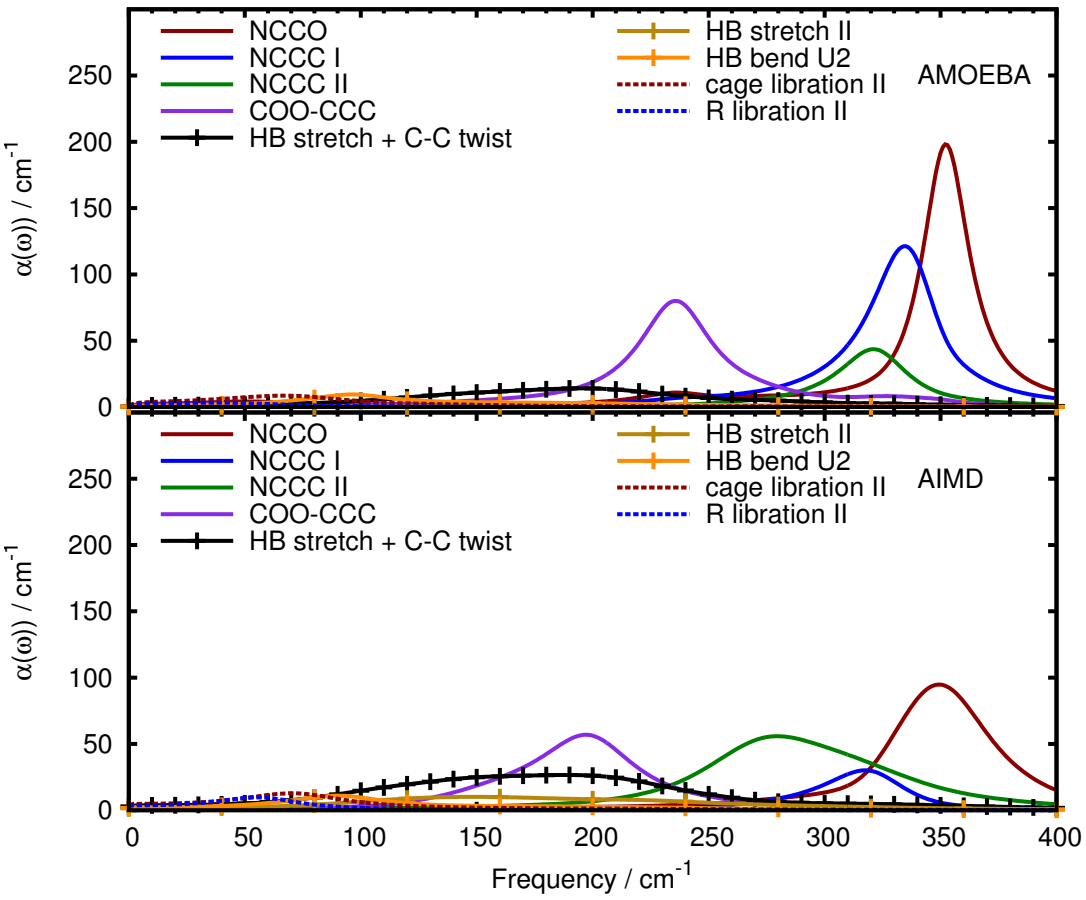


Figure 2.8: Mode-specific THz absorption spectra $\alpha_k(\omega)$ of valine based on AIMD and AMOEBA simulation data obtained within the SSC(+) approach. Only the THz modes with intensity greater than one wavenumber are shown

the THz spectra that report on the solvation dynamics of small zwitterionic amino acids in aqueous solution. Still, this polarizable force field leads to overstructured total THz absorption spectra of glycine and valine in water as judged by comparing to both AIMD and experimental results. At the level of the computed spectra, the cross-correlations between solute and solvent molecules seem to be less pronounced than in the AIMD simulations. This effect is particularly evident in the intensity of those mode-specific THz spectra that are dominated by strong couplings of the solute to the water hydrogen-bond network such as in case of the C–C twisting modes of both glycine and valine that are located in the 200 cm^{-1} region. The hindered translational motion, giving rise to cage rattling modes in THz spectra, is less clearly pronounced according to the analysis of the polarizable force field data when gauged with the AIMD mode displacement patterns. Despite such caveats at a detailed level of assessment, the overall AMOEBA performance bodes well for future

Table 2.1: Peak positions of mode-specific absorption spectra of glycine in water obtained within the SSC(+) approach with significant contribution to the THz spectrum (see Fig. 2.7) depending on system size.

Mode	AMOEBA 30 Wat	AMOEBA 253 Wat	AIMD 30 Wat
NCCO open/close	305	306	304
C-C twist + HB stretch	236	237	247
HB stretch I	214	177	210
HB stretch II	146	149	218
C_α out-of-plane	135	137	125
HB bend I	92	96	102
cage libration III	84	87	89
cage libration II	71	74	90
HB bend II	68	73	90
cage libration I	68	66	82
cage rattling II	63	64	73
cage rattling III	61	29	64
cage rattling I	50	25	62

Table 2.2: Peak positions of mode-specific absorption spectra of valine in water obtained within the SSC(+) approach with significant contribution to the THz spectrum (see Fig. 2.8) depending on system size.

Mode	AMOEBA 60 Wat	AMOEBA 256 Wat	AIMD 60 Wat
NCCO open/close	352	353	349
NCCC I	335	335	317
NCCC II	321	320	280
COO-CCC	236	239	197
C-C twist + HB stretch	196	201	189
HB stretch I	131	140	112
HB bend I	86	95	101
R libration I	75	79	85
HB bend-X3	63	76	83
cage libration I	52	57	68
Cage rattling	51	52	56
R libration II	38	40	57

THz studies on larger systems such as enzymes in water or extended aqueous interfaces with a focus on qualitative insights and trends given that the computational cost is much lower in comparison to AIMD simulations.

To arrive at these encouraging results, we have presented a straightforward approach to THz spectral analysis using the AMOEBA model by engineering additional charged pseudo-sites in a manner that allows the decomposition of the spectrum into mode-specific absorption

coefficients as was done earlier in the AIMD simulations. Importantly, this scheme relies an an additive "divide an conquer" idea based on functional groups or molecular fragments that can be readily transferred to much more complex molecular systems such as proteins or lipid membranes. Our approach to the calculation of THz spectra from AMOEBA approximately includes electronic polarization effects which are known to play a key role in determining the correct intensity modulations as a function of frequency and thus the overall lineshape function. Nonetheless further improvement in the methodology for decomposing the THz spectra is warranted for the polarizable force field since the computational approach of decomposing the total absorption spectrum into mode-specific cross sections based on the full charge current auto-correlation function as devised for AIMD simulations (where direct access to localized molecular orbitals via the Wannier centers and thus the full charge current are readily available) is ill-suited for the AMOEBA model (where no such purely electronic information is straightforwardly accessible). Since AMOEBA uses a point multipole representation of the permanent electrostatics and polarization contributions that are atom centered, this introduces both distortions of the modes and/or higher off-diagonal terms in the spectral decomposition than observed in AIMD. One future modification of the approach would replace the weighting of the modes by formal charges from AIMD for one which fully takes advantage of the AMOEBA point monopole, dipole, and induced dipoles in the weighting scheme also at the level of analysis.

Another source of future investigation is to better understand the limitations introduced by the lack of charge transfer in the AMOEBA model. Since charge transfer is certainly present in the AIMD calculations, and contributes to the intensity of the cross-correlations between the water molecules and the amino acid solutes, it is expected to play a significant role in the intensities of the modes [86]. Therefore introducing charge transfer into the AMOEBA force field will further improve the capabilities of describing the THz spectrum in solutions, and its decomposition into assigned modes of the dynamics, and will be the subject of future efforts.

Finally, the AIMD results were computed using the PBE density functional, which has both well established strengths and weaknesses, but has been well-validated against experimental THz spectra of glycine and valine aqueous solutions and neutron diffraction data of the solvation shell structure of glycine in water. Nonetheless there has been an explosion of new density functionals that show demonstrable improvements in properties ranging from binding and isomerization energies, barrier heights, through to thermochemistry that should be considered in future validation studies. Unfortunately, reliable calculations of THz spectra are computationally much more demanding than computing radial distribution functions or alike since on the order of 100 independent microcanonical AIMD runs based on uncorrelated initial conditions sampled from a long canonical AIMD simulation are typically required in order to converge lineshapes at such low frequencies and thus their subtle modulations which encode the desired molecular information after mode-specific decompositions. Furthermore, while raising the temperature or using quantum thermostats can be an *ad hoc* way to roughly emulate missing nuclear quantum effects, quantum delocalization contributions are difficult to predict *a priori* since they can also give rise to strengthening

hydrogen-bonding that counteracts the effect of raising temperature.

Chapter 3

Solvent Contributions in Kemp Eliminases

Note: This chapter is part of a collaborative project. The analysis has been performed in collaboration with MSc. Viren Pattni and Prof. Matthias Heyden, Max Planck Institut für Kohlenforschung and has been reproduced here with permission.

3.1 Introduction

Enzymes are molecular catalysts that have been highly optimized by evolution, and can increase the rate of a targeted biochemical reaction, anywhere between ∼10-20 orders of magnitude over the uncatalyzed reaction in solvent (typically water). In addition to steric positioning effects by residues in the catalytic pocket of an enzyme, a large component of their catalytic power also arises from the pre-organization of stabilizing electric fields for the transition state due to the enzyme scaffold, that is absent in bulk water environments [2, 87]. Dynamics of the enzyme residues are also important factors in catalysis, and have been discussed in detail in Chapt. 4. However, in addition to the protein itself, the solvent dipoles and their fluctuations can in principle also contribute to the enzyme catalytic activity[15, 16]. This might arise especially from the integrated effect of a hydration layer which behaves differently from the bulk in terms of structure, dipole orientation, and self-diffusion, due to the varied chemistry and topography immediate to the enzyme surface[18], and the possibility that these solvent alterations persist from the surface for more than one solvation layer [17, 88]. Hence, the solvation of the enzyme, in particular, the active site, would play an important role in determining the catalytic activity of the enzyme.

In this work we analyze the spatial decomposition of solvent entropy contributions to enzyme catalysis for the KE07 and KE70 enzymes. The KE07 enzyme was designed by creating an active site which could carry out the Kemp Elimination reaction, and engineering that active site into the TIM barrel scaffold of a thermostable imidazole-3-glycerolphosphate synthase from Thermatoga Maritima[22]. The target reaction for this enzyme is the Kemp

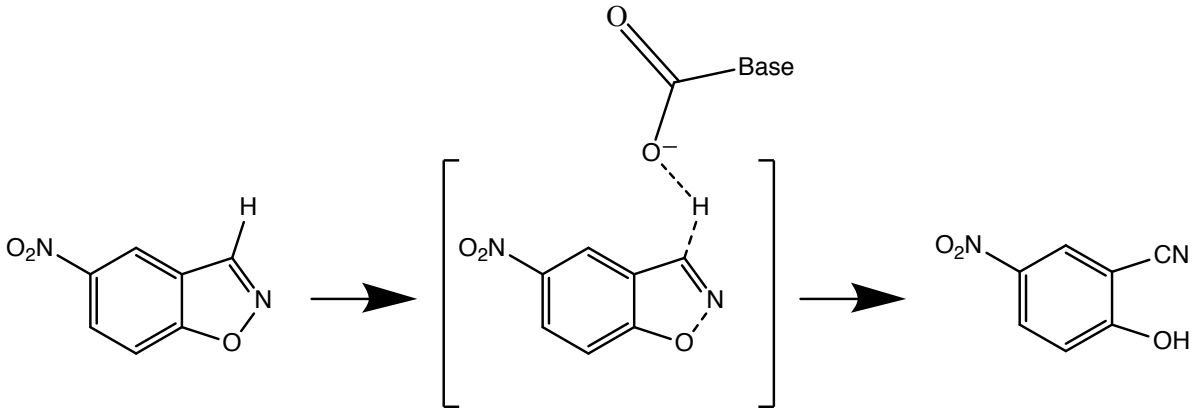
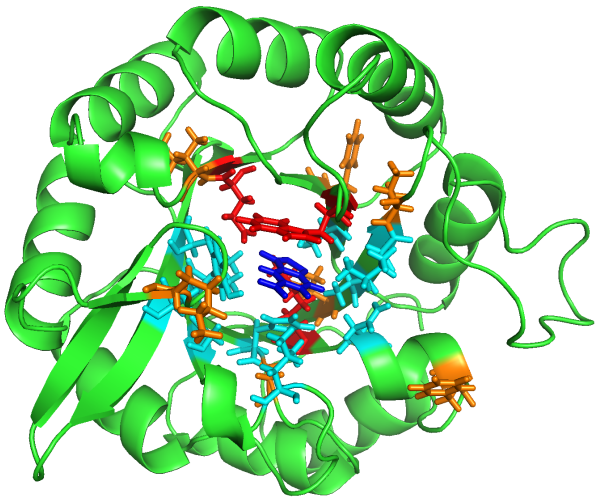
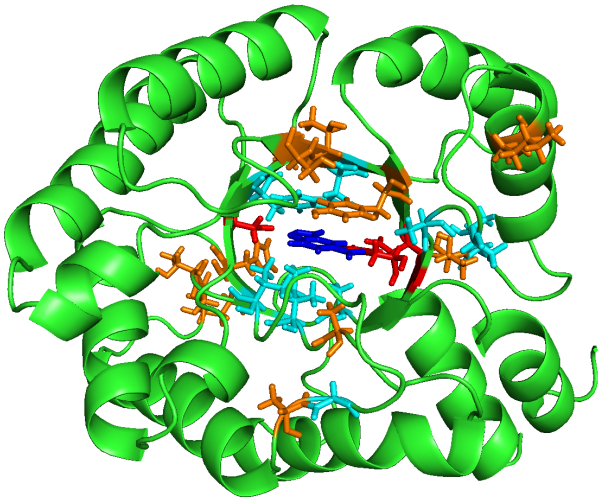


Figure 3.1: Kemp Elimination Reaction. This reaction involves the abstraction of a proton from the substrate, 5-nitrobenzisoxazole to give the cyanophenol product.



(a) KE07 Catalytic residues Trp50, Glu101, and Lys222. Lys222 is replaced as the catalytic residue as discussed in Sec. 3.3



(b) KE70 Catalytic residues His17, Asp45, Tyr48, and Ser138. Multiple designed residues, including the catalytic Ser138 were mutated from WT to R6-1.

Figure 3.2: Kemp Eliminase Design Structures and Catalytic Sites: Catalytic residues (Red), Mutated residues from WT to R7-2 (KE07) or R6-1 (KE70) (Orange), Designed residues (Cyan), Ligand (Blue)

Elimination reaction, which involves the extraction of a proton by a base, concerted with breaking a nitrogen-oxygen bond, leading to the formation of a cyanophenol product as in Fig 3.1. The active site of the enzyme is centrally located in the β-barrel as seen in Fig 3.2a.

KE70, on the other hand was created by engineering an active site into another TIM-barrel scaffold, i.e. the deoxyribose phosphate aldolase of *Escherichia Coli*[23]. The catalytic base here is a His-Asp dyad, i.e. a histidine polarized by an adjacent aspartic acid residue. His17-Asp48 form the dyad of residues involved in proton extraction, Tyr 48 acts as the π-stacking residue of the substrate and Ser138 acts as the H-bond donor in the stabilization of the negative charge being formed on the phenolic oxygen during the bond breaking.

We use a three-dimensional spatial resolution to analyze the change in solvent entropy in the neck of the barrel and on the surface of the protein relative to bulk solvent in order to see how that differs in the designed enzyme and the best laboratory directed evolution (LDE) for the two Kemp Eliminases. While both designed enzymes and their evolved variants show entropic signatures that differ from the bulk, we observe a very distinct solvation behavior of the two Kemp Eliminases, showing they utilize very distinct methodologies for carrying out their catalysis.

3.2 Methods

2PT Entropy Calculation

Thermodynamic properties of liquids cannot be accurately computed by harmonic or quasi-harmonic expressions. Hence, the 2PT(2 Phase Thermodynamic) approach provides an alternative approach where the liquid phase is approximated as a superposition of a solid and a gas, both phases for which simple models can be proposed to calculate thermodynamic quantities. This approach has been developed by Goddard and coworkers for a Lennard-Jones fluid[89], and further extended by them to water[90].

The 2PT approach partitions the VDOS(Ω) into a gas phase and a solid phase component.

$$\Omega(\omega) = \bar{\Omega}^g(\omega) + \bar{\Omega}^s(\omega) \tag{3.1}$$

where $\bar{\Omega}^s(\omega)$ is computed from a quantum harmonic oscillator model, which has zero density of states at zero frequency, and $\bar{\Omega}^g(\omega)$ is described in terms of a hard sphere model.

Any thermodynamic quantity can then be computed from a weighted sum of the two contributions, which we illustrate for the entropy S

$$S = \int_0^\infty d\omega \bar{\Omega}^s(\omega) W_S^s(\omega) + \int_0^\infty d\omega \bar{\Omega}^g(\omega) W_S^g \tag{3.2}$$

where $W_S^s(\omega)$ and W_S^g are the weights for the solid and gas phase for calculating entropy respectively.

As the solid component partition function uses the quantum harmonic oscillator(HO), its weighting function for entropy is given by:

$$W_S^s(\omega) \approx W_S^{HO}(\omega) = \frac{\beta h\omega}{exp(\beta h\omega) - 1} - \ln[1 - exp(-\beta h\omega)] \tag{3.3}$$

The gas component, on the other hand uses a hard sphere (HS) model, and the density of states distribution is given by:

$$\bar{\Omega}^g(\omega) \approx \bar{\Omega}^{HS}(\omega) = \frac{12N^g\alpha}{\alpha^2 + 4\pi^2\omega^2} \tag{3.4}$$

where $N^g = fN$ is the number of hard sphere particles in the system, with (f) defined as the fraction of modes in gas (i.e. fluid, as opposed to solid) phase, is a measure of the "fluidicity" of the system. α is the Enskog friction constant related to collisions between hard spheres. This can be computed by setting $\omega = 0$ in Eq. 3.4

$$\bar{\Omega}^{HS}(0) = \frac{12fN}{\alpha} \tag{3.5}$$

As the solid phase has no diffusivity, the solid phase VDOS has zero amplitude at zero frequency. Hence the zero frequency VDOS amplitude for the system comes only from the gas-like, i.e. hard-sphere component of the VDOS for the total system at zero frequency $\bar{\Omega}^{HS}(0) = \bar{\Omega}(0)$. Hence, we obtain an expression for α in terms of f and $\bar{S}(0)$. Substituting in Eq. 3.4, we get,

$$\bar{\Omega}^g(\omega) = \bar{\Omega}^{HS}(\omega) = \frac{\bar{S}(0)}{1 + \left[\frac{\pi\bar{S}(0)\omega}{6fN}\right]} \tag{3.6}$$

Thus, the gas phase is completely characterized by two parameters: the gas fraction or fluidicity f and density of states for the total system at zero frequency $\bar{\Omega}(0)$.

The fluidicity factor f is defined proportional to the diffusivity, to satisfy the requirements that at high temperature/low density $f \to 1$ and at high density $f \to 0$

$$f = \frac{D(T,\rho)}{D_0^{HS}(T,\rho;\sigma^{HS})} \tag{3.7}$$

where ρ is the density of the system and T is the temperature. $D(T,\rho)$ is the system diffusion coefficient determined using the Green-Kubo relation from the zero-frequency value of the system VDOS ($\bar{\Omega}(0)$) as derived in Ref: 89:

$$D(T,\rho) = \frac{\bar{S}(0)kT}{12mN} \tag{3.8}$$

where k is the Boltzmann constant, T is the temperature, and m and N are the mass and number of particles.

D_0^{HS} is the hard sphere diffusion coefficient determined in the zero pressure limit[91]

$$D_0^{HS}(T,\rho;\sigma^{HS}) = \frac{3}{8}\frac{1}{\rho(\sigma^{HS})^2}\left(\frac{kT}{\pi m}\right)^{\frac{1}{2}} \tag{3.9}$$

where σ^{HS} is the radius of the hard sphere particle which can be determined as outlined in Ref. 89 and ρ is the density of the system

The weight for the hard sphere component is given by:

$$W_S^g(\omega) = W_S^{HS}(\omega) = \frac{1}{3}\frac{\bar{S}^{HS}(\omega)}{k} \tag{3.10}$$

The detailed derivation of these quantities can be found in Refs. 89 and 90.

In this work, we estimate $\bar{\Omega}^s(\omega)$ from the total (liquid state) VDOS computed via simulation after subtracting off the hard sphere quantity $\bar{\Omega}^g(\omega)$. The translational VDOS is computed from the velocities by the Fourier transform of the mass-weighted velocity autocorrelation function,

$$\bar{\Omega}(\omega) \propto m\int_{\tau_{max}}^{0} dt\langle\vec{v}(0)\vec{v}(t)\rangle e^{(-i\omega t)} \tag{3.11}$$

where τ_{max} is the maximum time constant which we set to 1.6ps to capture the complete decay of the autocorrelation function of water motion with a frequency resolution of $20cm^{-1}$. A corresponding quantity is evaluated for the rotational VDOS, which is computed from the rotational velocities via a rigid rotor formulation. The nature of the equations that result for the rotational contribution to the entropy are analogous to those outlined below for the translational entropy, and are derived by Lin et al[90]. For brevity, only the derivations for translational entropy from Ref. 89 are briefly summarized here. Lin et al[90] have also shown that the vibrational contribution to the VDOS in a flexible water model is quite small($< 1\%$), and can hence be neglected.

The 2PT approach has been modified for spatial resolution (3D-2PT) by Heyden and coworkers[92, 93] to study solvent entropy around selected solutes. This approach decomposes the simulation box into a grid of voxels of specified dimensions, and all dimensions in this work are of uniform cubic volume, V of length 2.25 $\mathring{A}$ on a side. The entropy values are calculated for each voxel and are a weighted average, weighted by the number densities over the trajectories, to obtain a spatial distribution of the entropies being calculated. The number densities themselves are linearly averaged.

$$\langle S_i(\vec{r})\rangle = \frac{1}{\sum_j \langle\rho_i^j(\vec{r})\rangle}\sum_j \langle\rho_i^j(\vec{r})\rangle \cdot \langle S_i^j(\vec{r})\rangle \tag{3.12}$$

where i is the i-th voxel and j is the j-th trajectory. Only voxels with an occupation that exceeded densities of $0.2\rho_{bulk}$ were computed. $\langle\rho_i^j(\vec{r})\rangle$ and $\langle S_i^j(\vec{r})\rangle$ are number density and

entropy per water molecule for a single voxel i averaged over all frames of a single trajectory j, and $\langle S_i(\vec{r})\rangle$ is the entropy per water molecule for a single voxel i weighted averaged over all the trajectories. The total weighted entropy per water molecule for a selected volume, i.e. group of selected voxels is

$$\langle S_{volume}\rangle = \frac{\sum\limits_{k}\langle\rho(\vec{r_k})\rangle \cdot \langle S(\vec{r_k})\rangle}{\sum\limits_{k}\langle\rho(\vec{r_k})\rangle} \tag{3.13}$$

where k is the k-th voxel in the selected volume. We will also be interested in changes in the water entropy in a given region with respect to the bulk solvent,

$$\Delta S_{region} = \sum_{i}[V_i\langle\rho(\vec{r_i})\rangle(\langle S(\vec{r_i})\rangle - \langle S_{bulk}\rangle)] \tag{3.14}$$

where i sums over all the voxels of a given region. V_i is the volume of the i-th voxel. All voxels have identical volume, hence V_i is a constant. The uncertainties in these local entropy quantities is calculated via error propagation as

$$\sigma_{\Delta S_{region}} = V_i\sqrt{\sum_{i}[\rho^2(\vec{r_i})\sigma^2_{S(\vec{r_i})}] + (\langle S(\vec{r_i})\rangle - \langle S_{bulk}\rangle)^2\sigma^2_{\rho(\vec{r_i})}} \tag{3.15}$$

Parameterization

The parameters for the ligand were generated using the standard AMOEBA parameterization protocol[94]. The permanent atomic multipoles were calculated using the distributed multipole analysis from the DMA program by Stone[77]. This was based on single point calculations in *Gaussian*[78] using MP2/6-311G(1d,1p). The *poledit* program from *TINKER*[95] was used to rotate the multipoles from DMA into local frames compatible with the AMOEBA parameters files. *poledit* also defines the intramolecular polarization in accordance to Thole[72]. Because of the aromatic nature of this molecule, the whole molecule was assigned to a single polarization group, rather than distinct polarization group definitions. This gives an initial estimate of the multipole values. A further refinement is performed on these via a single point energy calculation in *Gaussian* using MP2/aug-cc-pVTZ. This is used to derive the electron density and construct the electrostatic potential on a grid outside the VdW envelope using *Cubegen*[78]. The *TINKER potential* program refines the atomic multipoles based on the quantum mechanics electrostatic potential. The initial multipole values obtained from the original DMA are not modified in the refinement step.

The structure of the ligand obtained from Ref: 21 was in the transition state (TS) form. For calculating the parameters for the TS state, this ligand structure was used, along with a carboxylate group to represent the electrostatic environment of the glutamate group. This structure consisted of the partially broken C-H bond, a partial triple C-N bond, and a partially broken N-O bond. For calculating the ligand parameters the ligand bound state, an optimization step was performed on the ligand. This restored the partially broken bonds to full bonds and restored the C-N bond to the original double bond form. The environment glutamate was not included in the parameterization of the ligand bound state because this state only consists of the bound form, and does not include any partial bond character with the glutamate.

Systems and Simulations

The KE07 and KE70 designed models for the apo state were obtained from the PDB database (PDB ID: 2RKX and 3NPU, respectively). The reactant EL state structures for KE07 and KE70 designs were obtained through the original KE07 and KE70 publication[21, 23]. The LDE evolved structures for both the apo and reactant states were modeled via homology modeling using MODELLER[96]. The substrate geometry and position of the active site for the $EL^{\dagger}$ state was the same as in the EL stat, with only the charges changed to reflect the transition state. All proteins in the apo, reactant and transition state were solvated in water boxes with water extending to a minimum of ~ 10 $\mathring{A}$ and a maximum of ~ 40 $\mathring{A}$ from the surface.

The solvated systems were minimized and equilibrated for 500ps in the NPT ensemble at 300K and 1atm. The structure with the most probable density was chosen as the starting point for a 3ns NVT trajectory at 300K and structures and velocities were saved every 100ps. These 30 snapshots were used to run NVE trajectories of 100ps each, for which coordinates, velocities and dipole moments were saved every 8 fs from the trajectories. VDOS (Eq. 3.11) were calculated for combined trajectories, each containing 300ps of NVE simulation, i.e. 37500 frames. In addition to the protein simulations, a pure water box with 4098 waters was also simulated following the same protocol to compute the entropy of pure bulk water.

All simulations were carried out using the AMOEBA force field[97] in OpenMM[98] on NVIDIA GPUs; details of the AMOEBA force field can be found in the corresponding publications. The 2PT entropy of bulk water has been calculated from the AMOEBA model to be 56.4 J/(K-mol), which is underestimated with respect to the experimental value of 69.9 J/(K-mol). However, this difference is mostly due to the approximate nature of the 2PT theory, and not the AMOEBA model, since it systematically underestimates the bulk water entropy relative to all other reported methods by ∼5-10 J/(K-mol) when analyzed across 5 different water models[90].

3.3 Results

Catalytic Site Geometry

The original design of KE07 was constructed with Glu101 to act as the catalytic base for proton extraction, Trp50 to contribute to steric stabilization as well as π-cloud delocalization with the ligand, and Lys222 to stabilize the negative charge that would be formed on the oxygen in the ligand as part of the bond breaking. However, we observe over the course of our MD trajectories, HIS201 is much closer to the oxygen in the ligand, and is likely forming the hydrogen bond for stabilizing the negative charge being formed on the ligand oxygen. This has been reported in multiple previous studies[99–101] as well. Over a sample WT MD trajectory, we observe that the average distance of the Lys222H_{ζ} from the ligand oxygen to be 5.93$\mathring{A}$, while the distance of the His201H_{ϵ} is 3.13$\mathring{A}$, as shown in Fig 3.3. Similar distances are also observed in R7. This shows that His201 is clearly a more likely candidate for stabilizing the negative charge being formed on the oxygen than Lys222. A previous study[99] has an indepth discussion of the alternative bonding Lys222 is involved in, with Ser48, Gly46 and Ile/Asp7. Hence we chose the catalytic site to be the combination of residues Trp50, Glu101 and His201, and performed the analysis for solvation from the center of the catalytic site of these three residues.

ResNum	Original	Final	Nature
7	Ile	Asp	Hydrophobic → Polar Acidic
12	Val	Met	Hydrophobic → Hydrophobic
77	Phe	Ile	Hydrophobic → Hydrophobic
102	Ile	Phe	Hydrophobic → Hydrophobic
146	Lys	Thr	Polar Basic → Polar Uncharged
202	Gly	Arg	Hydrophobic → Polar Basic
224	Asn	Asp	Polar Uncharged → Polar Acidic
229	Phe	Ser	Hydrophobic → Polar Uncharged

Table 3.1: Mutations between WT and R7-2 in KE07

The catalytic residues for KE70 (Fig. 3.4), do not undergo any such spatial rearrangement and no alternative residues is involved in the catalysis.

Active Site Waters

The 8 residues mutated between WT and R7 in KE07 are shown in Table 3.1. Out of these residues, 3 (12,77,102) do not show a change in nature, while 4 (7,202,224,229) become more polar and charged. This would lead to a higher interaction of these residues with waters, through electrostatics or hydrogen bonding. This is particularly true for the mutated residues located within the inner channel (7,202,224). As seen in Fig 3.5, in general, more waters fit in the first shell from the catalytic site in R7 as compared to WT. This also corresponds with the fact that the mutations in the channel significantly increase its hydrophilic nature, which would make the neck more favorable for a greater number of waters. The increase observed also agrees with a previous computational study[101] of the solvation of the various KE07 mutants, where an increased number of waters was reported for R7 as well.

The mutations for KE70 are listed in Table 3.2. Exactly contrary to KE07, mutated residues of KE70 which are located within the central channel, proximal to the catalytic site (48, 74, 138, 166, and 204), become uncharged or more hydrophobic from WT to R6. This corresponds with the sharp drop in number of waters between WT and R6 in the apo state as seen in Fig. 3.6. The free volume of the channel in KE70 in both WT and R6 is also smaller than that in KE07, and hence the presence of the ligand, removes almost all of the waters in the catalytic site of KE70.

Solvation Layer

The solvation layer of the protein extends from the surface into the bulk of the protein. The extent of the solvation shell has been under debate, where experiments like neutron scattering have suggested that this solvation shell could extend to 5-6$\mathring{A}$[102]. Other experiments that have studied THz spectroscopy report the presence of water-protein picosecond fluctuations extending upto 15-20$\mathring{A}$ from the surface of the protein[17]. We observe in Fig. 3.7 that the solvent entropy is increasing as a function of the distance from the surface of the protein sharply to $\sim$6$\mathring{A}$ and then

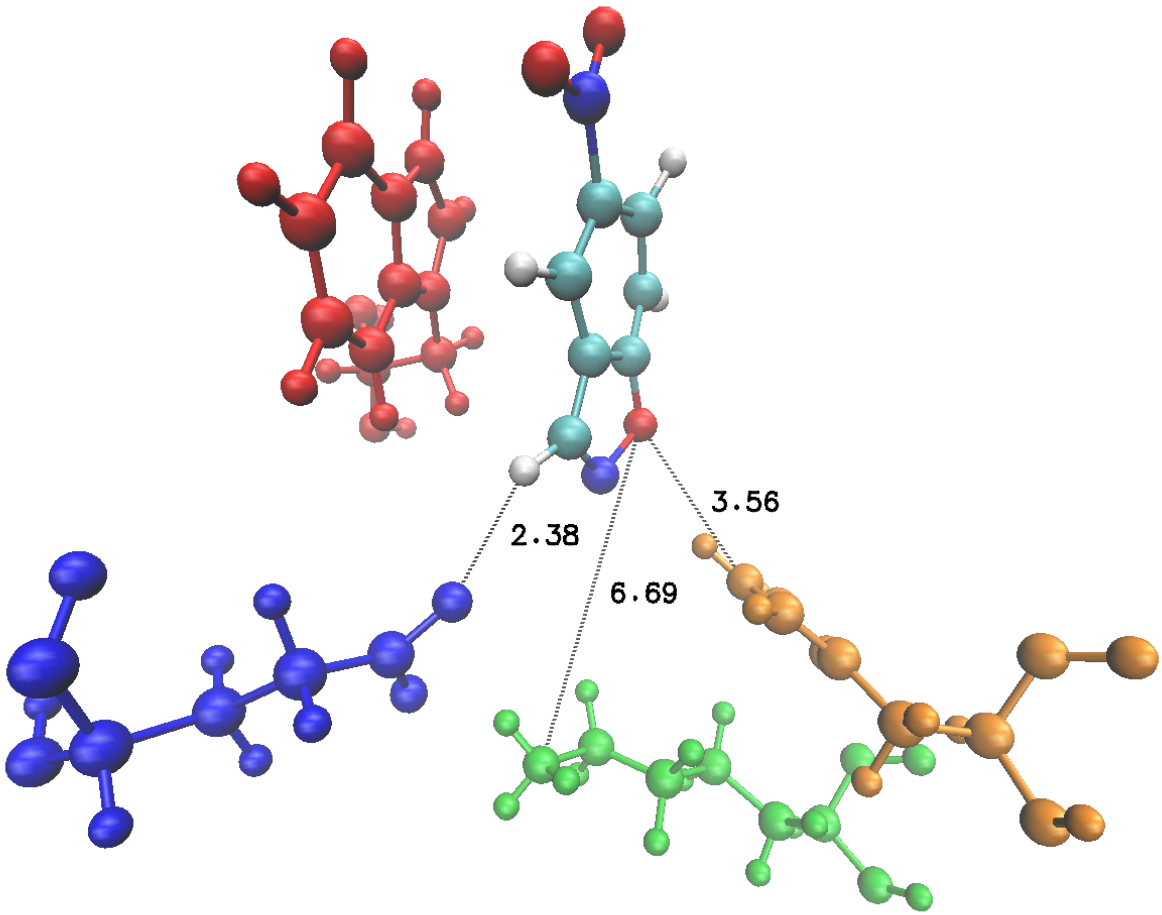


Figure 3.3: KE07 Catalytic Site: Trp50 (Red), Glu101 (Blue), His201 (Orange), Lys222 (Green). Lys222 (6.69$\mathring{A}$ away) in the incorrect configuration for stabilizing the negative charge being formed, while His201 (3.56$\mathring{A}$ away) can do the job.

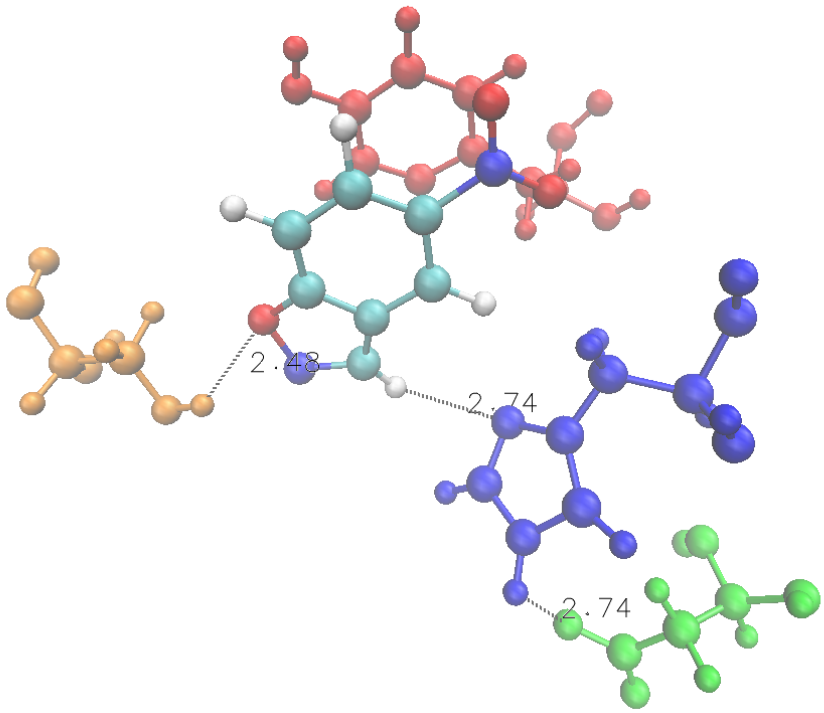


Figure 3.4: KE70 Catalytic Site: His17 (Blue), Asp45 (Green), Tyr48 (Red), Ser138 (Orange). Unlike KE07, this active site is maintained, and does not involve interactions with alternative residues.

ResNum	Original	Final	Nature
29	Lys	Asn	Polar Basic → Polar Uncharged
43	Thr	Asn	Polar Uncharged → Polar Uncharged
48	Tyr	Phe	Polar Uncharged → Hydrophobic
72	Trp	Cys	Hydrophobic → Polar Uncharged
74	Ser	Gly	Polar Uncharged → Hydrophobic
101	Gly	Ser	Hydrophobic → Polar Uncharged
138	Ser	Ala	Polar Uncharged → Hydrophobic
166	His	Asn	Polar Basic → Polar Uncharged
178	Ala	Ser	Hydrophobic → Polar Uncharged
197	Lys	Asn	Polar Basic → Polar Uncharged
198	Thr	Ile	Polar Uncharged → Hydrophobic
204	Ala	Val	Hydrophobic → Hydrophobic

Table 3.2: Mutations between WT and R7-2 in KE70

showing a slow approach to the bulk entropy value at $\sim 10\mathring{A}$ from the surface of the protein. This is in agreement with previous experimental results[102] which showed that the hydration shell based on neutron scattering extends to $\sim 6\mathring{A}$ from the surface of the protein. Proximity to the surface, with residues of different natures, gives rise to differing chemical potentials for voxels in the first layer, leading to intrinsic variability in the entropies very close to the surface of the protein.

The entropy of bulk water has been calculated from the AMOEBA model to be 56.4J/(K-mol).

Hydration Water Species Analysis

The entropies of waters in the hydration layer of the protein depend both on the topology of the neighbouring protein atoms, which would introduce steric constraints on the waters, as well as the chemical nature of the neighbouring residues, which would affect the chemical potential of the waters. Hence, hydration waters surrounding the protein show a large distribution of entropy values. As seen in Fig. 3.8, the entropy distribution shows that entropies of the waters are distributed between 35-60 $J/K - mol$ per water molecule.

As discussed in Sec. 3.2, the entropy is computed from harmonic oscillator and hard sphere contributions. Hence water molecules with differing vibrational features may have the same entropy values. Thus, vibrational features can be used to classify waters into different species based on their degrees of binding to the protein. Hydration water species are chosen in particular for analysis, in order to understand the various signatures of water behaviour around the protein as well as in the active site in particular. These waters are chosen by filtering for voxels with a number density of waters $>$ 1.1*bulk water density i.e. 1100 kg/m^3.

The vibrational densities of states show major signatures for H-bond bending at frequencies below 100 cm^{-1}, H-bond stretch at $\sim$ 200 cm^{-1} and librational motions between 300 to 1000cm^{-1} , as shown in Fig. 3.9. These three major signatures of the VDOS can hence be used to cluster the waters into various species based on their vibrational properties.

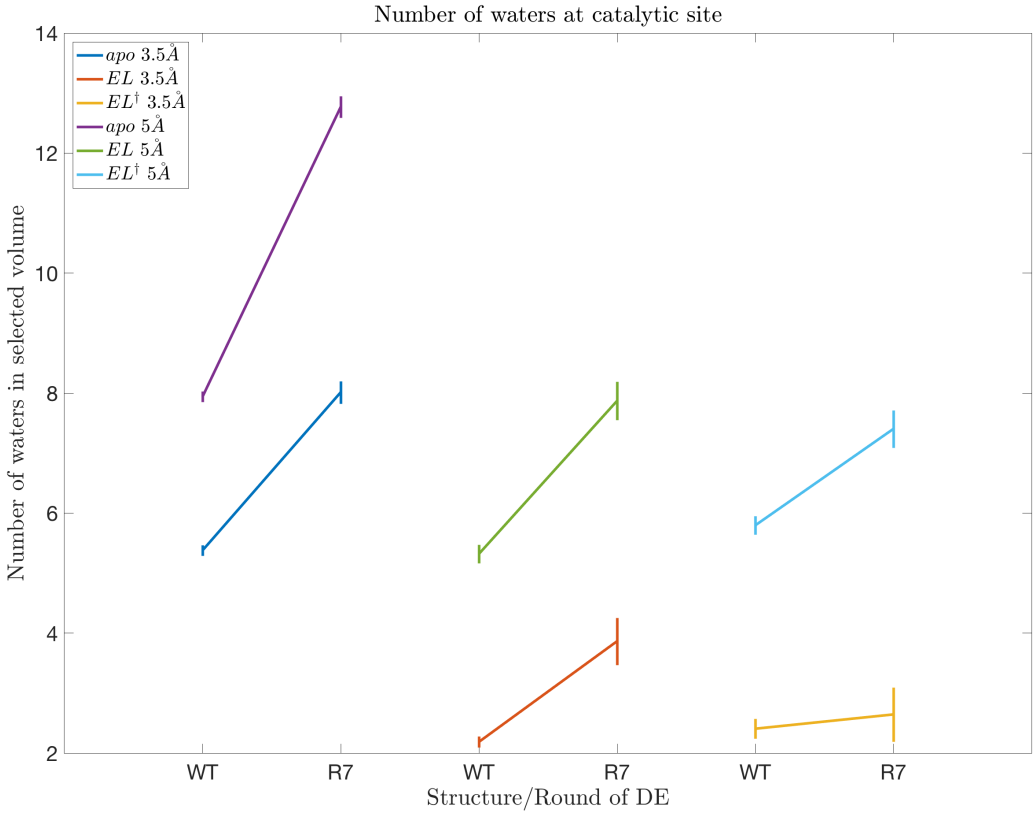


Figure 3.5: Average number of waters within 3.5$\mathring{A}$ and 5$\mathring{A}$ from the catalytic site for KE07. The waters are averaged over all trajectories for each of the systems.

For each voxel, the area under the VDOS curve for ranges 0-100 cm^{-1}, 100-200 cm^{-1}, 300-400 cm^{-1} are chosen as parameters to classify the waters. A k-means clustering algorithm was used to group the waters into clusters. For values of $k > 3$ the average VDOS of the identified clusters did not significantly differ from each other. Hence, three clusters were chosen to group the waters.

The total entropy distribution is thus further analyzed into different species of waters based on clustering using VDOS of the waters. The three clusters show progressively increasing entropies, as shown in Fig. 3.10. These can hence be classified into strongly bound (magenta), weakly bound (green), and unbound (yellow) waters, with progressively increasing entropies. The entropy of bulk water is also shown as a sharp peak, and the unbound waters display, as expected, an entropy very close to bulk waters.

The voxels are further clustered in the 3D space created by the values of the areas under the curves in these three ranges. The total area under the entire VDOS curve corresponds to the total degrees of freedom of each molecule. In this case, the total number of degrees of freedom are 6, i.e. 3 translational and 3 rotational (neglecting the vibrational degrees of freedom, as their contribution to the entropy is negligible). The values of the area under the curves in various frequency regimes

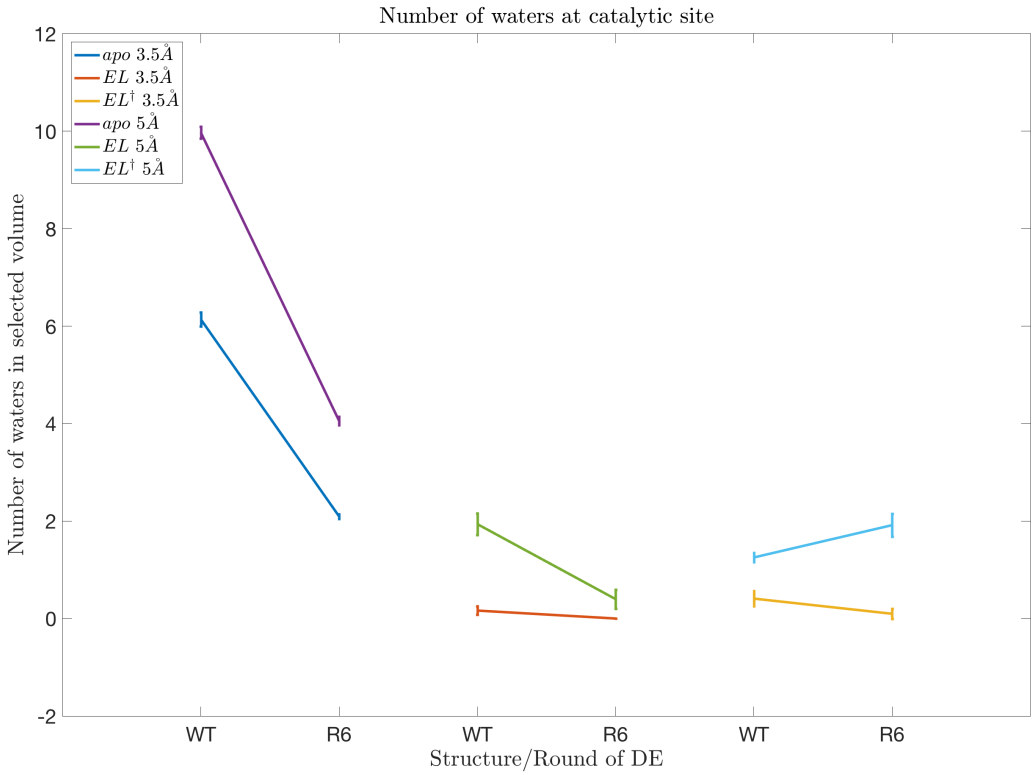


Figure 3.6: Average number of waters within 3.5$\mathring{A}$ and 5$\mathring{A}$ from the catalytic site for KE70. The waters are averaged over all trajectories for each of the systems.

are normalized such that the area under the entire curve corresponds to the 6 degrees of freedom. This clustering against the degrees of freedom can be seen in projections as shown in in Fig. 3.11.

Comparing waters in the proximity of catalytic residues (choosing all voxels $< 5\mathring{A}$ of the closest catalytic residue atom) across the various structures (Fig. 3.12), we can see that the waters confined in the central channel of the protein, despite all being relatively confined and constrained, shows a distribution mirroring the distribution of all hydration waters on the outer surfaces of the protein. In particular, comparing between the two KE07 variants, WT shows a significant shift to low entropy strongly bound waters from the weakly and unbound waters in the presence of the ligand, both in the EL and $EL^{\dagger}$ states. The fraction of strongly bound waters is mostly unchanged, while there is a shift from the unbound to the weakly bound waters in R7. Hence, overall we can see that the ligand binding increases the binding of the waters and hence shows an overall lowering of entropy as a result.

Performing a similar analysis of hydration water species for KE70, we see that in WT, as we transition from $apo \rightarrow EL \rightarrow EL^{\dagger}$, the presence of the ligand ($apo \rightarrow EL$) leads to an increase in the fraction of bound waters at the expense of unbound waters, while the transition from $EL \rightarrow EL^{\dagger}$ further leads to an increase in the fraction of weakly bound waters at the expense of unbound waters. Likewise in R6, both transitions ($apo \rightarrow EL$ and $EL \rightarrow EL^{\dagger}$) show an increase in the

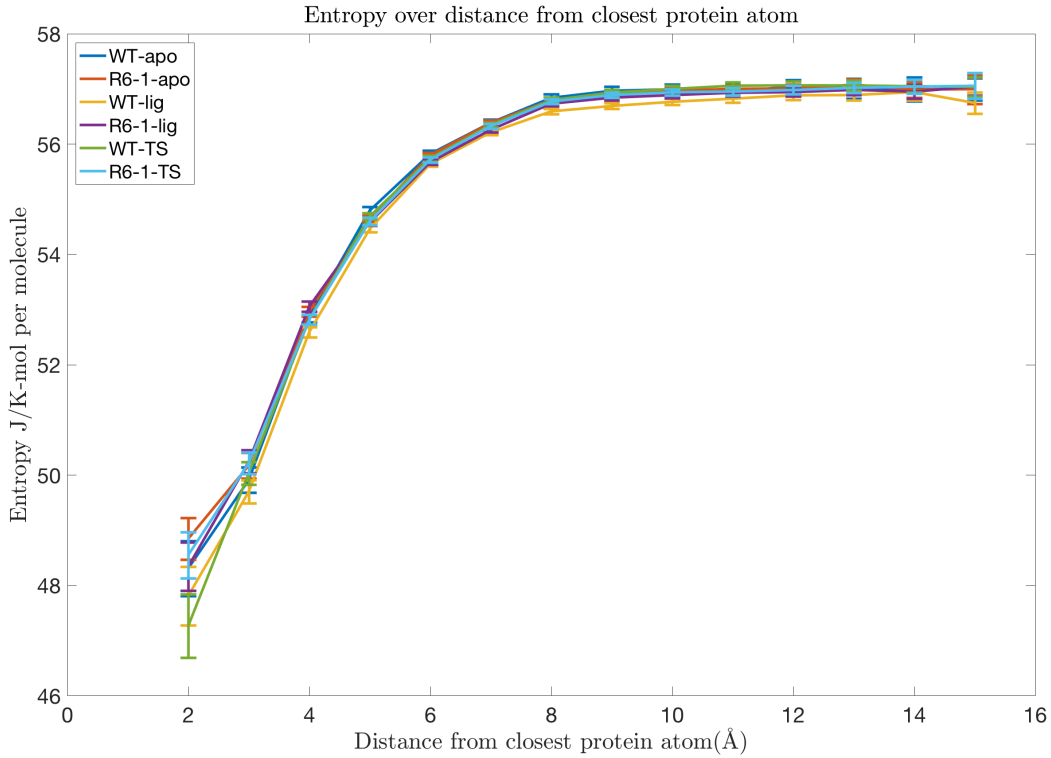


Figure 3.7: Solvent entropy per water molecule as a function of distance from the nearest protein surface for KE07.

fraction of weakly bound waters at the expense of the other two species.

As also evidenced by the number of waters at the catalytic site for KE70 (Fig. 3.6), the central volume of the catalytic site can accommodate very few waters in the states where the ligand is present (EL and $EL^{\dagger}$). While KE07 has >2 water molecules in each state within 3.5$\mathring{A}$(Fig. 3.5), KE70 has < 2 waters in both rounds even as far out as 5$\mathring{A}$ from the center of it's catalytic site. This suggests that the ligand pushes out all the unbound waters in both variants of this enzyme, only leaving in a much larger fraction of weakly bound waters in the better performing variant as compared to the design, where the bound waters dominate. In general, however, as opposed to KE07, KE70 relies on predominantly creating a hydrophobic catalytic site, which does not accommodate large number of waters. Hence the two enzymes show very distinct behaviors in the context of catalytic site hydration.

Entropic Contributions to Catalysis

The entropic contributions for waters located within the neck proximal to the catalytic site are computed. To compare commensurate volumes across rounds, spherical volumes of radii 3.5 $\mathring{A}$ and 5.0 $\mathring{A}$ are chosen from the center of the three catalytic residues, i.e. 50, 101 and 201 for KE07.

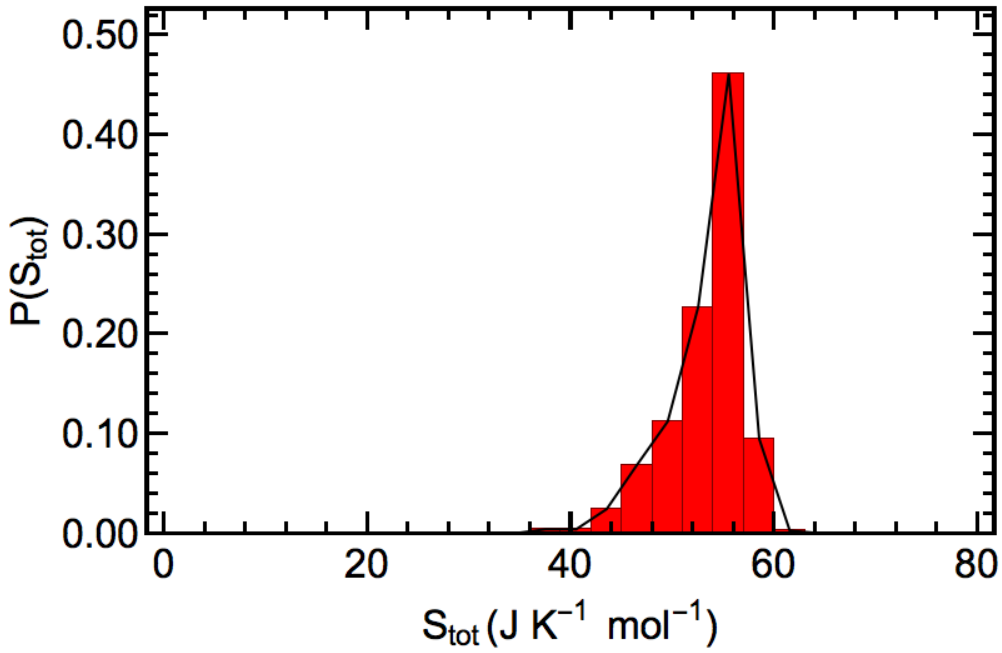


Figure 3.8: Entropy distribution for all waters within the hydration shell of the protein

	Catalytic Site Solvation Entropy -TS(kcal/mol)			
	Voxels $< 3.5\mathring{A}$		Voxels $< 5.0\mathring{A}$	
State	Designed Enzyme	Best LDE Variant	Designed Enzyme	Best LDE Variant
apo	-4.40 (0.05)	-6.14 (0.13)	-4.97 (0.04)	-8.04 (0.13)
EL	-1.26 (0.06)	-3.09 (0.3)	-2.59 (0.05)	-4.83 (0.23)
$EL^{\dagger}$	-1.48 (0.06)	-2.20 (0.4)	-3.41 (0.12)	-5.05 (0.26)

Table 3.3: Catalytic Site Solvation Entropy $-TS$ (kcal/mol) for volume within given radius from center of catalytic site in KE07

	Catalytic Site Solvation Entropy -TS(kcal/mol)			
	Voxels $< 3.5\mathring{A}$		Voxels $< 5.0\mathring{A}$	
State	Designed Enzyme	Best LDE Variant	Designed Enzyme	Best LDE Variant
apo	-0.90 (0.07)	0.50 (0.04)	-0.91 (0.06)	1.16 (0.04)
EL	-0.053 (0.03)	-	-0.26 (0.03)	0.074 (0.05)
$EL^{\dagger}$	-0.26 (0.1)	0.047 (0.05)	-0.54 (0.05)	0.14 (0.08)

Table 3.4: Catalytic Site Solvation Entropy $-TS$ (kcal/mol) for volume within given radius from center of catalytic site in KE70. The entropy values are very low in the ligand bound states due to the very low number of waters present in these EL and $EL^{\dagger}$

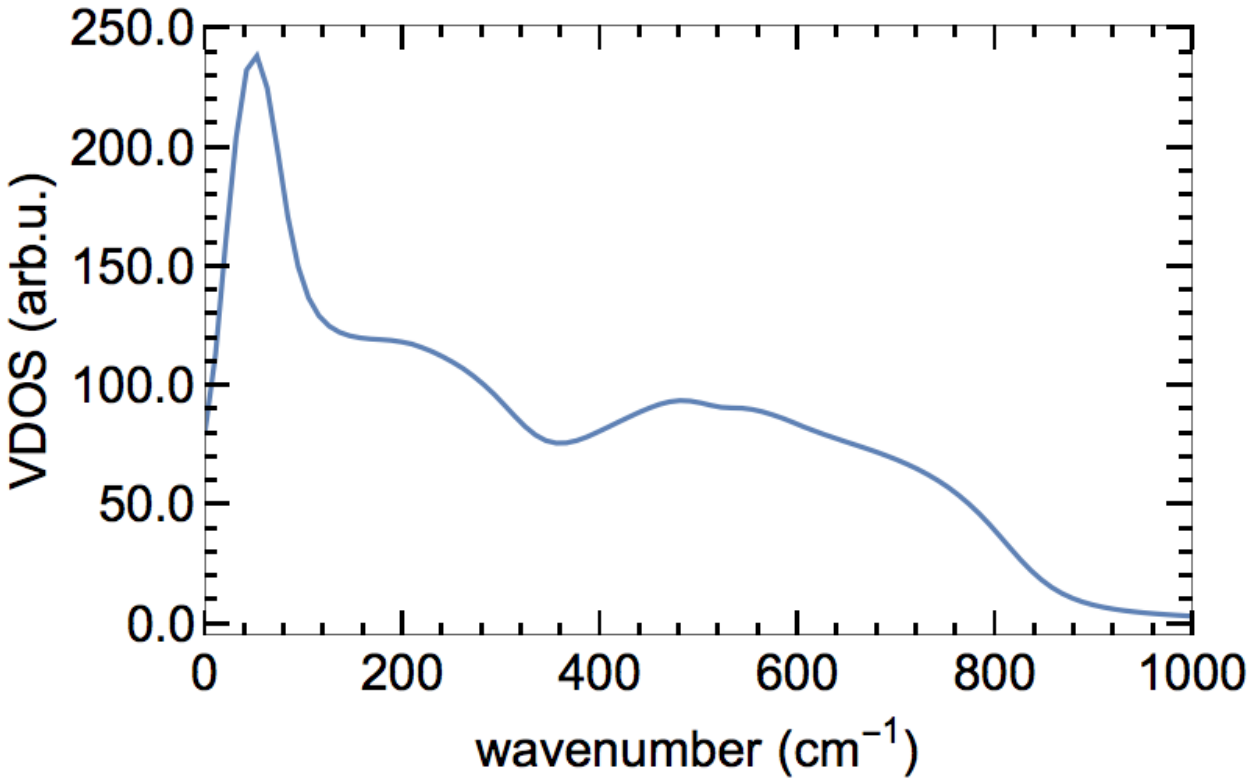


Figure 3.9: Vibrational density of states as a function of frequency.

Table 3.3 shows that the entropic change for catalytic waters in KE07 for R7 are smaller than WT for all states (APO, LIG, and TS). This shows that the total entropic effect of moving the selected number of water molecules from bulk to the neck containing the catalytic site leads to a greater loss of entropy corresponding to those water molecules in R7 than in WT. This contribution comes from a greater confinement of a larger number of water molecules in the first solvation shell for R7 than for WT. This can also be seen from the average number of water molecules in each of the systems averaged over all trajectories as shown in Figure 3.5.

Performing a similar analysis for KE70, we observe a very starkly different scenario. The inner barrel volume for KE70 is much smaller than that for KE07, as can be seen by comparing Figs. 3.5 and 3.6. In addition, as seen in Table 3.2, the mutations between KE70 Design and best LDE variant within the central barrel, residues 48, 74, 138, 166 and 204 in particular, become less charged or more hydrophobic. This makes the environment even more unfavorable for the waters. Very few of waters are trapped within a hydrophobic environment, and are hence unable to form hydrogen bonds with both nearby polar/charged residues (since they are mutated out) or with other waters (because there are very few waters in that site). This leads to very high entropies for these waters in the catalytic volume of the best LDE variant in KE70. The change in the nature of the surrounding residues and the corresponding compaction of volume available for water can be seen in Fig. 3.14. Fig. 3.14a shows the voxels in the design, where the voxels of interest are

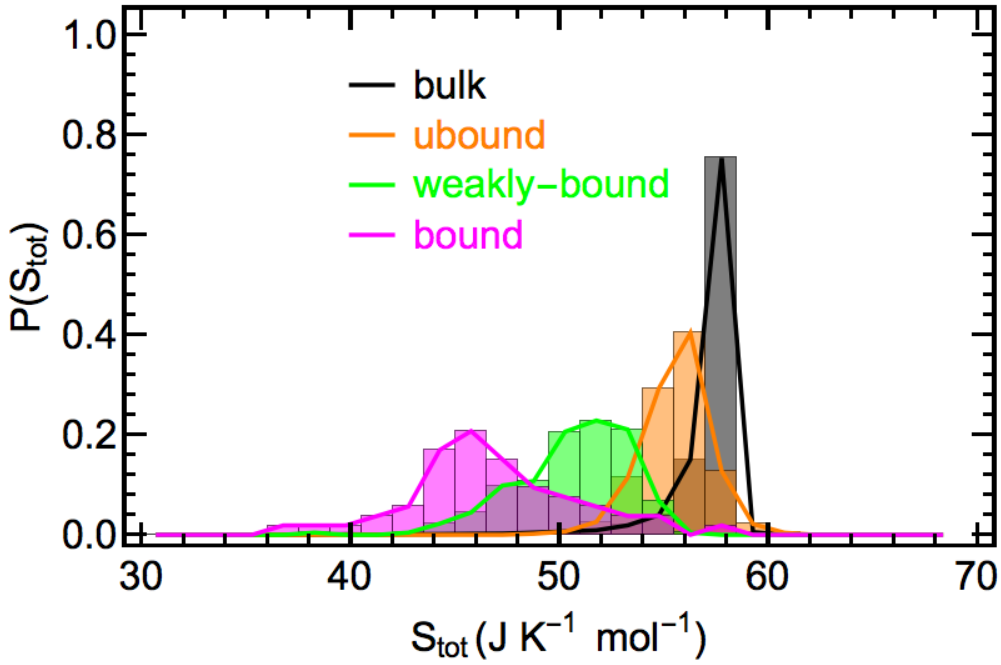
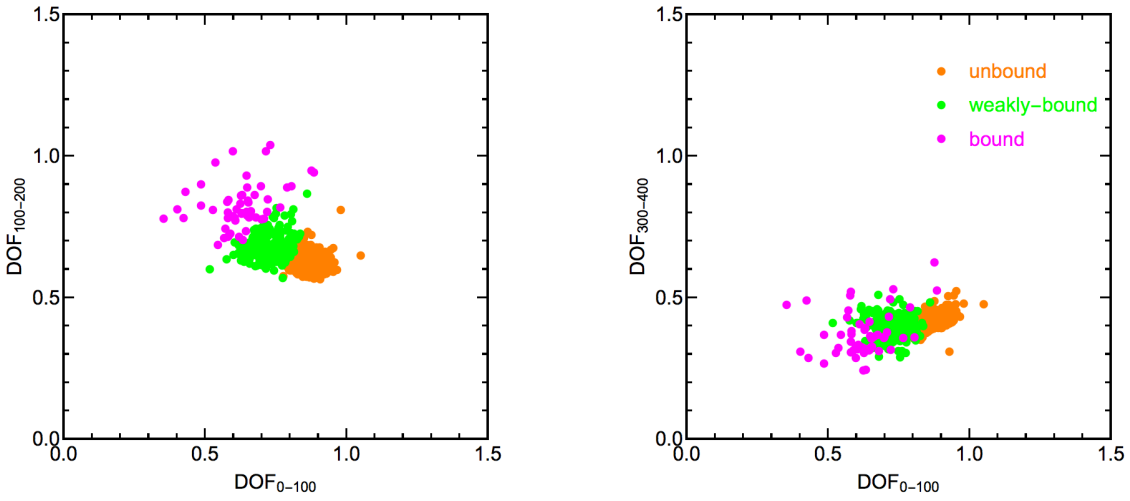


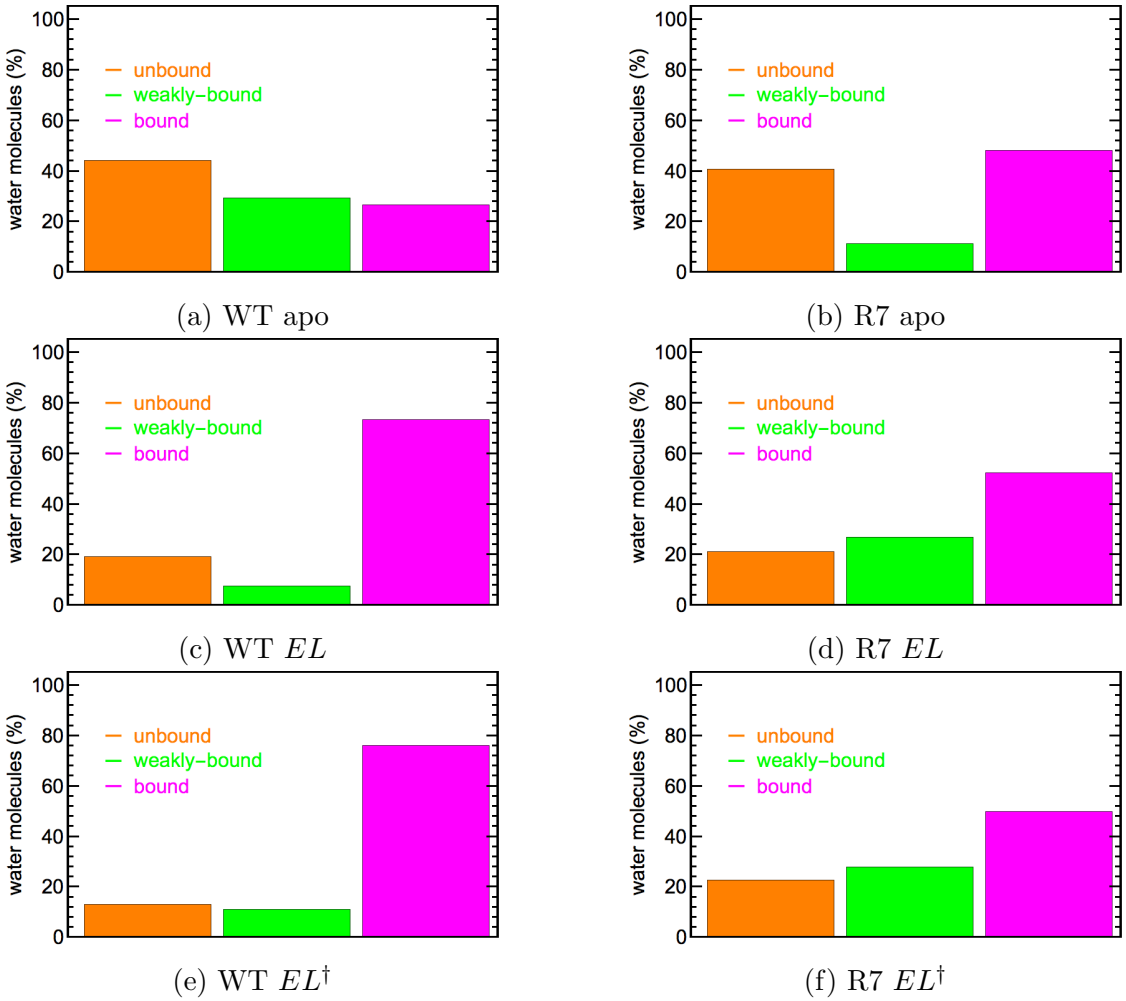
Figure 3.10: Species specific distribution of entropies over all waters



(a) Cluster Projection 1

(b) Cluster Projection 2

Figure 3.11: Projections of the clusters on the DOF_{0-100} - $DOF_{100-200}$ and DOF_{0-100} - $DOF_{300-400}$ planes



(a) WT apo

(b) R7 apo

(c) WT *EL*

(d) R7 *EL*

(e) WT $EL^{\dagger}$

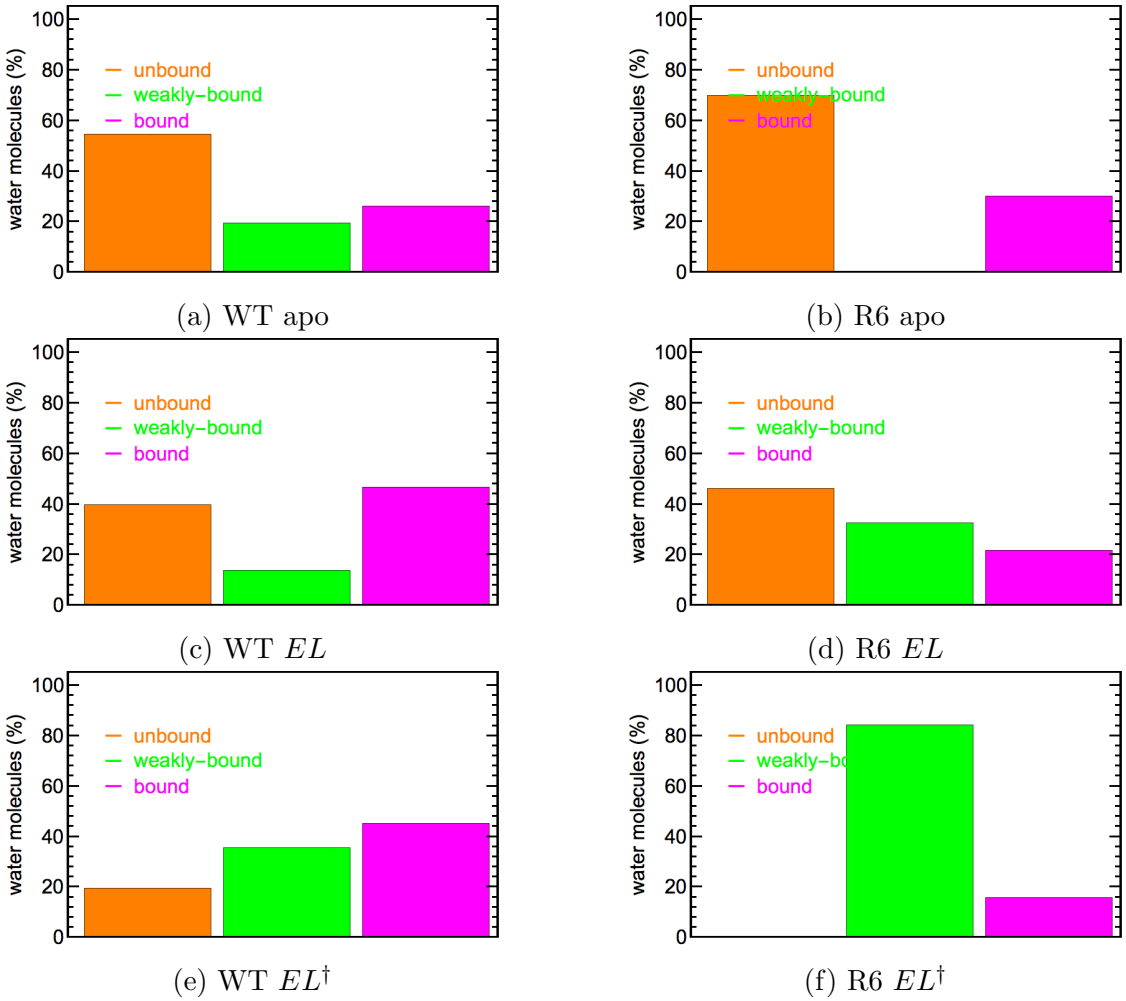
(f) R7 $EL^{\dagger}$

Figure 3.12: Distribution of water species within 5$\mathring{A}$ of catalytic site residues for KE07

surrounded by more charged residues, while in the best LDE variant Fig.3.14b, a much smaller number of voxels is available for waters, and very few polar/charged residues surround the voxels of interest. Hence, the waters in the best LDE variant for KE70 have very high entropy, in this case even higher than bulk (Table 3.4). Waters in the bulk form extended H-bond networks, which influence their motions. However, these isolated water molecules in the hydrophobic pocket of the best LDE variant in KE70 can neither bond to the surrounding residues (as they are hydrophobic), nor to any surrounding water molecules, as very few water molecules are present in this catalytic volume.

3.4 Conclusion

Previous work has shown the contribution of the changes resulting in the enzyme due to mutations going from the designed to evolved variants of the Kemp Eliminases to the increased catalytic activ-



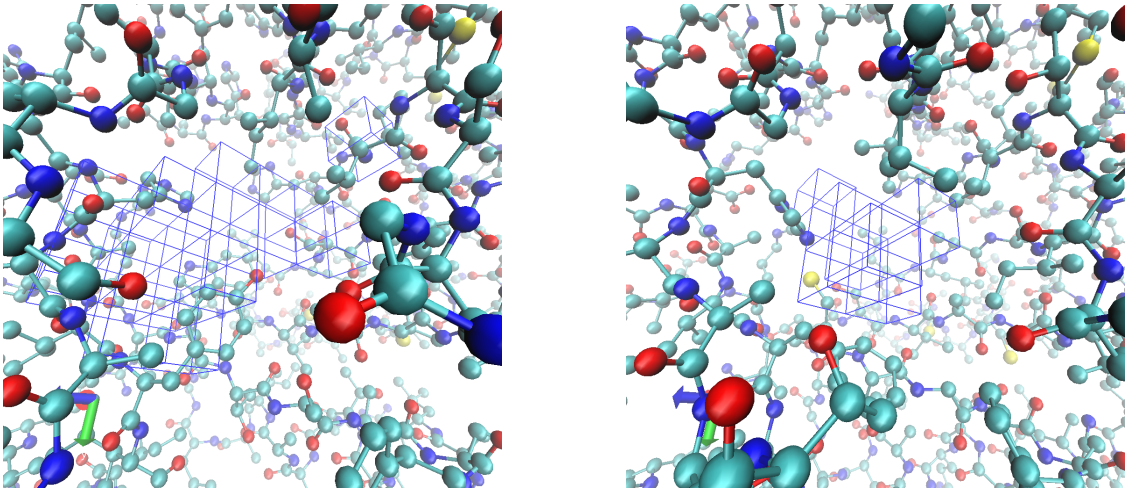
(a) WT apo (b) R6 apo

(c) WT *EL* (d) R6 *EL*

(e) WT $EL^{\dagger}$ (f) R6 $EL^{\dagger}$

Figure 3.13: Distribution of water species within 5$\mathring{A}$ of catalytic site residues for KE70

ity through residue side chain entropy[99] and protein electrostatics[87]. There has also been some analysis of the role of the solvent in mediating activity[101]. However, we present here a detailed analysis of the natures of the waters within the catalytic sites of two designed Kemp Eliminases, and highlight how mutations can change the nature and behavior of waters in the catalytic sites of the enzymes. This study also highlights how two enzymes designed from very similar scaffolds, and carrying out the same reaction, can operate through radically different solvation behaviours. In KE07, the mutations leading to the the mutations in the better performing enzyme make the inner channel more hydrophilic, thus opening it to a larger number of waters, thus boosting solvation of the catalytic site. Hence the presence of the waters aids the reaction. In KE70, on the other hand, the mutations in the best variant make the inner channel more hydrophobic, and thus eliminate the waters from the central channel. The catalysis in KE70 is hence, promoted by the central hydrophobic cage.



(a) WT apo (b) R6 apo

Figure 3.14: Voxels in the catalytic volume and surrounding residues in KE70

Chapter 4

Residue Correlation

4.1 Introduction

Naturally occurring enzymes are optimized for high catalytic activity by lowering the energy of activation in the transformation of reactants to products, resulting in ∼10-20 orders of magnitude increase in the rate compared to the uncatalyzed reaction. There are, however, many reactions for which naturally occurring enzyme catalysts are not available. Recently, there has been significant success in the design of enzymes that can catalyze new reactions that lack a natural biocatalyst, albeit with generally very low catalytic activity. The current state of the art for these de novo enzymes is to subject them to multiple rounds of laboratory directed evolution (LDE) to improve their activity, typically by 2-3 orders of magnitude.

Whether natural or designed, the catalytic activity of enzymes is directly dependent on the structure and energetics of the bound enzyme-substrate ES complex and the subsequent formation of the transition state $EL^{\dagger}$ via the classical Michaelis-Menten mechanism. Warshel et. al. have convincingly advanced the idea that electrostatic interactions from the scaffold and surrounding solvent is pre-organized in such a way that any reorganization costs for solvating the transition state has been eliminated, leading to its stabilization relative to the uncatalyzed reaction in water[2]. This primarily thermodynamic view is well supported by experiment.

However, other recent studies have suggested that protein dynamics is another major factor that can influence the catalytic activity[3–14]. Dynamical motions that contribute to catalysis can be defined as any motion that, when impeded, would reduce the ability of the enzyme to perform its catalytic function by its standard mechanism[8]. However, the definition of dynamics that is relevant for enzyme catalysis is variable among research groups, and can be defined as (1) the reactive flux through the transition state barrier, (2) the dynamical coupling of modes on different timescales, or (3) statistical fluctuations arising from entropic factors (the latter which is thus really thermodynamics).

Multiple studies of di-hydrofolate reductase (DHFR), triosephosphate isomerase (TIM), liver alcohol dehydrogenase (LADH) and other enzymes suggest that conformational motion arising from statistical fluctuations are mostly relevant to catalytic activity[4–14]. A recent study of two designed Kemp Eliminases, demonstrates that side chain motions captured through side chain entropy can contribute to transition state stabilization that improves under laboratory directed

evolution of the best evolved enzymes[99]. Other studies on DHFR have identified a network of residues[4] whose motions are coupled to the electron donor-acceptor catalytic reaction coordinate, since certain mutations which disrupt these coupled motions have been found to decrease the rate of hydride transfer catalyzed by DHFR[9, 10]. Similarly, Schwartz and coworkers have found that the correlated motions of certain residues with the protein promoting vibration important for hydrogen tunneling in horse LADH (eg. Val203) can be captured through the coupling of their velocities of motion[103]. For the Kemp Eliminases, recent work [99] found that when the definition of single site entropy was extended to describe mutual information i.e. the correlations of statistical fluctuations between residues - the LDE optimized enzymes showed higher mutual information in the transition state.

The focus of this study is to provide a mechanistic view of how these correlated motions derived from mutual information can be analyzed under a distance-distance time correlation function for the marginally performing designed Kemp Eliminase enzyme KE07[21, 22], and how these dynamical signatures change as the enzyme improves under LDE. In particular, we show that the mutations in successive rounds of directed evolution significantly change the dynamics of the surface loops in their correlated motions with the active site. The results offer the possibility that computational methods might not only provide good de novo enzyme leads, but take more direct part in their optimization, by bridging the structure-function relationship to more fully embrace the role of dynamics and flexibility for designing novel biocatalysis constructs.

The TIM barrel scaffold, into which KE07 has been designed, is an example of the $(\beta/\alpha)_8$ fold and consists of an inner barrel of β-sheets surrounded on the outside by α-helices, which are connected by 8 loops. This scaffold is the most common fold for enzymes, constituting approximately 10% of all known enzymes[104]. The active sites of these enzymes are located on the catalytic face of the barrel, consisting of the C-terminal end of the β-strands and the loops connecting these strands to the helices. The opposite face, which consists of loops which connect the α-helices to the N-termini of the β-strands, is important for maintaining fold integrity. The hinge residues connecting loops to the flanking b-strand and a-helix have also been shown to be important, since they can significantly alter the dynamics of the loops, and thus modify functionality[105]. The 8 loops on the catalytic face of the Triosephosphate Isomerase from *S. cervisiae* are shown in Fig. 4.1.

For the Triosephosphate Isomerase from *S. cervisiae*, which is a prototype example of the TIM barrel scaffold structure, the dynamics of a particular surface loop, namely Loop6, have been shown to be important for the catalysis[106]. The active site of this particular enzyme is located at the intersection of an inner b-strand with the loop, and two conformers of this loop show distinct mechanistic functions, where the open conformation facilitates the binding and release, while the closed conformation catalyzes the chemical reaction. Removal of the important Loop 6 shows an activity 5 orders of magnitude less than the original [107]. NMR experiments have also shown this loop motion to be part of the natural dynamics of the TIM protein[108]. Ligand binding also introduces a configurational change in loop 7 of a similar trypanosomal TIM[109].

Like the S. cervisiae TIM, the KE07 scaffold has 8 loops on the catalytic face. While the KE07 design active site itself is located in a different location, i.e. the center of the barrel, structural alignment of the two TIM scaffolds shows that we can draw correspondences between the loops of the two proteins. Corresponding loops in KE07 to the S. cervisiae TIM loops highlighted in Figure 1 are: Loop1 (residues 13-30), Loop2 (51-56), Loop3 (81-85), Loop4 (102-110), Loop5 (132-152),

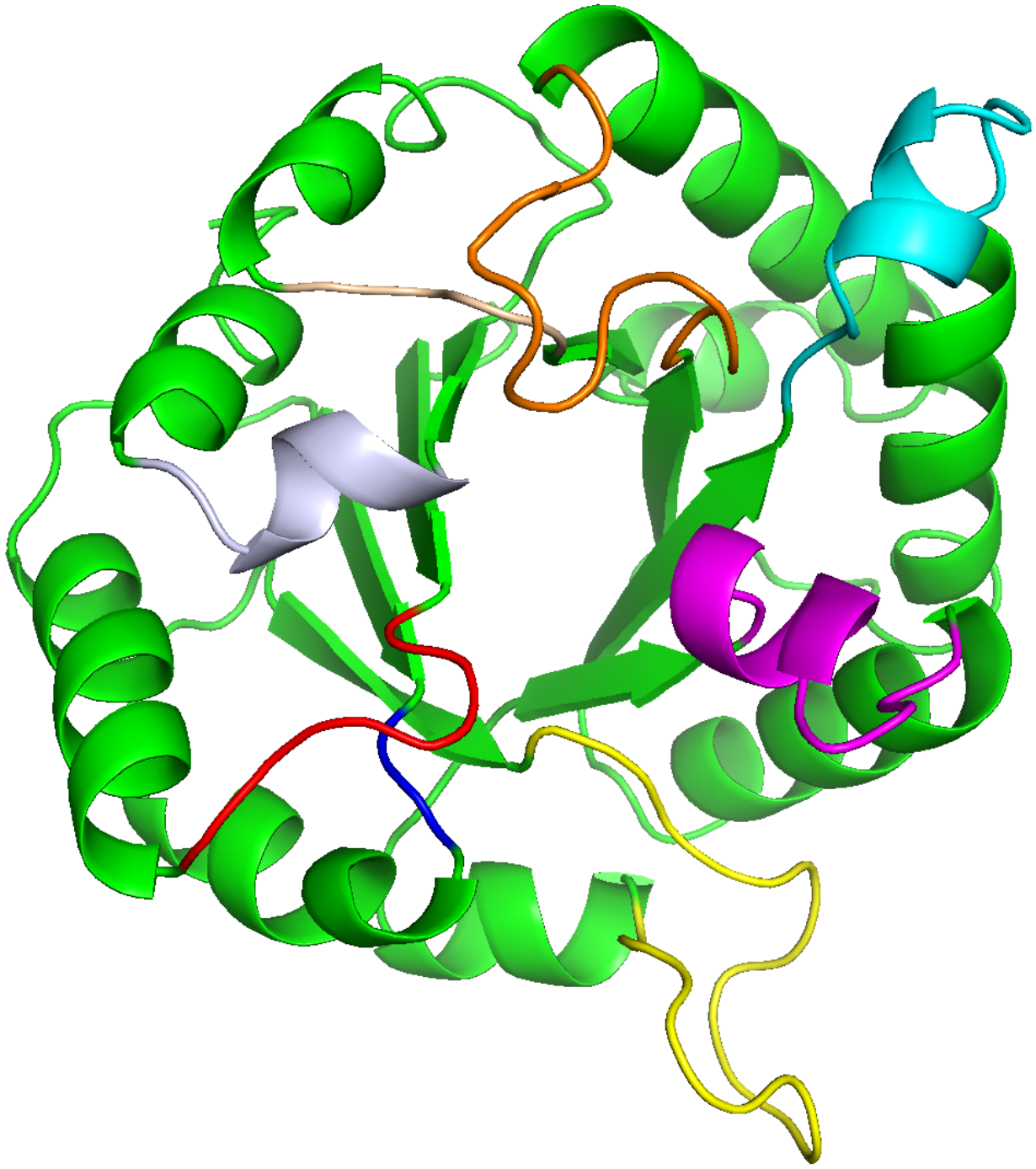


Figure 4.1: Loops in Triosephosphate Isomerase from *S. cervisiae* Loop1 (red), Loop2 (blue), Loop3 (yellow), Loop4 (magenta), Loop5 (cyan), Loop6 (orange), Loop7 (wheat), and Loop 8 (gray).

Loop6 (172-182), Loop7 (201-206), and Loop8 (229-232). What we show in this study that while the highly correlated motion of loops in natural TIM barrels is lost in the KE07 design, most of these correlated loop motions are restored in the best performing variant under LDE.

4.2 Methods

Crystal structures are available from the PDB for the apo states of the original KE07 design and rounds R4, R6 and R7.1 from LDE were used as starting states for the simulations; no crystal structures were available for the ligand bound state due to difficulties in crystallization, and thus we relied on the supplementary material from Rothlisberger and co-workers[21] to define the 5-nitrobenzisoxazole ligand bound state, which we found agrees well with the reported crystal structure of the 5-nitrobenzisoxazole ligand bound state of KE70. The apo structures for R2, R3, R5 and R7.2 as well as ligand bound structures for R2 through R7.2 were derived using MODELLER[96].

Each of the structures in the apo, ligand bound, and transition state was equilibrated using an established protocol, with resulting equilibrated systems being comprised of ∼10,000 TIP4P-EW water molecules, 6 Na+ ions added to neutralize overall charge. Simulations were run using the AMBER 13 package, using the pmemd module, with the protein modeled using the AMBER-ff99sb force field. Charge parameters for the ligand were obtained from Frushicheva et al.[110]. The system was simulated under NPT conditions using a weak barostat coupling at 1 bar and a temperature of 300K. A 2 fs timestep was used, and long range electrostatic interactions were calculated using the particle mesh ewald method. A 12 A cutoff was set for real space electrostatics and LJ interactions. Two independent trajectories of 100 ns were run starting with different random velocities and time correlation functions were calculated along each of the trajectories.

Distance-distance time correlation functions. The extent of correlations in motions of residues were quantified using distance-distance time correlation functions (DD-TCF) between pairs of residues of interest

$$C(t) = \langle D_{AB}(t) \cdot D_{BC}(t+\tau) \rangle \tag{4.1}$$

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} D_{AB}(t) \cdot D_{BC}(t+\tau) dt \tag{4.2}$$

$$= \frac{1}{T-\tau+1} \sum_{t=0}^{T} D_{AB}(t) \cdot D_{BC}(t+\tau) \tag{4.3}$$

where D_{IJ} is the distance between residues I and J. Schwartz et al.[103] have shown that the time cross correlation function of the projections of relative velocities of donor and acceptor with distal residues give a measure of the strength of the impact of the motion of the distal residue on the donor acceptor motion (Protein Promoting Vibration). We propose that the DD-TCFs are a similar measure of the dynamical interactions between various residues in the protein. The Fourier transform of the DD-TCF gives the power spectrum,

$$J(\omega) = \int_{-\infty}^{\infty} C(t) e^{-i\omega t} dt \tag{4.4}$$

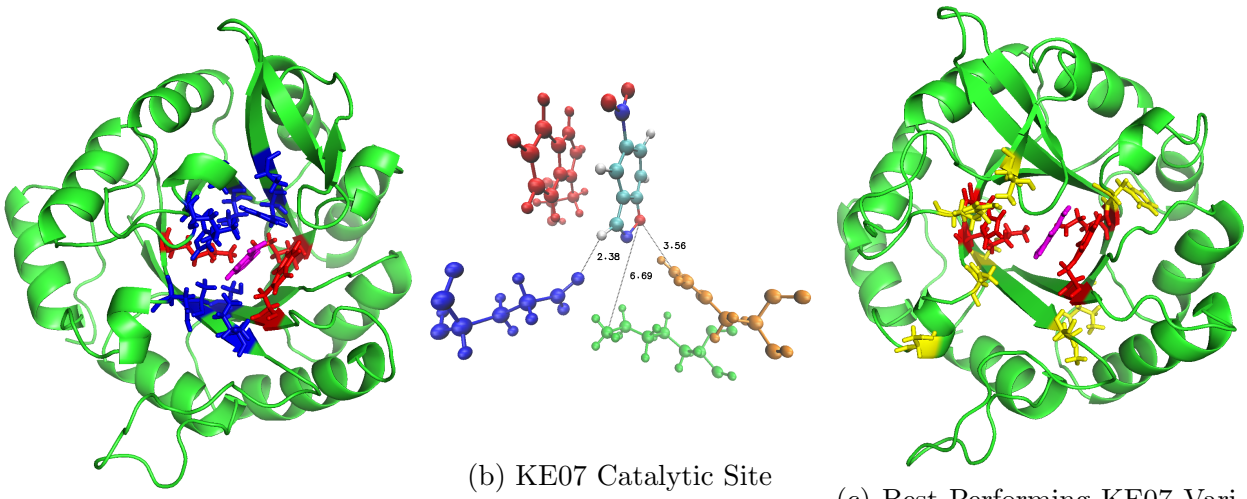
As the power spectrum gives the distribution of the power (or, in this context, the extent of the correlation) across various frequencies, the maximum amplitude is used to measure the strength

of dynamical coupling between the two motions of AB and BC. In accordance with the Wiener-Khinchin theorem, this power spectrum is the spectral decomposition of the time correlation. The DD-TCFs were calculated over different trajectory lengths to ensure convergence. The time correlation functions were calculated using frames saved from the MD trajectories every picosecond, with the largest amplitude from the spectral density Eq. 4.4 always located between 0-2 THz. Hence, at this time scale, the strength of the time correlation functions was concentrated at very low frequencies. As a result, comparing only the amplitudes at these low frequencies was used as a measure of the dynamical interactions between various residues in the protein. We hypothesize that these dynamical interactions are optimized by directed evolution in such a way as to better preorganize the enzyme in subsequent rounds of directed evolution so as to improve catalysis. Since these correlations are not calculated on the timescale of the actual reaction, they are not considered as coupled to the reaction coordinate and actually driving the reaction, the way Schwartz et al propose, but instead are optimizing dynamic interactions in order to preorganize the environment so as to allow more efficient catalysis.

Dynamical Coupling Analysis. Co-evolving residues within a protein are thought to be functionally related. Statistical coupling analysis has been used as a bioinformatics approach to analyze the thermodynamic relatedness of residues within proteins[111]; multiple sequence alignments of homologs of the protein of interest are used to identify residues that co-evolve with each other, and a tree can be constructed using a statistical energy function to determine the degree of relatedness of residues. Here we postulate that dynamical correlations, measured as the dominant amplitude using Eq. 4.4, can be used as similar metric to identify relatedness between functional motions of various residue within the enzyme.. A simple Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method is used to create networked trees, based on the degree of correlations computed between various pairs of residues, with the catalytic residues serving as the base of the tree. Residues which are proximal nodes on these trees have motions strongly correlated to each other. These relations between residues on the tree can be used as a metric to observe a network of interacting residues within the enzyme. The idea is to measure the network of coupled dynamical motions and their changes in each round of directed evolution, to understand the two orders of magnitude improvement in catalytic performance that is observed for KE07 at the end of the LDE process. A threshold of 1.25 is set as the distance between leaves for identifying the residues most highly correlated with each other in a network.

4.3 Results

We consider the de novo designed single step Kemp elimination reaction in the reported designed construct KE07[21] that involved multiple rationally positioned mutations of a TIM barrel scaffold needed to manifest a biologically active enzyme. Fig. 4.2a shows the original computational design of KE07 that introduced 13 mutations into the thermostable scaffold imidazole-3-glycerolphosphate synthase (PDB: 1THF). The design involved the introduction of the catalytic triad including Glu101 acting as the catalytic base that performs the proton abstraction from 5-nitrobenzisoxazole, Trp50 that stabilizes the positioning of the ligand through aromatic interactions, and Lys222 that stabilizes the negative charge on the ligand of the transition state complex. While the original design intended residues Trp50, Glu101 and Lys222 to act as the catalytic triad, we have previously found that



(a) Original designed KE07

(b) KE07 Catalytic Site

(c) Best Performing KE07 Variant R7.2

Figure 4.2: *Relevant active site residues and LDE mutations.* (a) Original designed KE07. Designed catalytic residues (50, 101, 222) in red, designed scaffold residues in blue. (b) Original design proposed Glu101 (blue) for proton extraction, Trp50 (red) for steric positioning and pi-cloud delocalization, and Lys222 (green) to stabilize the negative charge being formed on the oxygen. However, we observe that His201 (orange) is closer and probably performs that function of Lys222[99]. (c) Best performing evolved KE07 variant (R7.2) Catalytic residues in red, mutations from original design in yellow. Ligand in magenta in both.

His201 actually fulfills the hydrogen-bonding donor role of stabilizing the ligand O^- charge (Fig. 4.2b); such a change in the catalytic residue involved in catalysis has been observed in another Kemp Eliminase by Hilvert et al[112].

The original design was further optimized with LDE, which randomly introduced 10 additional satellite mutations[22], as shown in Fig. 4.2c (7 of which appeared in at least two DE rounds) and a single mutation of a designed active site residue that was changed in all rounds of mutagenesis (N224D). A majority of the mutations are located either on surface loops or at the boundaries between the loops on either the top or bottom of the enzyme. LDE primarily yielded improvements in the rate of catalysis, k_{cat}, by $\sim$75 fold, with little optimization of the binding affinity, K_M, which shows only $\sim$2.5 fold improvement between KE07 and LDE R7.2 (1.4 uM to 0.54 uM). Overall $\frac{k_{cat}}{K_M}$ improved by a factor of $\sim$200, exhibiting the necessity of combining the de novo computational design with LDE. The enzyme structure is found to be very similar from round to round of LDE, with an RMSD < 0.5 $\mathring{A}$ with respect to the original designed enyzme for the crystallized mutants (R4, R6, R7.1 apo). While the enzyme showed decrease in stability as measured by a decrease in melting temperature, the LDE screening assay is for the most active mutant, not the most stable mutant. Therefore we analyze the dynamical couplings of the scaffold residues to the catalytic triad residues Eq. 4.3 in the designed enzyme relative to the best LDE enzyme, R7.2.

First we study dynamical correlations of all possible pairs of scaffold residues A and C with respect to each catalytic residue in the catalytic triad, Cat. In particular these correlations are

Table 4.1: The residues identified in the networks are outlined for the ligand bound states.

Catalytic Pair	**Residues Identified in Networks**	
	Design	**R7.2**
50 101	48, 50	176, 202
50 201	200, 2, 195	54, 84
101 201	2, 195	176, 81, 177

calculated as the DD-TCFs

$$C(t) = \langle D_{A-Cat}(t) \cdot D_{Cat-C}(t+\tau) \rangle \tag{4.5}$$

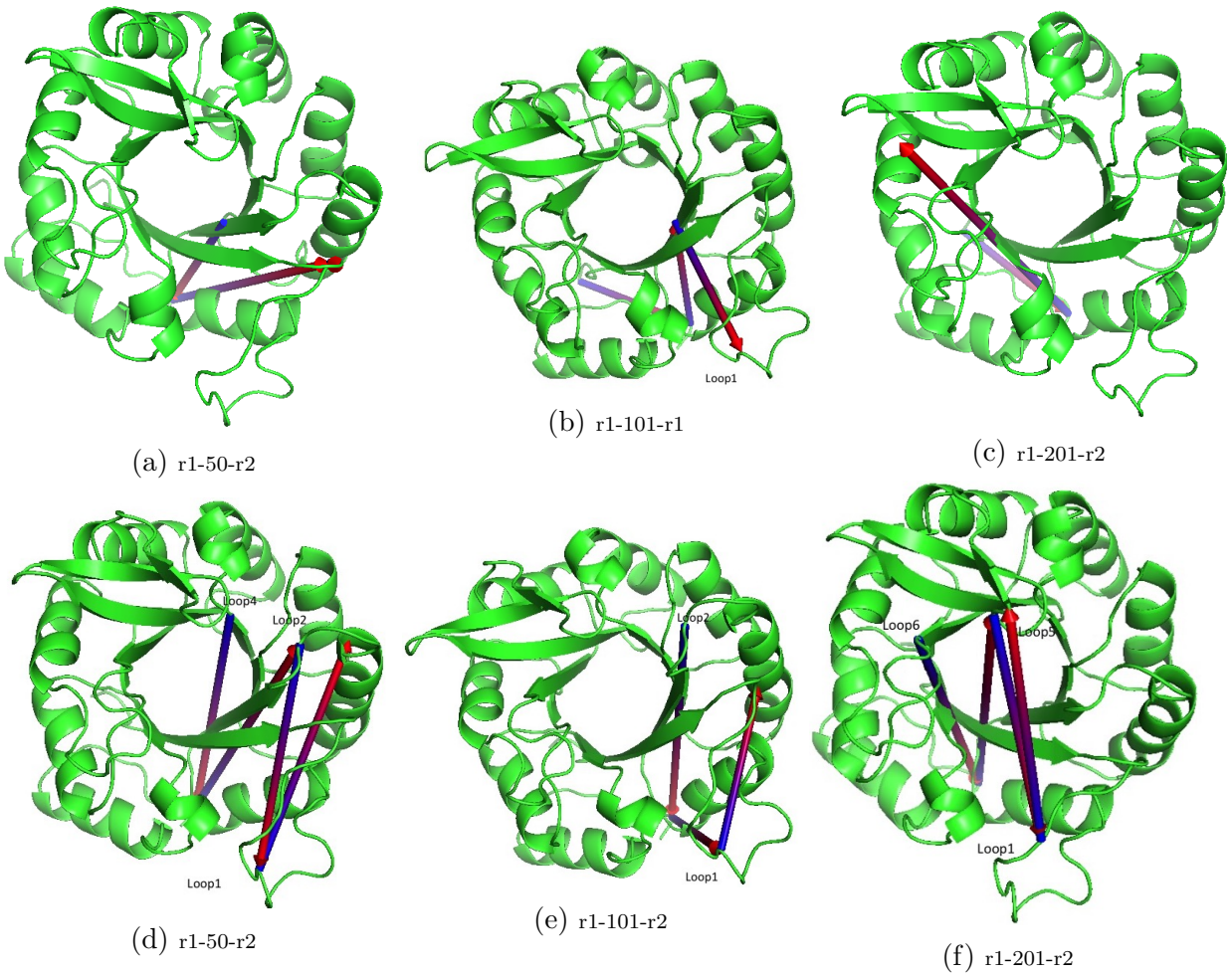
The entire set of residue-residue correlation matrices have been used to construct inter-residue trees based on the standard UPGMA approach (see method section), creating trees that give an ordered sequence of residues that are closely connected to each other through their degree of dynamic coupling. Fig. 4.3 shows that new highly correlated motions between the catalytic center and loops Loop1, Loop2, Loop3, and Loop6 are evident in the best LDE enzyme R7.2 that were not significantly observed in the original designed enzyme in the ligand bound state. These important correlated loop motions from the original scaffold are broken by the design process, which introduces 13 design mutations in the enzyme, including one in Loop6 (Asp176Leu) and one at the start of Loop2 (Leu50Trp). These strongly correlated loop motions appear necessary for holding the ligand in place in the active site within the central β-barrel. We found that out of 8 independent trajectories of the original designed enzyme, starting from the same initial structure with different randomized velocities, 2 of the trajectories showed the ligand became unbound in the active site located in the central b-barrel. However in R7.2 the ligand remained bound in all MD simulations. We suggest that some of the low catalytic activity is because the enzyme is unable to bind and hold the substrate in place as efficiently as the subsequent rounds of directed evolution, and that absent correlated loop motions further restrict the catalytic step.

Next we consider a definition of the DD-TCF and the corresponding power spectrum based on all KE07 residues, A, with the BC pairs of catalytic residues Trp50-Glu101, Trp50-His201, and Glu101-His201.

$$C(t) = \langle D_{A-Cat1}(t) \cdot D_{Cat1-Cat2}(t+\tau) \rangle \tag{4.6}$$

This correlation analysis is extended by considering a network of correlated motions that are identified by first selecting two catalytic residues, and identifying the most strongly correlated scaffold residue motion to this catalytic pair, and then determining the next scaffold residue whose motion is most strongly correlated with the motion of the previous scaffold residue and one of the catalytic residues. This is continued till the network is closed, and one of the residues already in the network is identified again. All networks we observe are of very short length (4-5 residues), but do show additional dynamically proximal clusters of residues highly correlated with the catalytic residues, again involving the TIM barrel loops for the high performing R7.2.

The most highly correlated networks are provided in Table I for the ligand bound states for the KE07 design and the best evolved R7.2 enzyme.



(a) r1-50-r2 (b) r1-101-r1 (c) r1-201-r2

(d) r1-50-r2 (e) r1-101-r2 (f) r1-201-r2

Figure 4.3: Residues with highly correlated motions w.r.t catalytic residues, selected based on distance on the tree constructed using magnitudes of correlation. (a-c)WT (d-f)R7.2

Again it is evident that the designed enzyme shows no correlated loop motions, whereas the network analysis again identifies 4 loops involved in correlated motions in R7.2. Unlike the first part of the analysis involving residue-residue correlations (Eq. 4.5), Loop1 is not identified in the network, but residue 202 of Loop 7 is identified instead. In this case Loop 6 and Loop 7 are identified as having correlated dynamics in R7.2. Another prominent cluster is comprised of residues 54, 84 which correlates Loop2 and Loop3, and finally cluster 81, 176, 177 that couples Loop 3 and Loop 6 (Fig. 4.4). Combining both analyses, we can identify 5 catalytic face loops involved in strongly correlated motions with respect to the active site in the best evolved enzyme that was destroyed in the original design.

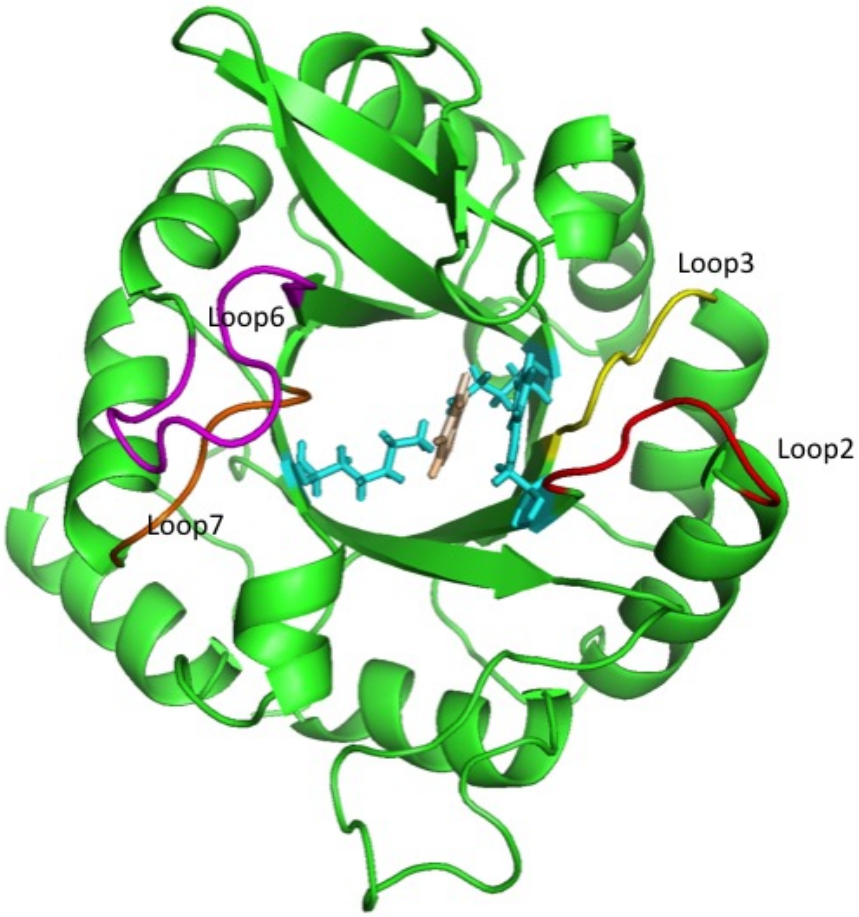


Figure 4.4: Key loops identified in R7.2 by a network analysis starting from catalytic residues and moving outwards finding the highest correlated residues for each subsequent selected pair. Catalytic residues (cyan), Ligand (wheat), Loop2 (red), Loop3(yellow), Loop6(magenta), Loop7(orange)

4.4 Conclusion

In this work we have introduced a distance-distance time correlation function analysis to show that the original *de novo* design of the Kemp Eliminase enzyme KE07 [21] destroyed vital scaffold motions that are correlated with the active site residues. However we find that these dynamical motions are restored by the changes in sequence that are introduced through multiple rounds of LDE, which correlates with the observed improvements in functional performance for KE07 [21–23]. In particular, correlated motions of the surface loops of the TIM barrel scaffold ubiquitously known to be important for catalysis of all naturally occurring enzymes based on the TIM barrel fold, are disrupted by the design mutations introduced in the WT design of KE07. However, LDE mutations are shown to restore the natural benefits of correlations in the best performing variant, suggesting that, like the original scaffold, those loop motions are important for catalysis,

in particular holding the ligand in place, and aiding in the catalytic step. The results offer the possibility that computational methods might not only provide good de novo enzyme leads, but take more direct part in their optimization, by bridging the structure-function relationship to more fully embrace the role of dynamics and flexibility for designing novel biocatalysis constructs.

Bibliography

[1] Xiyun Zhang and K. N. Houk. "Why Enzymes Are Proficient Catalysts Beyond the Pauling Paradigm". In: *Acc. Chem. Res.* 38(5) (2005), pp. 379–385.

[2] Arieh Warshel et al. "Electrostatic Basis for Enzyme Catalysis". In: *Chem. Rev.* 106 (2006), pp. 3210–3235.

[3] Zachary D. Nagel and Judith P. Klinman. "A 21st century revisionist's view at a turning point in enzymology". In: *Nat. Chem. Biol.* 5 (2009), pp. 543–550.

[4] Pratul K. Agarwal et al. "Network of coupled promoting motions in enzyme catalysis". In: *Proc. Nat. Acad. Sci.* 99(5) (2002), pp. 2794–2799.

[5] Nicholas Boekelheide, Romelia Salomon-Ferrer, and Thomas F. III Miller. "Dynamics and dissipation in enzyme catalysis". In: *Proc. Nat. Acad. Sci.* 108(39) (2011), 1615916163.

[6] Sharon Hammes-Schiffer and Stephen J. Benkovic. "Relating Protein Motion to Catalysis". In: *Ann. Rev. Biochem.* 75 (2006), pp. 519–541.

[7] Dynamic personalities of proteins. "Henzler-Wildman, Katherine and Kern, Dorothee". In: *Nature* 450 (2007), pp. 964–972.

[8] Vishal C. Nashine, Sharon Hammes-Schiffer, and Stephen J. Benkovic. "Coupled motions in enzyme catalysis". In: *Curr. Opin. Chem. Biol.* 14 (2010), pp. 644–651.

[9] Jennifer L. Radkiewicz and Charles L. III Brookes. "Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase". In: *J. Am. Chem. Soc.* 122 (2000), pp. 225–231.

[10] H. Rod Thomas, Jennifer L. Radkiewicz, and Charles L. III Brookes. "Correlated motion and the effect of distal mutations in dihydrofolate reductase". In: *Proc. Nat. Acad. Sci.* 100(12) (2003), pp. 6980–6985.

[11] Jory Z. Ruscio et al. "The Influence of Protein Dynamics on the Success of Computational Enzyme Design". In: *J. Am. Chem. Soc.* 131 (2009), 1411114115.

[12] Audrey Tousignant and N. Pelletier Joelle. "Protein Motions Promote Catalysis". In: *Chem. & Biol.* 11 (2004), pp. 1037–1042.

[13] Yan Wang, Rebecca B. Berlow, and Patrick J. Loria. "Role of Loop-Loop Interactions in Coordinating Motions and Enzymatic Function in Triosephosphate Isomerase". In: *Biochemistry* 48 (2009), pp. 4548–4556.

[14] Kim F. Wong et al. "Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase". In: *Proc. Nat. Acad. Sci.* 102(19) (2005), pp. 6807–6812.

[15] P. W. Fenimore et al. "Bulk-solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and fluctuations". In: *Proc. Nat. Acad. Sci.* 101(40) (2004), pp. 14408–14413.

[16] H. Frauenfelder et al. "Protein folding is slaved to solvent motions". In: *Proc. Nat. Acad. Sci.* 103(42) (2006), pp. 15469–15472.

[17] Simon Ebbinghaus et al. "An extended dynamical hydration shell around proteins". In: *Proc. Nat. Acad. Sci.* 104(52) (2007), pp. 20749–20752.

[18] Margaret E. Johnson et al. "Hydration Water Dynamics Near Biological Interfaces". In: *J. Phys. Chem. B* 113 (2009), pp. 4082–4092.

[19] Matthias Heyden and Martina Havenith. "Combining THz Spectroscopy and MD simulations to study protein-hydration coupling". In: *Methods* 52 (2010), pp. 74–83.

[20] Jian Sun et al. "Understanding THz Spectra of Aqueous Solutions: Glycine in Light and Heavy Water". In: *J. Am. Chem. Soc.* 136 (2014), pp. 5031–5038.

[21] D. Röthlisberger et al. "Kemp elimination catalysis by computational enzyme design". In: *Nature* 453 (2008), pp. 190–195.

[22] Olga Khersonsky et al. "Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series". In: *J. Mol. Biol.* 396 (2010), pp. 1025–1042.

[23] Olga Khersonsky et al. "Optimization of the *In-Silico*-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution". In: *J. Mol. Biol.* 407 (2011), pp. 391–412.

[24] Alexander* Esser et al. "Mode specific THz spectra of solvated amino acids using the AMOEBA polarizable force field". In: *Phys. Chem. Chem. Phys.* 19 (2017), pp. 5579–5590.

[25] U. Heugen et al. "Solute-induced retardation of water dynamics probed directly by terahertz spectroscopy". In: *Proc. Nat. Acad. Sci.* 103 (2006), pp. 12301–12306.

[26] Chenfeng Zhang and Stephen M. Durbin. "Hydration-Induced Far-Infrared Absorption Increase in Myoglobin". In: *J. Phys. Chem. B* 110 (2006), pp. 23607–23613.

[27] Ismael A. Heisler and Stephen R. Meech. "Low-Frequency Modes of Aqueous Alkali Halide Solutions: Glimpsing the Hydrogen Bonding Vibration". In: *Science* 327.5967 (2010), pp. 857–860.

[28] K. J. Tielrooij et al. "Cooperativity in Ion Hydration". In: *Science* 328.5981 (2010), pp. 1006–1009.

[29] K. J. Tielrooij et al. "Anisotropic Water Reorientation around Ions". In: *J. Phys. Chem. B* 115.43 (2011), pp. 12638–12647. DOI: 10.1021/jp206320f.

[30] Moran Grossman et al. "Correlated structural kinetics and retarded solvent dynamics at the metalloprotease active site". In: *Nat. Struct. Mol. Biol.* 18 (2011), pp. 1102–1108.

[31] D. Marx and J. Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, Cambridge, 2009.

[32] Matthias Heyden et al. "Dissecting the THz spectrum of liquid water from first principles via correlations in time and space". In: *Proc. Nat. Acad. Sci.* 107 (2010), pp. 12068–12073.

[33] M. Heyden et al. "Understanding the Origins of Dipolar Couplings and Correlated Motion in the Vibrational Spectrum of Wat". In: *J. Phys. Chem. Lett.* 3 (2012), pp. 2135–2140.

[34] Maciej Śmiechowski, Harald Forbert, and Dominik Marx. "Spatial decomposition and assignment of infrared spectra of simple ions in water from mid-infrared to THz frequencies: Li+(aq) and F(aq)". In: 139.1, 014506 (2013). URL: http://scitation.aip.org/content/aip/journal/jcp/139/1/10.1063/1.4812396.

[35] Maciej Śmiechowski et al. "Solvation shell resolved THz spectra of simple aqua ions - distinct distance- and frequency-dependent contributions of solvation shells". In: 17 (13 2015), pp. 8323–8329.

[36] S. D. Ivanov, A. Witt, and D. Marx. "Theoretical spectroscopy using molecular dynamics". In: *Phys. Chem. Chem. Phys.* 15 (2013), pp. 10270–10299.

[37] Michiel Sprik and Michael L. Klein. "A polarizable model for water using distributed charge sites". In: *J. Chem. Phys.* 89 (1988), p. 7556.

[38] Daniel Van Belle et al. "Molecular dynamics simulation of polarizable water by an extended Lagrangian method". In: *Mol. Phys.* 77 (1992), pp. 239–255.

[39] Jay W. Ponder and David A. Case. "Force fields for protein simulations". In: *Adv. Prot. Chem.* 66 (2003), pp. 27–85.

[40] Guillaume Lamoureux, Alexander D. MaxKerell Jr., and Benoit Roux. "A simple polarizable model of water based on classical Drude oscillators". In: *J. Chem. Phys.* 119 (2003), p. 5185.

[41] Pengyu Ren and Jay P. Ponder. "Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation". In: *J. Phys Chem. B* 107 (2003), pp. 5933 –5947.

[42] George A. Kaminski et al. "Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from ab Initio Quantum Chemistry". In: *J. Phys. Chem. A* 108 (2004), pp. 621–627.

[43] S Patel and Charles L. III Brookes. "CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations". In: *J. Comput. Chem.* 25 (2004), pp. 1–15.

[44] Edward Harder et al. "Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field based on the classical Drude oscillator". In: *J. Chem. Theory Comput.* 2.6 (2006), pp. 1587–1597.

[45] Jean-Philip Piquemal, Hilaire Chevreau, and Nohad Gresh. "Toward a separate reproduction of the contributions to the Hartree-Fock and DFT intermolecular interaction energies by polarizable molecular mechanics with the SIBFA potential". In: *J. Chem. Theory Comput.* 3.3 (2007), pp. 824–837.

[46] Jay W Ponder et al. "Current status of the AMOEBA polarizable force field". In: *J. Phys. Chem. B* 114.8 (2010), pp. 2549–2564.

[47] Revati Kumar et al. "A second generation distributed point polarizable water model". In: *J. Chem. Phys.* 132.1 (2010), p. 014309.

[48] Pedro E. M. Lopes et al. "Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator". In: *J. Chem. Theo. Comput.* 9 (2013), pp. 5430–5449.

[49] O. N. Demerdash, E. H. Yap, and Teresa Head-Gordon. "Advanced potential energy surfaces for condensed phase simulation". In: *Annu. Rev. Phys. Chem.* 65 (2014), pp. 149–174.

[50] Alex Albaugh et al. "Advanced Potential Energy Surfaces for Molecular Simulation". In: *J. Phys. Chem. B* 120(37) (2016).

[51] Pengyu Ren, Chuanjie Wu, and Jay W. Ponder. "Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules". In: *J. Chem. Theo. Comp.* 7 (2011), pp. 3143–3161.

[52] Lee-Ping Wang et al. "Systematic Improvement of a Classical Molecular Model of Water". In: *J. Phys. Chem. B* 117 (2013), pp. 9956–9972.

[53] Marie L. Laury et al. "Revised Parameters for the AMOEBA Polarizable Atomic Multipole Water Model". In: *J. Phys. Chem. B* 119 (2015), pp. 9423–9437.

[54] Maciej Śmiechowski et al. "Correlated Particle Motion and THz Spectral Response of Supercritical Water". In: 116 (2 2016), p. 027801. DOI: 10.1103/PhysRevLett.116.027801. URL: http://link.aps.org/doi/10.1103/PhysRevLett.116.027801.

[55] Paul N Day et al. "An effective fragment method for modeling solvent effects in quantum mechanical calculations". In: *J. Chem. Phys.* 105.5 (1996), pp. 1968–1986.

[56] Claude Millot et al. "Revised anisotropic site potentials for the water dimer and calculated properties". In: *J. Phys. Chem. A* 102.4 (1998), pp. 754–770.

[57] Mark A. Freitag et al. "Evaluation of charge penetration between distributed multipolar expansions". In: *J. Chem. Phys.* 112.17 (2000), pp. 7300–7306.

[58] J.-P. Piquemal, N. Gresh, and C. Giessner-Prettre. "Improved formulas for the calculation of the electrostatic contribution to the intermolecular interaction energy from multipolar expansion of the electronic distribution". In: *J. Phys. Chem. A* 107.48 (2003), pp. 10353–10359.

[59] Riccardo Chelli et al. "Polarization response of water and methanol investigated by a polarizable force field and density functional theory calculations: Implications for charge transfer". In: *J. Chem. Phys.* 122.7 (2005), p. 074504.

[60] Lyudmila.V. Slipchenko and Mark S. Gordon. "Electrostatic energy in the effective fragment potential method: Theory and application to benzene dimer". In: *J. Comput. Chem.* 28.1 (2007), pp. 276–291.

[61] Anthony J Stone. "Electrostatic damping functions and the penetration energy". In: *J. Phys. Chem. A* 115.25 (2011), pp. 7017–7027.

[62] Bo Wang and Donald G Truhlar. "Screened electrostatic interactions in molecular mechanics". In: *J. Chem. Theory Comput.* 10.10 (2014), pp. 4480–4487.

[63] Qiantao Wang et al. "General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field". In: *J. Chem. Theo. Comp.* 11.6 (2015), pp. 2609–2618.

[64] Christophe Narth et al. "Scalable improvement of SPME multipolar electrostatics in anisotropic polarizable molecular mechanics using a short-range penetration correction up to quadrupoles". In: *J. Comp. Chem.* 37.5 (2016), pp. 494–506.

[65] Alexis J. Lee and Steven W. Rick. "The effects of charge transfer on the properties of liquid water". In: *J. Chem. Phys.* 134.18 (2011), p. 184507.

[66] Marielle Soniat and Steven W. Rick. "The effects of charge transfer on the aqueous solvation of ions". In: *J. Chem. Phys.* 137.4 (2012), p. 044511.

[67] Marielle Soniat and Steven W. Rick. "Charge transfer effects of ions at the liquid water/vapor interface". In: *J. Chem. Phys.* 140.18 (2014), p. 184703.

[68] Marielle Soniat and Steven W. Rick. "Charge transfer models of zinc and magnesium in water". In: *J. Chem. Theo. Comp.* 11.4 (2015), pp. 1658–1667.

[69] Marielle Soniat, Revati Kumar, and Steven W. Rick. "Hydrated proton and hydroxide charge transfer at the liquid/vapor interface of water". In: *J. Chem. Phys.* 143.4 (2015), p. 044702.

[70] Marielle Soniat et al. "Ion association in aqueous solution". In: *Fluid Phase Equilib.* 407 (2016), pp. 31–38.

[71] A. J. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, 1996.

[72] B. T. Thole. "Molecular polarizabilities calculated with a modified dipole interaction". In: *Chem. Phys.* 59 (1981), pp. 341–350.

[73] Thomas A. Halgren. "The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters". In: *J. Am. Chem. Soc.* 114 (1992), pp. 7827–7843.

[74] Yue Shi et al. "Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins". In: *J. Chem. Theo. Comp.* 9 (2013), pp. 4046–4063.

[75] Lee-Ping Wang, Jiahao Chen, and Troy Van Voorhis. "Systematic parameterization of polarizable force fields from quantum chemistry data". In: *J. Chem. Theo. Comp.* 9 (2013), pp. 452–460.

[76] Yihan Shao et al. "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package". In: *Mol. Phys.* 113.2 (2015), pp. 184–215. ISSN: 0026-8976. DOI: 10.1080/00268976.2014.952696. URL: http://www.tandfonline.com/doi/abs/10.1080/00268976.2014.952696.

[77] A. J Stone. "Distributed Multipole Analysis: Stability for Large Basis Sets". In: *J. Chem. Theo. Comp.* 1 (2005), pp. 1128–1132.

[78] M. J. Frisch et al. "Gaussian09 Revision E.01". In: (). Gaussian Inc. Wallingford CT 2009.

[79] E. Scoppola et al. "Water-peptide site-specific interactions: a structural study on the hydration of glutathione". In: *Biophys. J.* 106 (2014), pp. 1701–1709.

[80] J. VandeVondele et al. In: *Comp. Phys. Comm.* 167 (2005), pp. 103–128.

[81] J. Sun et al. "Glycine in aqueous solution: solvation shells, interfacial water, and vibrational spectroscopy from ab initio molecular dynamics". In: *J. Chem. Phys.* 133 (2012), pp. 114508–1–10.

[82] Rafael Ramírez et al. "Quantum corrections to classical time-correlation functions: Hydrogen bonding and anharmonic floppy modes". In: 121.9 (2004), pp. 3973–3983. DOI: http://dx.doi.org/10.1063/1.1774986. URL: http://scitation.aip.org/content/aip/journal/jcp/121/9/10.1063/1.1774986.

[83] S. D. Ivanov et al. "Quantum–induced symmetry breaking explains IR spectra of CH_5^+ isotopologues". In: *Nat. Chem.* 2 (2010), pp. 298–302.

[84] Gerald Mathias and Marcel D. Baer. "Generalized Normal Coordinates for the Vibrational Analysis of Molecular Dynamics Simulations". In: *J. Chem. Theo. Comp.* 7 (2011), pp. 2028–2039.

[85] Gerald Mathias et al. "Infrared Spectroscopy of Fluxional Molecules from (ab Initio) Molecular Dynamics: Resolving Large-Amplitude Motion, Multiple Conformations, and Permutational Symmetries". In: *J. Chem. Theo. Comp.* 8 (2012), pp. 224–234.

[86] Eloy Ramos-Cordoba, Daniel S. Lambrecht, and Martin Head-Gordon. "Charge-transfer and the hydrogen bond: Spectroscopic and structural implications from electronic structure calculations". In: *Farad. Discuss.* 150 (2011), pp. 345–362.

[87] Asmit Bhowmick, Sudhir C. Sharma, and Teresa Head-Gordon. "The Importance of the Scaffold for de Novo Enzymes: A Case Study with Kemp Eliminase". In: *J. Am. Chem. Soc.* (2017).

[88] Matthew J. DiTucci, Sven Heiles, and Evan R. Williams. "Role of Water in Stabilizing Ferricyanide Trianion and Ion-Induced Effects to Hydrogen-Bonding Water Network at Long Distance". In: *J. Am. Chem. Soc.* 137(4) (2015), pp. 1650–1657.

[89] Shiang-Tai Lin, Mario Blanco, and William A. III Goddard. "The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids". In: *J. Chem. Phys.* 119(22) (2003), pp. 11792–11805.

[90] Shiang-Tai Lin, Prabal K. Maiti, and William A. III Goddard. "Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations." In: *J. Phys. Chem. B.* 114(24) (2010), pp. 8191–8198.

[91] A. A. McQuarrie. *Statistical Mechanics*. Harper & Row, New York, 1976.

[92] Viren Pattni et al. "Distinct protein hydration water species defined by spatially resolved spectra of intermolecular vibrations". In: *Submitted* ().

[93] Rasmus Persson et al. "Signatures of solvation thermodynamics in spectra of intermolecular vibrations". In: *Submitted* ().

[94] Pengyu Ren, Chuanjie Wu, and Jay W. Ponder. "Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules". In: *J. Chem. Theo. Comp.* 7 (2011), pp. 3143–3161.

[95] Jay W. Ponder. *Tinker–Software Tools for Molecular Design*. Tinker 7.1. St. Louis, MO: Washington University, 2015.

[96] B. Webb and A. Sali. "Comparative Protein Structure Modeling Using Modeller". In: *Current Protocols in Bioinformatics* 5.6 (2014), pp. 1–32.

[97] Yue Shi et al. "Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins". In: *J. Chem. Theo. Comp.* 9 (2013), pp. 4046–4063.

[98] P. Eastman et al. "OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation". In: *J. Chem. Theor. Comput.* 9(1) (2013), pp. 461–469.

[99] Asmit Bhowmick et al. "The role of side chain entropy and mutual information for improving the *de novo* design of Kemp eliminases KE07 and KE70". In: *Phys. Chem. Chem. Phys.* 18 (2016), pp. 19386–19396.

[100] A. N. Alexandrova et al. "Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination." In: *J. Am. Chem. Soc.* 130(47) (2008), pp. 15907–15915.

[101] G. Kiss et al. "Evaluation and ranking of enzyme designs". In: *Protein Sci.* 19 (2010), pp. 1960–1773.

[102] Stefania Perticaroli et al. "Description of Hydration Water in Protein (Green Fluorescent Protein) Solution". In: *J. Am. Chem. Soc.* 139 (2017), pp. 1098–1105.

[103] Stavros Caratzoulas, Joshua S. Mincer, and Steven D. Schwartz. "Identification of a Protein-Promoting Vibration in the Reaction Catalyzed by Horse Liver Alcohol Dehydrogenase Stavros". In: *J. Am. Chem. Soc.* 124 (2002), pp. 3270–3276.

[104] Adriàn Ochoa-Leyva et al. "Protein Design through Systematic Catalytic Loop Exchange in the $(\beta/\alpha)8$ Fold". In: *J. Mol. Biol.* 387 (2009), pp. 949–964.

[105] Adriàn Ochoa-Leyva et al. "Exploring the StructureFunction Loop Adaptability of a $(\beta/\alpha)8$-Barrel Enzyme through Loop Swapping and Hinge Variability". In: *J. Mol. Biol.* 411 (2011), pp. 143–157.

[106] Sharon Rozovsky and Ann E. McDermott. "The Time scale of the Catalytic Loop Motion in Triosephosphate Isomerase". In: *J. Mol. Bio.* 310 (2001), pp. 259–270.

[107] David L. Pompliano, Anusch Peyman, and Jeremy R. Knowles. "Stabilization of a Reaction Intermediate as a Catalytic Device: Definition of the Functional Role of the Flexible Loop in Triosephosphate Isomerase". In: *Biochemistry* 29(13) (1990), pp. 3186–3194.

[108] John C. Williams and Ann E. McDermott. "Dynamics of the Flexible Loop of Triosephosphate Isomerase: The Loop Motion Is Not Ligand Gated". In: *Biochemistry* 34 (1995), pp. 8309–8319.

[109] Francesca Massi, Chunyu Wang, and Arthur G. III Palmer. "Solution NMR and Computer Simulation Studies of Active Site Loop Motion in Triosephosphate Isomerase". In: *Biochemistry* 45(36) (2006), pp. 10787–10794.

[110] Maria P. Frushicheva et al. "Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase". In: *Proc. Nat. Acad. Sci.* 107(39) (2010), pp. 16869–16874.

[111] Steve W. Lockless and Rama Ranganathan. "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families". In: *Science* 286 (1999), pp. 295–299.

[112] Rebecca Blomberg et al. "Precision is essential for efficient catalysis in an evolved Kemp eliminase". In: *Nature* 503(7476) (2013), pp. 418–421.

Appendix A

Supplementary information for Chapter 2

Structures

Figure A.1 shows the reference structures used for the mode decomposition where each atom is labeled.

Radial and Angular Distribution Functions

The radial distribution functions (RDFs) of valine in water show the radial arrangement of water molecules around the protonated amino and deprotonated carboxyl groups. Akin to the results shown for glycine in water in the main text the original AMOEBA parametrization leads to an

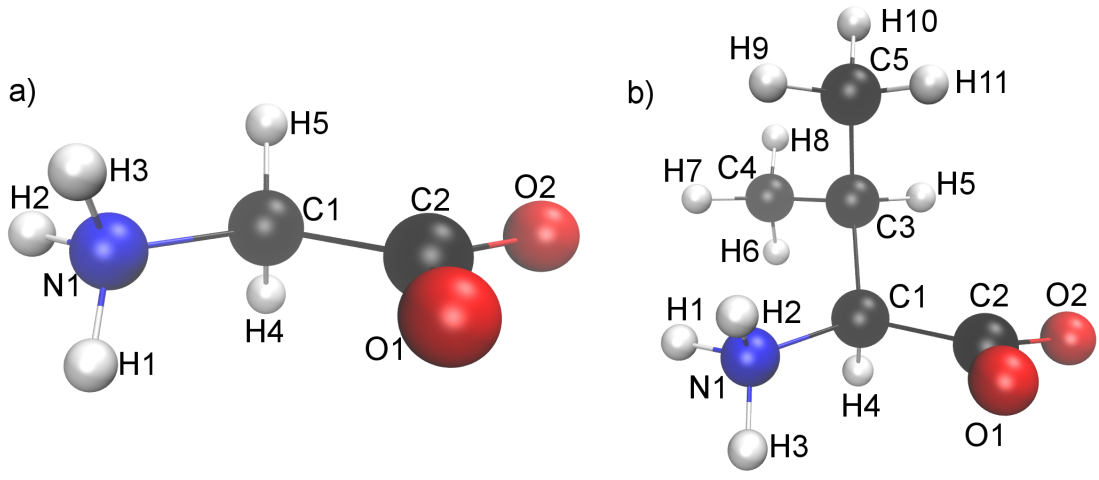


Figure A.1: Reference structures of (a) glycine-only and (b) valine-only atom labels.

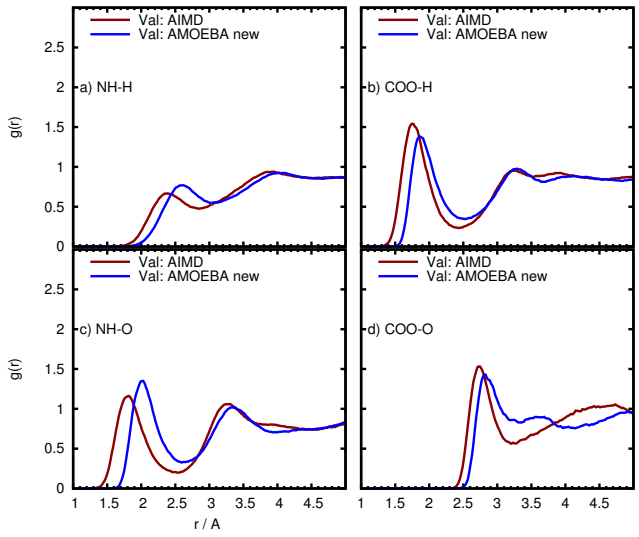


Figure A.2: Radial distribution functions between the protonated amino and deprotonated carboxyl groups of valine in water with respect to the water molecules.

outward shift compared to AIMD also in case of valine which is largely corrected when using the modified AMOEBA model that is shown here; note that the corresponding AIMD RDFs of glycine are overall in reasonable accord with experimental data as demonstrated in the main text and thus are expected to serve as benchmarks also for valine.

Angular distribution functions for glycine and valine are shown in Fig. A.3 and Fig. A.4, respectively. The angles as obtained from the original AMOEBA model nicely match for both glycine and valine those of the corresponding angles obtained from AIMD so that no optimization was carried out in the modified model.

The dihedral distribution functions are shown in Fig. A.5 for glycine and Fig. A.6 for valine. Overall the dihedral angles show similar tendencies to the ones from AIMD but are generally too stiff in AMOEBA.

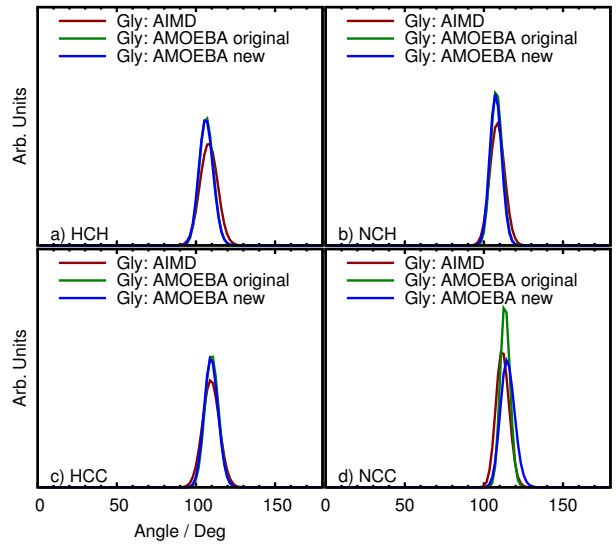


Figure A.3: Angular distribution functions for glycine. (a) H4-C1-H5, (b) N1-C1-H4/H5, (c) H4/H5-C1-C2, (d) N1-C1-C2.

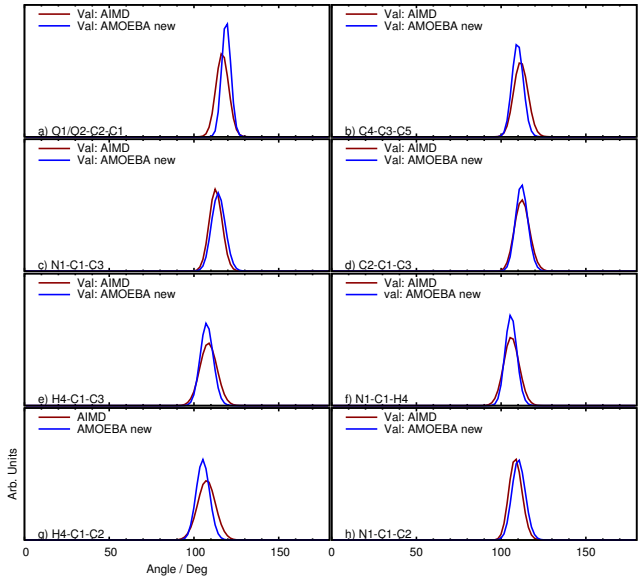


Figure A.4: Angular distribution functions for Valine. (a) O1/O2-C2-C1, (b) C4-C3-C5, (c) N1-C1-C3, (d) C2-C1-C3, (e) H4-C1-C3, (f) N1-C1-H4, (g) H4-C1-C2, (h) N1-C1-C2.

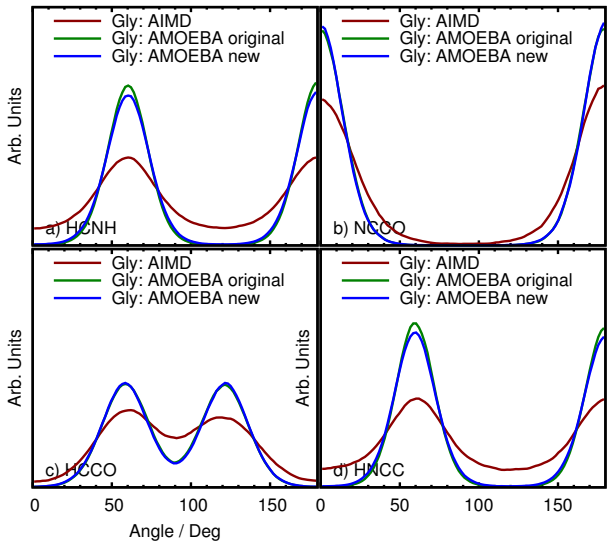


Figure A.5: Dihedral distribution functions for glycine. (a) HCNH, (b) NCCO, (c) HCCO, (d) HNCC.

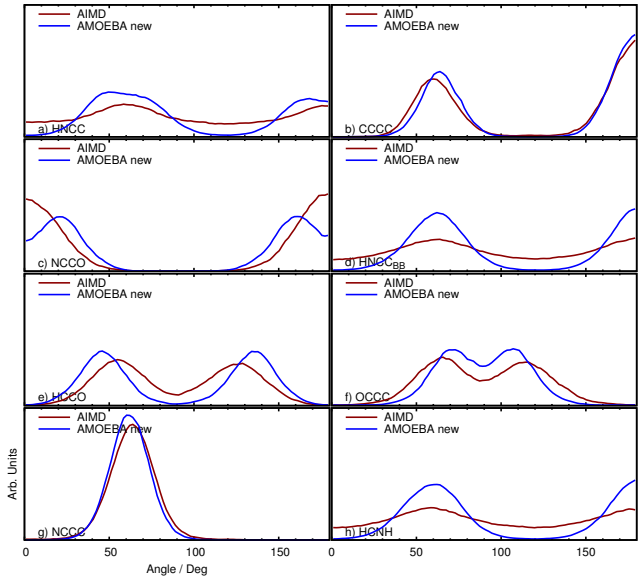


Figure A.6: Dihedral distribution functions for valine. (a) amide H and water H, (b) carboxyl O and water H, (c) amide H and water O, (d) carboxyl O and water O.

Modes of Glycine According to SSC(+) Analysis

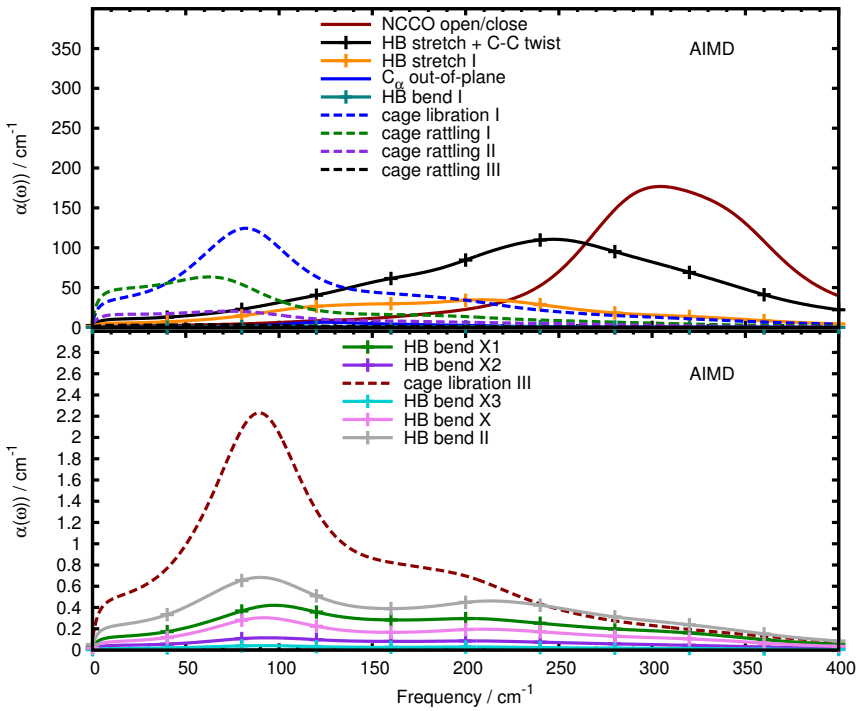


Figure A.7: THz mode intensities of glycine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AIMD solvated by 30 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

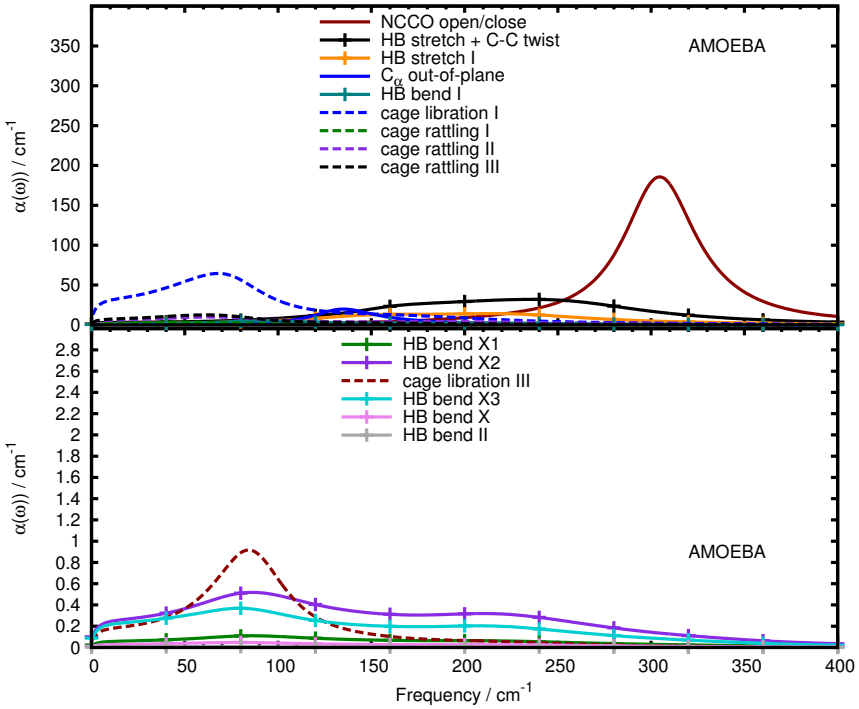


Figure A.8: THz mode intensities of glycine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AMOEBA solvated by 30 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

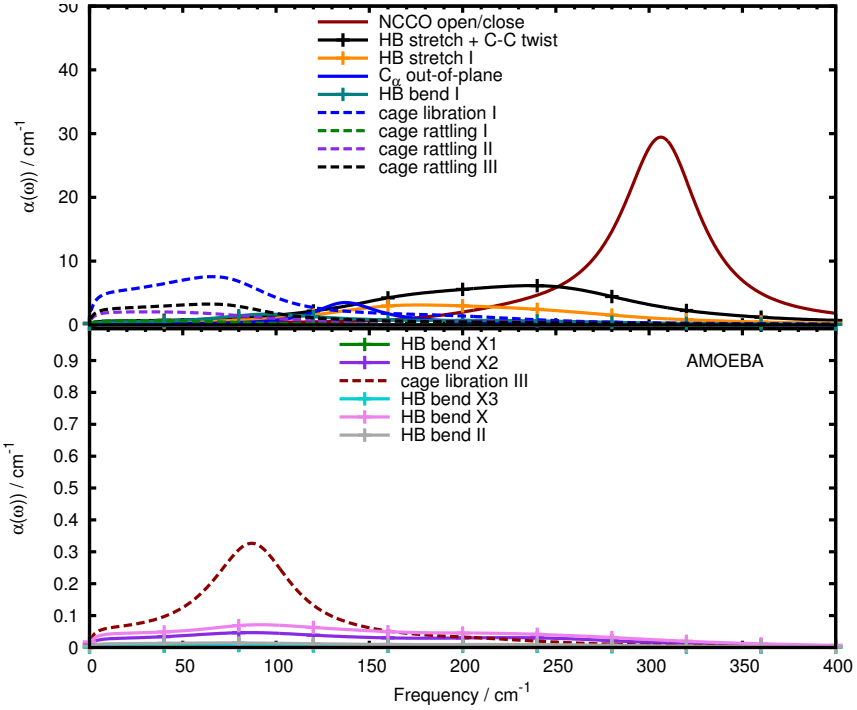


Figure A.9: THz mode intensities of glycine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AMOEBA solvated by 256 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

Modes of Valine According to SSC(+) Analysis

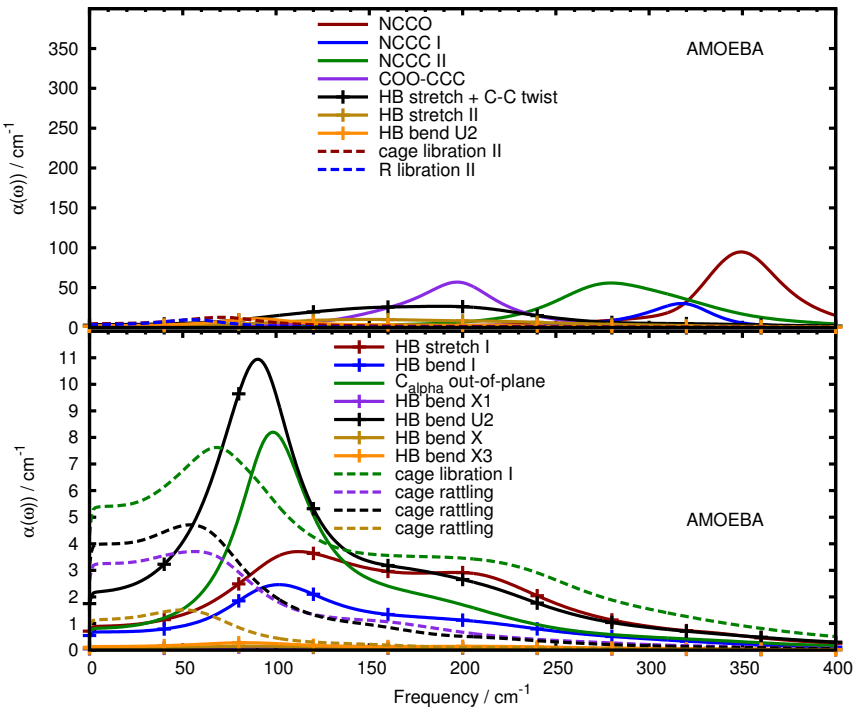


Figure A.10: THz mode intensities of valine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AIMD solvated by 60 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

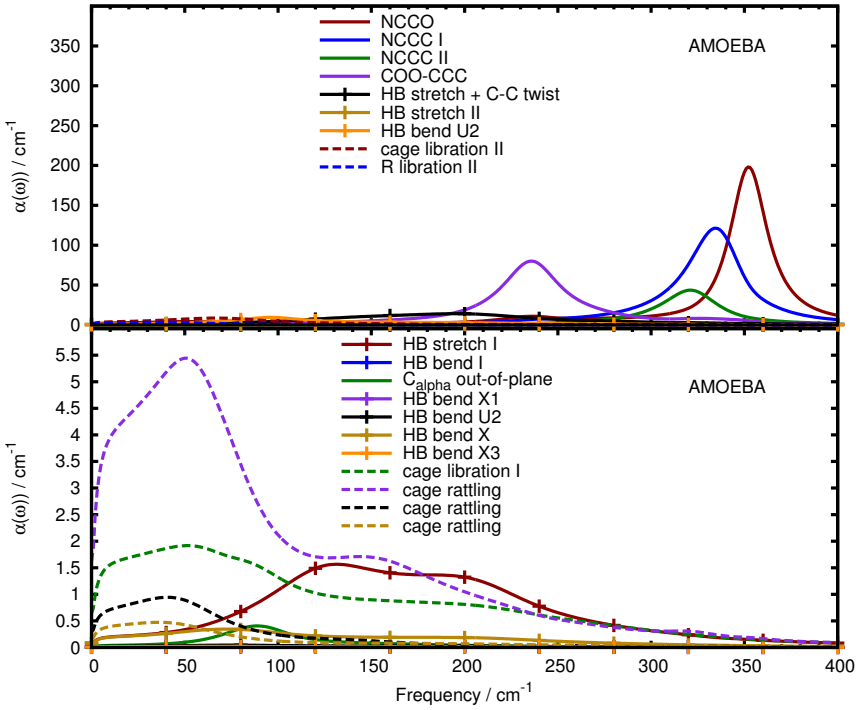


Figure A.11: THz mode intensities of valine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AMOEBA solvated by 60 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

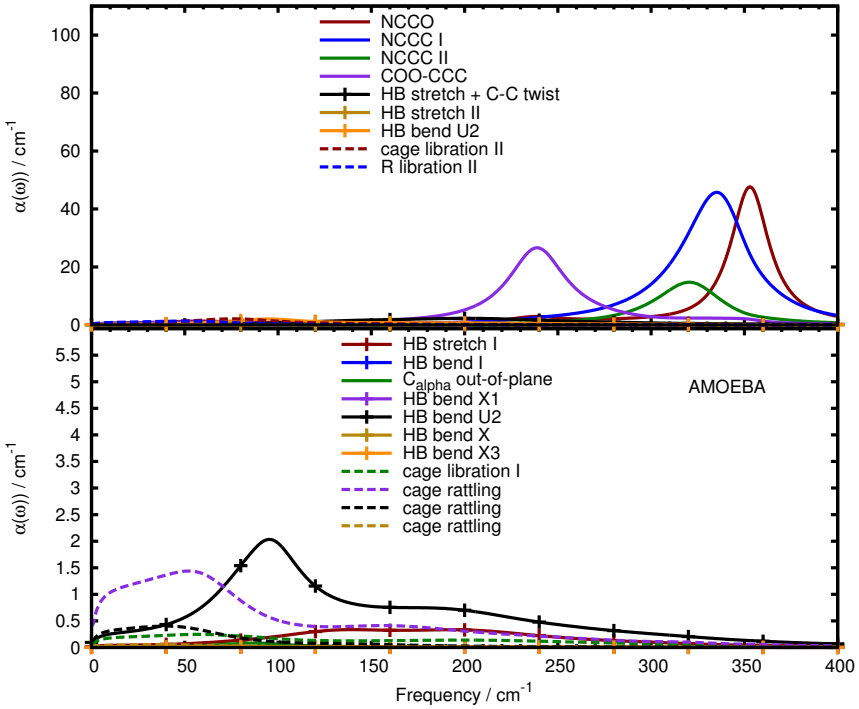
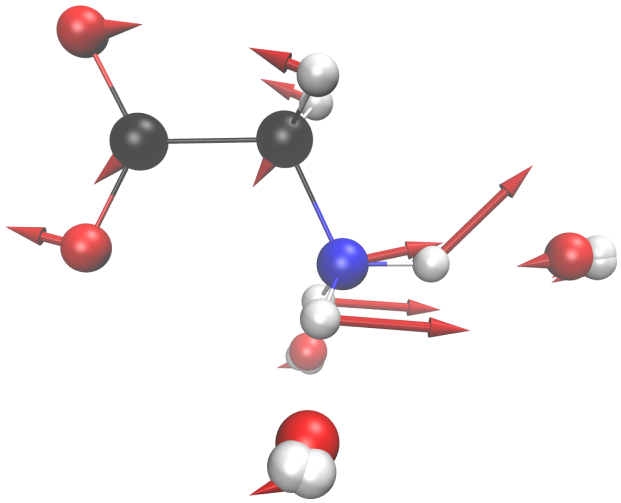
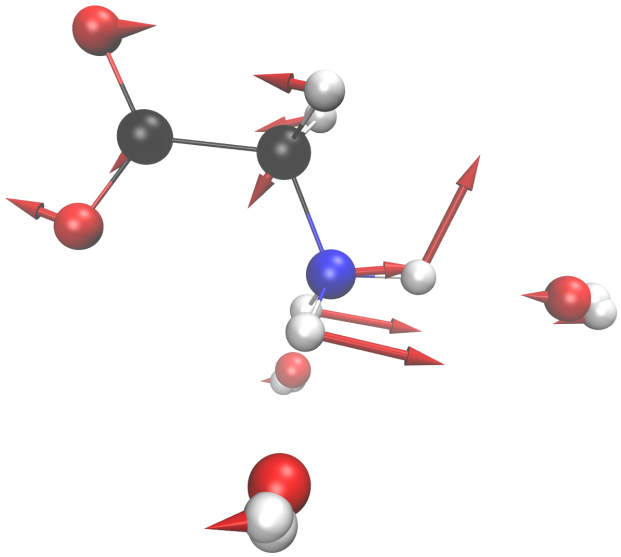


Figure A.12: THz mode intensities of valine within the supermolecular solvation complex at the amino group (SSC(+)) computed via AMOEBA solvated by 256 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

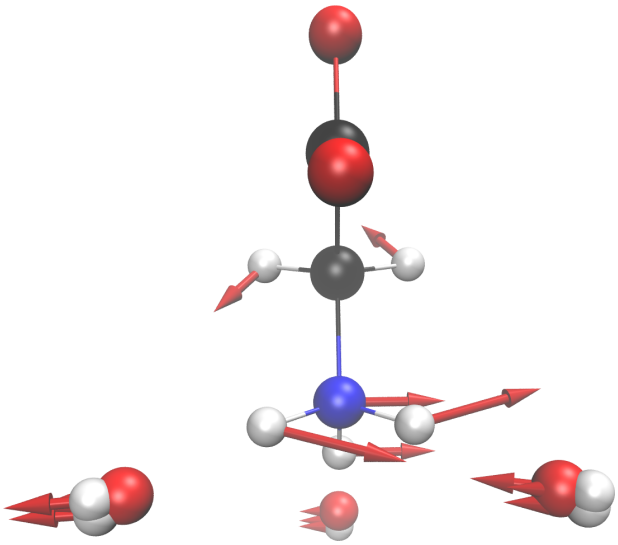
Mode Displacement Vectors of Glycine



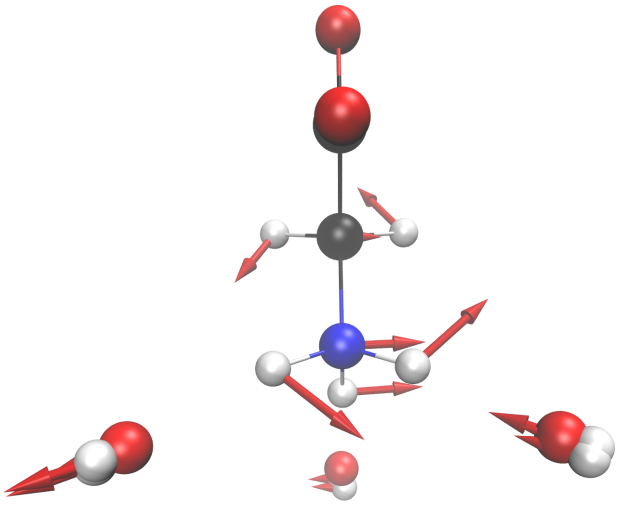
AIMD: NCCO 304 cm^{-1}



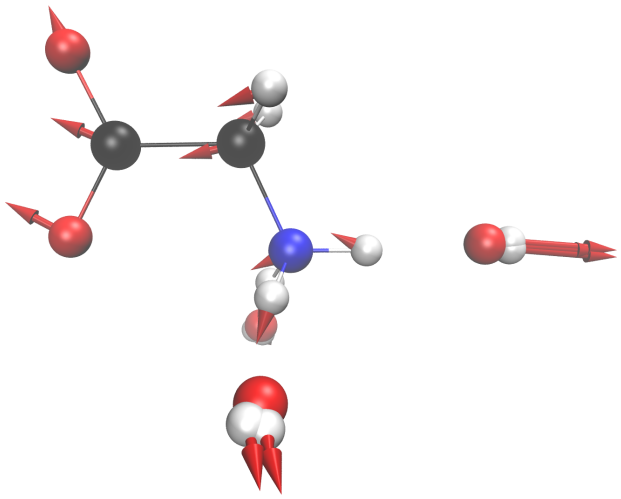
AMOEBA: NCCO 305 cm^{-1}



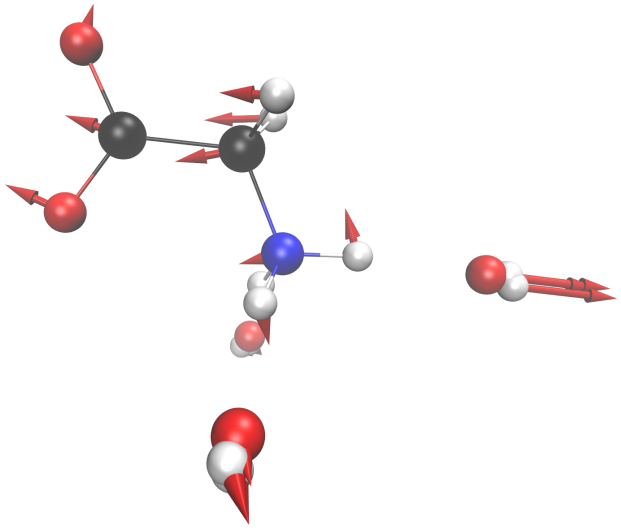
AIMD: HB-stretch-+-CC-Twist 247 cm^{-1}



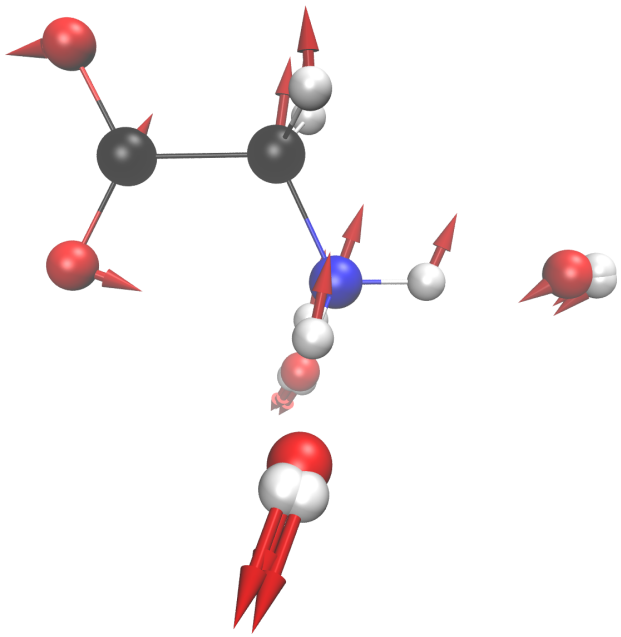
AMOEBA: HB-stretch-+-CCtwist 236 cm^{-1}



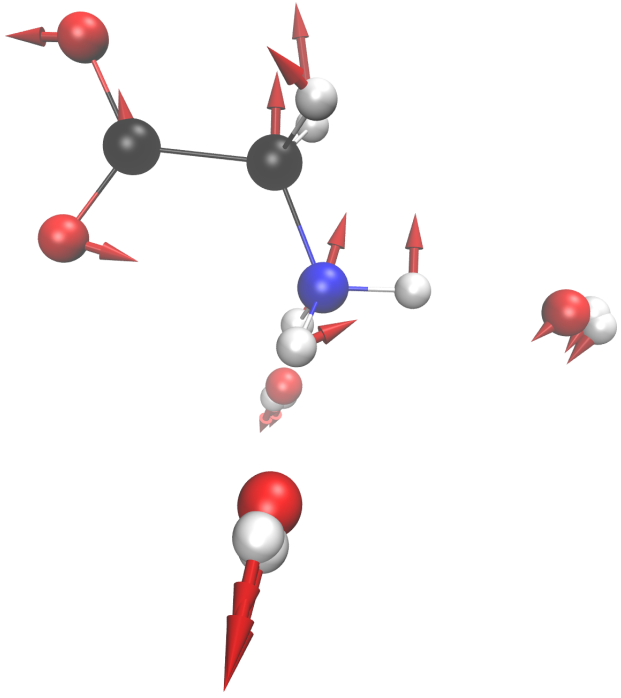
AIMD: HB-stretch-II 218 cm^{-1}



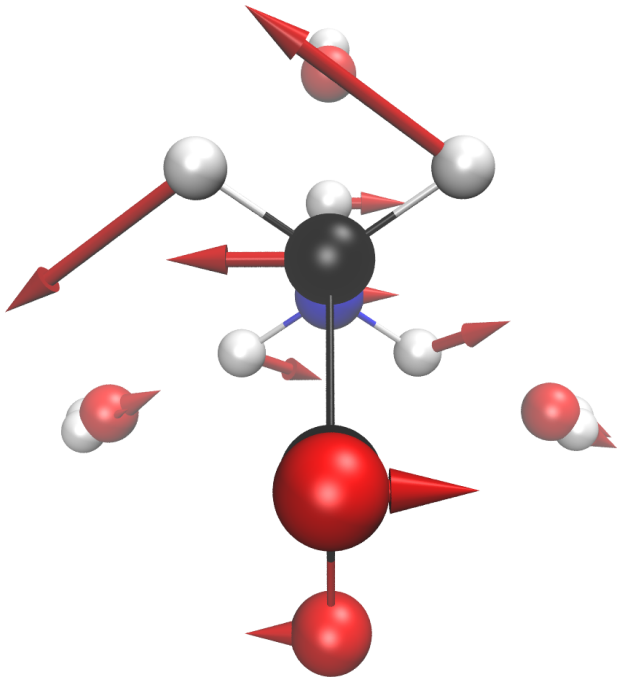
AMOEBA: HB-stretch-II 146 cm^{-1}



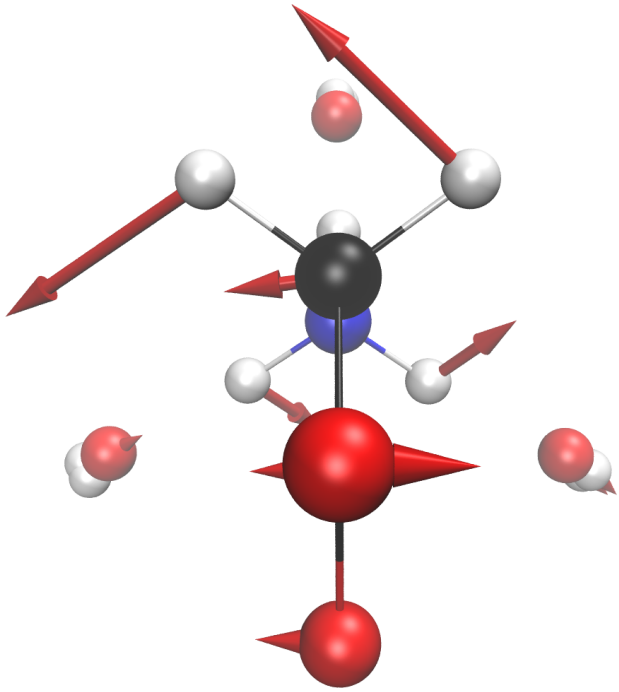
AIMD: HB-stretch-I 210 cm^{-1}



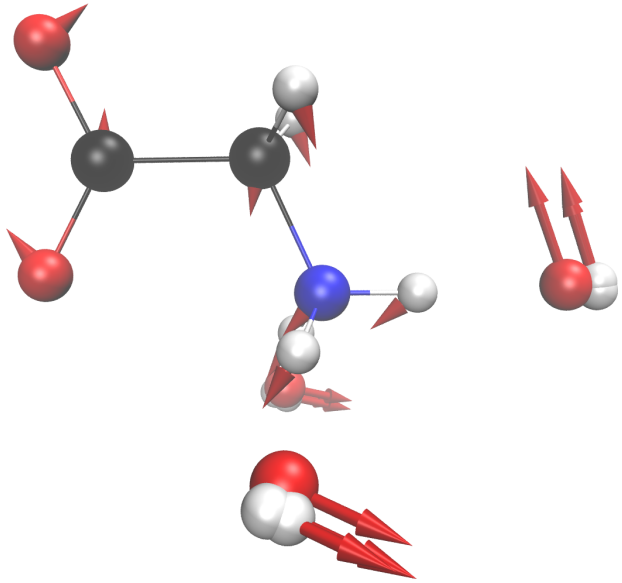
AMOEBA: HB-stretch-I 214 cm^{-1}



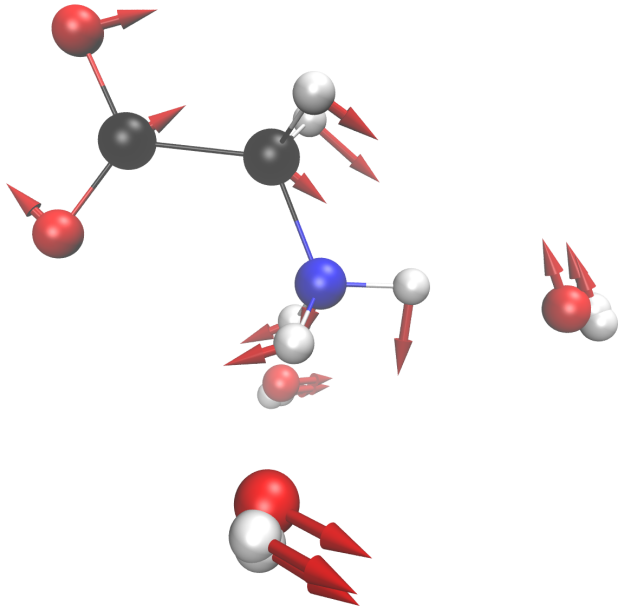
AIMD: C-alpha-oop 125 cm^{-1}



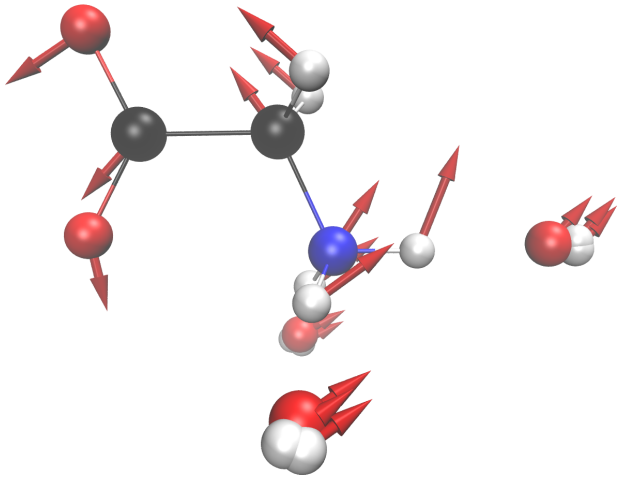
AMOEBA: C-alpha-oop 135 cm^{-1}



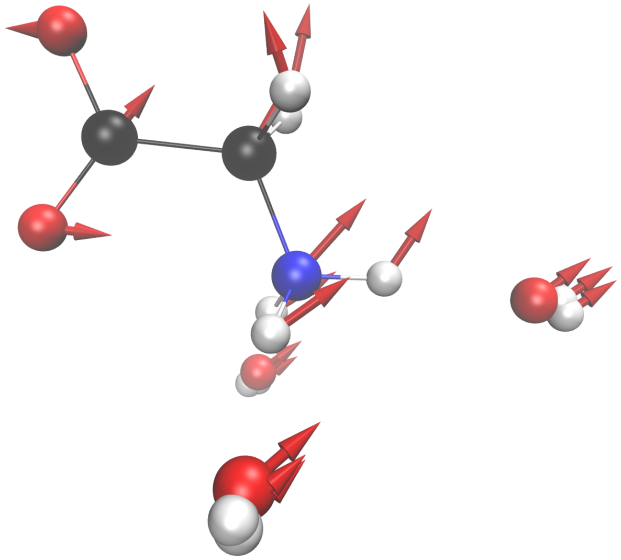
AIMD: HB-bend-I 102 cm^{-1}



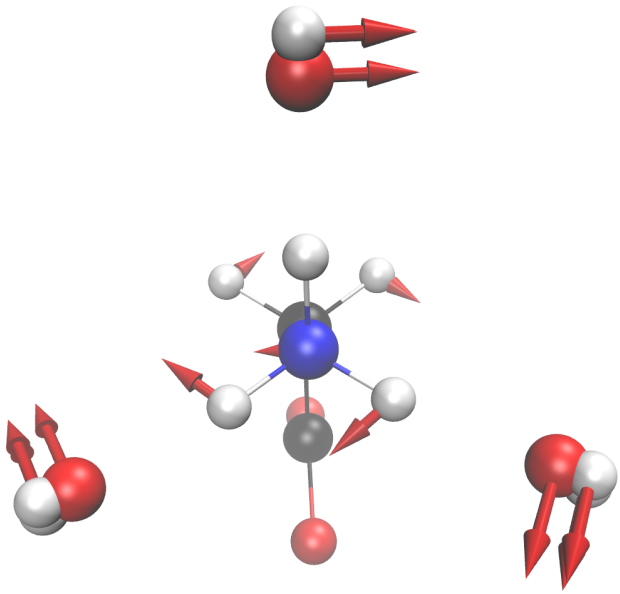
AMOEBA: HB-bend-I 92 cm^{-1}



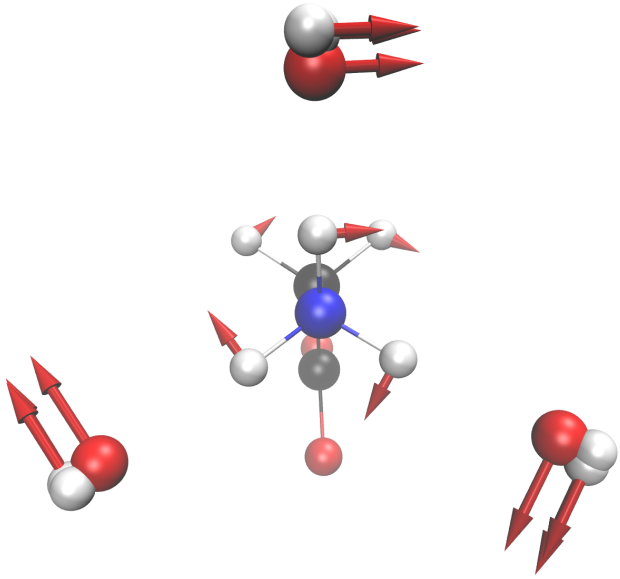
AIMD: cage-libration-II 90 cm^{-1}



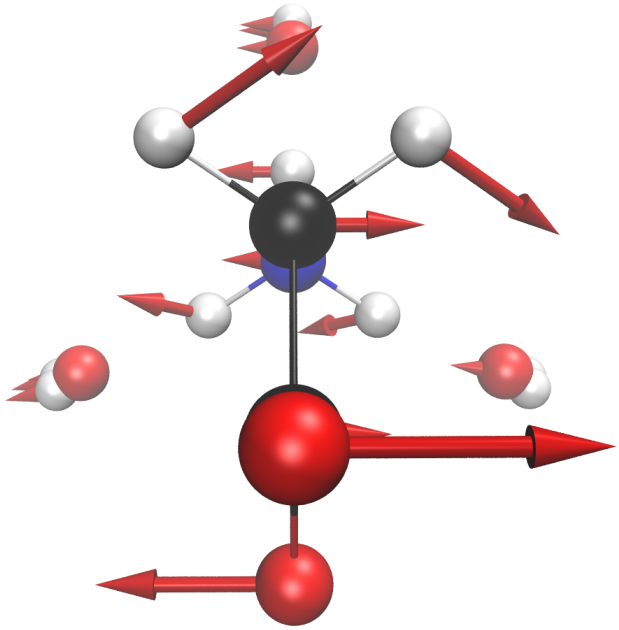
AMOEBA: cage-libration-II 71 cm^{-1}



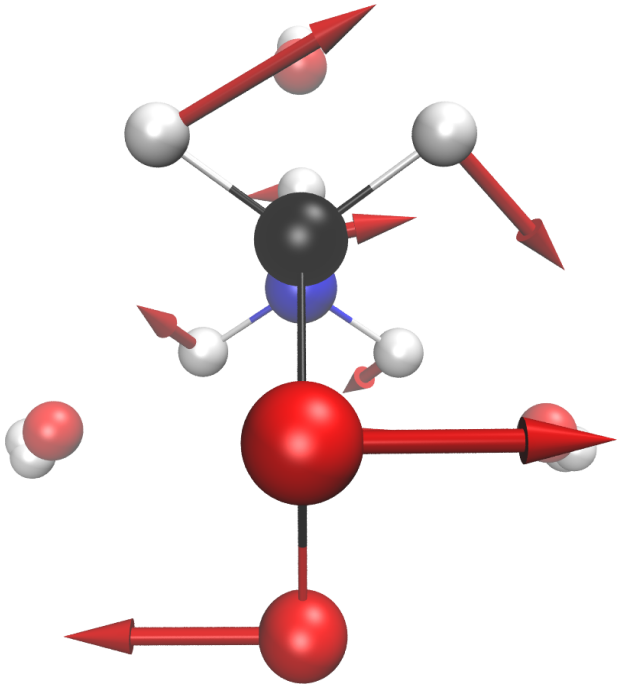
AIMD: HB-bend-II 90 cm^{-1}



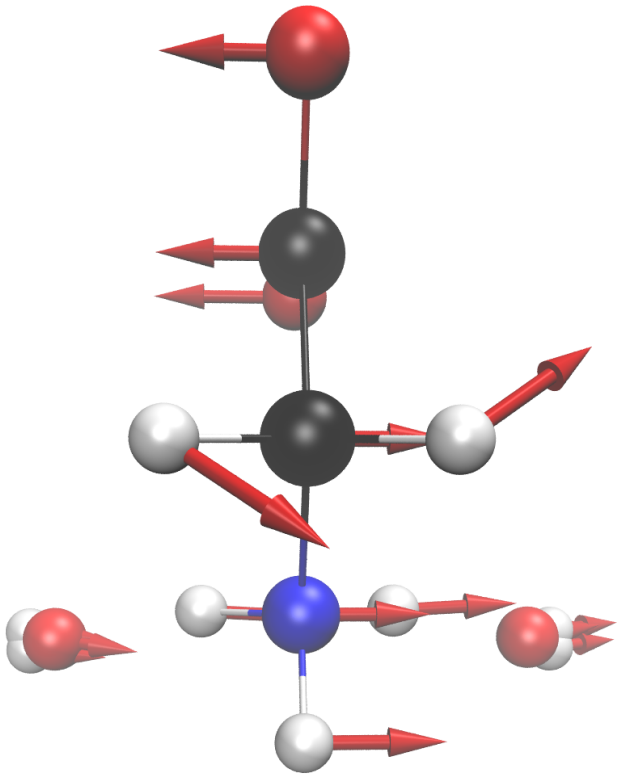
AMOEBA: HB-bend-II 68 cm^{-1}

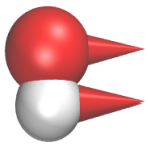


AIMD: cage-libration-III 89 cm^{-1}

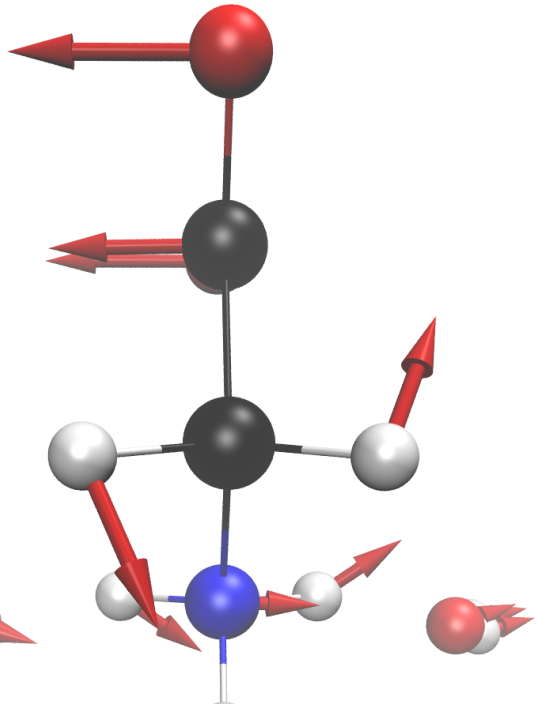


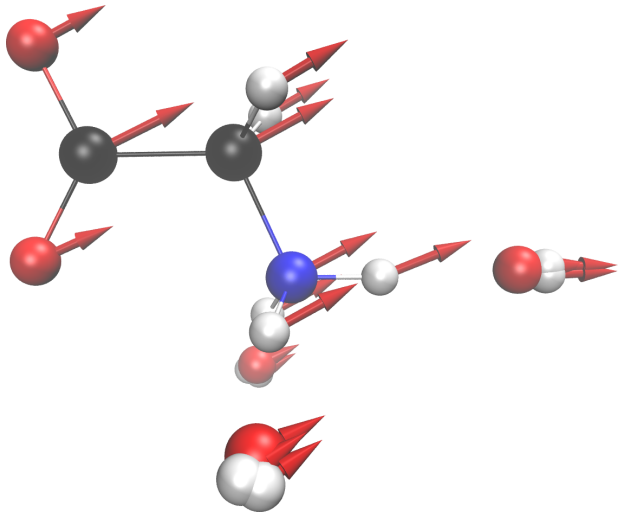
AMOEBA: cage-libration-III 84 cm^{-1}



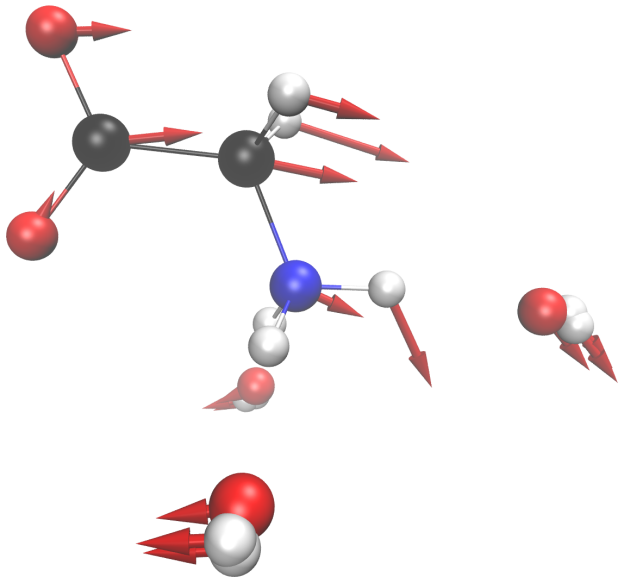


AIMD: cage-libration-I 82 cm^{-1}

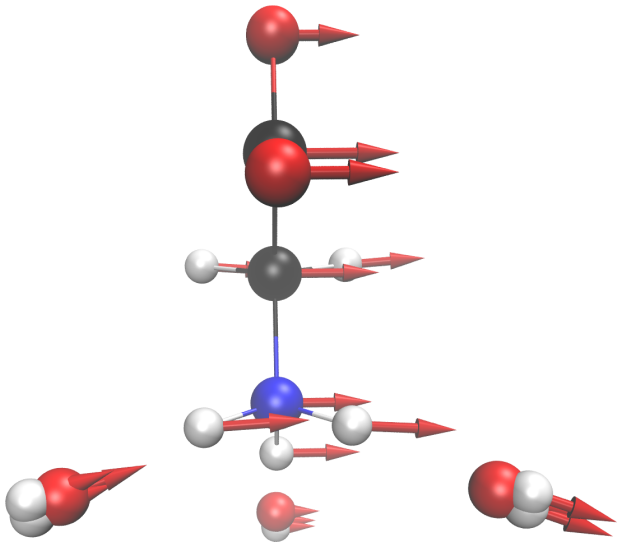




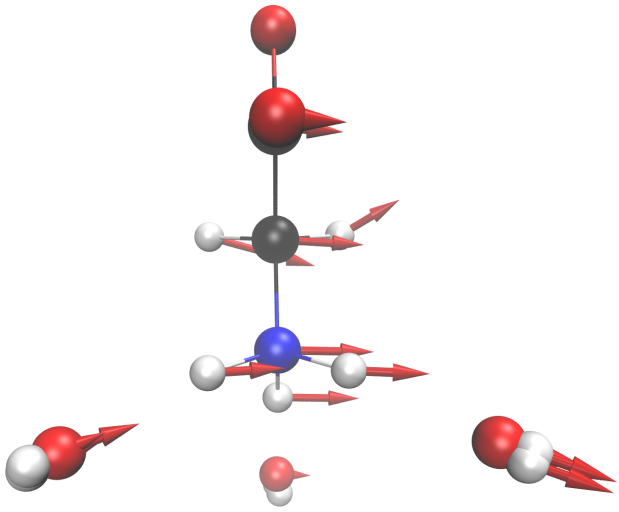
AIMD: cage-rattling-II 73 cm^{-1}



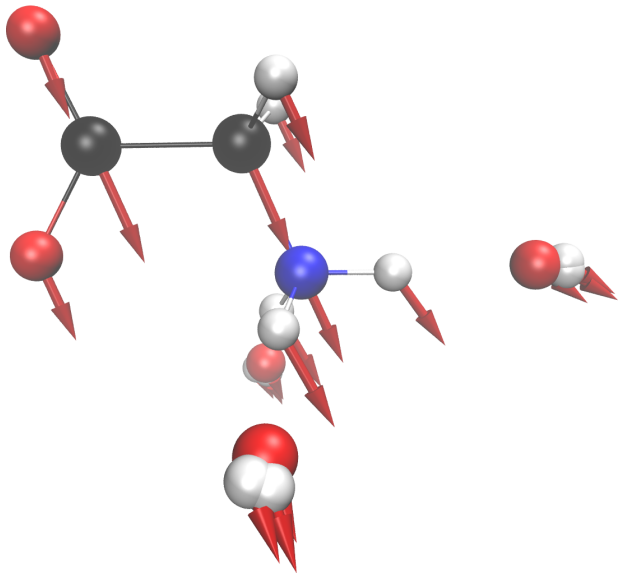
AMOEBA: cage-rattling-II 63 cm^{-1}



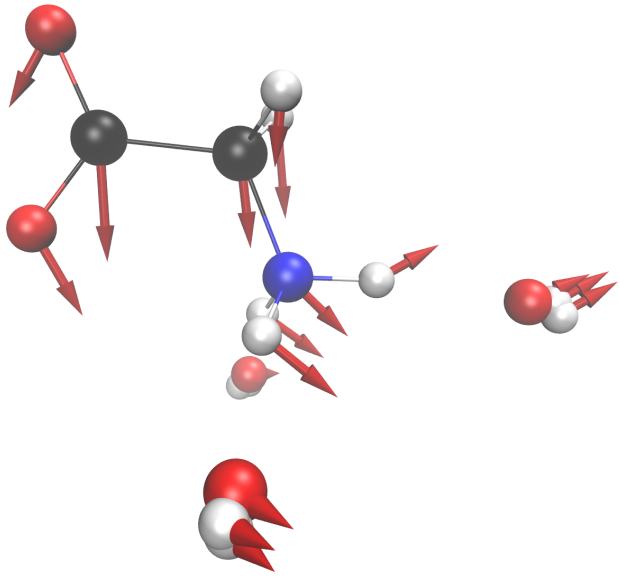
AIMD: cage-rattling-III 64 cm^{-1}



AMOEBA: cage-rattling-III 61 cm^{-1}

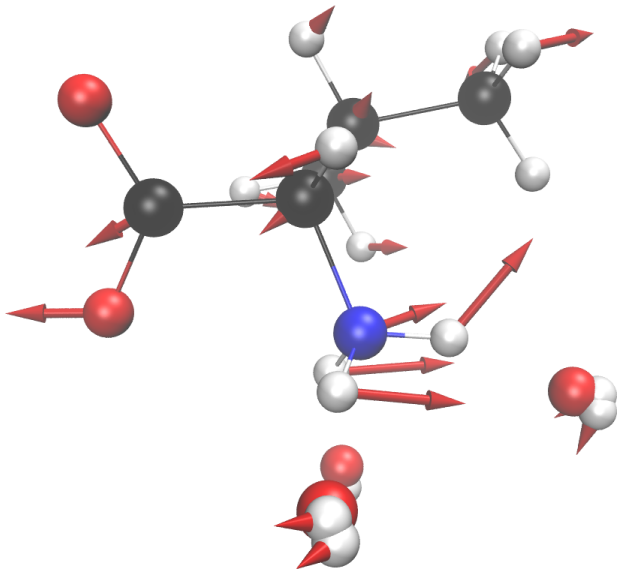


AIMD: cage-rattling-I 62 cm^{-1}

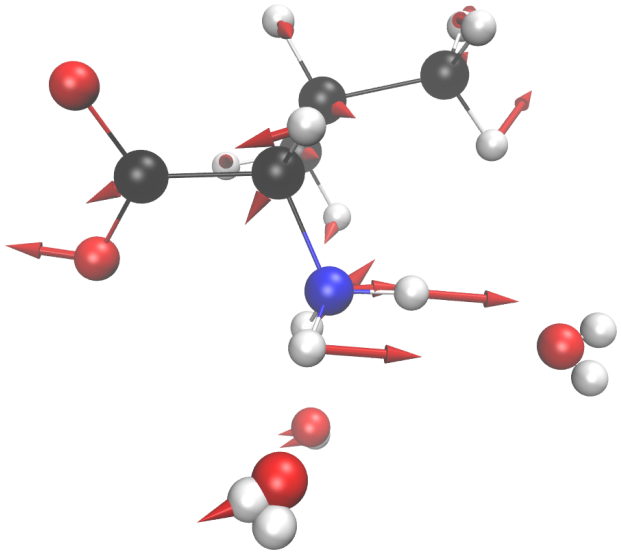


AMOEBA: cage-rattling-I 50 cm^{-1}

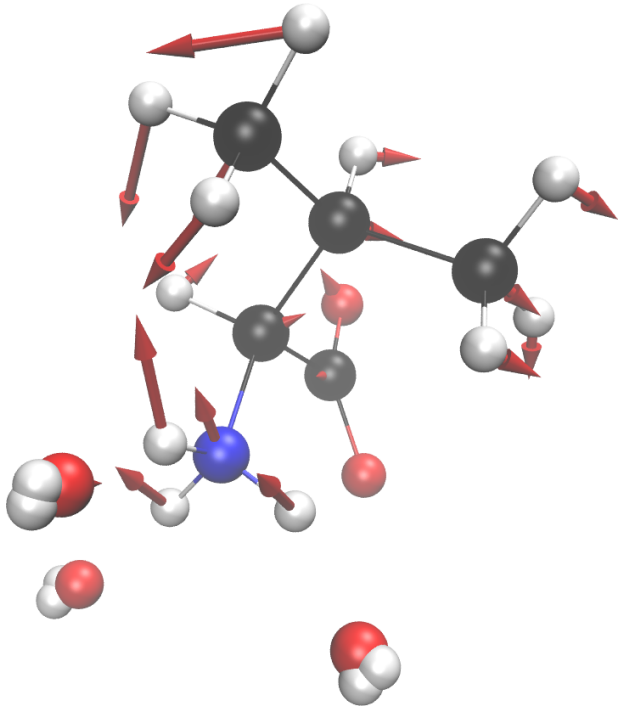
Mode Displacement Vectors of Valine



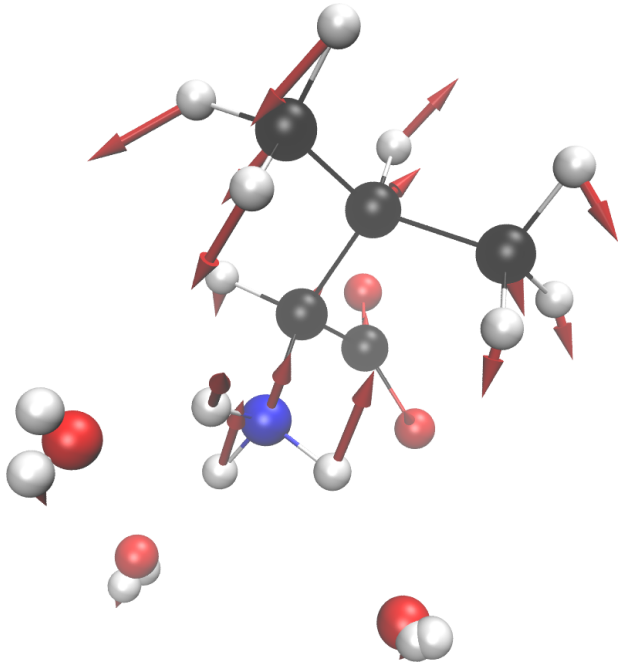
AIMD: NCCO 349



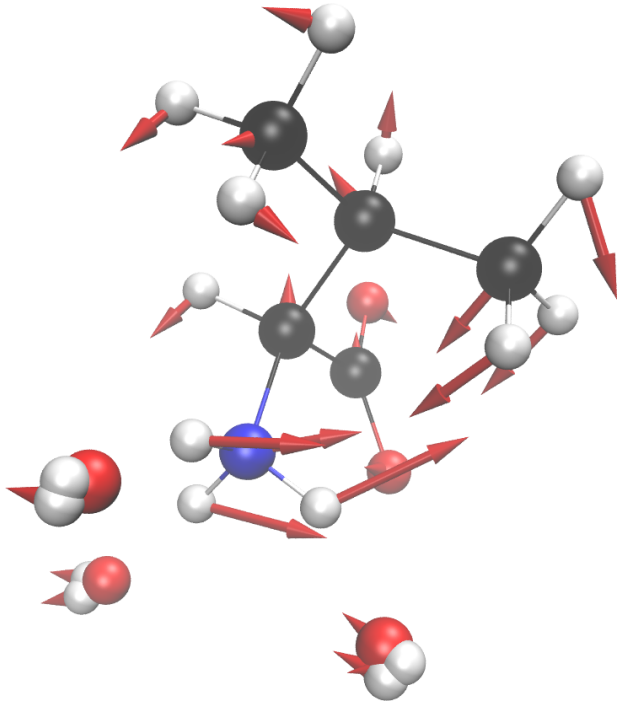
AMOEBA: NCCO 352



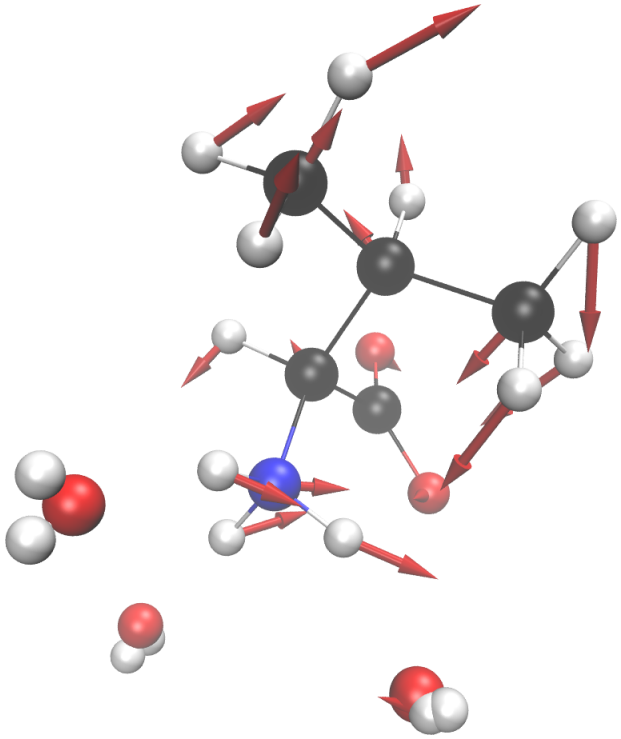
AIMD: NCCC-I 317



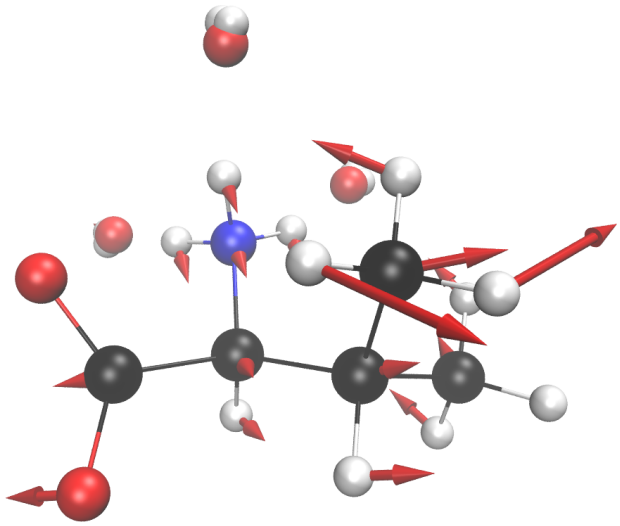
AMOEBA: NCCC-I 335



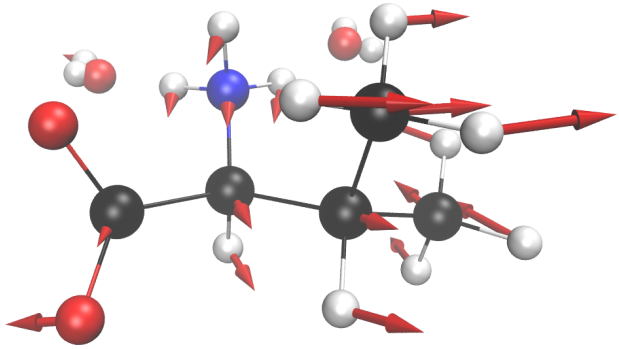
AIMD: NCCC-II 280



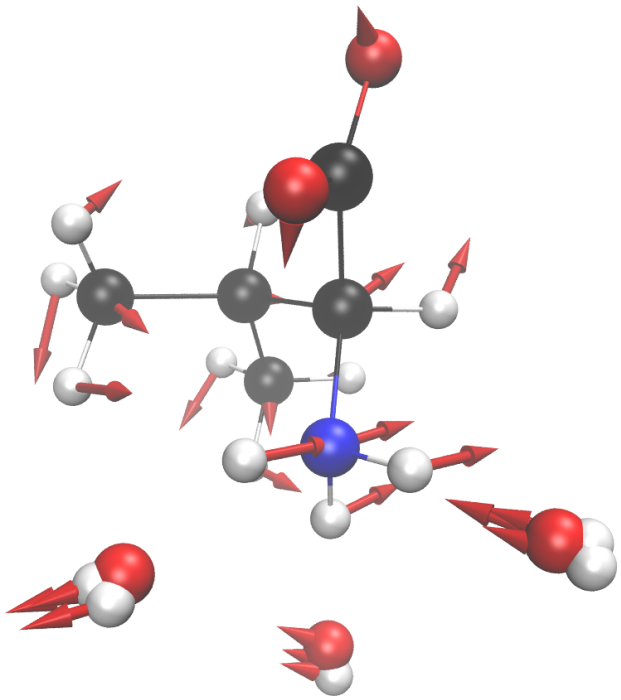
AMOEBA: NCCC-II 321



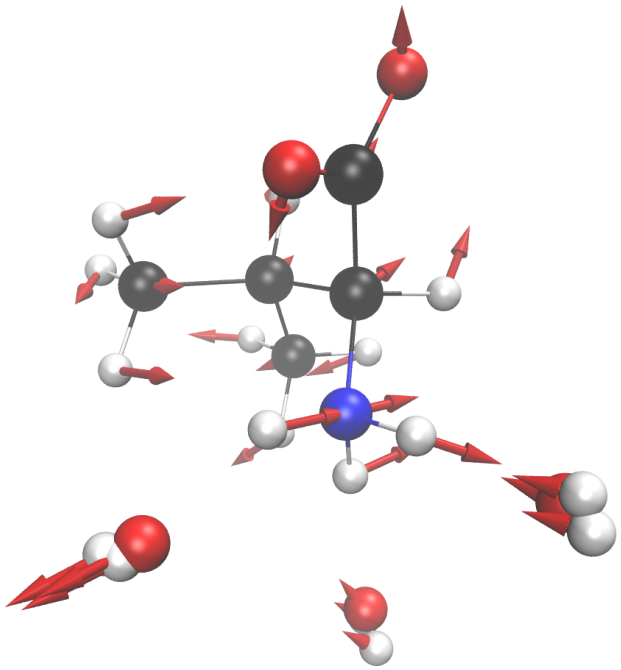
AIMD: COO-CCC 197



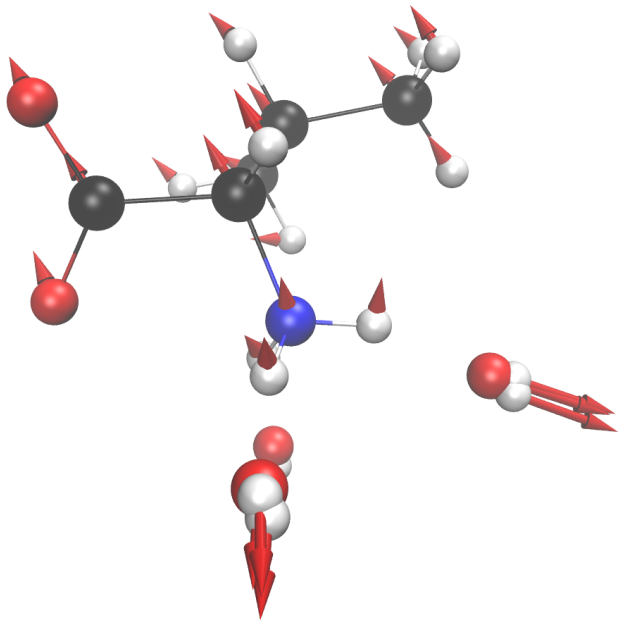
AMOEBA: COO-CCC 236



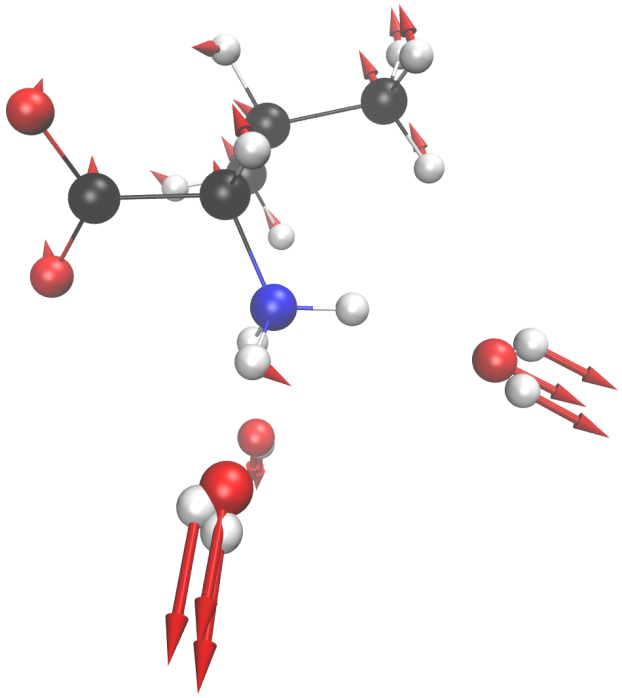
AIMD: CC-Twist-+-HB-Twist 189



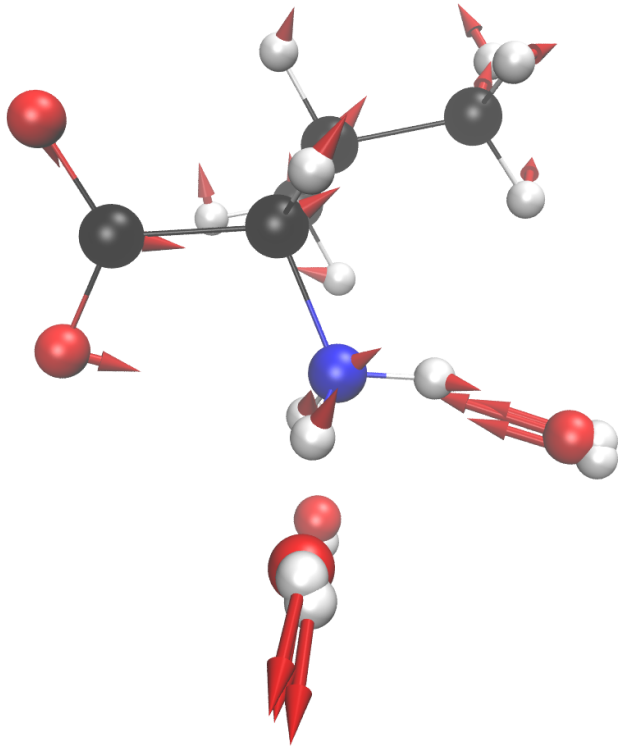
AMOEBA: CC-Twist-+-HB-stretch 196



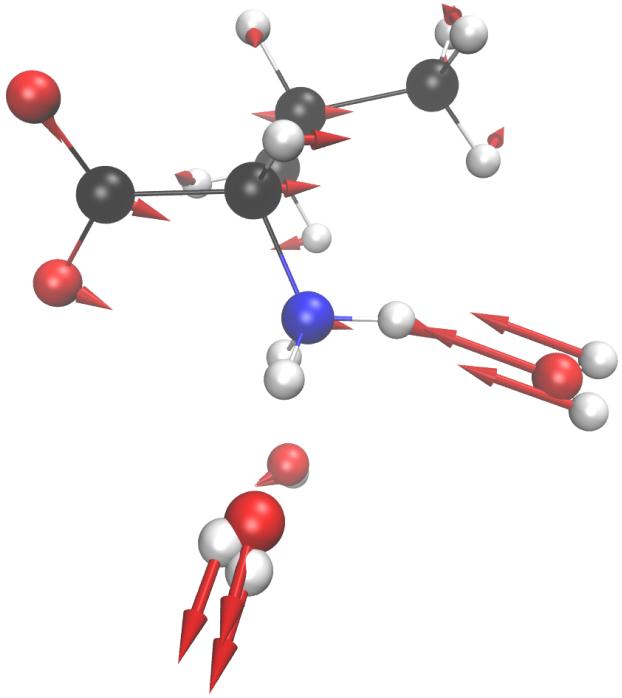
AIMD: HB-stretch-II 148



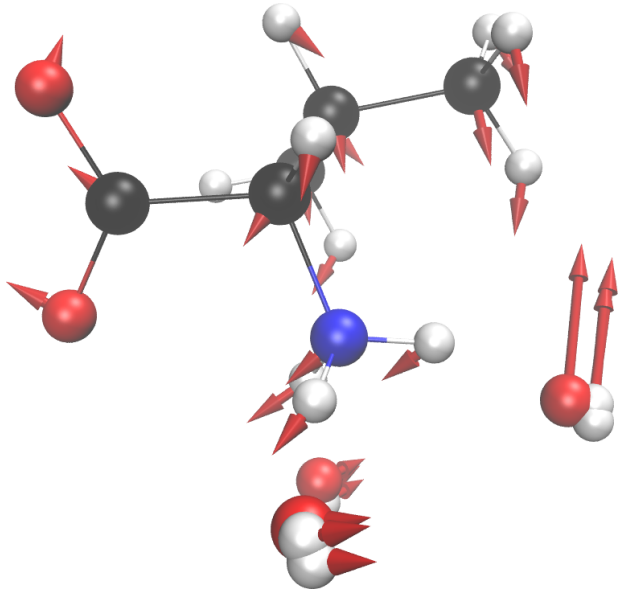
AMOEBA: HB-stretch-II 164



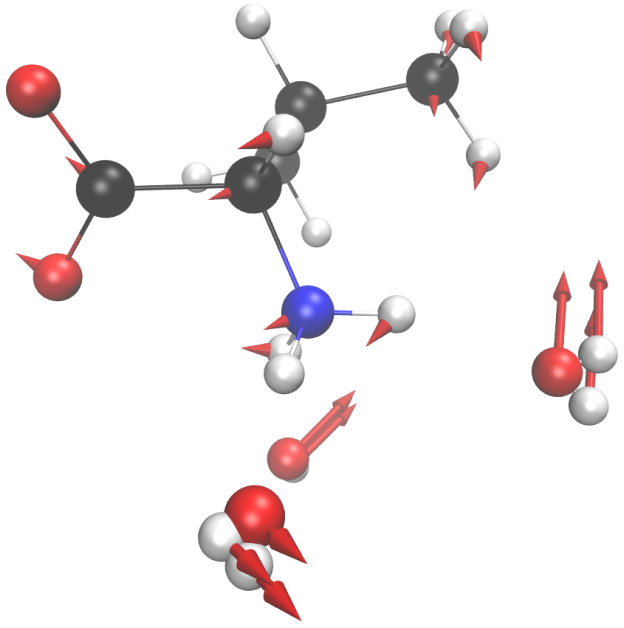
AIMD: HB-stretch-I 112



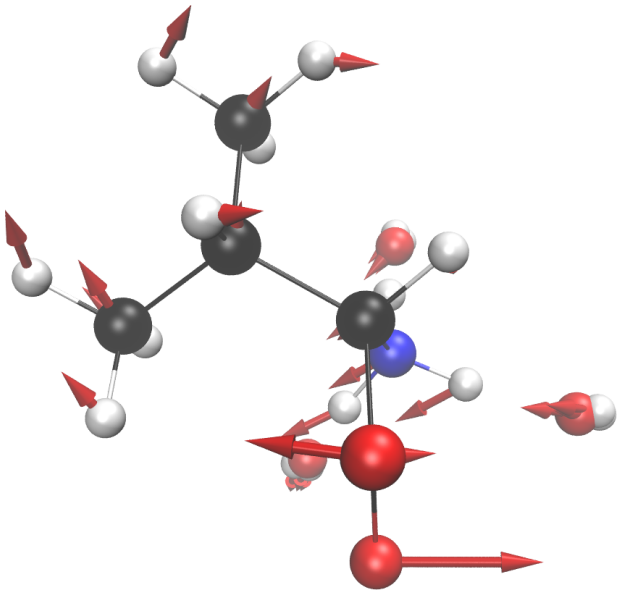
AMOEBA: HB-stretch-I 131



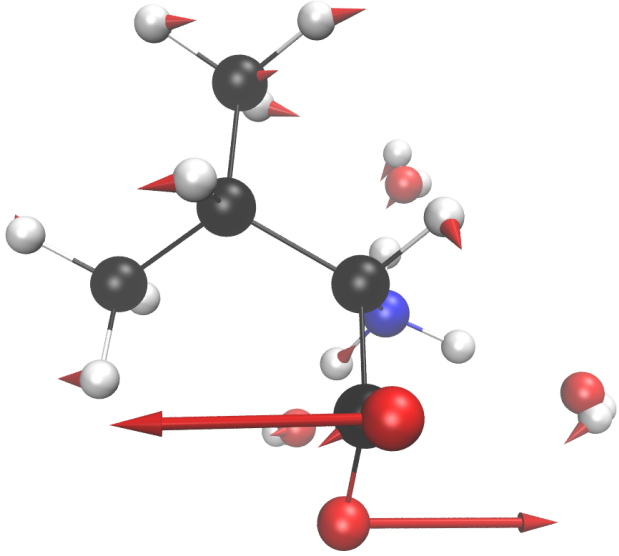
AIMD: HB-bend-I 101



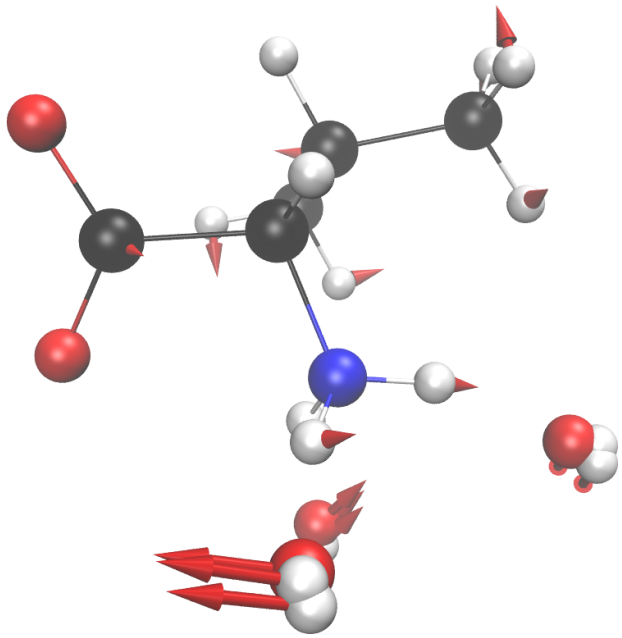
AMOEBA: HB-bend-I 86



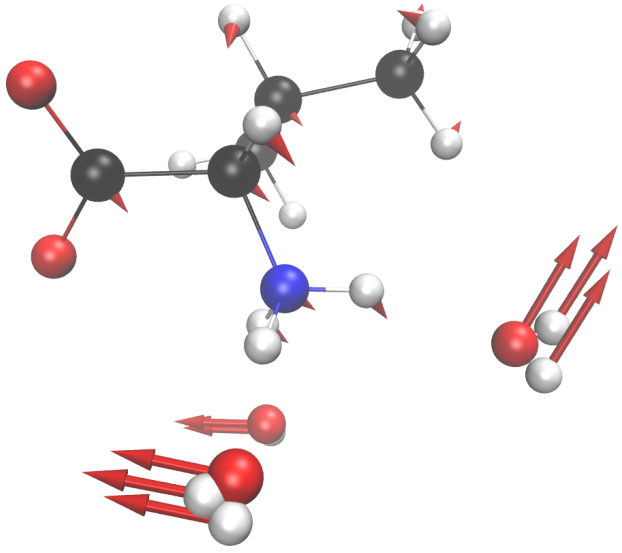
AIMD: Ca-oop-deformed 98



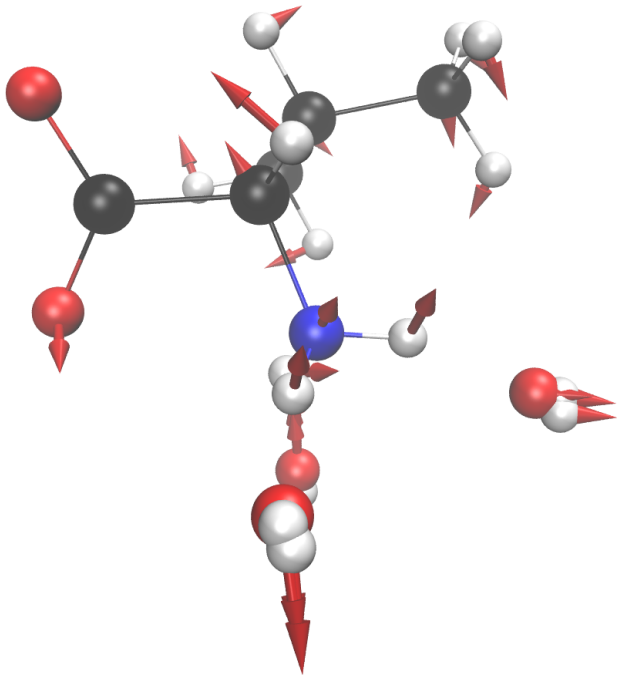
AMOEBA: Caoop 89



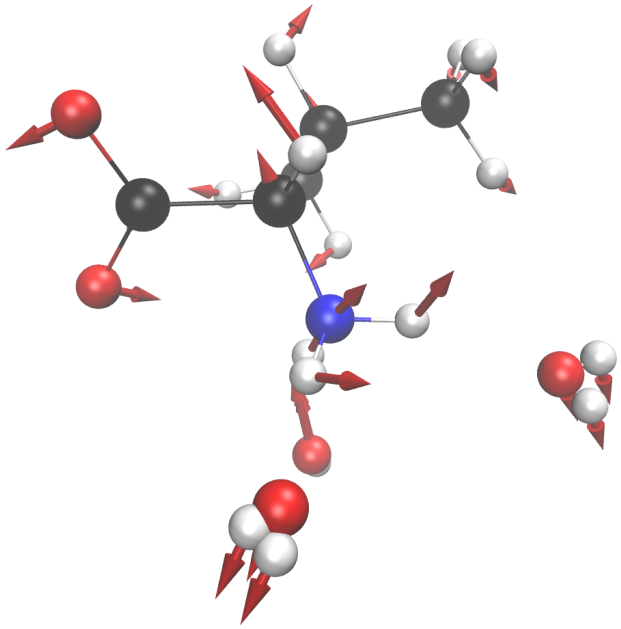
AIMD: HB-bend-X1 94



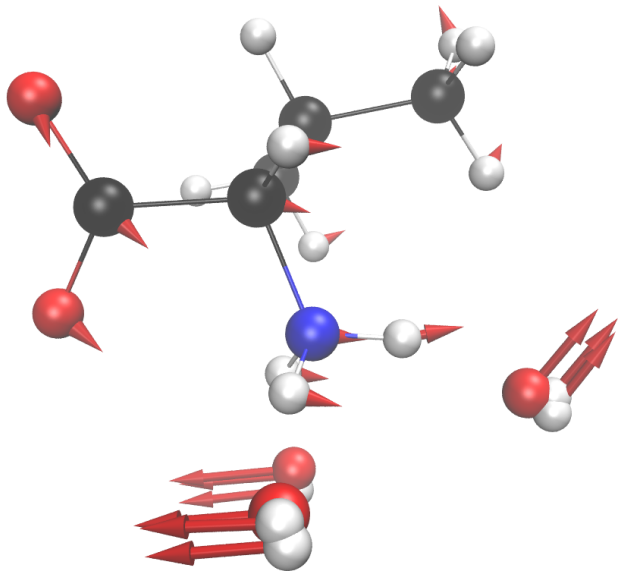
AMOEBA: HB-bend-X1 80



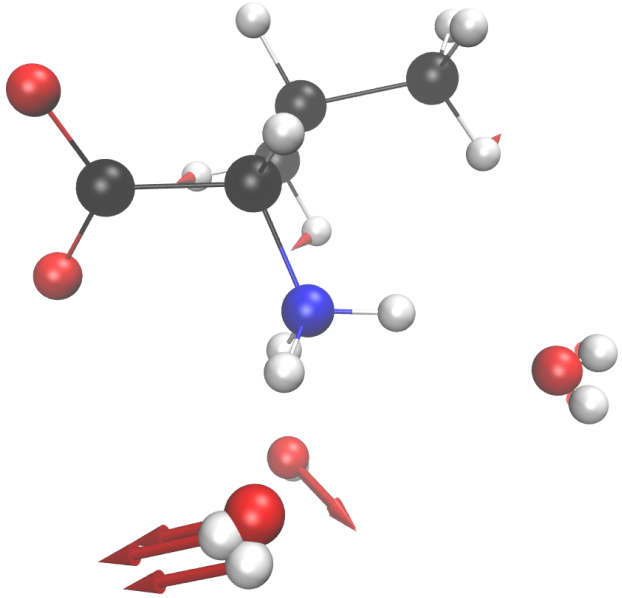
AIMD: HB-bend-U2 90



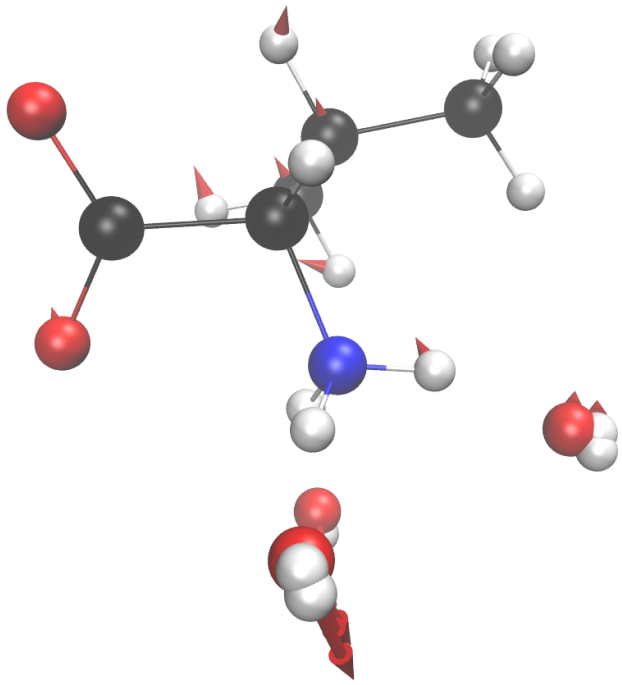
AMOEBA: HB-bend-U2 96



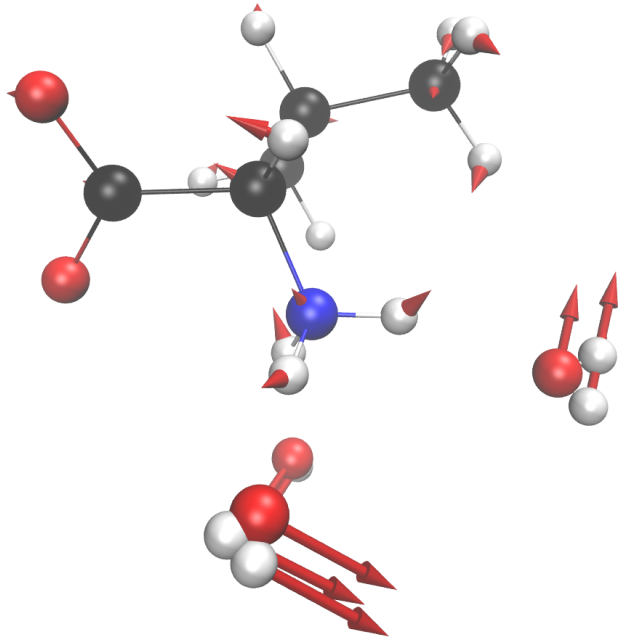
AIMD: HB-bend-U1 88



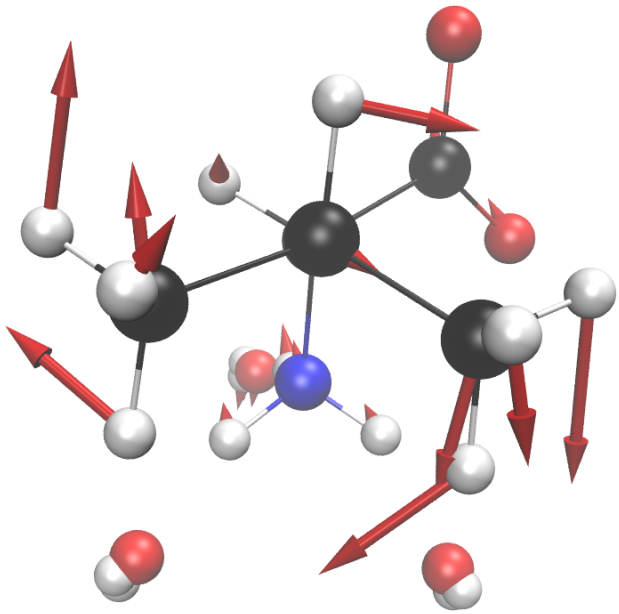
AMOEBA: HB-bend-U1 75



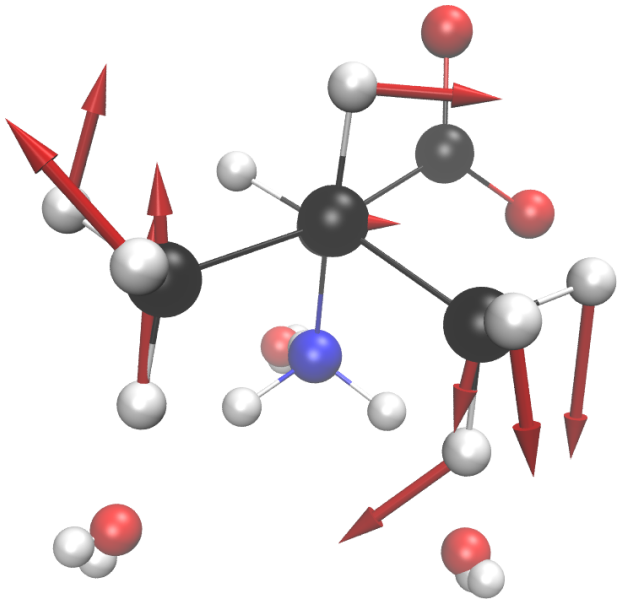
AIMD: HB-bend-X 87



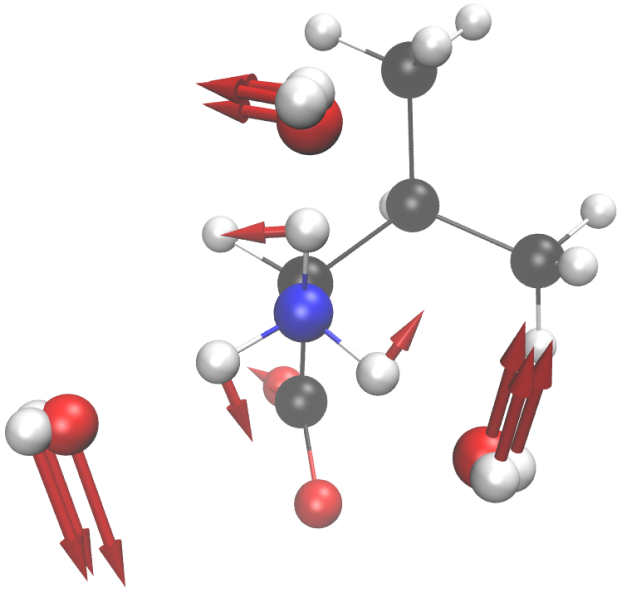
AMOEBA: HB-bend-X 73



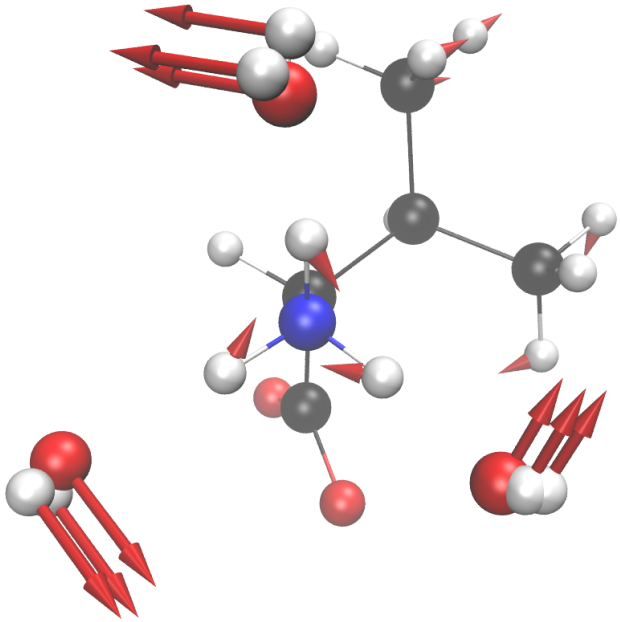
AIMD: R-libration-I 85



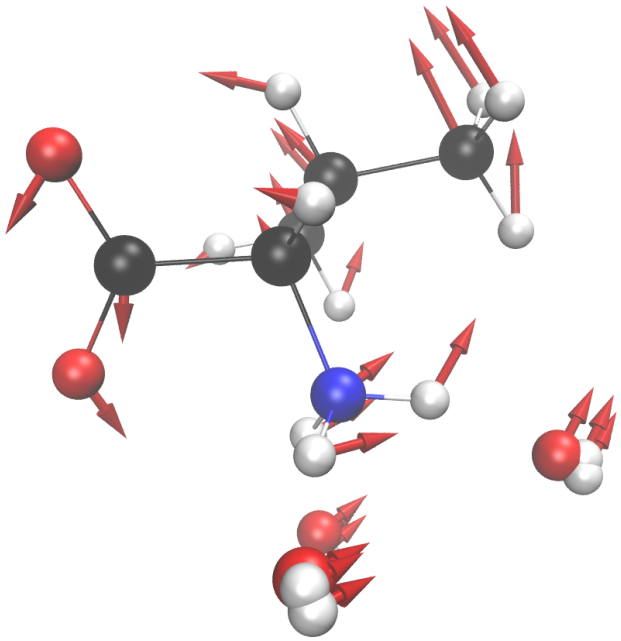
AMOEBA: R-libration-I 75



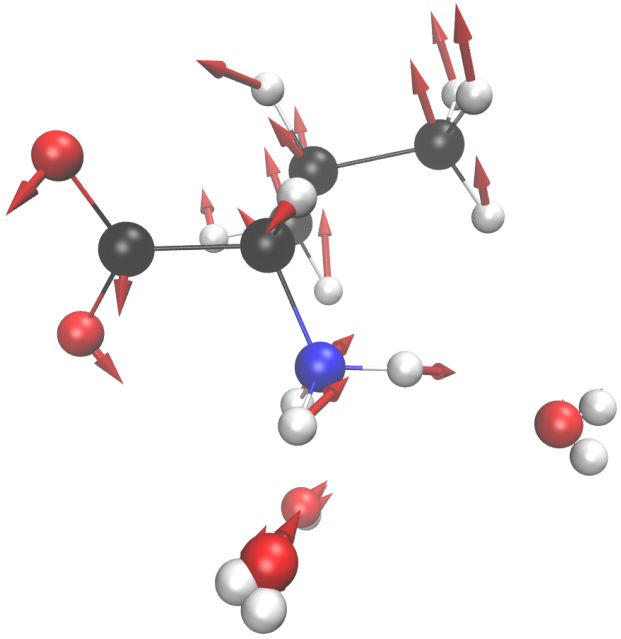
AIMD: HB-bend-X3 83



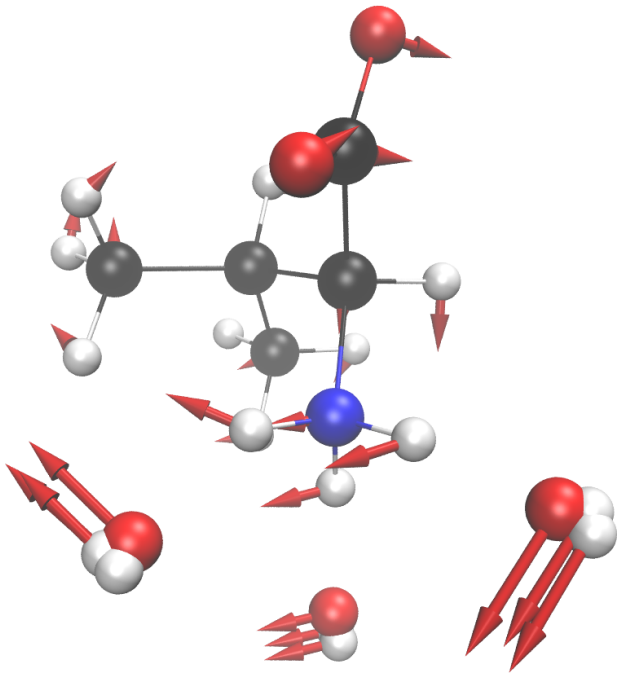
AMOEBA: HB-bend-X3 63



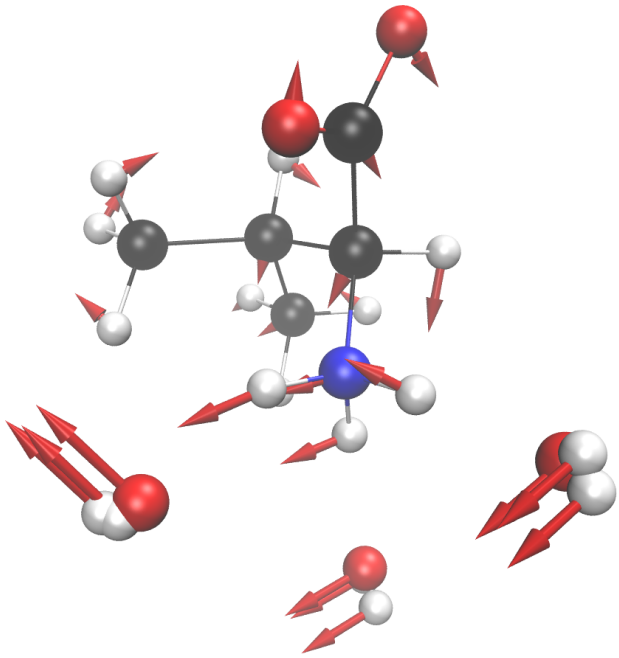
AIMD: cage-libration-II 70



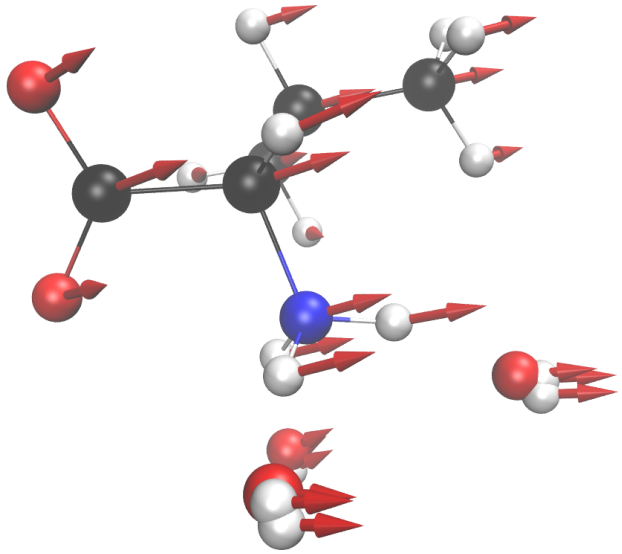
AMOEBA: cage-libration-II 67



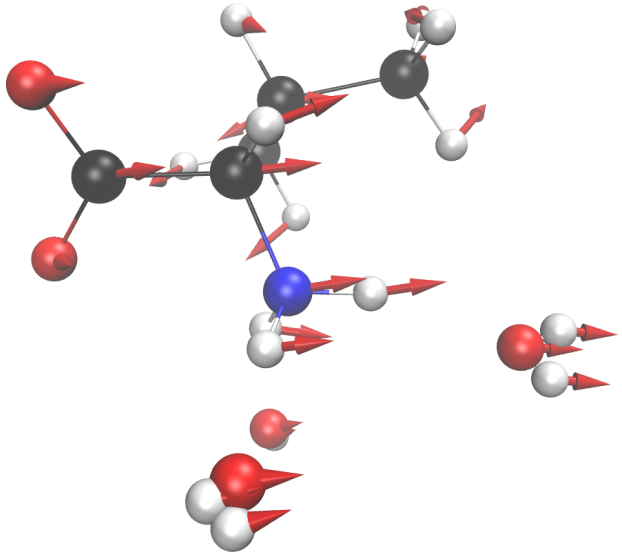
AIMD: cage-libration-I 68



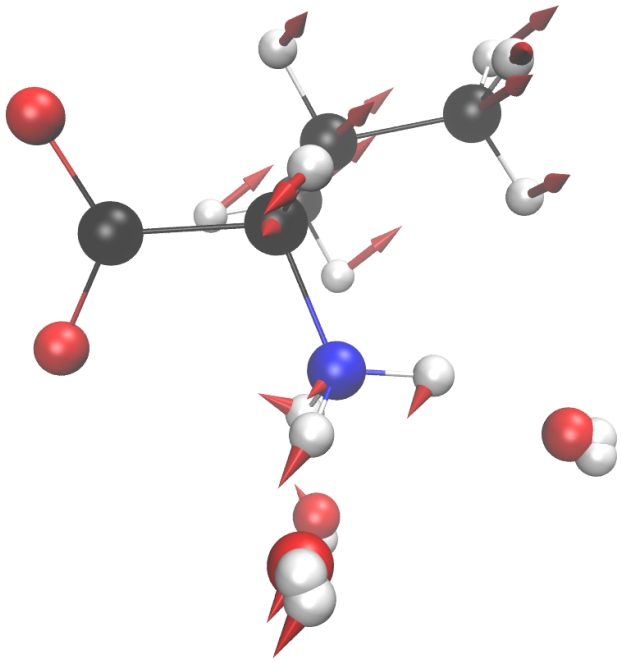
AMOEBA: cage-libration-I 52



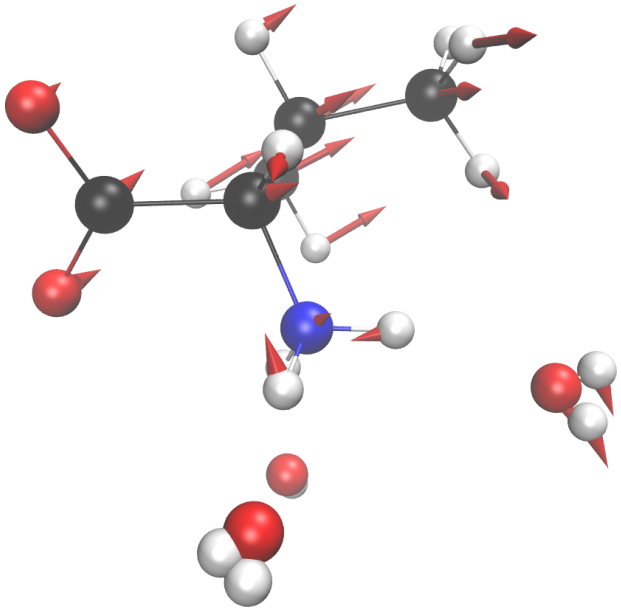
AIMD: cage-rattling-II 56



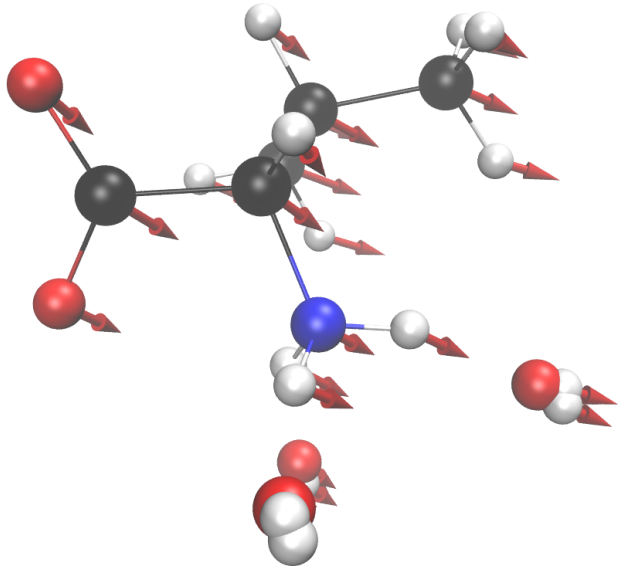
AMOEBA: cage-rattling-def 51



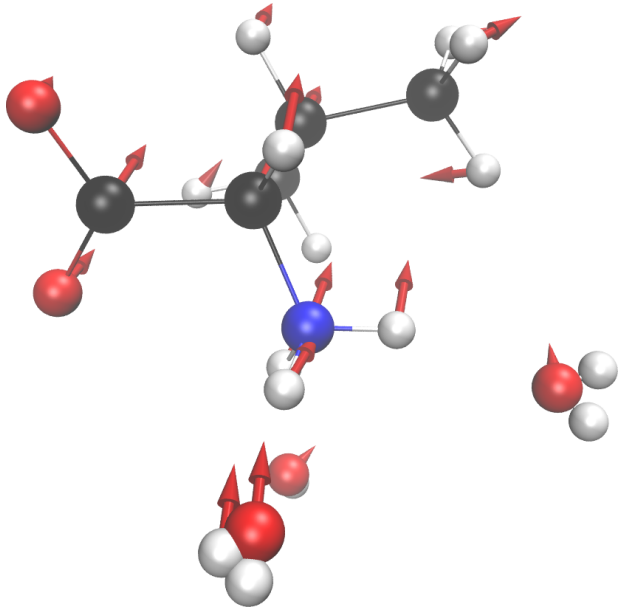
AIMD: R-libration-II 55



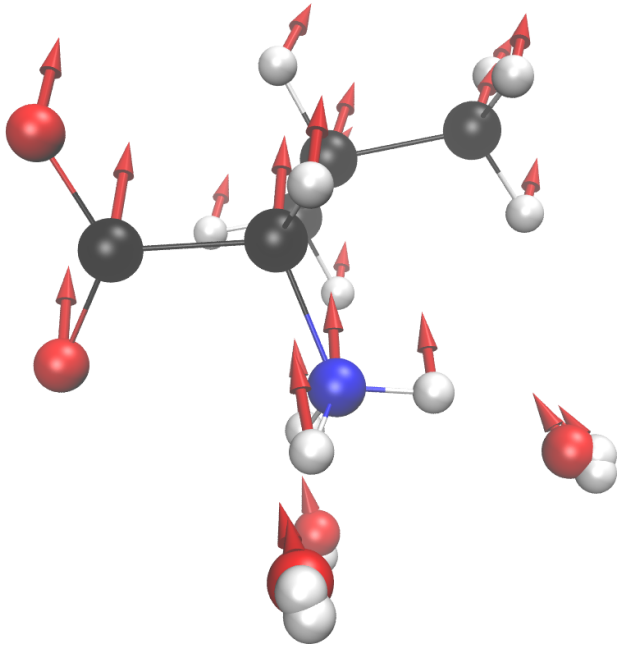
AMOEBA: R-libration-II 38



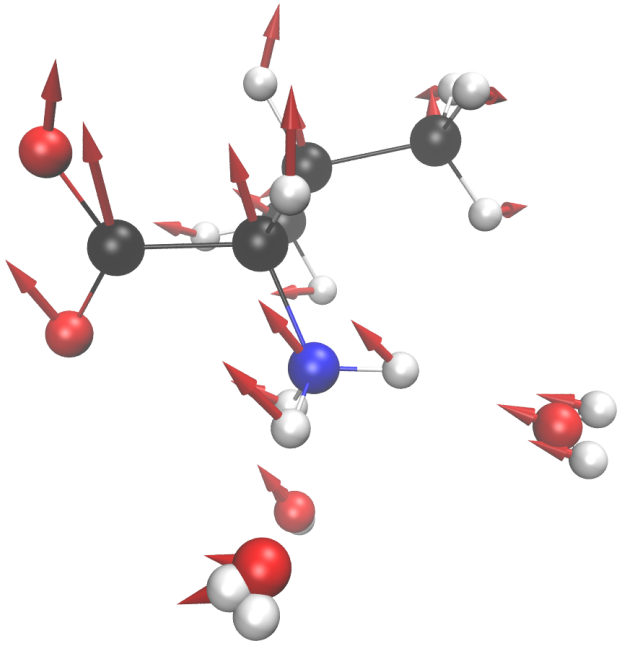
AIMD: cage-rattling-I 54



AMOEBA: cage-rattling-def 41



AIMD: cage-rattling-III 50



AMOEBA: cage-rattling-def 36

Modes of Glycine in the SSC(-)

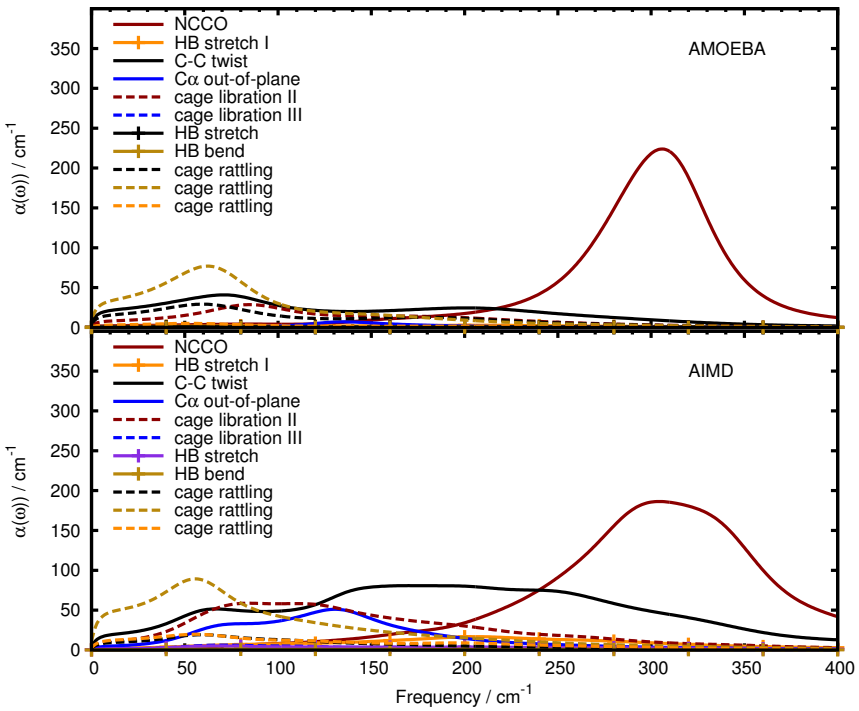


Figure A.47: THz mode intensities of glycine within the supermolecular solvation complex at the carboxylate group (SSC(-)) computed via AMOEBA (top) and AIMD (bottom) solvated by 30 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.

Modes of Valine in the SSC(-)

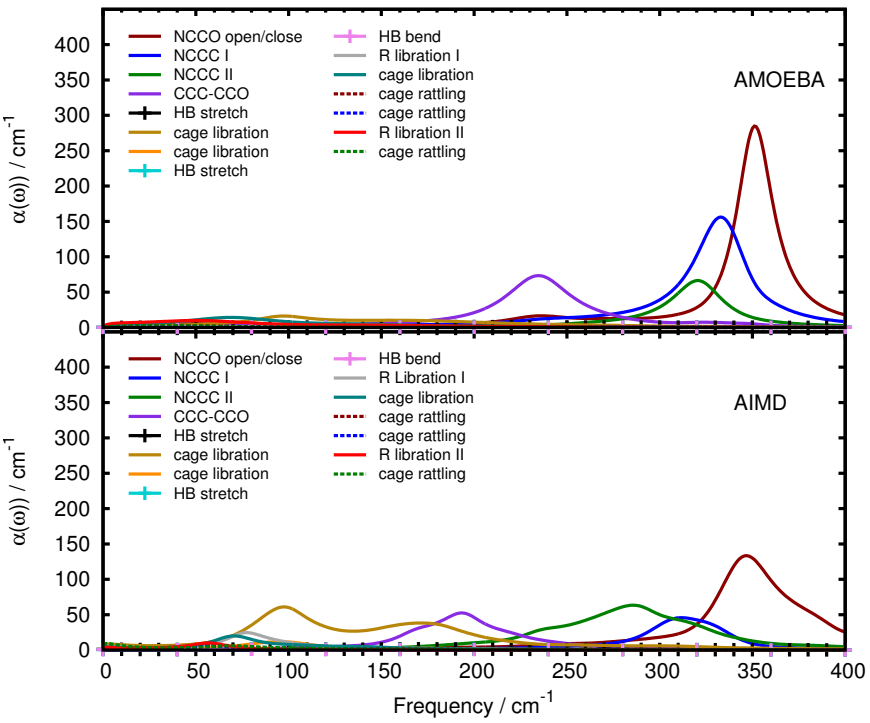


Figure A.48: THz mode intensities of valine within the supermolecular solvation complex at the carboxylate group (SSC(-)) computed via AMOEBA (top) and AIMD (bottom) solvated by 60 water molecules. The top panel shows high THz intensities, the bottom shows low intensities.