THE UNIVERSITY OF CHICAGO

PROTEIN CONFORMATIONAL VARIATION STUDIED BY AMIDE I INFRARED SPECTROSCOPY AND COMPUTATIONAL SPECTROSCOPY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

CHI-JUI FENG

CHICAGO, ILLINOIS

MARCH 2021

Contents

List of Figures

List of Tables

Acknowledgments

The ride through graduate school has been incredibly challenging in all aspects, and this thesis would not have been finished without the support from many wonderful people. First, I want to thank my advisor, Andrei Tokmakoff, for his support, encouragement, being approachable, and giving freedom to let me try all sort of ideas. Providing all of these really helps me become a more and more independent and competent person of doing scientific research.

I would definitely not continue doing research without my undergrad advisor, Yuan-Chung Cheng. I felt extremely grateful to have him along the way for being enthusiastic, keeping me motivated, giving me a lot of freedom to explore different things while at the same time being able to remind me of the big pictures, imposing high standards in a good way. I was impressed about how excited he was when we first ever simulated two-dimensional electronic spectroscopy (2DES) of water-soluble chlorophyll binding protein (WSCP) and how excellent the agreement it turned out with head-to-head comparison with Donatas Zigmantas' experimental 2D population-time-dependent spectra. This is part of the reason why I am still aiming for doing solid research whenever possible and trying to be altruistic. I enjoyed the moment when someone got genuinely interested in research, and felt excited about what can be learnt. This really shapes me into the person who I am.

I enjoyed my time in the Tokmakoff group with my lab mates. To Mike Reppert for assistance and commiseration in computational stuff, for showing high standards and critical thinking, and for good food. To Rajib Biswas for enjoyable break back to the old days on all sort of discussions, and for genuine support and commiseration. To Memo Carpenter as my great friend for being an optimistic and enthusiastic colleague, for bookcamps on nonlinear topics,

Mukamelian stuff, and stochastic processes, for taquería runs with Lukas Whaley-Mayda, for trying vegan food with our vegan crew: Kade Head-Marsden, Julia Murphy, and Elle Rathbun, and for my favorite Bulbasaur in Pokémon Go. To Paul Sanstead for being patient on teaching me experimental stuff so that I felt comfortable about doing experiments even with the fact that I was trained to be a theorist, for commiseration and support on rough circumstances I experienced, and for working together collaboratively at the later stage on insulin project. To Nick Lewis, and Lukas Whaley-Mayda for showing different aspects of setting high bars, and conveying scientific stuff clearly and concisely. To Brennan Ashwood for being a great colleague working in the big lab, for open and free scientific discussions, for being altruistic, and for playing Switch games together along with Ram Itani, Nick Lewis, and Yumin Lee. To Ram Itani for sharing positive thoughts, commiseration, for having nice food together with many friends, and for grocery runs. To Luis Busto de Moner for being proactive, enthusiastic and positive, and for showing passion to work on computational insulin project together. To all of my folks for not showing bothered by the "Chi-Jui's walk", and always willing to discuss all sort of matters.

I really enjoyed the short but pleasant time in the theory group office and enjoyed my time in the GCIS building. To Adam Antoszewski for enjoyable discussions and working together productively. To Erik Thiede and Bodhi Vani for discussions on the insulin project. To Sam Greene in the Berkelbach group for being a really nice neighbor in the theory group office, for sharing life experience, and for enjoyable vegan foods. To John Phillips for incredible efforts to keep the building function properly and smoothly, and for keep lab conditions at the high standards to ensure our productivity.

I would also like to thank friends who supported me outside the lab, and made me think I am not really alone through the graduate school. To Ginny Wei for fully supporting me through

my enjoyable TA gap year prior to Chicago and the early stage in the graduate school, for motivating me to become a vegan, and to dig through Laozi and Krishnamurti, which result in a rock-solid belief and a philosophical ground to sustain myself, to be internally consistent, and to be kind to others as much as I can afford. To my great friends, Ko-Lan Tsung, and An-Hsuan Hsieh for commiseration during the TA year, and for long-term accompany even though we are scattered around the Midwest over the years. To Po-Chieh Ting as a considerate great friend since undergrad for helping me have a jump start in Chicago, for accompany through rough times, for vegan foods, and for genuine and valuable long-term friendship. To Hang Yin for deep and philosophical teatimes over all sort of subjects, for long-term friendship from China. To Polina Navotnaya as a good friend for commiseration, for trying new restaurants together, and for accompany when necessary. To Kade Head-Marsden for vegan muffin, chocolates, and puzzles, for letting me know Mr. Pokee, and for the teatime with Memo Carpenter. To Hung-Tzu Chang for enjoyable discussions in the theory group office back in Taiwan, and for the bookcamp on classical electrodynamics. To Mu-Chieh Chang and Yu-Shih Lin for always being welcoming people during their time in Chicago. To my vegan German friend Marie Weiel for long-term friendship from Germany, for support, encouragement, and different perspectives of life, and for motivating me to learn German.

Finally, I would like to express my gratitude to Shou-Ting Hsieh for irreplaceable support and accompany since COVID-19, for Taiwanese snacks during quarantine time, for fun chats over Zoom, and for Pokémon Go. I would not have made my journey on scientific research this far without the support from my parents, and they are always available for me to chat with, and provide incredible support.

Funding

I want to thank the National Institutes of Health, and the University of Chicago Research Computing Center for their generous support towards my degree and the work that is presented in this thesis.

Abstract

Protein structure-function relationship has been rethought over the past decades to account for conformational variation and its functional role. Coupled-folding and binding process in biomolecular recognition is a manifestation of such conformational disorder and heterogeneous conformational ensemble. The ensemble nature and the coupled dynamics nature between protein-water interactions, conformational fluctuations, and folding/binding events require high structural and temporal resolution, which create major experimental challenges. To address these challenges, this thesis presents a combined experimental and computational approach using two-dimensional infrared (2D IR) spectroscopy and computational spectroscopy. 2D IR spectroscopy offers sub-picosecond temporal resolution to probe protein structural variation. Site-specific structural information can be achieved by introducing isotope-labeling on selected amide groups, and computational spectroscopy that translates protein structures into an IR spectrum.

To validate this approach, peptide-water interactions are studied on dialanine using 2D IR spectroscopy, and computational IR spectra predicted from molecular dynamics simulations. Amide I frequency fluctuation and vibrational energy relaxation is found to have the common origin of effective fluctuating forces due to water hydrogen bond dynamics. A chemical exchange process is also observed experimentally on the order of tens of ps, and is predicted to be coupled peptide-water motions through computational analysis.

The idea of estimating conformational ensemble from isotope-labeled IR spectroscopy is tested on trialanine, which is well-characterized for its conformational variation. A Bayesian ensemble refinement scheme is developed for direct characterization of conformational ensemble against experimental IR spectroscopy. Isotope-edited 2D IR spectroscopy is found to provide a

stringent constraint on the conformational distribution, and it returns consistent ensembles across different force fields and water models. The dominant factor influencing the quality of the ensemble refinements is the systematic frequency uncertainty from spectroscopic maps, but it can be significantly reduced by incorporating 2D IR spectra in addition to the Fourier-transform IR spectroscopy. This Bayesian ensemble refinement method with IR spectroscopy provide an effective approach to determine complex protein conformational ensemble.

One of the model systems to understand coupled-folding and binding processes is association of human insulin monomer. During association, partially disordered B-chains from each monomer form an inter-monomer β-sheet in a native dimer crystal structure. However, the dimer conformational characterization still needs investigation. Conformational ensemble of insulin dimer is characterized using site-specific isotope-edited 2D IR spectroscopy, Markov State Models (MSMs), and amide I computational spectroscopy. Isotope-edited IR spectroscopy indicates an additional spectroscopic species other than the native dimer, and the distribution of these species can be influenced by tuning the ionic strength. The MSM predicted this additional conformation states of twisted dimer that exhibits a ~55° rotation of the native dimer interface, resulting in shifting the β-sheet registry and reorganizing its sidechain packing. Computational spectroscopy of the twisted dimer consistently accounts for the additional spectroscopic species. The presence of twisted conformations suggests a potential kinetic intermediate along the homodimer association and/or multi-pathway nature. This study provides additional insight on the conformational distribution of dimer and establishes a refined molecular picture of describing coupled folding and binding process in insulin monomer association.

Chapter 1

Introduction

1.1 Conformational Disorder in Proteins

Proteins are versatile macromolecules that undergo a variety of non-equilibrium biological processes in living systems, including muscle motion, metabolism, reproduction, *etc*. Understanding the role of proteins in biological function and the detailed mechanism is one of the central questions in biology. Since the development of protein crystallography in 1950s[4] and solution protein NMR spectroscopy in early 1970s[6] that probe the molecular structures in detail, connections between three-dimensional folded structures of proteins and biological functions have been increasingly recognized. These structural tools provide an effective approach to build foundation of detailed structural understanding of proteins that aids physical and biological interpretation of the biological processes, leading to the famous concept of structure-function relationship.[7]

Such relationship implicitly assumes the one-to-one mapping between a stable protein structure and a specific biological function. However, proteins are in essence dynamic molecules that can exhibit ultrafast motions from picosecond, angstrom-scale vibrations to slow conformational rearrangement and folding/unfolding beyond μs time scale. Dynamic proteins motions are inherently coupled to its own static structures. Also, environmental factors such as

protein-water interactions and thermal fluctuations also mediate conformational changes of proteins.[8-11] All of these factors contribute to the conformational disorder of proteins, and protein structure-function relationships have been rethought over the past two decades to account for the functional role of conformational disorder, which appears in intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs).[12]

Energetically, IDP or IDRs in proteins exhibit a variety of thermally accessible conformers undergoing rapid structural fluctuations or activated interconversion kinetics between free-energy basins on a complex free energy landscape in contrast to single native conformation of folded protein staying in a free energy basin (Fig. 1.1).[13-15] Such IDPs and proteins with IDRs have been observed to be involved in many biological processes including regulation, signaling, and coupled-folding and binding to functional partners.[12, 16] The dynamic nature of conformational interconversions in IDPs and proteins with IDP means that conformational variation and conformational dynamics cannot be decoupled, and that the structural characterization requires an ensemble description. These create major challenges on both experiments and simulations.

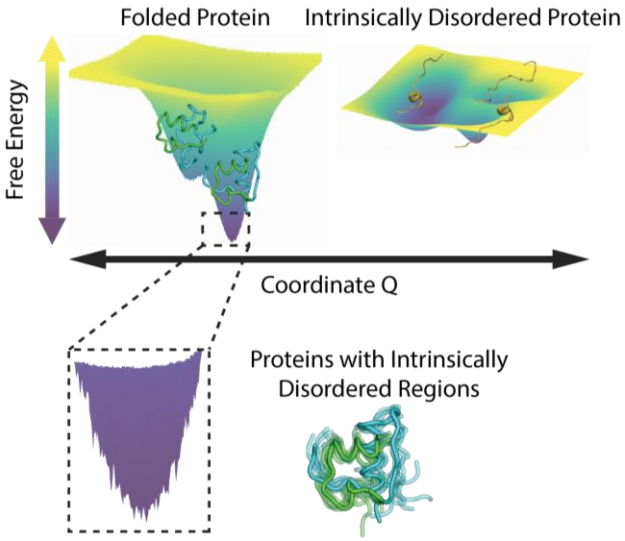

Figure 1.1: Schematic free energy surface of a folded protein, an intrinsically disordered protein and a protein with intrinsically disordered regions. Representative structure are obtained from MD simulations with the starting structures of insulin monomer (PDB: 2JV1) and S-peptide (PDB: 1RNU).

First, a great number of distinct conformers in the protein conformational ensemble usually leads to overlapping or even averaged signals in experimental measurements, which complicates the interpretation. Inherently the ensemble nature requires high structural resolution of the experimental probe, but most of the experiments suffer from ensemble-averaged observables that have limited sensitivity to the conformational disorder. Computationally, it is also challenging for atomistic simulations to sample transitions of relevant conformers adequately, which is usually rare events due to the gap between computationally accessible time scales and the time scales of conformational dynamics. Enhanced sampling techniques[17-18] and molecular kinetic models such as Markov State Models[19-20] have been applied to address rare event problems in conformational dynamics.

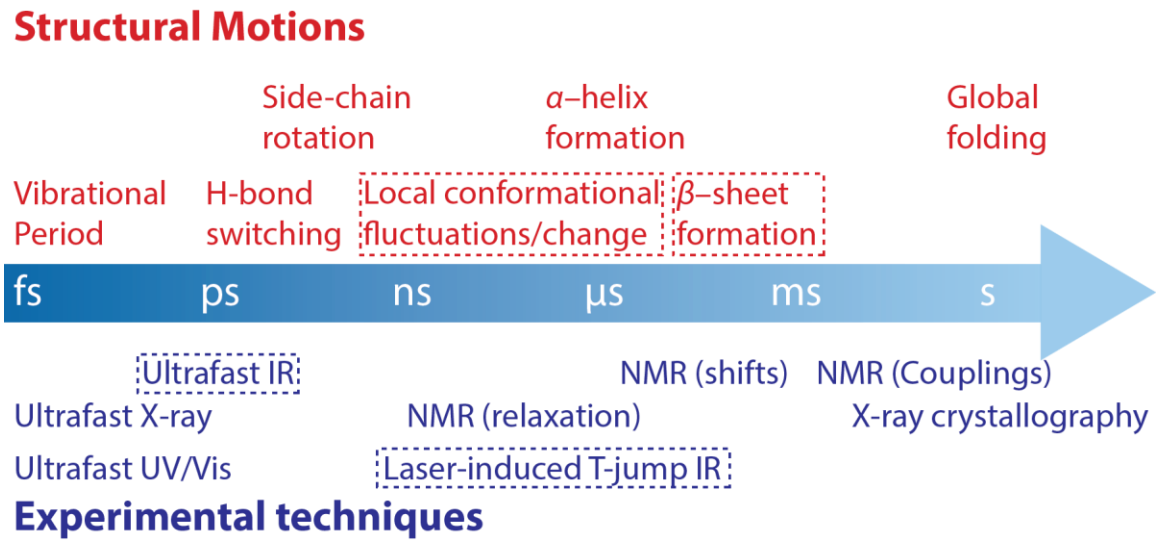

Figure 1.2: Time scales of structural motions and time resolution experimental techniques.

Second, coupling between fast conformational fluctuations and slow activated biological motions such as folding events span multiple decades in time (Fig. 1.2), which experimentally requires high temporal resolution. Additionally, hierarchical dynamics has been observed in conformational dynamics and interconversions that formation of hydrogen bonds and water solvation mediates the slow conformational interconversions.[8-11, 21-24] A wide span of time scales impose a challenge to biomolecular structural tools. Traditional structural tools such as NMR spectroscopy has only limited time resolution to measure chemical shifts and J-coupling due to coalescence time scale of ms.[25-26] Spectroscopy with lower resolution such as (infrared) IR spectroscopy, and circular dichroism (CD) spectroscopy provides structural information about secondary structures, but site-specific information is difficult to be extracted. Experiments utilizing labels such as Förster resonance energy transfer microscopy (FRET) provides sensitivity about distances and potentially orientation of the labeled pairs to infer structural and dynamical behaviors, but a somewhat more detailed description requires exhaustive labeling.[27]

Recent efforts have been focused on developing frameworks of determining conformational ensemble of disordered proteins using a combination of multiple experimental probes and computational modeling, and developing structural probes with high intrinsic time resolution. An increasing popular approach of ensemble refinement, which weights existing ensemble populations from simulations against experimental data, has proven successful for facilitating the inference of protein conformational ensembles consistent with experiments,[28-29] including developing different frameworks such as maximum entropy principle,[30-34] Bayesian statistics,[35-37] and biological applications against experimental data such as NMR,[29, 38] SAXS,[39-40] and IR spectroscopy.[41-42] The work in this thesis exploits the capability of isotope-edited two-dimensional infrared (2D IR) spectroscopy as a probe to investigate protein conformational disorder and dynamics, including dynamics of peptide-water interactions, development of ensemble refinement scheme for 2D IR spectroscopy, and investigating conformational heterogeneity of insulin dimer.

1.2 Infrared Spectroscopy as a Probe to Protein Structure and Dynamics

Infrared (IR) spectroscopy provides a direct route to investigate vibrations of a biomolecule, which is inherently sensitive to the chemical structures and surrounding environments. In particular, 2D IR spectroscopy brings ultrafast sub-ps temporal resolution to investigate protein conformational distribution that is essentially static. [41, 43-52] In this section, the basic information content in 2D IR spectroscopy and amide I vibration as a protein structural probe will be introduced.

1.2.1 Two-Dimensional Infrared Spectroscopy

Two-dimensional infrared (2D IR) spectroscopy utilizes ultrafast IR pulses for studying molecular vibrational dynamics with sub-ps temporal resolution (Fig. 1.3). The first pump pulse pair excites the vibration of a molecule, which allows an interferometric measurement during τ_1 to extract information about the vibrational excitation. The duration between the first two pump pulses and the subsequent probe defines the waiting time (τ_2), which can be used to interrogate vibrational dynamics. The frequency of the signal emitted after the probe pulse is spatial dispersed onto an array detector, and the resulting spectra as a function of τ_1 can be Fourier-transformed to a 2D IR spectrum.

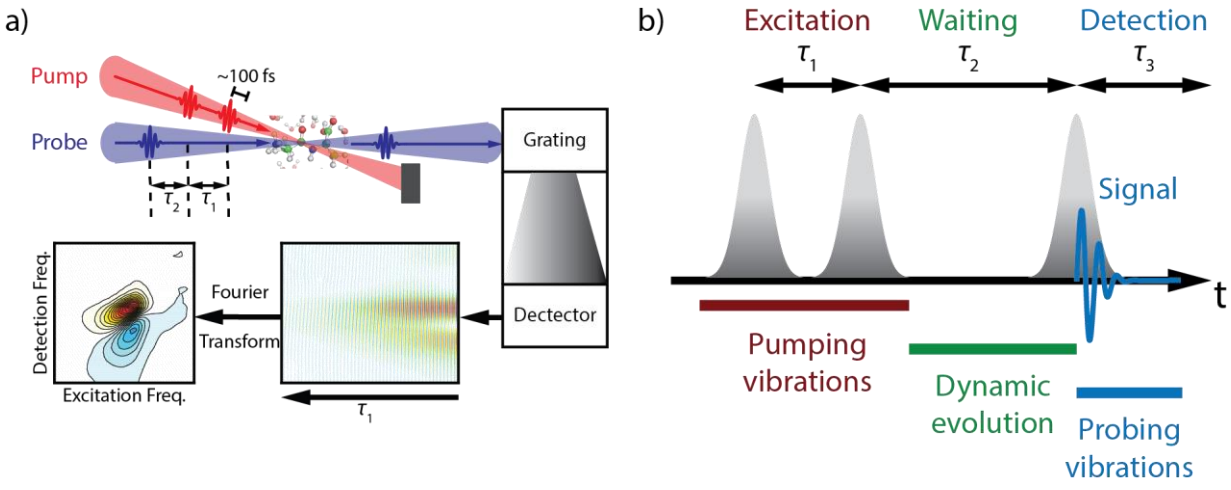

Figure 1.3: (a) Schematic representation of 2D IR measurement (b) Pulse sequence of 2D IR spectroscopy.

A 2D IR spectrum is a correlation map between the vibrational excitation frequency and detection frequency, and thus provides more information than the traditional Fourier-transform (FT) IR spectra, such as a 2D lineshape that reflects the underlying broadening mechanisms and cross-peaks indicating coupling between vibrational modes. Each resonance in the 2D spectrum is

a positive/negative (red/blue) doublet (Fig. 1.3a), which represent the ground state bleach (GSB) of the 0–1 quantum transition and excited state absorption (ESA) from the 1–2 quantum transition. The detection frequencies of these transitions differ as a result of the anharmonicity of vibrations.

Dynamical processes that can be observed using 2D IR spectroscopy includes vibrational relaxation, spectral diffusion, vibrational energy transfer (VET) and chemical exchange processes.[43, 53] Spectral diffusion and vibrational relaxation can be illustrated in Fig. 1.4. The excited vibration can experience frequency fluctuation caused by interactions with the surrounding environment such that the resulting spectrum rounds out along the waiting time τ_2. Characterization of the frequency fluctuation can be performed using the center line slope (CLS) method.[54-55] The CLS decay is a proxy to the underlying frequency-frequency correlation function (FFCF) of the vibration. Also, the vibration experiences relaxation along waiting time as shown in Fig. 1.4b.

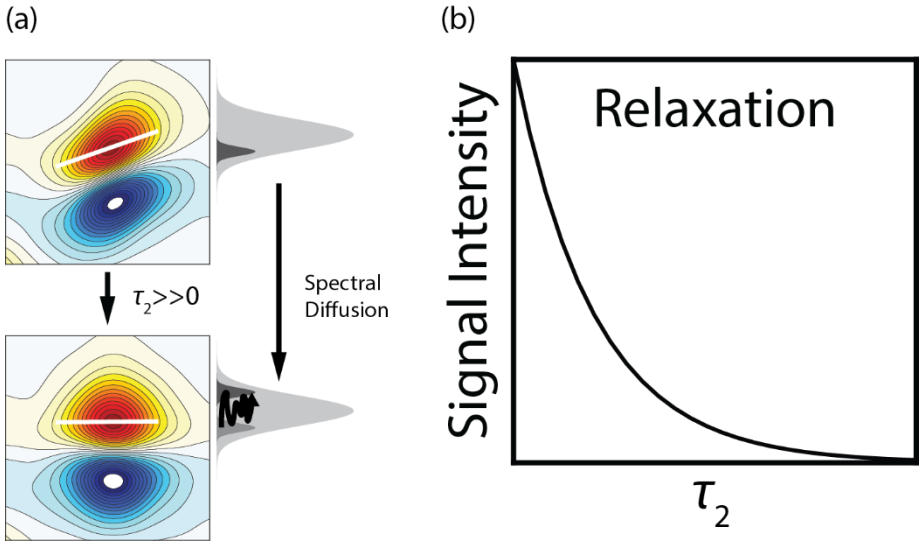

Figure 1.4: (a) Schematic illustration of spectral diffusion. White line indicates the center line slope (b) Schematic illustration of the vibrational relaxation as a function of waiting time τ_2.

Additional features of the 2D spectra can be found in the cross peaks, which probes vibrational coupling, energy transfer between coupled vibrations, and chemical exchange processes.[43] Two coupled vibrations with different diagonal frequencies exhibit off-diagonal cross-peak doublets at zero waiting time (Fig. 1.5). Increasing waiting time leads to the intensity growth of the cross peaks, which can be interpreted as the energy transferring from one vibration to the other (VET), or the vibration undergoes changes on the environments or chemical processes to have the same frequency as the other (chemical exchange). Manipulating the polarization of the pulse sequence can be used to infer the underlying orientations of the coupled vibrations. For instance in Fig. 1.5b, a coupled vibration with perpendicular orientation has preferentially enhanced cross-peak intensities when the perpendicular polarization of the pulses (ZZYY) is used.

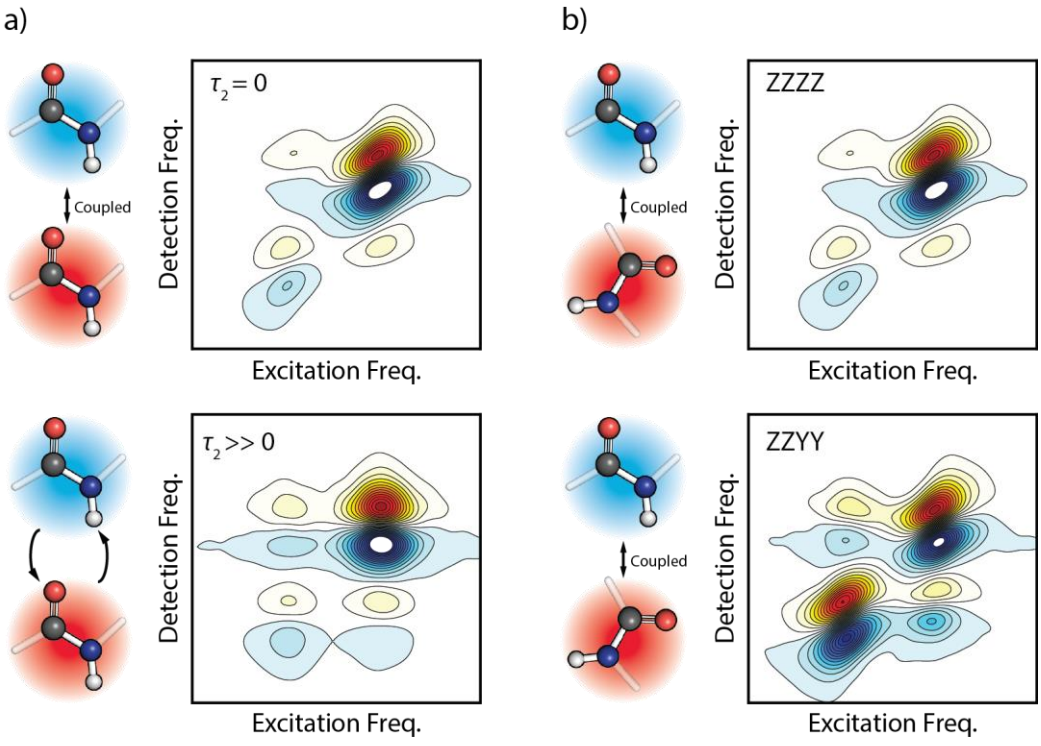

Figure 1.5: (a) Schematic illustration of vibrational energy transfer and chemical exchange in the parallel (ZZZZ) polarization. (b) Schematic illustration of the orientational preference on coupled vibrations using both ZZZZ and ZZYY polarization.

All of the information can help us interrogate the molecular dynamics of the protein vibrations. The extensive use of such information can be found in Chapter 6 for the investigation of peptide-water interaction dynamics. The vibrational probe to protein structures and dynamics throughout the thesis is amide I vibration, which is introduced below.

1.2.2 Amide I Spectroscopy

Amide I vibration is one of the backbone amide group vibration that consists of primarily C=O stretch with N–H wagging, with the typical frequency ranging from 1600–1700 cm^{-1} (Fig. 1.6). Amide I vibration has been a robust, label-free probe to protein structures,[56-57] and applied to analyze secondary structural content.[3, 57-58] The sensitivity to secondary structures comes from sub-Å sensitivity and distinct orientational dependence to the hydrogen bonds,[59-60] as well as delocalization of the amide I vibrations across the protein backbone. Such exciton-like energetic structure leads to the capability of sensing spatial arrangement of nearby amide oscillators. Hence, different secondary structures exhibit distinct frequency distributions and spectral lineshape on the amide I infrared spectrum.[2, 56]

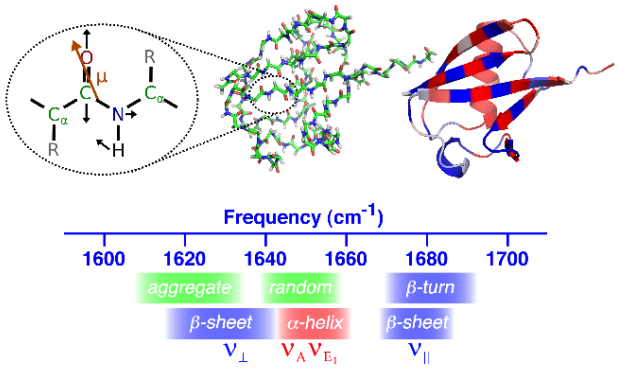

Figure 1.6: Amide I spectroscopy of proteins: (top) atom displacements and transition dipole moment associated with amide I vibrations in a single amide unit, along with the contributions of different units to a normal mode in Ubiquitin. The amplitude and phase of the vibration is encoded on the grayscale intensity: dark, light are 180 degrees out of phase. (bottom) Representation of experimentally-observed IR bands corresponding to different structural motifs. Taken from Figure 2 of Ref. 2. Copyright 2013 CRC Press.

The most distinct spectral difference comes from α-helix and β-sheet vibrational modes. As an illustration of sensitivity to protein secondary structural content, Fig. 1.7 shows a progression of IR spectra changing from myoglobin, ubiquitin, to conconavalin A. A helical structure (myoglobin) shows a featureless absorption band centered around 1640 cm^{-1}. In contrast, β-sheet exhibits two distinct peaks with the intense $\nu_{\perp}$ mode around 1630–1640 cm^{-1} and the $\nu_{\parallel}$ mode near 1680 cm^{-1}, demonstrating the capability of amide I vibration to distinguish secondary structures. However, the short vibrational lifetime of amide I vibration (1.0–1.3 ps[61-62]) lowers the structural resolution with highly congested amide I spectra and inhibits detailed structural assignments.[63]

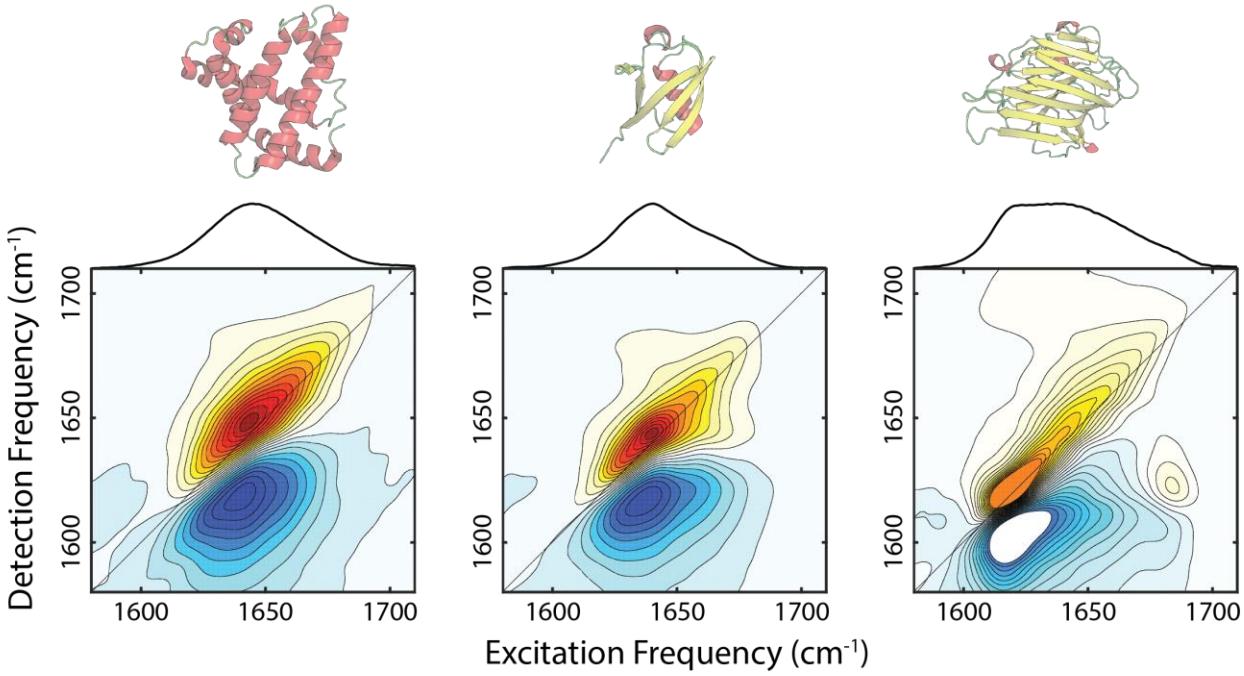

Figure 1.7: Amide I linear and 2D IR spectra of myoglobin, ubiquitin, and concanavalin A. Protein structures are shown as cartoon for reference (PDB: 1MBO, 1UBQ, and 1JBC). Data taken from Ref. 3.

Such limitation of structural interpretation can be mitigated by the use of site-specific isotope labeling.[64-65] In this approach, isotope labels such as ^{13}C or $^{13}C^{18}O$ are introduced into the specific amide group. In the classical harmonic oscillator picture, isotope labeling increases the reduced mass so that effectively it can induce a frequency shift of either -40 cm (^{13}C) or -65 cm^{-1} ($^{13}C^{18}O$). The frequency shift decouples the labeled amide group from the rest to isolate the labeled vibration and obtain structural information. Fig. 1.8 shows an example of incorporating site-specific isotope label on β-hairpin peptide TrpZip2 (TZ2).[46, 66] Upon isotopic substitution around the turn, the labeled spectrum shows two distinct peaks, which can be assigned to different peptide configurations: solvent exposed configuration and intact backbone hydrogen-bonded configuration. This label reveals conformational heterogeneity of TZ2 and cannot be observed in the unlabeled spectrum. The structural sensitivity of amide I vibrational and site-specific isotope labeling has also assisted in addressing conformational distributions and dynamics of peptides and

proteins,[41, 47, 67-68] helix-coil transition dynamics,[68-70] and ion-permeation mechanism of ion channels.[49, 71]

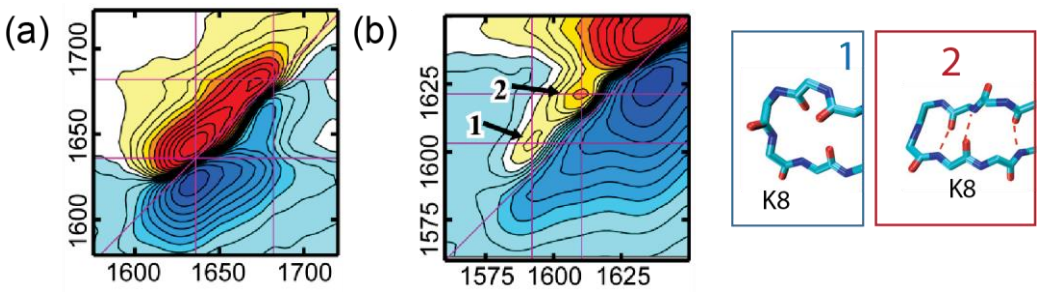

Figure 1.8: Site-specific isotope-labeling in TrpZip2 (TZ2). (a) 2D IR spectrum of unlabeled TZ2. (b) ^{13}C-labeled 2D IR spectrum of K8-labeled TZ2, which shows two distinct peaks corresponding to solvent-exposed environment (1) and intact backbone hydrogen bond environment (2), respectively. Figure modified from Figure 3 of Ref. 5. Copyright 2016 Annual Reviews.

Direct quantitative comparison of atomistic protein structures and IR experiments is now possible using computational amide I spectroscopy.[5] This method can be used to predict traditional IR absorption and 2D IR spectra for simulated conformational distributions drawn from MD trajectories or structure-based models such as Markov State Models (MSMs), [20, 72] providing a unique route to investigate structure-spectrum correlations. Specifically, amide I vibrational frequencies can be predicted to high accuracy using spectroscopic "maps" that relate frequency to the local electrostatic potential or electric field from MD force fields at specific amide carbonyl sites.[60, 73-79] Similarly, maps for vibrational coupling between different amide I vibrations are used to calculate the interaction of multiple backbone amide groups.[80-83] These maps have reached the point of predicting amide I spectroscopic observables to a high level of accuracy with 2 cm^{-1} frequency uncertainty and provided a direct way to structurally interpret experimental IR spectra.[79] This approach has additional power when interpreting site-specific isotope-edited IR spectroscopy using computational spectroscopy with MD simulations and MSMs.

1.3 Insulin Dimer Dissociation

Since the discovery of insulin, which is responsible for the regulation of blood glucose level, it has revolutionized the treatment of diabetes, and insulin has been actively studied for decades.[84] Insulin is a 51-residue peptide composed of two disulfide-bonded chains with 21 residues on chain A and 30 residues on chain B. It has three α-helical segments, a β-turn located from B20Gly to B23Gly, and the B-chain C-terminus ranging from B24Phe to B30Thr, which is involved in the association to its dimeric form (Fig. 1.9a). Insulin dimer is the key structural element for forming the hexamer with the presence of Zn^{2+} for storage, whereas monomeric form is the biologically active state in blood for binding to insulin receptor and subsequent signaling and regulation.

Structurally, B-chain C-terminus is known to be conformationally disordered depending on mutations and solution environments.[26, 85-91] Upon dimerization, this IDR folds and binds into a well-defined inter-monomer β-sheet in the native dimer structure stabilized by hydrogen bonding and packing of the aromatic triplet of B24Phe, B25Phe, and B26Tyr,[91-94] which is an example of coupled-folding and binding processes. This B-chain C-terminus is also involved in the insulin receptor binding, exhibiting detachment, a significant dihedral rotation on B24Phe, and hinging motion upon recognition with the insulin receptor,[95-98] which may share similar conformational transitions as in the dimer dissociation.[1] Also, amide I IR spectroscopy is demonstrated to be sensitive to dimer dissociation due to distinct spectral change associated with the inter-monomer β-sheet,[94, 99] and an efficient synthesis approach of site-specific isotope-labeled human insulin has been developed.[100] Insulin has become a good model system to study conformational variation and coupled-folding and binding processes.

Dynamically, a conceptual model for the coupled-folding and binding process in insulin dimerization is shown in Fig. 1.9b. Limiting pathways of monomer association includes the induced-fit pathway where the nearby monomers form transient contacts and then the folding within the monomer occurs, or the conformational selection pathway where the monomers rearrange to conformations close to the native structures and then bind. More commonly, association of protein with conformational disorder regions is multi-pathway such that both limiting mechanisms or many other trajectories can be adopted.[27, 101] Fundamental questions regarding the insulin homodimer dissociation or association can be categorized into the structure problem and the dynamics problem: (1) What is the equilibrium conformational ensemble of the monomer/dimer state? (2) What are the reaction coordinates that describe the association/dissociation dynamics?

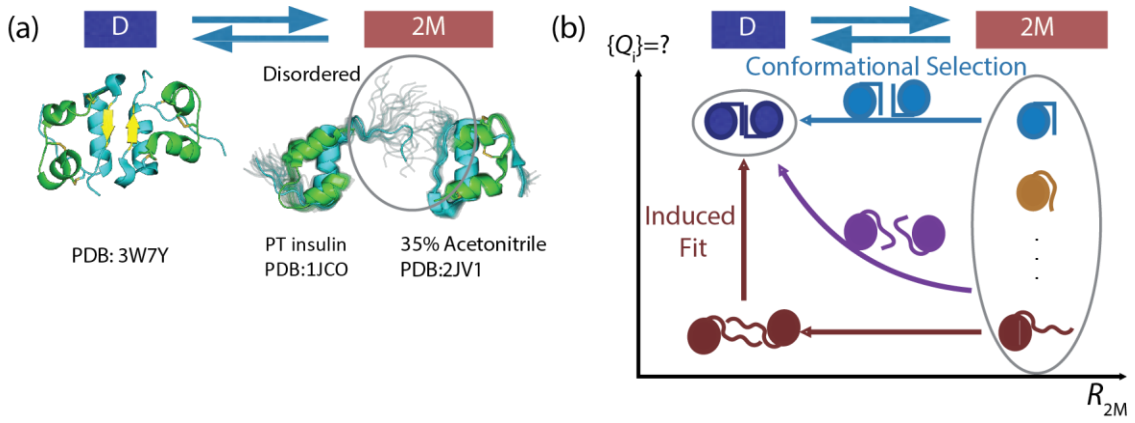

Figure 1.9: (a) Representative structures of insulin in dimer dissociation process (PDB: 3W7Y, 1JCO, and 2JV1) (b) Cartoon scenarios of insulin dimerization pathways, including limiting cases of induced fit, and conformational selection, as well as the diagonal pathway(s) of binding on the fly of folding.

1.3.1 Conformational Characterization of Insulin

The experimental conformational characterization of wild-type (WT) human insulin monomer is surprisingly challenging. WT insulin dimer has a small dissociation constant (~μM) at typical experimental conditions.[102] High-resolution structural tools such as NMR spectroscopy requires mM concentration for detectable signals, where insulin is predominantly in the dimeric state. In addition, insulin is notorious for irreversible fibril formation at destabilizing conditions for the dimer, such as low pH, high temperature, high ionic strength, character of ions and agitation,[103-106] which may account for the fact that quite many structural studies of insulin monomer have been done on various mutants or chemical modifications.[85, 87-89, 107-111] Most of the mutants are designed to maintain biological potency and at the same time remain conformations close to the native state, but mutation also biases the free energy landscape of the monomer such that the resulting conformational distribution can be significantly altered. For instance, mutation of B24 residue from Phe to Gly leads to completely disordered B-chain C-terminus.[85] Another example is the PT insulin that swaps the residues of B28Pro and B29Thr exhibits fully extended conformations in contrast to compact native structure of the monomer.[26, 88] All of these factors hinder the conformation characterization of WT insulin monomer.

Interestingly, structural studies have focused almost entirely on the conformational characterization in the monomeric state, with the current viewpoint that the dimer structure resembles published crystal structures. However, recent simulation study on the mutant dimer, showed that mutation of B24Phe to Gly resulted in additional dimer conformations including strongly interacting dimer and weakly interacting dimer, which involves conformational change between B10His and B13Glu, and increased solvation of the dimer interface.[112] A detailed description of the dimer conformational distribution still requires investigation, in particular

accounting for changes of solvation environment such as pH, ionic strength, and temperature that can mediate dimer conformational changes and in turn dynamics of dimer dissociation and association.[113-114]

1.3.2 Kinetics and Dynamics of Dimer Dissociation/Association

An early kinetic study of the insulin dimerization utilized stopped-flow technique and capacitor-discharge temperature-jump (T-jump) experiment that had μs heating time.[115] The observed kinetics behaved as two-state kinetics even though it was claimed that faster kinetics may not be resolved due to limitation of the experiments. Laser-induced T-jump amide I IR spectroscopy of bovine insulin showed non-two-state kinetic behavior.[99] Prior to the full dimer dissociation on the order of hundreds of μs, the additional kinetic component was observed in the time scale of 5–150 μs, which was assigned to conformational rearrangement of the monomers in the dimeric state. Time-resolved X-ray solution scattering experiment also suggest additional kinetic intermediates although the detailed assignment differ from the amide I IR spectroscopy. These experiments qualitatively show additional conformational dynamics other than full dissociation; the structural information is nonetheless limited.

Recent computational studies have made significant progress on structural interpretations of the dissociation/association. Atomistic MD simulations of insulin association suggests that successful association events occur when the encounter complex already exhibits similar orientations as in the native complex state.[116] When non-native contacts are made, dissociation and re-association occurs instead of remaining spatial proximity and extensive search around the protein surface. Additionally, the monomer conformation exhibits little conformational variation. Bagchi and co-workers utilized metadynamics to compute the free energy as a function of the

number of inter-monomer contacts (n_{MM}) and center-of-mass distance between the two monomers.[117] The minimum energy path suggests a single pathway of dissociation in which n_{MM} decrease drastically before the separation of the two monomers, and that the intra-molecular rearrangement of B24Phe and B26Tyr is coupled to the progress of dissociation. Antoszeski et al. performed a combination of Replica Exchange MD and computational amide I spectroscopy to characterize free energy of dimer dissociation using additional structural collective variables of inter-facial contacts on B-chain α-helix and β-sheet.[1] Diverse and energetically similar pathways were found (Fig. 1.10 or Figure 5 in Ref. 1), suggesting the multipathway nature in the dimer dissociation. One limiting pathway (α-pathway) of the dissociation involves solvation on the interfacial α-helices first and the subsequent detachment of the B-chain C-terminus, which is also appearing in the receptor binding process, whereas the other limiting pathway (β-pathway) solvates the interfacial β-strands first and exhibits little detachment. The solvation of these interfacial residues were proposed to correlate with spectroscopic changes on isotope-edited IR spectroscopy, serving as structure-based models that can be experimentally tested.[1]

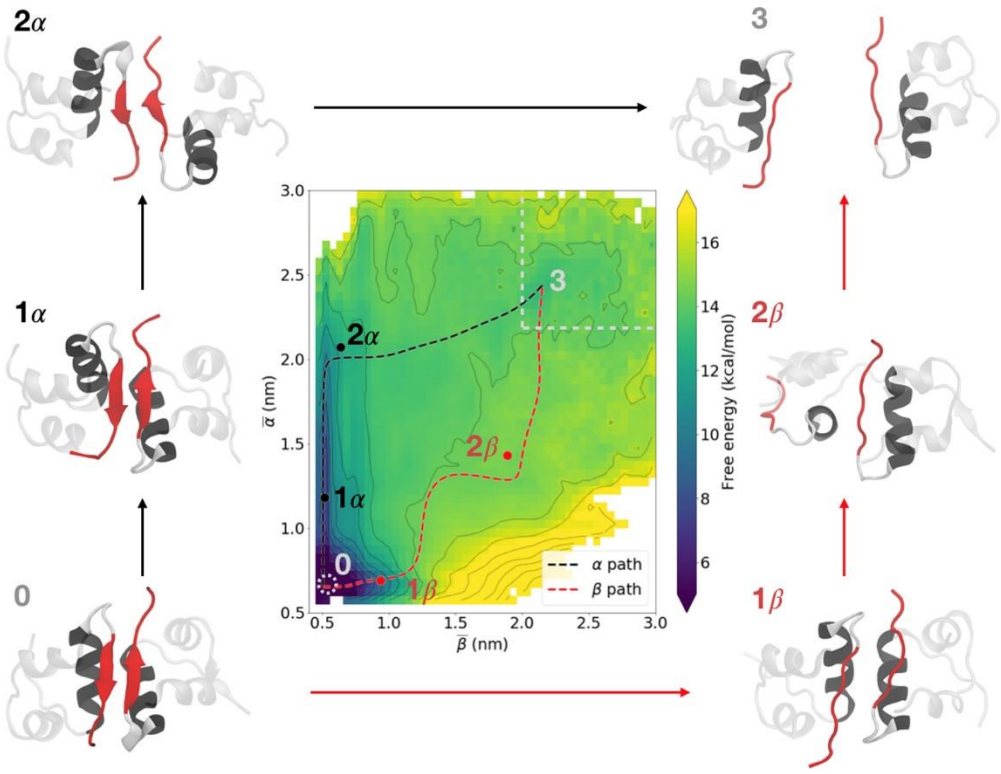

Figure 1.10: Potential of mean force (PMF) as a function of the average distance of α contact ($\bar{\alpha}$) and β contact ($\bar{\beta}$). Limiting mean free energy paths in which the interfacial α or β contacts break first are indicated by black and red dashed lines, respectively. Representative structures corresponding to the marked points along the paths are labeled and shown adjacent to the PMF. The dimer is marked by a dotted white circle, and the monomeric state is marked by a dotted white box. Contour lines are every 2 k_BT. The color scale is capped at both the upper and lower ends to more clearly show the variation in the partially dissociated regime. Figure taken from Figure 5 of Ref. 1. Copyright 2020 The American Chemical Society.

1.4 Thesis Outline

The content of this thesis is summarized below. Chapter 2 covers the theory of linear and non-linear IR spectroscopy, which is a well-established theory for analyzing the optical signals of non-linear spectroscopy. This chapter will not be exhaustive to go through every detail of the theoretical derivation but still discuss relevant assumptions and how to connect molecular response to optical signals. The framework discussed will be used throughout the thesis for both

experimental analysis and computational spectroscopic simulations. Chapter 3 contains the details of computational spectroscopy. This chapter provides the foundation for spectroscopic maps, methods to calculate non-linear response functions, and recent success on the application of 2D IR spectral simulations feasible for large protein systems such as insulin dimer. Chapter 4 consists of experimental characterization of human insulin. This chapter will present a brief summary of IR spectroscopy used for peptides and proteins, and recent progress on experimental characterization of human insulin that makes equilibrium and transient IR experiments feasible for isotope-edited insulin. Chapter 5 provides a comprehensive collections of analysis methods tied to experimental and computational protein IR spectroscopy. This chapter includes thermodynamic models that are useful for insulin homodimer dissociation and unfolding processes, maximum entropy method of reconstructing physically meaning pure components from experimental data, and ensemble refinement frameworks to infer the most consistent computational conformational ensembles against experimental information.

Chapter 6 presents a joint experimental and computational work of the dynamics interaction between dialanine and water with the focus on how water mediates conformational exchange of the peptide. The results show detailed spectroscopic and computational analysis on waiting-time dependent amide I 2D IR spectroscopy and present a common solvent description of vibrational frequency fluctuations, vibrational energy relaxation, and chemical exchange processes from coupled water and peptide motions. Chapter 7 presents a proof-of-concept study on applying ensemble refinement to trialanine against isotope-edited amide I spectroscopy. This chapter provides systematic evaluations of potential factors that can influence the quality of ensemble refinement on experimental data and computational models. Chapter 8 presents a computational study on characterizing conformational heterogeneity of insulin dimer in aqueous solution using

Markov State Models. We found an additional twisted conformation not observed in the previous studies on insulin dimer, and used computational spectroscopy to predict potential experiments to distinguish such conformation from native dimer. Chapter 9 presents an experimental study on characterizing conformational equilibrium of insulin dimer using isotope-edited amide I spectroscopy. Isotope labels show the additional spectroscopic consistent with the twisted dimer presented in the previous chapter, and ion parameters such as ionic strength is the key factor influencing the conformational equilibrium between native and twisted dimer. These two chapters laid foundation of the dimer conformational ensemble for analyzing future kinetic studies on dimer dissociations in the future.

1.5 References

1. Antoszewski, A.; Feng, C. J.; Vani, B. P.; Thiede, E. H.; Hong, L.; Weare, J.; Tokmakoff, A.; Dinner, A. R., Insulin Dissociates by Diverse Mechanisms of Coupled Unfolding and Unbinding. *J Phys Chem B* **2020**, 5571-87.
2. Baiz, C.; Reppert, M.; Tokmakoff, A., An Introduction to Protein 2D IR Spectroscopy. In *Ultrafast Infrared Vibrational Spectroscopy*, Fayer, M. D., Ed. Taylor & Francis: New York, 2013; pp 361-404.
3. Baiz, C. R.; Peng, C. S.; Reppert, M. E.; Jones, K. C.; Tokmakoff, A., Coherent two-dimensional infrared spectroscopy: quantitative analysis of protein secondary structure in solution. *Analyst* **2012,** *137* (8), 1793-9.
4. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958,** *181* (4610), 662-6.
5. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
6. Allerhand, A.; Doddrell, D.; Glushko, V.; Cochran, D. W.; Wenkert, E.; Lawson, P. J.; Gurd, F. R., Conformation and segmental motion of native and denatured ribonuclease A in solution. Application of natural-abundance carbon-13 partially relaxed Fourier transform nuclear magnetic resonance. *J Am Chem Soc* **1971,** *93* (2), 544-6.
7. Petsko, G. A.; Ringe, D., *Protein structure and function*. New Science Press: 2004.
8. Cheung, M. S.; Garcia, A. E.; Onuchic, J. N., Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci U S A* **2002,** *99* (2), 685-90.

9. Head-Gordon, T., Minimalist models for protein folding and design. *Current Opinion in Structural Biology* **2003,** *13* (2), 160-167.
10. Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G., Water in protein structure prediction. *Proc Natl Acad Sci U S A* **2004,** *101* (10), 3352-7.
11. Bagchi, B., *Water in Biological and Chemical Processes*

From Structure and Dynamics to Function. Cambridge University Press 2013.
12. Tompa, P., Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* **2012,** *37* (12), 509-16.
13. Papoian, G. A., Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proc Natl Acad Sci U S A* **2008,** *105* (38), 14237-8.
14. Burger, V.; Gurry, T.; Stultz, C., Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **2014,** *6* (10), 2684-2719.
15. Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M., Controlling entropy to tune the functions of intrinsically disordered regions. *Curr Opin Struct Biol* **2014,** *26*, 62-72.
16. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M., Classification of intrinsically disordered regions and proteins. *Chem Rev* **2014,** *114* (13), 6589-631.
17. Bernardi, R. C.; Melo, M. C. R.; Schulten, K., Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta* **2015,** *1850* (5), 872-877.
18. Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q., Enhanced sampling in molecular dynamics. *J Chem Phys* **2019,** *151* (7), 070902.
19. Chodera, J. D.; Noe, F., Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* **2014,** *25*, 135-44.
20. Husic, B. E.; Pande, V. S., Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018,** *140* (7), 2386-2396.
21. Buchenberg, S.; Schaudinnus, N.; Stock, G., Hierarchical Biomolecular Dynamics: Picosecond Hydrogen Bonding Regulates Microsecond Conformational Transitions. *J Chem Theory Comput* **2015,** *11* (3), 1330-6.
22. Papoian, G. A.; Ulander, J.; Wolynes, P. G., Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc* **2003,** *125* (30), 9170-8.
23. Levy, Y.; Onuchic, J. N., Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct* **2006,** *35*, 389-415.
24. Lim, V. I.; Curran, J. F.; Garber, M. B., Hydration shells of molecules in molecular association: A mechanism for biomolecular recognition. *J Theor Biol* **2012,** *301*, 42-8.
25. Bryant, R. G., The NMR time scale. *Journal of Chemical Education* **1983,** *60* (11), 933.
26. Bocian, W.; Sitkowski, J.; Bednarek, E.; Tarnowska, A.; Kawecki, R.; Kozerski, L., Structure of human insulin monomer in water/acetonitrile solution. *J Biomol NMR* **2008,** *40* (1), 55-64.
27. Kim, J. Y.; Chung, H. S., Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science* **2020,** *368* (6496), 1253-1257.
28. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M., Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **2017,** *42*, 106-116.
29. Rieping, W.; Habeck, M.; Nilges, M., Inferential structure determination. *Science* **2005,** *309* (5732), 303-6.

30. Pitera, J. W.; Chodera, J. D., On the Use of Experimental Observations to Bias Simulated Ensembles. *J Chem Theory Comput* **2012,** *8* (10), 3445-51.
31. Roux, B.; Weare, J., On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys* **2013,** *138* (8), 084107.
32. Cavalli, A.; Camilloni, C.; Vendruscolo, M., Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J Chem Phys* **2013,** *138* (9), 094112.
33. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K., Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* **2014,** *10* (2), e1003406.
34. Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noe, F., Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc Natl Acad Sci U S A* **2017,** *114* (31), 8265-8270.
35. Beauchamp, K. A.; Pande, V. S.; Das, R., Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys J* **2014,** *106* (6), 1381-90.
36. Xiao, X.; Kallenbach, N.; Zhang, Y., Peptide Conformation Analysis Using an Integrated Bayesian Approach. *J Chem Theory Comput* **2014,** *10* (9), 4152-4159.
37. Brookes, D. H.; Head-Gordon, T., Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J Am Chem Soc* **2016,** *138* (13), 4530-8.
38. Fisher, C. K.; Huang, A.; Stultz, C. M., Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc* **2010,** *132* (42), 14919-27.
39. Rozycki, B.; Kim, Y. C.; Hummer, G., SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011,** *19* (1), 109-16.
40. Shevchuk, R.; Hub, J. S., Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput Biol* **2017,** *13* (10), e1005800.
41. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
42. Sethi, A.; Anunciado, D.; Tian, J.; Vu, D. M.; Gnanakaran, S., Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics. *Chem Phys* **2013,** *422*.
43. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press: 2011.
44. Baiz, C. R.; Reppert, M.; Tokmakoff, A., An Introduction to Protein 2D IR Spectroscopy. *Ultrafast Infrared Vibrational Spectroscopy* **2013**, 361-403.
45. Woutersen, S.; Hamm, P., Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J Phys Chem B* **2000,** *104* (47), 11316-11320.
46. Smith, A. W.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A.; Roy, S.; Jansen, T. L.; Knoester, J., Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J Phys Chem B* **2010,** *114* (34), 10913-24.
47. Baiz, C. R.; Tokmakoff, A., Structural disorder of folded proteins: isotope-edited 2D IR spectroscopy and Markov state modeling. *Biophys J* **2015,** *108* (7), 1747-1757.
48. Feng, Y.; Huang, J.; Kim, S.; Shim, J. H.; MacKerell, A. D., Jr.; Ge, N. H., Structure of Penta-Alanine Investigated by Two-Dimensional Infrared Spectroscopy and Molecular Dynamics Simulation. *J Phys Chem B* **2016,** *120* (24), 5325-39.
49. Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D.; Skinner, J. L.; Perozo, E.; Roux, B.; Valiyaveetil,

F. I.; Zanni, M. T., Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016,** *353* (6303), 1040-1044.
50. Ghosh, A.; Ostrander, J. S.; Zanni, M. T., Watching Proteins Wiggle: Mapping Structures with Two-Dimensional Infrared Spectroscopy. *Chem Rev* **2017,** *117* (16), 10726-10759.
51. Buchanan, L. E.; Dunkelberger, E. B.; Tran, H. Q.; Cheng, P. N.; Chiu, C. C.; Cao, P.; Raleigh, D. P.; de Pablo, J. J.; Nowick, J. S.; Zanni, M. T., Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. *Proc Natl Acad Sci U S A* **2013,** *110* (48), 19285-90.
52. Lomont, J. P.; Ostrander, J. S.; Ho, J. J.; Petti, M. K.; Zanni, M. T., Not All beta-Sheets Are the Same: Amyloid Infrared Spectra, Transition Dipole Strengths, and Couplings Investigated by 2D IR Spectroscopy. *J Phys Chem B* **2017,** *121* (38), 8935-8945.
53. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR Spectroscopy: Molecular Structure and Dynamics in Solution. *The Journal of Physical Chemistry A* **2003,** *107* (27), 5258-5279.
54. Kwak, K.; Park, S.; Finkelstein, I. J.; Fayer, M. D., Frequency-frequency correlation functions and apodization in two-dimensional infrared vibrational echo spectroscopy: a new approach. *J Chem Phys* **2007,** *127* (12), 124503.
55. Fenn, E. E.; Fayer, M. D., Extracting 2D IR frequency-frequency correlation functions from two component systems. *J Chem Phys* **2011,** *135* (7), 074502.
56. Barth, A.; Zscherp, C., What vibrations tell us about proteins. *Q Rev Biophys* **2002,** *35* (4), 369-430.
57. Dousseau, F.; Pezolet, M., Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods. *Biochemistry* **1990,** *29* (37), 8771-9.
58. Dong, A.; Huang, P.; Caughey, W. S., Protein secondary structures in water from second-derivative amide I infrared spectra. *Biochemistry* **1990,** *29* (13), 3303-8.
59. Hamm, P.; Lim, M.; Hochstrasser, R. M., Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *The Journal of Physical Chemistry B* **1998,** *102* (31), 6123-6138.
60. Torii, H., Amide I Vibrational Properties Affected by Hydrogen Bonding Out-of-Plane of the Peptide Group. *J Phys Chem Lett* **2015,** *6* (4), 727-33.
61. Hamm, P.; Lim, M.; DeGrado, W. F.; Hochstrasser, R. M., The two-dimensional IR nonlinear spectroscopy of a cyclic penta-peptide in relation to its three-dimensional structure. *Proc Natl Acad Sci U S A* **1999,** *96* (5), 2036-41.
62. Feng, C. J.; Tokmakoff, A., The dynamics of peptide-water interactions in dialanine: An ultrafast amide I 2D IR and computational spectroscopy study. *J Chem Phys* **2017,** *147* (8), 085101.
63. Bredenbeck, J.; Hamm, P., Peptide structure determination by two-dimensional infrared spectroscopy in the presence of homogeneous and inhomogeneous broadening. *The Journal of Chemical Physics* **2003,** *119* (3), 1569-1578.
64. Torres, J.; Kukol, A.; Goodman, J. M.; Arkin, I. T., Site-specific examination of secondary structure and orientation determination in membrane proteins: The peptidic13C?18O group as a novel infrared probe. *Biopolymers* **2001,** *59* (6), 396-401.
65. Decatur, S. M., Elucidation of residue-level structure and dynamics of polypeptides via isotope-edited infrared spectroscopy. *Acc Chem Res* **2006,** *39* (3), 169-75.
66. Jones, K. C.; Peng, C. S.; Tokmakoff, A., Folding of a heterogeneous beta-hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc Natl Acad Sci U S A* **2013,** *110* (8), 2828-33.

67. Woutersen, S.; Hamm, P., Isotope-edited two-dimensional vibrational spectroscopy of trialanine in aqueous solution. *The Journal of Chemical Physics* **2001,** *114* (6), 2727-2737.
68. Meuzelaar, H.; Marino, K. A.; Huerta-Viga, A.; Panman, M. R.; Smeenk, L. E.; Kettelarij, A. J.; van Maarseveen, J. H.; Timmerman, P.; Bolhuis, P. G.; Woutersen, S., Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations. *J Phys Chem B* **2013,** *117* (39), 11490-501.
69. Huang, C. Y.; Getahun, Z.; Zhu, Y.; Klemke, J. W.; DeGrado, W. F.; Gai, F., Helix formation via conformation diffusion search. *Proc Natl Acad Sci U S A* **2002,** *99* (5), 2788-93.
70. Tucker, M. J.; Abdo, M.; Courter, J. R.; Chen, J.; Brown, S. P.; Smith, A. B., 3rd; Hochstrasser, R. M., Nonequilibrium dynamics of helix reorganization observed by transient 2D IR spectroscopy. *Proc Natl Acad Sci U S A* **2013,** *110* (43), 17314-9.
71. Stevenson, P.; Gotz, C.; Baiz, C. R.; Akerboom, J.; Tokmakoff, A.; Vaziri, A., Visualizing KcsA conformational changes upon ion binding by infrared spectroscopy and atomistic modeling. *J Phys Chem B* **2015,** *119* (18), 5824-31.
72. Noe, F.; Rosta, E., Markov Models of Molecular Kinetics. *J Chem Phys* **2019,** *151* (19), 190401.
73. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.
74. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.
75. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
76. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
77. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.
78. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
79. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
80. Torii, H.; Tasumi, M., Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *Journal of Raman Spectroscopy* **1998,** *29* (1), 81-86.
81. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
82. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
83. Hayashi, T.; Mukamel, S., Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides. *J Phys Chem B* **2007,** *111* (37), 11032-46.
84. Vecchio, I.; Tornali, C.; Bragazzi, N. L.; Martini, M., The Discovery of Insulin: An Important Milestone in the History of Medicine. *Front Endocrinol (Lausanne)* **2018,** *9*, 613.

85. Hua, Q. X.; Shoelson, S. E.; Kochoyan, M.; Weiss, M. A., Receptor binding redefined by a structural switch in a mutant human insulin. *Nature* **1991,** *354* (6350), 238-41.
86. Ludvigsen, S.; Roy, M.; Thogersen, H.; Kaarsholm, N. C., High-resolution structure of an engineered biologically potent insulin monomer, B16 Tyr-->His, as determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **1994,** *33* (26), 7998-8006.
87. Olsen, H. B.; Ludvigsen, S.; Kaarsholm, N. C., Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry* **1996,** *35* (27), 8836-45.
88. Keller, D.; Clausen, R.; Josefsen, K.; Led, J. J., Flexibility and bioactivity of insulin: an NMR investigation of the solution structure and folding of an unusually flexible human insulin mutant with increased biological activity. *Biochemistry* **2001,** *40* (35), 10732-40.
89. Ludvigsen, S.; Olsen, H. B.; Kaarsholm, N. C., A structural switch in a mutant insulin exposes key residues for receptor binding. *J Mol Biol* **1998,** *279* (1), 1-7.
90. Kosinova, L.; Veverka, V.; Novotna, P.; Collinsova, M.; Urbanova, M.; Moody, N. R.; Turkenburg, J. P.; Jiracek, J.; Brzozowski, A. M.; Zakova, L., Insight into the structural and biological relevance of the T/R transition of the N-terminus of the B-chain in human insulin. *Biochemistry* **2014,** *53* (21), 3392-402.
91. Zoete, V.; Meuwly, M.; Karplus, M., A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. *J Mol Biol* **2004,** *342* (3), 913-29.
92. Baker, E. N.; Blundell, T. L.; Cutfield, J. F.; Cutfield, S. M.; Dodson, E. J.; Dodson, G. G.; Hodgkin, D. M.; Hubbard, R. E.; Isaacs, N. W.; Reynolds, C. D.; et al., The structure of 2Zn pig insulin crystals at 1.5 A resolution. *Philos Trans R Soc Lond B Biol Sci* **1988,** *319* (1195), 369-456.
93. Jørgensen, A. M. M.; Kristensen, S. M.; Led, J. J.; Balschmidt, P., Three-dimensional solution structure of an insulin dimer. *Journal of Molecular Biology* **1992,** *227* (4), 1146-1163.
94. Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy. *Phys Chem Chem Phys* **2010,** *12* (14), 3579-88.
95. Menting, J. G.; Yang, Y.; Chan, S. J.; Phillips, N. B.; Smith, B. J.; Whittaker, J.; Wickramasinghe, N. P.; Whittaker, L. J.; Pandyarajan, V.; Wan, Z. L.; Yadav, S. P.; Carroll, J. M.; Strokes, N.; Roberts, C. T., Jr.; Ismail-Beigi, F.; Milewski, W.; Steiner, D. F.; Chauhan, V. S.; Ward, C. W.; Weiss, M. A.; Lawrence, M. C., Protective hinge in insulin opens to enable its receptor engagement. *Proc Natl Acad Sci U S A* **2014,** *111* (33), E3395-404.
96. Croll, T. I.; Smith, B. J.; Margetts, M. B.; Whittaker, J.; Weiss, M. A.; Ward, C. W.; Lawrence, M. C., Higher-Resolution Structure of the Human Insulin Receptor Ectodomain: Multi-Modal Inclusion of the Insert Domain. *Structure* **2016,** *24* (3), 469-76.
97. Gutmann, T.; Kim, K. H.; Grzybek, M.; Walz, T.; Coskun, U., Visualization of ligand-induced transmembrane signaling in the full-length human insulin receptor. *J Cell Biol* **2018,** *217* (5), 1643-1649.
98. Weis, F.; Menting, J. G.; Margetts, M. B.; Chan, S. J.; Xu, Y.; Tennagels, N.; Wohlfart, P.; Langer, T.; Muller, C. W.; Dreyer, M. K.; Lawrence, M. C., The signalling conformation of the insulin receptor ectodomain. *Nat Commun* **2018,** *9* (1), 4420.
99. Zhang, X. X.; Jones, K. C.; Fitzpatrick, A.; Peng, C. S.; Feng, C. J.; Baiz, C. R.; Tokmakoff, A., Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J Phys Chem B* **2016,** *120* (23), 5134-45.

100. Dhayalan, B.; Fitzpatrick, A.; Mandal, K.; Whittaker, J.; Weiss, M. A.; Tokmakoff, A.; Kent, S. B., Efficient Total Chemical Synthesis of (13) C=(18) O Isotopomers of Human Insulin for Isotope-Edited FTIR. *Chembiochem* **2016,** *17* (5), 415-20.
101. Gianni, S.; Dogan, J.; Jemth, P., Coupled binding and folding of intrinsically disordered proteins: what can we learn from kinetics? *Curr Opin Struct Biol* **2016,** *36*, 18-24.
102. Antolikova, E.; Zakova, L.; Turkenburg, J. P.; Watson, C. J.; Hanclova, I.; Sanda, M.; Cooper, A.; Kraus, T.; Brzozowski, A. M.; Jiracek, J., Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J Biol Chem* **2011,** *286* (42), 36968-77.
103. Sluzky, V.; Klibanov, A. M.; Langer, R., Mechanism of insulin aggregation and stabilization in agitated aqueous solutions. *Biotechnol Bioeng* **1992,** *40* (8), 895-903.
104. Nielsen, L.; Khurana, R.; Coats, A.; Frokjaer, S.; Brange, J.; Vyas, S.; Uversky, V. N.; Fink, A. L., Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. *Biochemistry* **2001,** *40* (20), 6036-46.
105. Whittingham, J. L.; Scott, D. J.; Chance, K.; Wilson, A.; Finch, J.; Brange, J.; Guy Dodson, G., Insulin at pH 2: Structural Analysis of the Conditions Promoting Insulin Fibre Formation. *Journal of Molecular Biology* **2002,** *318* (2), 479-490.
106. Haas, J.; Vohringer-Martinez, E.; Bogehold, A.; Matthes, D.; Hensen, U.; Pelah, A.; Abel, B.; Grubmuller, H., Primary steps of pH-dependent insulin aggregation kinetics are governed by conformational flexibility. *Chembiochem* **2009,** *10* (11), 1816-22.
107. Hua, Q. X.; Weiss, M. A., Comparative 2D NMR studies of human insulin and des-pentapeptide insulin: sequential resonance assignment and implications for protein dynamics and receptor recognition. *Biochemistry* **1991,** *30* (22), 5505-15.
108. Knegtel, R. M.; Boelens, R.; Ganadu, M. L.; Kaptein, R., The solution structure of a monomeric insulin. A two-dimensional 1H-NMR study of des-(B26-B30)-insulin in combination with distance geometry and restrained molecular dynamics. *Eur J Biochem* **1991,** *202* (2), 447-58.
109. Sorensen, M. D.; Led, J. J., Structural details of Asp(B9) human insulin at low pH from two-dimensional NMR titration studies. *Biochemistry* **1994,** *33* (46), 13727-33.
110. Jorgensen, A. M.; Olsen, H. B.; Balschmidt, P.; Led, J. J., Solution structure of the superactive monomeric des-[Phe(B25)] human insulin mutant: elucidation of the structural basis for the monomerization of des-[Phe(B25)] insulin and the dimerization of native insulin. *J Mol Biol* **1996,** *257* (3), 684-99.
111. Zakova, L.; Kletvikova, E.; Lepsik, M.; Collinsova, M.; Watson, C. J.; Turkenburg, J. P.; Jiracek, J.; Brzozowski, A. M., Human insulin analogues modified at the B26 site reveal a hormone conformation that is undetected in the receptor complex. *Acta Crystallogr D Biol Crystallogr* **2014,** *70* (Pt 10), 2765-74.
112. Raghunathan, S.; El Hage, K.; Desmond, J. L.; Zhang, L.; Meuwly, M., The Role of Water in the Stability of Wild-type and Mutant Insulin Dimers. *J Phys Chem B* **2018,** *122* (28), 7038-7048.
113. Wicky, B. I. M.; Shammas, S. L.; Clarke, J., Affinity of IDPs to their targets is modulated by ion-specific changes in kinetics and residual structure. *Proc Natl Acad Sci U S A* **2017,** *114* (37), 9882-9887.
114. Boreikaite, V.; Wicky, B. I. M.; Watt, I. N.; Clarke, J.; Walker, J. E., Extrinsic conditions influence the self-association and structure of IF1, the regulatory protein of mitochondrial ATP synthase. *Proc Natl Acad Sci U S A* **2019,** *116* (21), 10354-10359.

115. Koren, R.; Hammes, G. G., A kinetic study of protein-protein interactions. *Biochemistry* **1976,** *15* (5), 1165-71.
116. Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M.; Shaw, D. E., Atomic-level characterization of protein-protein association. *Proc Natl Acad Sci U S A* **2019,** *116* (10), 4244-4249.
117. Banerjee, P.; Mondal, S.; Bagchi, B., Insulin dimer dissociation in aqueous solution: A computational study of free energy landscape and evolving microscopic structure along the reaction pathway. *J Chem Phys* **2018,** *149* (11), 114902.

Chapter 2

Theory of Nonlinear Spectroscopy

2.1 Introduction

Spectroscopy is a study of interactions between matter (usually referred to as the sample or the system) and incident light. Because of the nature of spectroscopy as a probe to the materials, we always measure some changes of the light to infer the molecular properties, which inevitably requires a model or some theoretical descriptions. As an example, measuring the absorption spectrum of a molecule provides information such as wavelength of the absorption peak(s) and the corresponding absorbance. Based on the Golden rule,[1-2] the peak wavelengths inform us about excited state energy levels of the molecule whereas the associated absorbance is proportional to oscillator strength of the transition, which is quite a remarkable connection between non-equilibrium absorption and the equilibrium molecular properties. A decent understanding of the theory of nonlinear spectroscopy is necessary to appreciate how such connection is developed, as well as obtaining a physical intuition for the spectroscopic response, in particular multi-dimensional spectroscopy such as 2D IR spectroscopy.

In principle, the fundamental aspects of light-matter interactions come from quantum Liouville equation for the quantum system that is described by the Hamiltonian $\hat{H}$ and the density operator $\hat{\rho}$

$$\frac{\partial \hat{\rho}(t)}{\partial t} = \frac{-i}{\hbar}\left[\hat{H}(t), \hat{\rho}(t)\right] \tag{2.1}$$

where the field description in the quantum Liouville equation is obtained from Maxwell's equations describing the electromagnetic radiation (in Gaussian units)

$$\nabla \cdot \boldsymbol{E}(\boldsymbol{r}, t) = 4\pi\sigma \tag{2.2}$$

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{r}, t) = 0 \tag{2.3}$$

$$\nabla \times \boldsymbol{E}(\boldsymbol{r}, t) = -\frac{1}{c}\frac{\partial \boldsymbol{B}(\boldsymbol{r}, t)}{\partial t} \tag{2.4}$$

$$\nabla \times \boldsymbol{B}(\boldsymbol{r}, t) = \frac{4\pi}{c}\boldsymbol{J}(\boldsymbol{r}, t) + \frac{1}{c}\frac{\partial \boldsymbol{E}(\boldsymbol{r}, t)}{\partial t} \tag{2.5}$$

In Eqns. (2.2)–(2.5), $\boldsymbol{E}(t)$ and $\boldsymbol{B}(t)$ are the electric field and the magnetic field given the position vector $\boldsymbol{r}$ and time t, respectively. They are related to the density variables such as the charge density σ, and the current density $\boldsymbol{J}$. c corresponds to the speed of light. Gaussian units are chosen for convenience and we are not going to compute the exact value of fields throughout the thesis. Also, it is possible to convert Gaussian units to SI units.[3]

In practice, however, it is quite involved to obtain a quantitative and detailed description of how molecular properties connect to the measured spectroscopic signals. The thorough discussion of the theory is worth more than an entire book[4-6] that goes through important approximations, assumptions, and various experimental approaches to measure the signal. All of these are impossible to be included in this chapter. This chapter, however, is aiming for the goals listed below.

1. Building the connection between the measured spectroscopic signal to the equilibrium molecular response in both linear spectroscopy and third-order nonlinear spectroscopy such as transient absorption (TA) spectroscopy, and 2D IR spectroscopy.
2. Discussing relevant assumptions and approximations so that one can know where these apply.
3. Developing the set of equations for spectroscopic simulations of amide I spectroscopy, which will be discussed in more detail in Chapter 3.
4. Discussing important spectroscopic phenomena using simple models to build physical intuition.

I would like make some additional comments before digging into the rest of this chapter. First, the treatment above for describing the light-matter interaction is already a semi-classical description, where the material is treated quantum mechanically whereas the electromagnetic radiation is treated classically.[4] For the purpose of interpreting spectroscopic responses, it already offers a reasonable description. Second, the notation of nonlinear spectroscopy is still not yet consistent throughout the multidimensional spectroscopy community for all sort of reasons. For example, I still remember back to CMDS conference 2018 in Seoul, Korea. At check-in, everyone received a short survey about what is your convention for representing a 2D spectrum like placing ω_1 as the horizontal axis or vertical axis, and supposedly the result would be announced at the end. For some reason, it never happened officially. Here, I will use the notation which I found the most sensible and convenient for me.

2.2 Maxwell's Equations and Light-Matter Interaction

2.2.1 Maxwell's Equations in Vacuum

Maxwell's equations in the Eqns. (2.2)–(2.5) describe the time evolution of the fields in vaccum, $\boldsymbol{E}(t)$ and $\boldsymbol{B}(t)$, in total of 6 inter-dependent coordinate variables that intimately depend on the spatial coordinates and time (4 variables). Therefore, Maxwell's equations are still uniquely determined with these variables. More often, Maxwell's equations are written in the potential form by introducing a scalar potential $\phi(\boldsymbol{r},t)$ and a vector potential $\boldsymbol{A}(\boldsymbol{r},t)$

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= -\nabla\phi(\boldsymbol{r},t) - \frac{1}{c}\frac{\partial \boldsymbol{A}(\boldsymbol{r},t)}{\partial t} \\ \boldsymbol{B}(\boldsymbol{r},t) &= \nabla \times \boldsymbol{A}(\boldsymbol{r},t) \end{aligned} \tag{2.6}$$

in which $\boldsymbol{A}(\boldsymbol{r},t)$ has the same direction of $\boldsymbol{E}(\boldsymbol{r},t)$. One can see that the Eqns. (2.3) and (2.4) are automatically satisfied with the introduction of these potential variables. However, the potential variables are not uniquely determined and it is required to specify additional constraints or initial/boundary conditions (choosing a gauge).[7] A common choice of gauge is Coulomb gauge:

$$\nabla \cdot \boldsymbol{A}(t) = 0 \tag{2.7}$$

Recognizing that

$$\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} \tag{2.8}$$

and that the values of the charge density σ and current density $\boldsymbol{J}$ is zero in vacuum, we can obtain the following relation from Eqn. (2.2)

$$\nabla^2 \phi(t) = 0 \tag{2.9}$$

which is the Laplace's equation. Since it is in vacuum, one can set the potential to zero.

From the Eqns. (2.5), (2.6)–(2.8), we can obtain the following

$$\nabla\left(\nabla\cdot\boldsymbol{A}(\boldsymbol{r},t)+\frac{1}{c}\frac{\partial\phi(\boldsymbol{r},t)}{\partial t}\right)=\frac{4\pi}{c}\boldsymbol{J}(\boldsymbol{r},t)-\frac{1}{c^2}\frac{\partial\boldsymbol{A}(\boldsymbol{r},t)}{\partial t}+\nabla^2\boldsymbol{A}(\boldsymbol{r},t) \tag{2.10}$$

In vacuum, the potential can be set to zero and there is no current density, meaning that

$$\nabla^2\boldsymbol{A}(\boldsymbol{r},t)=\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}(\boldsymbol{r},t)}{\partial t^2} \tag{2.11}$$

Eqn. (2.11) turns out to be a classical wave equation, in which the vector potential propagates at the speed of light *c*, and the plane wave solution is

$$\boldsymbol{A}(\boldsymbol{r},t)=\boldsymbol{A}_0\left[e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}+\text{c.c.}\right] \tag{2.12}$$

Coulomb gauge gives

$$\nabla\boldsymbol{A}(\boldsymbol{r},t)=0=-2\boldsymbol{k}\cdot\boldsymbol{A}_0\sin(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t) \tag{2.13}$$

meaning that the wavevector $\boldsymbol{k}$ is perpendicular to the amplitude vector $\boldsymbol{A}_0$. We can obtain the electromagnetic field

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r},t)&=-\frac{1}{c}\frac{\partial\boldsymbol{A}(\boldsymbol{r},t)}{\partial t}=-\frac{\omega}{c}A_0\hat{\boldsymbol{a}}\sin(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)\\ \boldsymbol{B}(\boldsymbol{r},t)&=\nabla\times\boldsymbol{A}(\boldsymbol{r},t)=-A_0(\boldsymbol{k}\times\hat{\boldsymbol{a}})\sin(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)\end{aligned} \tag{2.14}$$

where $\hat{\boldsymbol{a}}$ is a unit vector along the same direction as $\boldsymbol{A}_0$. Eqn. (2.14) indicates that the electric field propagates along the same direction as the wavevector $\boldsymbol{k}$, but the amplitude is orthogonal to the wavevector. Maxwell's equation in vacuum gives the plane wave solution, which will be useful for later discussions.

2.2.2 Light-Matter Interaction

When an electric field $\boldsymbol{E}(\boldsymbol{r},t)$ interacts with a sample, macroscopically it can create a time-dependent electric polarization density $\boldsymbol{P}(\boldsymbol{r},t)$, and this induced polarization $\boldsymbol{P}(\boldsymbol{r},t)$ radiates a signal

electromagnetic field, which carries information of the sample. Measuring the signal field can help us understand some physical properties of the sample. There are two fundamental key questions associated with this induced polarization. (1) How does the incident electric field interact with the sample? (2) What is the relation between the induced polarization and the incident electric field?

For the first question, from the microscopic point of view, the corresponding semiclassical Hamiltonian of a charged particle with its charge q interacting with the electromagnetic field is given as[7]

$$\begin{aligned}\hat{H} &= \frac{1}{2m}\left(\hat{\boldsymbol{p}} - \frac{q}{c}\boldsymbol{A}(\boldsymbol{r},t)\right)^2 + \hat{V}(\boldsymbol{r}) \\ &= \left(\frac{-\hbar^2}{2m}\nabla^2 + \hat{V}(\boldsymbol{r})\right) + \frac{i\hbar q}{2mc}\left(\nabla\boldsymbol{A}(\boldsymbol{r},t) + \boldsymbol{A}(\boldsymbol{r},t)\nabla\right) + \frac{q^2}{2mc^2}\boldsymbol{A}^2(\boldsymbol{r},t) \\ &= \hat{H}_0 + \hat{H}_{\text{int}}\end{aligned} \tag{2.15}$$

where $\hat{H}_0$ is the Hamiltonian of the charged particle without the interaction with the electromagnetic field, and $\hat{H}_{\text{int}}$ describes the interaction between the electromagnetic field and the charged particle. Under the weak field assumption and Coulomb gauge, the $\boldsymbol{A}^2$ term is negligible and the first term in $\hat{H}_{\text{int}}$ is zero, so that

$$\hat{H}_{\text{int}} \approx \frac{i\hbar q}{2mc}\left(\nabla\boldsymbol{A}(\boldsymbol{r},t) + \boldsymbol{A}(\boldsymbol{r},t)\nabla\right) = \frac{-q}{2mc}\left(\boldsymbol{A}(\boldsymbol{r},t)\cdot\hat{\boldsymbol{p}}\right) \tag{2.16}$$

2.2.3 Electric-Dipole Approximation

Plugging Eqn. (2.12) into (2.16), we can obtain

$$\hat{H}_{\text{int}} = -\frac{q}{2mc}A_0 e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\hat{\boldsymbol{a}}\cdot\hat{\boldsymbol{p}} + \text{c.c} \tag{2.17}$$

Knowing that the magnitude of the wavevector is given as

$$k = \frac{2\pi}{\lambda} \tag{2.18}$$

In the regime of visible and IR light, the wavelength is pretty long compared to the size of the particle. The inner product between the wavevector and the coordinate vector becomes pretty small, meaning that $e^{i\mathbf{k}\cdot\mathbf{r}} \approx 1$, which is the point-dipole approximation.

$$\hat{H}_{\text{int}} \approx -\frac{q}{2mc} A_0 e^{-i\omega t} \hat{\boldsymbol{a}} \cdot \hat{\boldsymbol{p}} + \text{c.c} \tag{2.19}$$

It is seemingly difficult to evaluate the momentum operator. However, recognizing that the momentum operator and the position operator are intimately related through the commutation relation:

$$\hat{\boldsymbol{p}} = \frac{im}{\hbar}\left[\hat{H}_0, \hat{\boldsymbol{r}}\right] \tag{2.20}$$

We can investigate the interaction Hamiltonian in the eigenbasis of the system Hamiltonian $\hat{H}_0$:

$$\begin{aligned} \hat{H}_0 \left|k\right\rangle &= E_k \left|k\right\rangle \\ \hat{H}_0 \left|l\right\rangle &= E_l \left|l\right\rangle \end{aligned} \tag{2.21}$$

In Eqn. (2.21), $\left|k\right\rangle$ and $\left|l\right\rangle$ are the energy eigenstates of the system Hamiltonian. The corresponding matrix elements of the interaction Hamiltonian can be written as

$$\left\langle k\right|\hat{H}_{\text{int}}\left|l\right\rangle = \frac{i}{\hbar c} q A_0 \hat{\boldsymbol{a}} \sin(\omega t) \left\langle k\right|\hat{H}_0 \hat{\boldsymbol{r}} - \hat{\boldsymbol{r}}\hat{H}_0\left|l\right\rangle \tag{2.22}$$

By introducing the frequency of the energy gap between state *k* and *l*,

$$\omega_{kl} = \frac{E_k - E_l}{\hbar} \tag{2.23}$$

and the electric dipole operator,

$$\hat{\boldsymbol{\mu}} = q\hat{\boldsymbol{r}} \tag{2.24}$$

the interaction Hamiltonian is finally written into the well-known electric dipole Hamiltonian

$$\begin{aligned}\left\langle k\left|\hat{H}_{\text{int}}\right|l\right\rangle &= -\left(\frac{E_k - E_l}{\hbar\omega}\right)\left\langle k\left|q\hat{\boldsymbol{r}}\right|l\right\rangle\left(-\frac{\omega}{c}iA_0\hat{\boldsymbol{a}}\sin\left(\omega t\right)\right) \\ &= -\frac{\omega_{kl}}{\omega}\left\langle k\left|\hat{\boldsymbol{\mu}}\right|l\right\rangle\cdot\boldsymbol{E}\left(\boldsymbol{r},t\right)\end{aligned} \tag{2.25}$$

$$\hat{H}_{\text{int}} \approx -\hat{\boldsymbol{\mu}}\cdot\boldsymbol{E}\left(\boldsymbol{r},t\right) \tag{2.26}$$

The final approximation made here is the resonance condition, which is almost always true due to quantized nature of the system so that transition occurs only when the field frequency matches the frequency difference between *k* and *l*. For a system with many charged particles, the corresponding electric dipole operator is

$$\hat{\boldsymbol{\mu}} = \sum_i q_i\hat{\boldsymbol{r}}_i \tag{2.27}$$

The connection between macroscopic polarization density and the microscopic electric dipole is given as

$$\boldsymbol{P}\left(\boldsymbol{r},t\right) = \left\langle\hat{\boldsymbol{\mu}}\right\rangle = \text{Tr}\left(\hat{\boldsymbol{\mu}}\hat{\rho}\right) \tag{2.28}$$

Our treatment results in quite a simple electric-dipole Hamiltonian that only describes the interaction between incident electromagnetic field and the material. This electric dipole Hamiltonian has some physical interpretations. First, the diagonal elements of the electric dipole Hamiltonian are zero since $\omega_{kl} = 0$ when $k = l$. Second, this electric dipole Hamiltonian induces transition between energy eigenstates. Third, with the resonance condition, the frequency of the electric field matches the energy gap between the eigenstates, meaning that the electric field can be used to measure the energy. In summary, this semiclassical treatment of quantum particle with classical field involves weak field approximation, and electric dipole approximation. This electric dipole Hamiltonian is essentially the most important Hamiltonian for describing the response function later in the chapter.

2.2.4 Generation of a Signal Field in Dielectric Medium

For the second question that seeks the relation between incident electric field and the induced polarization, we can revisit Eqn. (2.10)

$$\nabla\left(\nabla\cdot\boldsymbol{A}(\boldsymbol{r},t)+\frac{1}{c}\frac{\partial\phi(\boldsymbol{r},t)}{\partial t}\right)=\frac{4\pi}{c}\boldsymbol{J}(\boldsymbol{r},t)-\frac{1}{c^2}\frac{\partial\boldsymbol{A}(\boldsymbol{r},t)}{\partial t}+\nabla^2\boldsymbol{A}(\boldsymbol{r},t) \tag{2.29}$$

Now, the situation is different since we are treating the dielectric medium instead of the vacuum, meaning that the current density $\boldsymbol{J}(\boldsymbol{r},t)$ may not be zero. In fact,[8]

$$\boldsymbol{J}(\boldsymbol{r},t)=\boldsymbol{J}_{\mathrm{DC}}(\boldsymbol{r},t)+\frac{\partial\boldsymbol{P}(\boldsymbol{r},t)}{\partial t} \tag{2.30}$$

in which the first term describes the current density from the direct current (DC) component, Coulomb gauge is still applicable, but we need to know additional information of the time derivative of the scalar potential. Instead, we apply the Lorenz gauge,

$$\nabla\cdot\boldsymbol{A}(\boldsymbol{r},t)+\frac{1}{c}\frac{\partial\phi(\boldsymbol{r},t)}{\partial t}=0 \tag{2.31}$$

Please note that in classical electrodynamics, potential is more like a mathematical tool, and choosing a gauge should not change the eventual relation of measurable physical observables. The following equation can be derived

$$\Box\boldsymbol{A}(\boldsymbol{r},t)=\left[\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]\boldsymbol{A}(\boldsymbol{r},t)=-\frac{4\pi}{c}\boldsymbol{J}(\boldsymbol{r},t) \tag{2.32}$$

Recognizing that the current density can be created by time-dependent induced polarization from the sample as in Eqn. (2.30) and that the gradient of the scalar potential ϕ is zero, the time derivative of Eqn. (2.32) gives

$$\Box \boldsymbol{E}(\boldsymbol{r},t)=\left[\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]\boldsymbol{E}(\boldsymbol{r},t)=\frac{4\pi}{c^2}\frac{\partial^2 \boldsymbol{P}(\boldsymbol{r},t)}{\partial t^2} \tag{2.33}$$

This inhomogeneous wave equation demonstrates that the incident electric field is intimately related to the induced polarization in the sample. Plugging the plane wave solution from Eqn. (2.14) also gives the plane wave solution for time-dependent induced polarization, but the phase is shifted by a factor of π/2. Also, the induced polarization that comes from collective induced dipoles can in turn generate the signal electric field that adds linearly to the total electric field. The polarization can be expanded in Taylor series

$$\begin{aligned}\boldsymbol{P}(\omega)&=\sum_{n=1}\boldsymbol{P}^{(n)}(\omega)=\sum_{n=1}\chi^{(n)}\boldsymbol{E}^n\\&=\boldsymbol{P}^{(1)}(\omega)+\boldsymbol{P}^{(\mathrm{NL})}(\omega)\\\boldsymbol{P}^{(\mathrm{NL})}(\omega)&=\sum_{n\geq 2}\boldsymbol{P}^{(n)}(\omega)\end{aligned} \tag{2.34}$$

Where $\chi^{(n)}$ is the n^{th}-order susceptibility in frequency domain that is intrinsic to the properties of the system, and the corresponding inverse Fourier-transformed time domain response function will be discussed below. $\boldsymbol{P}^{(\mathrm{NL})}(\omega)$ is referred to as the nonlinear electric polarization density.[8]

Now, we aim to investigate the properties of the signal field generated from the sample. From Eqn. (2.33), it is possible to show in the frequency domain, the relation between the electric field and the nonlinear polarization density is the following[8-9]

$$\left[\nabla^2-\frac{\varepsilon(\omega)}{c^2}\frac{\partial^2}{\partial t^2}\right]\boldsymbol{E}(\boldsymbol{r},t)=\frac{4\pi}{c^2}\frac{\partial^2 \boldsymbol{P}^{(\mathrm{NL})}(\boldsymbol{r},t)}{\partial t^2} \tag{2.35}$$

Or alternatively,

$$\nabla\times\nabla\times\boldsymbol{E}(\boldsymbol{r},\omega)-\frac{\omega^2}{c^2}\varepsilon(\omega)\boldsymbol{E}(\boldsymbol{r},\omega)=\frac{4\pi\omega^2}{c^2}\boldsymbol{P}^{(\mathrm{NL})}(\boldsymbol{r},\omega) \tag{2.36}$$

in which $\varepsilon(\omega)$ is the dielectric constant of the material. Physically, the induced nonlinear polarization mixes the frequencies and the wavevectors of the incident electric fields, and adds the resulting signal field to the total electric field propagating through the sample. The nonlinear polarization is written as

$$\boldsymbol{P}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)=\sum_{n\geq 2}\boldsymbol{P}_0^{(n)}\left(\omega\right)e^{i(\boldsymbol{k}_p\cdot\boldsymbol{r}-\omega_p t)} \tag{2.37}$$

$$\begin{aligned}\boldsymbol{k}_p&=\sum_{i=1}^{n}z_i\boldsymbol{k}_i\\ \omega_p&=\sum_{i=1}^{n}z_i\omega_i\\ z_i&=\pm 1\end{aligned} \tag{2.38}$$

Eqn. (2.38) shows the wavevector matching condition.

As a further step, separating the electric field into the longitudinal component and the traverse component gives

$$\nabla^2\boldsymbol{E}_{\perp}\left(\boldsymbol{r},\omega\right)+k^2\left(\omega\right)\boldsymbol{E}_{\perp}\left(\boldsymbol{r},\omega\right)=-4\pi k_0^2\left(\omega\right)\boldsymbol{P}_{\perp}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right) \tag{2.39}$$

$$\boldsymbol{E}_{\parallel}\left(\boldsymbol{r},\omega\right)=\frac{4\pi}{\varepsilon\left(\omega\right)}\boldsymbol{P}_{\parallel}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right) \tag{2.40}$$

in which $k_0\left(\omega\right)=\omega^2/c^2$ and $k\left(\omega\right)=\omega^2\varepsilon\left(\omega\right)/c^2$. One can easily see the longitudinal component in Eqn. (2.40) does not propagate. We assume the electric field is a plane wave propagating along the *z* direction, which takes the form similar as in Eqn. (2.14) in vacuum

$$\boldsymbol{E}\left(\boldsymbol{r},\omega\right)=\boldsymbol{E}_0\left(\boldsymbol{r},\omega\right)e^{ikz} \tag{2.41}$$

$\boldsymbol{E}_0\left(\boldsymbol{r},\omega\right)$ is the field envelope that most likely vary slowly along the *z* direction. Simplifying Eqn. (2.39) gives

$$\nabla_{\perp}^{2}\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)+\left(\frac{\partial^{2}\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)}{\partial z^{2}}+i2k\left(\omega\right)\frac{\partial\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)}{\partial z}\right)=-4\pi k_{0}^{2}\left(\omega\right)\boldsymbol{P}_{\perp}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)e^{-ikz} \tag{2.42}$$

Additional approximations are needed to make some progress. The first approximation applied here is the plane wave approximation, in which the electric field does not change along the *x* and *y* directions. This indicates that the Laplacian in the *x*-*y* plane is zero, or

$$\nabla_{\perp}^{2}\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)=0 \tag{2.43}$$

The second approximation is the slow-varying envelope approximation,[8-9] which will be true when

$$\left|\frac{\partial^{2}\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)}{\partial z^{2}}\right|\ll\left|2k\left(\omega\right)\frac{\partial\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)}{\partial z}\right| \tag{2.44}$$

This approximation is a reasonable approximation when the spectrum of the signal is narrow-banded (narrow-band approximation).

$$\frac{\partial\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)}{\partial z}=i\frac{2\pi\omega}{cn\left(\omega\right)}\boldsymbol{P}_{\perp}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)e^{-ikz} \tag{2.45}$$

Usually, the nonlinear polarization can also be written as

$$\boldsymbol{P}_{\perp}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)=\boldsymbol{P}_{0}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)e^{ik_{P}z} \tag{2.46}$$

in which the envelope does not depend on *z* and k_P depends on the electric fields. Assuming the depletion of electric field is negligible (neglecting pump depletion), Eqn. (2.45) can be simply integrated as[4]

$$\begin{aligned}\boldsymbol{E}_{0}\left(\boldsymbol{r},\omega\right)&=\boldsymbol{E}_{0}\left(x,y,z=0,\omega\right)+i\frac{4\pi\omega}{cn\left(\omega\right)}\boldsymbol{P}_{0}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)e^{i\Delta kz/2}\frac{\sin\left(\Delta kz/2\right)}{\Delta k}\\&=\boldsymbol{E}_{0}\left(x,y,z=0,\omega\right)+i\frac{2\pi\omega z}{cn\left(\omega\right)}\boldsymbol{P}_{0}^{(\mathrm{NL})}\left(\boldsymbol{r},\omega\right)e^{i\Delta kz/2}\mathrm{sinc}\left(\Delta kz/2\right)\end{aligned} \tag{2.47}$$

where Δk is defined as

$$\begin{aligned} \Delta k &= k_P - k \\ \left|k_P\right| &= \omega_P \frac{n\left(\omega_P\right)}{c} \end{aligned} \tag{2.48}$$

When Δk is large, the field created by the nonlinear polarization is negligible so that only the original field component survives. When Eqn. (2.48) is equal to 0 (phase-matching condition), the electric field that traverses through the sample along the z direction, with the travel length l is given by

$$\begin{aligned} \boldsymbol{E}\left(x, y, z=l, \omega\right) &= \boldsymbol{E}_0\left(x, y, z=l, \omega\right) e^{ik_p l} + i\frac{2\pi\omega l}{cn\left(\omega\right)} \boldsymbol{P}^{(\mathrm{NL})}\left(x, y, z=l, \omega\right) \\ &= \boldsymbol{E}_0\left(x, y, z=l, \omega\right) e^{ik_p l} + \boldsymbol{E}_{\mathrm{sig}}\left(x, y, z=l, \omega\right) \end{aligned} \tag{2.49}$$

Eqn. (2.49) indicates that the field passing through the sample consists of the original electric field that is not depleted due to neglecting pump depletion, and the signal field radiated by the sample.

2.3 Formalism of Nonlinear Response Function

One of the central questions in spectroscopy is how we can relate the observed spectroscopic signals to molecular properties. Absorption and emission are intrinsically irreversible and non-equilibrium processes, but spectroscopy indeed informs us about the equilibrium observables such as electronic structure, molecular structure and conformation, *etc*. Also, many simulations are performed at equilibrium to predict non-equilibrium experimental observables. What is the connection between non-equilibrium experiments and equilibrium observables?

In 1931, Onsager stated the reciprocal relation that "*the average regression of fluctuations will obey the same law as the corresponding microscopic irreversible processes*".[10] It was later appreciated and proved by Callen and Welton in 1951, recognized as the well-known fluctuation-

dissipation theorem (FDT).[11] A general formalism of the response theory was later derived by Ryogo Kubo in 1957 to describe various phenomena including electronic transport problems.[12] This framework, however, was much later used as a common theoretical framework all sort of condensed-phase nonlinear ultrafast spectroscopy such as transient grating, photon-echo spectroscopy, coherent anti-Stokes Raman spectroscopy, *etc*.[4, 13-15] In the modern language, the quantitative description of Onsager regression hypothesis is stated as follows.

$$\frac{\Delta\bar{A}(t)}{\Delta\bar{A}(0)} = \frac{\langle \delta A(0)\delta A(t)\rangle}{\langle \delta A^2\rangle} \tag{2.50}$$

in which $\Delta A(t)$ is the deviation of an observable A from the equilibrium at time t

$$\Delta\bar{A}(t) = \bar{A}(t) - \langle A\rangle = \delta\bar{A}(t) \tag{2.51}$$

δA(t) is the spontaneous fluctuation around the equilibrium at time t given as

$$\delta A(t) = A(t) - \langle A\rangle \tag{2.52}$$

As one can see, it connects the non-equilibrium relaxation to the equilibrium spontaneous fluctuations, which is indeed a special case of the linear response, and this section summarizes the general formalism of nonlinear response function and how this framework can be connected from molecular correlation function to spectroscopic response we observe.

2.3.1 General Formalism

Suppose an equilibrium system is interacting with an external perturbation such as electromagnetic field described below

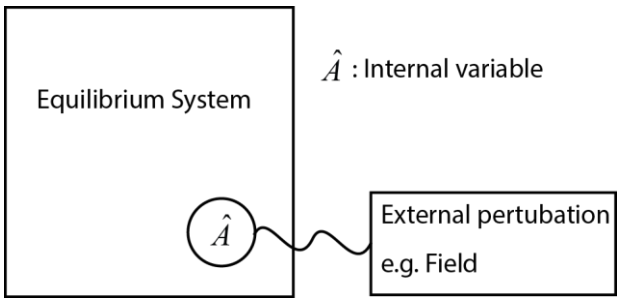

Figure 2.1: The setup of an equilibrium system weakly coupled to an external perturbation.

The Hamiltonian of the system and this external perturbation is

$$\hat{H}(t) = \hat{H}_0 + V(t) = \hat{H}_0 - f(t) \cdot \hat{A} \tag{2.53}$$

where $f(t)$ is the time-dependent action of the external perturbation, and since the action is weak enough, the deviation from the equilibrium is linear in the internal variable $\hat{A}$. Here, $f(t)$ is a real valued function that is zero for $t \leq 0$ to ensure causality, and that the system is allowed to be prepared at the equilibrium before the application of such time-dependent action. Please also note that here the setup is already implying a semi-classical description, in which the system is treated quantum mechanically whereas the time-dependent action is treated classically as we did previously. For our application to molecular spectroscopy, this semi-classical approach is adequate enough for capturing relevant spectroscopic responses.

Given the system Hamiltonian, the density operator that describes the initially equilibrated system is

$$\hat{\rho}_0 = \frac{e^{-\beta \hat{H}_0}}{\mathrm{Tr}\left\{ e^{-\beta \hat{H}_0} \right\}} \tag{2.54}$$

which is diagonal in the basis of $\hat{H}_0$ eigenstates and commutes to $\hat{H}_0$. In the interaction picture, the density operator evolves under the quantum Liouville equation

$$\frac{\partial \hat{\rho}_I(t)}{\partial t} = \frac{-i}{\hbar}\left[\hat{V}_I(t), \hat{\rho}_I(t)\right] \tag{2.55}$$

where $\hat{\rho}_I(t)$ and $\hat{V}_I(t)$ are the operators in the interaction picture, which take the form

$$\hat{\rho}_I(t) = \hat{u}_0^{\dagger}(t)\hat{\rho}(t)\hat{u}_0(t) = e^{\frac{i}{\hbar}\hat{H}_0 t}\hat{\rho}(t)e^{-\frac{i}{\hbar}\hat{H}_0 t} \tag{2.56}$$

and

$$\hat{V}_I(t) = \hat{u}_0^{\dagger}(t)\hat{V}(t)\hat{u}_0(t) \tag{2.57}$$

$\hat{u}_0(t)$ is the unperturbed time-evolution operator. One can formally apply the integration on Eqn. (2.55) to get an integral equation

$$\hat{\rho}_I(t) = \hat{\rho}_0 - \frac{i}{\hbar}\int_0^t dt_1\left[\hat{V}_I(t_1), \hat{\rho}_I(t_1)\right] \tag{2.58}$$

Performing the perturbative expansion similar as the Dyson expansion gives

$$\hat{\rho}_I(t) = \sum_{n=0}^{\infty}\left(\frac{-i}{\hbar}\right)^n \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \left[\hat{V}_I(t_1), \left[\hat{V}_I(t_2), \cdots \left[\hat{V}_I(t_n), \hat{\rho}_0\right]\right]\right] \tag{2.59}$$

Eqn. (2.59) is an exact expression, but in practice one cannot evaluate the infinite number of terms, and the time-ordering in Eqn. (2.59), meaning that $0 \le t_n \le t_{n-1} \le \cdots \le t_2 \le t_1 \le t$.

Now, we aim to calculate an observable *B*(*t*) described by the corresponding quantum operator $\hat{B}$, which does not depend on time in the Schrödinger picture. Evaluating the expectation value of $\hat{B}$ gives

$$\begin{aligned}\left\langle \hat{B}(t)\right\rangle &= \mathrm{Tr}\left\{\hat{\rho}_I(t)\hat{B}_I(t)\right\} \\ &= \sum_{n=0}^{\infty}\left(\frac{-i}{\hbar}\right)^n \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \mathrm{Tr}\left\{\hat{B}_I(t)\left[\hat{V}_I(t_1), \left[\hat{V}_I(t_2), \cdots \left[\hat{V}_I(t_n), \hat{\rho}_0\right]\right]\right]\right\}\end{aligned} \tag{2.60}$$

From Eqn. (2.53), we can obtain

$$\begin{aligned}\left\langle \hat{B}(t)\right\rangle = \sum_{n=0}^{\infty}\left(\frac{-i}{\hbar}\right)^{n}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t_1} dt_2 \cdots \int_{-\infty}^{t_{n-1}} dt_n \, f(t_1) f(t_2)\cdots f(t_n) \\ \times \mathrm{Tr}\left\{\hat{B}_I(t)\left[\hat{A}_I(t_1),\left[\hat{A}_I(t_2),\cdots\left[\hat{A}_I(t_n),\hat{\rho}_0\right]\right]\right]\right\}\end{aligned} \tag{2.61}$$

where the extension of the lower limit of the integration from 0 to -∞ originates from $f(t) = 0 \; \forall \; t \leq 0$. Instead of using the absolute time, using the time interval between adjacent time variables provided below:

$$\begin{aligned}\tau_1 &= t_{n-1} - t_n \\ &\vdots \\ \tau_{n-1} &= t_1 - t_2 \\ \tau_n &= t - t_1\end{aligned} \tag{2.62}$$

one can get

$$\begin{aligned}\left\langle \hat{B}(t)\right\rangle = \sum_{n=0}^{\infty}\left(\frac{-i}{\hbar}\right)^{n}\int_{0}^{\infty} d\tau_n \int_{0}^{\infty} d\tau_{n-1}\cdots\int_{0}^{\infty} d\tau_1 f(t-\tau_n) f(t-\tau_{n-1}-\tau_n)\cdots f(t-\tau_1-\cdots-\tau_n) \\ \times \mathrm{Tr}\left\{\hat{B}_I(t)\left[\hat{A}_I(t-\tau_n),\left[\hat{A}_I(t-\tau_{n-1}-\tau_n),\cdots\left[\hat{A}_I(t-\tau_1-\cdots-\tau_n),\hat{\rho}_0\right]\right]\right]\right\}\end{aligned} \tag{2.63}$$

Knowing that the trace is invariant under cyclic permutation, and that the unperturbed time-evolution operator commutes with the equilibrium density operator, one can eventually obtain

$$\begin{aligned}\left\langle \hat{B}(t)\right\rangle = \sum_{n=0}^{\infty}\left(\frac{-i}{\hbar}\right)^{n}\int_{0}^{\infty} d\tau_n \int_{0}^{\infty} d\tau_{n-1}\cdots\int_{0}^{\infty} d\tau_1 f(t-\tau_n) f(t-\tau_{n-1}-\tau_n)\cdots f(t-\tau_1-\cdots-\tau_n) \\ \times \mathrm{Tr}\left\{\hat{B}_I(\tau_1+\tau_2+\cdots+\tau_n)\left[\hat{A}_I(\tau_1+\tau_2+\cdots+\tau_{n-1}),\cdots\left[\hat{A}_I(0),\hat{\rho}_0\right]\right]\right\}\end{aligned} \tag{2.64}$$

In Eqn. (2.64), the resulting trace is independent of the external perturbation $f(t)$, meaning that the trace contains the intrinsic molecular response of the system. This property allows us to understand the molecular properties from a given form of the external perturbation, and to predict the average

time-dependent observable under arbitrary form of the perturbation once this intrinsic "response function" is known. By defining a response function

$$R^{(n)}\left(\tau_1,\tau_2,\cdots\tau_n\right)=\Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\cdots\Theta\left(\tau_n\right)\left(\frac{-i}{\hbar}\right)^n \times\mathrm{Tr}\left\{\hat{B}_I\left(\tau_1+\tau_2+\cdots+\tau_n\right)\left[\hat{A}_I\left(\tau_1+\tau_2+\cdots+\tau_{n-1}\right),\cdots\left[\hat{A}_I\left(0\right),\hat{\rho}_0\right]\right]\right\} \tag{2.65}$$

$\Theta(\tau)$ is the step function that ensures causality when the integration extended to -∞, and this constraint leads to the Kramers-Kronig relation that connects the real part and the imaginary part of the linear susceptibility in frequency domain.[16] However, application of Kramers-Kronig relation to nonlinear response function is not trivial.[17] Note that this response function is a real function since $i\left[\hat{A},\hat{B}\right]$ is Hermitian. By performing repeated cyclic permutation on Eqn. (2.65), one can arrive the form

$$\begin{aligned}R^{(n)}\left(\tau_1,\tau_2,\cdots\tau_n\right)&=\Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\cdots\Theta\left(\tau_n\right)\left(\frac{-i}{\hbar}\right)^n\\&\times\mathrm{Tr}\left\{\left[\left[\hat{B}_I\left(\tau_1+\tau_2+\cdots+\tau_n\right),\hat{A}_I\left(\tau_1+\tau_2+\cdots+\tau_{n-1}\right)\right]\cdots,\hat{A}_I\left(0\right)\right]\hat{\rho}_0\right\}\\&=\Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\cdots\Theta\left(\tau_n\right)\left(\frac{-i}{\hbar}\right)^n\\&\times\left\langle\left[\left[\hat{B}_I\left(\tau_1+\tau_2+\cdots+\tau_n\right),\hat{A}_I\left(\tau_1+\tau_2+\cdots+\tau_{n-1}\right)\right]\cdots,\hat{A}_I\left(0\right)\right]\right\rangle_0\end{aligned} \tag{2.66}$$

in which the ensemble average in the interaction picture over the initial equilibrium is performed. Finally, one can rewrite Eqn. (2.63) as

$$\left\langle\hat{B}(t)\right\rangle=\sum_{n=0}^{\infty}\int_0^{\infty}d\tau_n\int_0^{\infty}d\tau_{n-1}\cdots\int_0^{\infty}d\tau_1 f\left(t-\tau_n\right)\cdots f\left(t-\tau_1-\cdots-\tau_n\right)R^{(n)}\left(\tau_1,\tau_2,\cdots,\tau_n\right) \tag{2.67}$$

Eqn. (2.67) is in essence the formula representing the theory of nonlinear response, which is a general framework to describe any weakly perturbative system. The key physical interpretation

here is that the non-equilibrium response of *B* to the external perturbation *V* is fully characterized by an equilibrium time correlation function, which is the famous fluctuation-dissipation theorem. One physical insight of the FDT is that the spontaneous fluctuation and the non-equilibrium relaxation comes from the same physical origin. For example, friction of a particle experience in a solution can be determined by the diffusion coefficient, which originates from the molecular interactions between the particle and the surrounding solvent. Another example related to spectroscopy is the (non-equilibrium) pure dephasing can be characterized by (equilibrium) de-correlation of the fluctuating frequency.[4] In essence, both phenomena are rooted in the coupling between the chromophore and the surrounding bath.

By convention, each order that contributes to $\left\langle \hat{B}(t) \right\rangle$ is referred to as the n^{th}-order response. For instance, the first-order response is written as

$$\left\langle \hat{B}(t) \right\rangle = \int_0^\infty d\tau_1 f(t-\tau_1) R^{(1)}(\tau_1) \tag{2.68}$$

which can be regarded as the convolution between time-dependent external perturbation and the first-order response function. The third-order response is

$$\left\langle \hat{B}(t) \right\rangle = \int_0^\infty d\tau_3 \int_0^\infty d\tau_2 \int_0^\infty d\tau_1 f(t-\tau_3) f(t-\tau_2-\tau_3) f(t-\tau_1-\tau_2-\tau_3) R^{(3)}(\tau_1,\tau_2,\tau_3) \tag{2.69}$$

The Eqns. (2.68) and (2.69) form the basis for describing the linear and nonlinear spectroscopy throughout this thesis.

2.3.2 Application of Response Theory to Infrared Spectroscopy

For the application to IR spectroscopy, the external perturbation is the interaction between the incident electromagnetic field with the dielectric sample. The corresponding perturbation is to first order described by the electric-dipole Hamiltonian in Eqn. (2.26)

$$\hat{V}(t) = -\boldsymbol{E}(t)\cdot\hat{\boldsymbol{\mu}} = -\sum_{\alpha=1}^{3} E_\alpha(t)\hat{\mu}_\alpha \tag{2.70}$$

In Eqn. (2.70), $\hat{\boldsymbol{\mu}}$ is the electric dipole operator of the system that corresponds to the internal variable in Fig. 2.1, and α is the index running through the spatial coordinates of the system. The physical observable is the system polarization vector

$$P(\boldsymbol{r},t) = \left\langle \hat{\boldsymbol{\mu}}(t) \right\rangle \tag{2.71}$$

in which $\boldsymbol{r}$ is the position of the electric field evaluated within the sample. Applying the nonlinear response function in Eqn. (2.67), the resulting component of the system polarization vector can be written as

$$\begin{aligned} P_{\alpha_{n+1}}(\boldsymbol{r},t) = \sum_{n=1}^{\infty}\sum_{\alpha_n=1}^{3}\cdots\sum_{\alpha_1=1}^{3}\int_0^\infty d\tau_n\int_0^\infty d\tau_{n-1}\int_0^\infty d\tau_1 E_{\alpha_n}(\boldsymbol{r},t-\tau_n)\cdots E_{\alpha_1}(\boldsymbol{r},t-\tau_1-\cdots-\tau_n) \\ \times R^{(n)}_{\alpha_1\cdots\alpha_n}(\tau_1,\tau_2,\cdots,\tau_n) \end{aligned} \tag{2.72}$$

with the n^{th}-order response tensor in the interaction picture defined as

$$\begin{aligned} R^{(n)}_{\alpha_1\cdots\alpha_{n+1}}(\tau_1,\tau_2,\cdots\tau_n) = \Theta(\tau_1)\Theta(\tau_2)\cdots\Theta(\tau_n)\left(\frac{-i}{\hbar}\right)^n \\ \times\left\langle\left[\left[\hat{\mu}_{\alpha_{n+1}}(\tau_1+\tau_2+\cdots+\tau_n),\hat{\mu}_{\alpha_n}(\tau_1+\tau_2+\cdots+\tau_{n-1})\right]\cdots,\hat{\mu}_{\alpha_1}(0)\right]\right\rangle_0 \end{aligned} \tag{2.73}$$

One important note is that the n^{th}-order response tensor corresponds to the n+1-point dipole time correlation function evaluated under the initial equilibrium ensemble. This expression is quite

general for IR spectroscopy and also applicable to other frequency ranges and perturbation strengths. The corresponding n^{th}-order polarization can be simply taken from Eqn. (2.72).

$$P^{(n)}_{\alpha_{n+1}}(\boldsymbol{r},t)=\sum_{\alpha_n=1}^{3}\cdots\sum_{\alpha_1=1}^{3}\int_0^\infty d\tau_n\int_0^\infty d\tau_{n-1}\int_0^\infty d\tau_1 E_{\alpha_n}(\boldsymbol{r},t-\tau_n)\cdots E_{\alpha_1}(\boldsymbol{r},t-\tau_1-\cdots-\tau_n) \times R^{(n)}_{\alpha_1\cdots\alpha_{n+1}}(\tau_1,\tau_2,\cdots,\tau_n) \tag{2.74}$$

Similarly, one can write down explicitly the first-order polarization and the third-order polarization as

$$P^{(1)}_{\alpha_2}(\boldsymbol{r},t)=\sum_{\alpha_1=1}^{3}\int_0^\infty d\tau_1 E_{\alpha_1}(\boldsymbol{r},t-\tau_1)R^{(1)}_{\alpha_1\alpha_2}(\tau_1) \tag{2.75}$$

$$P^{(3)}_{\alpha_4}(\boldsymbol{r},t)=\sum_{\alpha_1,\alpha_2,\alpha_3=1}^{3}\int_0^\infty d\tau_3\int_0^\infty d\tau_2\int_0^\infty d\tau_1 E_{\alpha_3}(\boldsymbol{r},t-\tau_3)E_{\alpha_2}(\boldsymbol{r},t-\tau_1-\tau_2)E_{\alpha_1}(\boldsymbol{r},t-\tau_1-\tau_2-\tau_3) \times R^{(3)}_{\alpha_1\alpha_2\alpha_3\alpha_4}(\tau_1,\tau_2,\tau_3) \tag{2.76}$$

The more detailed connection to the linear and nonlinear IR spectroscopy will be discussed later.

2.4 Connection to Experimental Spectroscopy

The general form of the response function in Eqn. (2.73) consists of the nested commutators and all possible combination of vector components, meaning that the computation of the n^{th}-order response function is cumbersome since it seemingly requires considering 3^{n+1} elements of the response tensor, and addition expansion of the commutators. However, this challenging situation can be significantly simplified when taking physical constraints of the experimental setup into account, including the symmetry properties of the sample in isotropic solution, wavevector matching, resonance condition, *etc*. This section summarizes these physical

constraints and provides a connection to the experimental IR spectroscopy and spectroscopic simulations.

2.4.1 Symmetry in Isotropic Solution

The response tensor itself is one of the intrinsic properties of the material, meaning that the symmetry possessed by the system also applies to the response tensor. Symmetry allows us to identify non-zero elements and therefore significantly simplify the computation. Typically, solution spectroscopy is almost always performed with an isotropic solution, meaning that the physical properties have identical values with every possible direction. Also, the inversion symmetry holds for isotropic solution, meaning the value at (*x*,*y*,*z*) is equal to the value at (-*x*,-*y*,-*z*). Additionally, the reflection through a plane in the isotropic gives identical values for any physical observable.

For instance, the first-order response has 9 elements, including $R_{XX}^{(1)}$, $R_{XY}^{(1)}$, $R_{XZ}^{(1)}$, $R_{YX}^{(1)}$, $R_{YY}^{(1)}$, $R_{YZ}^{(1)}$, $R_{ZX}^{(1)}$, $R_{ZY}^{(1)}$, and $R_{ZZ}^{(1)}$ in the laboratory frame. Taking a reflection place on the *xy* plane will reflect the *z* component so that $z \rightarrow -z$. At the same time, the observable has the inversion symmetry such that

$$R_{XY}^{(1)}(t) = -R_{XY}^{(1)}(t) = 0 \tag{2.77}$$

Similarly, one can find that all of the non-diagonal elements mush vanish, and that all of the diagonal elements are equal to each other thanks to the isotropic property.

$$R^{(1)}(t) = R_{XX}^{(1)}(t) = R_{YY}^{(1)}(t) = R_{ZZ}^{(1)}(t) \tag{2.78}$$

Therefore, only one element survives and has non-zero element in isotropic solution, which greatly simplifies the problem. Also, the symmetry implies that the wavevector of the signal is the same

as the wavevector incident field, which is automatically heterodyned. On the flip side, linear spectroscopy in isotopic solution cannot investigate the orientational properties of a molecule.

For even-order response, all of the elements contain odd number of vector components (n+1). Inversion operations always leads to at least a single unpaired vector component, leading to vanishing all of the possible tensorial elements. In other words, even-order responses are always zero in isotropic solution. Only odd-order responses have non-vanishing values in isotropic solution. The next simplest odd-order response, third-order response, has only four non-vanishing elements

$$R^{(3)}_{\alpha\alpha\alpha\alpha}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\beta\beta\beta\beta}\left(\tau_1,\tau_2\tau_3\right) \tag{2.79}$$

$$R^{(3)}_{\alpha\alpha\beta\beta}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\beta\beta\alpha\alpha}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\alpha\alpha\gamma\gamma}\left(\tau_1,\tau_2\tau_3\right) \tag{2.80}$$

$$R^{(3)}_{\alpha\beta\beta\alpha}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\beta\alpha\alpha\beta}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\alpha\gamma\gamma\alpha}\left(\tau_1,\tau_2\tau_3\right) \tag{2.81}$$

$$R^{(3)}_{\alpha\beta\alpha\beta}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\beta\alpha\beta\alpha}\left(\tau_1,\tau_2\tau_3\right)=R^{(3)}_{\alpha\gamma\alpha\gamma}\left(\tau_1,\tau_2\tau_3\right) \tag{2.82}$$

in which α, β, γ are distinct indices not equal to each other.

2.4.2 Linear Spectroscopy

To get additional physical insight of the response function to incident electric field, we firstly evaluate the linear response in detail

$$\begin{aligned} R^{(1)}\left(\tau_1\right) &= \Theta\left(\tau_1\right)\left(\frac{-i}{\hbar}\right)\left\langle\left[\hat{\mu}_I\left(\tau_1\right),\hat{\mu}_I\left(0\right)\right]\right\rangle_0 \\ &= \Theta\left(\tau_1\right)\left(\frac{2}{\hbar}\right)\mathrm{Im}\left\langle\hat{\mu}_I\left(\tau_1\right)\hat{\mu}_I\left(0\right)\right\rangle_0 \end{aligned} \tag{2.83}$$

As the first attempt, we can expand the dipole time correlation function in the energy eigenbasis as in Eqn. (2.21) assuming the system is a closed, isolated system with known eigenstates,

$$\begin{aligned}\hat{H}_0|a\rangle &= E_a|a\rangle \\ \hat{H}_0|b\rangle &= E_b|b\rangle\end{aligned} \tag{2.84}$$

And the initial state is prepared at the thermal equilibrium so that the probability of finding the state *a*, P_a, follows the Boltzmann distribution

$$\begin{aligned}P_a &= \langle a|\hat{\rho}_0|a\rangle = e^{-\beta E_a} \\ \beta &= 1/k_{\mathrm{B}}T\end{aligned} \tag{2.85}$$

Typically, in our experiments, the energy gap between the lowest excited state and the ground state is much larger than the thermal energy $k_{\mathrm{B}}T$ at room temperature. For instance, the amide I vibration has the frequency around 1650 cm^{-1}. In contrast, thermal energy at 300 K is only 208.5 cm^{-1}. The corresponding Boltzmann factor

$$e^{-\beta\hbar\omega} \approx e^{-1650/208.5} = 3.66\times10^{-4} \tag{2.86}$$

meaning that the contribution to the response function from initially populated excited states is negligible. Then the dipole time correlation function evaluated in this eigenbasis can be written as

$$\begin{aligned}\left\langle \hat{\mu}_I(\tau_1)\hat{\mu}_I(0)\right\rangle_0 &= \sum_a P_a \langle a|\hat{\mu}_I(\tau_1)\hat{\mu}_I(0)|a\rangle \\ &= \sum_{a,b} P_a \langle a|\hat{\mu}_I(\tau_1)|b\rangle\langle b|\hat{\mu}_I(0)|a\rangle \\ &= \sum_{a,b} P_a |\mu_{ab}|^2 e^{-i\omega_{ba}\tau_1}\end{aligned} \tag{2.87}$$

$$\begin{aligned}R^{(1)}(\tau_1) &= \Theta(\tau_1)\left(\frac{2}{\hbar}\right)\sum_{a,b} P_a |\mu_{ab}|^2 \sin(\omega_{ab}\tau_1) \\ &= \Theta(\tau_1)\left(\frac{1}{\hbar}\right)\sum_{ab} P_a |\mu_{ab}|^2 \int d\omega e^{i\omega t}\left[\delta(\omega-\omega_{ab})-\delta(\omega+\omega_{ba})\right]\end{aligned} \tag{2.88}$$

Eqn. (2.87) gives quite a bit physical insight of how the sample responses to the incident electric field. In order to have the transition between the state *k* and the state *l*, the frequency of the electric field has to be resonant with the energy gap between these two states including the absorption,

$\delta(\omega-\omega_{ab})$, and stimulated emission, $\delta(\omega+\omega_{ba})$, which guarantees the approximation made in the Eqns. (2.25)–(2.26). Also, it implies the conservation of energy between the incident photon energy and the transition energy, albeit here the field is treated classically. The magnitude of the response at ω_{ab} is proportional to the squared transition dipole moment, or also called oscillator strength. The corresponding response function has zero magnitude at the zero time, and the phase of the oscillation is $\pi/2$ shifted compared to the incident electric field. Physically, the response cannot be instantaneous right when the field starts interacting with the sample, which is formally consistent with the causality. Also note that the linear response is in essence the Golden rule.

Although it seems trivial for linear response, here summarize the conditions for the response function:

1. Wavevector matching: The emitted signal is required to occur with the following condition

$$\boldsymbol{k}_{\mathrm{sig}} = \boldsymbol{k}_1 \tag{2.89}$$

$$\omega_{\mathrm{sig}} = \omega_1 \tag{2.90}$$

2. Phase matching: The wavevector of the signal is intimately related to the frequency of the signal via Eqn. (2.48), meaning that

$$\left|\boldsymbol{k}_{\mathrm{sig}}\right|^2 = \left(\omega_{\mathrm{sig}}\right)^2 \frac{n^2\left(\omega_{\mathrm{sig}}\right)}{c^2} \tag{2.91}$$

   In this situation, since $\boldsymbol{k}_1$ and ω_1 already satisfies its own phase matching condition, the signal also satisfies the condition.

3. Causality: No signal can be generated until the incident field arrives at the sample, which is guaranteed by the step function $\Theta(\tau_1)$.

4. Resonance condition: The response function has the characteristic oscillatory frequency ω_{kl}, which matches the frequency of the field in order to have such response. It is also referred to as the rotating wave approximation (RWA) with more detail in the third order response function.

For calculating the linear polarization, one can see from the time domain description of polarization in Eqn. (2.75) is indeed a convolution between the electric field and the response function, meaning that in the frequency domain

$$P^{(1)}(\omega) = \tilde{E}^{(1)}(\omega)\chi^{(1)}(\omega) \tag{2.92}$$

where $\chi^{(1)}(\omega)$ is the linear susceptibility, or the Fourier transform of the linear response function. Computationally, the form of the field in frequency domain can be determined based on the experimental setup. Therefore, the polarization can be determined exactly once the linear susceptibility is computed. However, due to the nature of causality, higher-order nonlinear polarization does not enjoy the simple convolution relation with the incident fields, and computation of nonlinear polarization may require time-consuming computation of the direct multiple time integrations, even though some calculation methods are proposed to deal with this finite pulse effect.[18-19] Also, efficient non-perturbative approach that propagates the system with the total Hamiltonian including the interaction Hamiltonian is present in the literature.[20-21]

Most of the time for computational spectroscopy, a rather draconic but useful approximation, impulsive limit, can be made for convenience of numerically computation. Impulsive limit means the incident pulse is essentially a delta function in time, or infinite band in frequency:

$$E(t) = E_0 \delta(t) \tag{2.93}$$

such that

$$\begin{aligned} P^{(1)}(t) &= \int_0^\infty d\tau_1 E(t-\tau_1) R^{(1)}(\tau_1) \\ &= E_0 R^{(1)}(t) \end{aligned} \tag{2.94}$$

Or trivially in frequency

$$P^{(1)}(\omega) = E_0 \chi^{(1)}(\omega) \tag{2.95}$$

Apparently, this limiting behavior will break down when the pulse duration in time is getting longer and even worse when pulse overlap happens between different pulses for non-linear spectroscopy such that the time ordering becomes ambiguous and different pathway contributions appear.[22-23] Throughout this thesis, however, all of the numerical computations on nonlinear spectroscopy imply the impulsive limit, meaning that response functions are computed to simulate a spectrum for comparison to experimental data without accounting for the finite pulse effect. Additional discussions can be found in Ref. 4.

As another step toward analyzing the more complicated third-order response function, here shows a double-sided Feynman diagram for the linear response in Fig. 2.2. The reading of the diagram is along the time axis from the bottom to the above. Initially, the system is prepared at the thermal equilibrium so that the population state of $|a\rangle\langle a|$ is prepared, with the associated probability P_a. At time zero, the first interaction between the pulse represented by the dipole operator $\hat{\mu}_I$ and the system induces the transition of the ket side to the state *b*, creating a coherence state $|b\rangle\langle a|$. The arrow pointing from the bottom left to the ladder corresponds to the absorption. During τ_1, this coherence state has an oscillatory phase $e^{-i\omega_{ba}\tau_1}$ with the frequency ω_{ba}, and can

create a signal emission, $\hat{\mu}_J$, represented as the wavy arrow, and eventually back to a population state, though in this case back to the same ground state. One can also see that the complex conjugate behaves as the same diagram but mirror imaged. Therefore, only the diagram on the left will be shown explicitly.

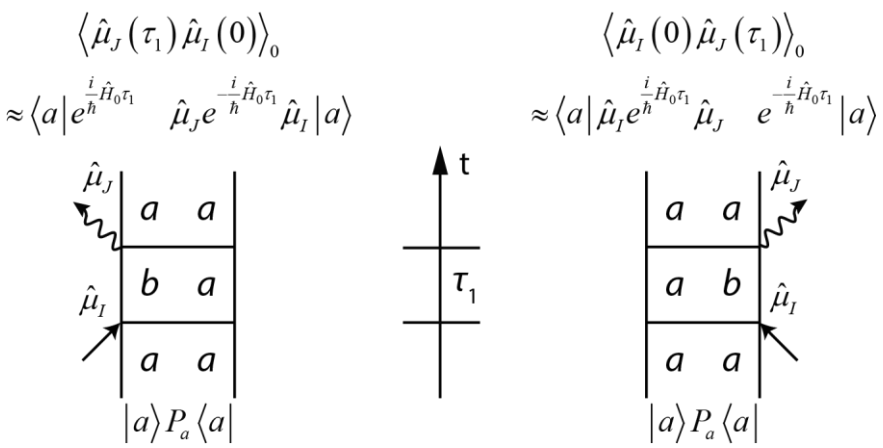

Figure 2.2: Diagrammatic representation for the linear response function.

The rules of drawing diagrams are listed below:[4]

1. Time runs from the bottom to the top. Each diagram represents the time evolution of the element of density operator. The left hand side corresponds to the ket side, and the right hand side corresponds to the bra side.
2. Interactions with the light are represented by the arrows as in Fig. 2.2. The signal emission represents the polarization $P^{(n)}_{\alpha_{n+1}}(t) = \langle \hat{\mu}_{\alpha_{n+1}}(t) \rangle$, which is represented by a wavy arrow or dashed arrow.

3. An arrow pointing to the right represents an electric field with $e^{i\mathbf{k}\cdot\mathbf{r}-i\omega t}$ and an arrow pointing to the left corresponds to an electric field with $e^{-i\mathbf{k}\cdot\mathbf{r}+i\omega t}$. The signal arrow is emitting to the left by convention.
4. An arrow pointing towards the ladder corresponds to absorption or excitation where as an arrow pointing away corresponds to stimulated emission or de-excitation, which is a manifestation of the rotating wave approximation (RWA). For instance, the arrow pointing in the diagram but the transition that goes from higher energy state to lower energy state cannot survive in the RWA.
5. The diagram has a sign of $(-1)^m$, where m is the number of interactions from the right. This sign originates from the ordering of the commutator.
6. After signal emission, the diagram must end in the population state.

Using this diagrammatic approach, it is possible to more easily study physical interpretation of the nonlinear response functions.

2.4.3 Conditions for Third-Order Response

We can extend similar analysis in the linear response function to the third order response function. The third order response function is given from Eqn. (2.73)

$$R^{(3)}_{IJKL}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\left(\frac{-i}{\hbar}\right)^3 \times\left\langle\left[\left[\left[\hat{\mu}_L(\tau_1+\tau_2+\tau_3),\hat{\mu}_K(\tau_1+\tau_2)\right],\hat{\mu}_J(\tau_1)\right],\hat{\mu}_I(0)\right]\right\rangle_0 \tag{2.96}$$

Before jumping into the eigenstate description, we first look at the behavior of the third-order polarization given the electric fields

$$\boldsymbol{E}^{(n)}(t)=\boldsymbol{E}_0^{(n)}(t)e^{iz_n\boldsymbol{k}_n\cdot\boldsymbol{r}-i\omega_n t} \tag{2.97}$$

where $z_n = \pm 1$, and $\boldsymbol{E}_0^{(n)}$ is the envelope. Then the third-order polarization will behave like

$$\begin{aligned}P_L^{(3)}(\boldsymbol{r},t)\sim & \sum_{z_1,z_2,z_3}^{z_1+z_2+z_3>0}\sum_{I,J,K=1}^{3} e^{i(z_1\boldsymbol{k}_1+z_2\boldsymbol{k}_2+z_3\boldsymbol{k}_3)\cdot\boldsymbol{r}}e^{-i(z_1\omega_1+z_2\omega_2+z_3\omega_3)t}\\ &\times\int_0^\infty d\tau_3 e^{i(z_1\omega_1+z_2\omega_2+z_3\omega_3)\tau_3}E_{0,K}^{(3)}(t-\tau_3)\\ &\times\int_0^\infty d\tau_2 e^{i(z_1\omega_1+z_2\omega_2)\tau_2}E_{0,J}^{(2)}(t-\tau_3-\tau_2)\\ &\times\int_0^\infty d\tau_2 e^{i(z_1\omega_1)\tau_1}E_{0,I}^{(1)}(t-\tau_3-\tau_2-\tau_1)R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3)\end{aligned} \tag{2.98}$$

One can see that the polarization wavevector and the frequency will satisfy

$$\begin{aligned}\boldsymbol{k}_{\text{sig}} &= z_1\boldsymbol{k}_1+z_2\boldsymbol{k}_2+z_3\boldsymbol{k}_3\\ \omega_{\text{sig}} &= z_1\omega_1+z_2\omega_2+z_3\omega_3\end{aligned}\quad, z_n=\pm 1 \tag{2.99}$$

with the constraint of $z_1+z_2+z_3>0$ to make sure the signal field propagates forward in time, or in certain circumstances the wavevector propagates the same direction as in one of the input wavevectors. Also, in order to make sure the time integrals not vanishing, the third order response function will approximately take the form

$$R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3)\sim e^{-i(z_1\omega_1)\tau_1}e^{-i(z_1\omega_1+z_2\omega_2)\tau_2}e^{i(z_1\omega_1+z_2\omega_2+z_3\omega_3)\tau_3} \tag{2.100}$$

which is called rotating wave approximation, meaning that we only investigate the (near) resonance frequency components. Otherwise, there will be oscillatory frequency of $2\omega_n$ such that the time integral will be negligible.

Now, looking into the details of the third-order response function given in Eqn. (2.96). Expanding the nested commutator results in 8 terms with 4 complex conjugate pairs, which can be written as the following

$$
\begin{aligned}
R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\{ & \\
&+\left\langle \hat{\mu}_J(\tau_1)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_I(0)\right\rangle_0 \\
&+\left\langle \hat{\mu}_I(0)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_J(\tau_1)\right\rangle_0 \\
&+\left\langle \hat{\mu}_I(0)\hat{\mu}_J(\tau_1)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_K(\tau_1+\tau_2)\right\rangle_0 \\
&+\left\langle \hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_J(\tau_1)\hat{\mu}_I(0)\right\rangle_0 \}
\end{aligned}
\tag{2.101}
$$

For the ease of further discussion, the third-ordered response function can be separated into 4 distinct Liouville pathways below:

$$
\begin{aligned}
R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3) &= \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\sum_{N=I}^{IV} R_{IJKL}^{(N)}(\tau_1,\tau_2,\tau_3) \\
R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3) &= \left\langle \hat{\mu}_J(\tau_1)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_I(0)\right\rangle_0 \\
R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3) &= \left\langle \hat{\mu}_I(0)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_J(\tau_1)\right\rangle_0 \\
R_{IJKL}^{(III)}(\tau_1,\tau_2,\tau_3) &= \left\langle \hat{\mu}_I(0)\hat{\mu}_J(\tau_1)\hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_K(\tau_1+\tau_2)\right\rangle_0 \\
R_{IJKL}^{(IV)}(\tau_1,\tau_2,\tau_3) &= \left\langle \hat{\mu}_L(\tau_1+\tau_2+\tau_3)\hat{\mu}_K(\tau_1+\tau_2)\hat{\mu}_J(\tau_1)\hat{\mu}_I(0)\right\rangle_0
\end{aligned}
\tag{2.102}
$$

Within these terms, wavevector matching condition leads to additional 4 distinct types of experiments:

1. Rephasing (R) and Photon Echo: $\boldsymbol{k}_{\mathrm{sig}} = -\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$
2. Nonrephasing (NR) and Transient Grating: $\boldsymbol{k}_{\mathrm{sig}} = +\boldsymbol{k}_1 - \boldsymbol{k}_2 + \boldsymbol{k}_3$
3. Double quantum coherence (DQC): $\boldsymbol{k}_{\mathrm{sig}} = +\boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}_3$
4. Third harmonic generation (THG): $\boldsymbol{k}_{\mathrm{sig}} = +\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$

For the rephasing and nonrephasing pathways, the phase matching condition (Eqn. (2.91)) is satisfied when the first two pulses are identical, which in practice is totally feasible with a pump pulse pair. Similarly, the double quantum coherence pathway satisfies the phase-matching condition when the last two pulses are identical. As a side note, in a typical 2D IR experiment that

uses the same source of laser for the pulse train, if the timing between the second pulse and the third pulse is ambiguous or swapped due to inaccurate timing and the finite pulse effect, then the double quantum coherence pathway can contribute to the response. Third harmonic generation is usually highly unfeasible using the same pulse, which requires

$$n(\omega) \approx n(3\omega) \tag{2.103}$$

given a common material that has frequency-dependent index of refraction. However, it is possible with a birefringent material whose refractive index depends on the polarization and incident angle of the pulse. Similar argument holds for second harmonic generation as well.

As an additional note, qualitatively under the rotating wave approximation and pulses with the same frequency ω, each of the pathway takes the following form

$$R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3) \sim e^{i\omega\tau_1}e^{-i\omega\tau_3} \tag{2.104}$$

$$R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3) \sim e^{-i\omega\tau_1}e^{-i\omega\tau_3} \tag{2.105}$$

$$R_{IJKL}^{(\mathrm{DQC})}(\tau_1,\tau_2,\tau_3) \sim e^{-i\omega\tau_1}e^{-i2\omega\tau_2}e^{i\omega\tau_3} \tag{2.106}$$

$$R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3) \sim e^{-i\omega\tau_1}e^{-i2\omega\tau_2}e^{i3\omega\tau_3} \tag{2.107}$$

One can see that with proper timing between τ_1 and τ_3 not equal to zero, the rephasing pathway can refocus the response (photon echo), whereas the non-resphasing pathway only possesses oscillatory behavior and potentially dephasing when accounting for interactions to the bath.

To summarize, the conditions of the third-order response are pretty much the same as seen in the linear response:

1. Wavevector matching: The emitted signal is required to occur with the following condition

$$\begin{aligned} \boldsymbol{k}_{\text{sig}} &= z_1\boldsymbol{k}_1 + z_2\boldsymbol{k}_2 + z_3\boldsymbol{k}_3 \\ \omega_{\text{sig}} &= z_1\omega_1 + z_2\omega_2 + z_3\omega_3 \end{aligned} \quad , z_n = \pm 1 \tag{2.108}$$

2. Phase matching:

$$\left|\boldsymbol{k}_{\text{sig}}\right|^2 = \left(\omega_{\text{sig}}\right)^2 \frac{n^2\left(\omega_{\text{sig}}\right)}{c^2} \tag{2.109}$$

The practical detail of the phase matching condition depends on the pathways of interest.

3. Causality: No signal can be generated until the incident field arrives at the sample, which is guaranteed by the step functions $\Theta(\tau_1)$, $\Theta(\tau_2)$, and $\Theta(\tau_3)$.

4. Resonance condition: To survive the time integration, the incident pulses need to be resonant with transition frequency of the material.

2.4.4 Eigenstate Description of the Third-Order Response

For 2D IR spectroscopy and transient absorption IR spectroscopy, rephasing and nonrephasing pathways contribute to the spectroscopic signal. For brevity, the following discussion will only focus on these two pathways in detail. In linear response of the thermally equilibrated sample, a single pulse gives rise to absorption events containing a conjugate pair of Liouville pathway. However, in the third-order response, the combination of pulses acting on the bra/ket side results in potentially 8 distinct terms, but the constraints of the initial state at thermal equilibrium and the energy gap much larger than the thermal energy physically reduces into a more-tractable number of terms with insightful physical meaning.

To start with, assuming the eigenstates of the closed, isolated system is known,

$$\hat{H}_0\left|a\right\rangle = E_a\left|a\right\rangle \tag{2.110}$$

and we use indices *a*, *b*, *c*, *d* to keep track of different eigenstates. Expanding Eqn. (2.102) gives

$$R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3) = \sum_{a,b,c,d} P_a \mu_I^{ba} \mu_J^{ad} \mu_K^{dc} \mu_L^{cb} e^{-i\omega_{ba}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \tag{2.111}$$

Similarly,

$$R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3) = \sum_{a,b,c,d} P_a \mu_I^{ad} \mu_J^{ba} \mu_K^{dc} \mu_L^{cb} e^{-i\omega_{ad}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \tag{2.112}$$

$$R_{IJKL}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3) = \sum_{a,b,c,d} P_a \mu_I^{ad} \mu_J^{dc} \mu_K^{ba} \mu_L^{cb} e^{-i\omega_{ad}\tau_1} e^{-i\omega_{ac}\tau_2} e^{-i\omega_{bc}\tau_3} \tag{2.113}$$

$$R_{IJKL}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3) = \sum_{a,b,c,d} P_a \mu_I^{ba} \mu_J^{cb} \mu_K^{dc} \mu_L^{ad} e^{-i\omega_{ba}\tau_1} e^{-i\omega_{ca}\tau_2} e^{-i\omega_{ba}\tau_3} \tag{2.114}$$

Now, we can apply the diagrammatic approach as in the linear response function shown in Fig. 2.3. The black diagrams correspond to the Eqns. (2.111)–(2.114). With the constraint of wavewector matching, $\Delta\omega \gg k_{\mathrm{B}}T$ as in Eqn. (2.86), resonance conditions, and the rules of the diagram, the corresponding rephasing and nonrephasing Liouville pathways turn out to only have three distinct pathways shown on the right of Fig. 2.3.

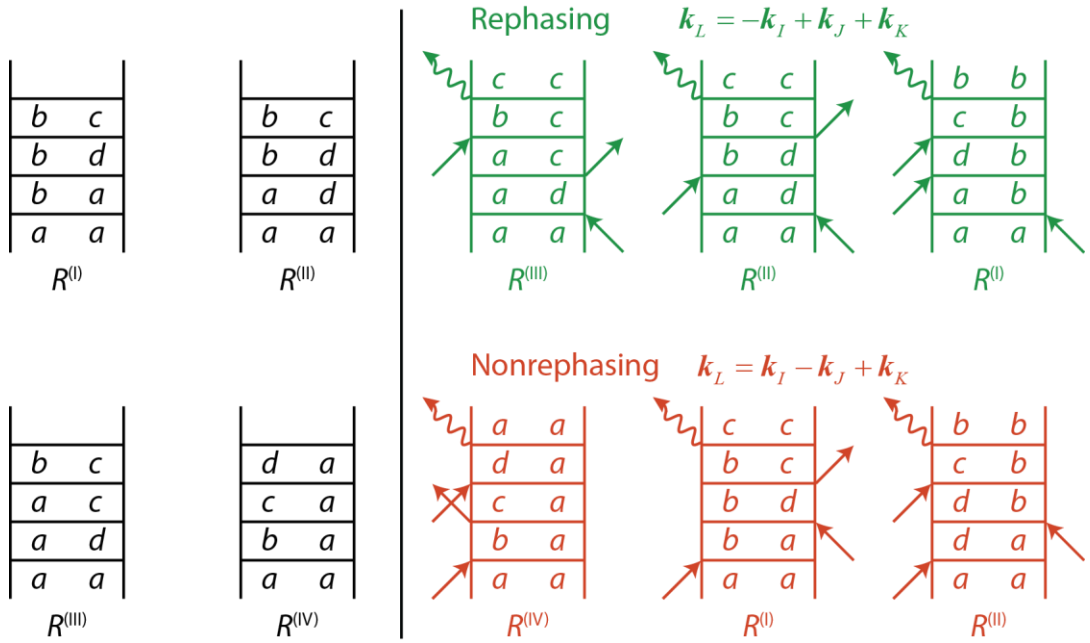

Figure 2.3: Diagrammatic representation for the third-order rephasing and nonrephasing response function.

In other words, the response function of the rephasing pathway and the non-rephasing pathway can be written as

$$\begin{aligned} R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\sum_{a,b,c,d}^{\mathrm{RWA}} P_a \{ \\ +\mu_I^{ad}\mu_J^{dc}\mu_K^{ba}\mu_L^{cb} e^{-i\omega_{ad}\tau_1} e^{-i\omega_{ac}\tau_2} e^{-i\omega_{bc}\tau_3} \\ +\mu_I^{ad}\mu_J^{ba}\mu_K^{dc}\mu_L^{cb} e^{-i\omega_{ad}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \\ -\mu_I^{ba}\mu_J^{ad}\mu_K^{dc}\mu_L^{cb} e^{-i\omega_{ba}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \} \end{aligned} \tag{2.115}$$

$$\begin{aligned} R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\sum_{a,b,c,d}^{\mathrm{RWA}} P_a \{ \\ +\mu_I^{ba}\mu_J^{cb}\mu_K^{dc}\mu_L^{ad} e^{-i\omega_{ba}\tau_1} e^{-i\omega_{ca}\tau_2} e^{-i\omega_{ba}\tau_3} \\ +\mu_I^{ba}\mu_J^{ad}\mu_K^{dc}\mu_L^{cb} e^{-i\omega_{ba}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \\ -\mu_I^{ad}\mu_J^{ba}\mu_K^{dc}\mu_L^{cb} e^{-i\omega_{ad}\tau_1} e^{-i\omega_{bd}\tau_2} e^{-i\omega_{bc}\tau_3} \} \end{aligned} \tag{2.116}$$

RWA indicates the enumeration is performed only for pathways satisfying the rotating wave approximation. Or more generally, the original response function in Eqn. (2.101) can be significantly simplified as

$$R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\{ R_{IJKL}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3)+R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)-R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)\} \tag{2.117}$$

$$R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\mathrm{Im}\{ R_{IJKL}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3)+R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)-R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)\} \tag{2.118}$$

Note that the Eqns. (2.117)–(2.118) have to satisfy all of the previous approximations and constraints, meaning wavevector matching, rotating wave approximation, *etc*.

To illustrate the distinct physical meaning of these allowed pathways, we used a simple anharmonic oscillator model that has only three states

$$\hat{H}_0=\hbar\sum_{a=0}^{2}\omega_a|a\rangle\langle a|,\quad \omega_a=\begin{cases}0, & a=0\\ \omega, & a=1\\ 2\omega-\Delta, & a=2\end{cases} \tag{2.119}$$

with the harmonic frequency ω, and anharmonicity Δ. The corresponding pathways are shown in Fig. 2.4. The first pathway, $R_{IJKL}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3)$ in rephasing and $R_{IJKL}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3)$ in nonrephasing shows the depletion of the ground state (ground state bleach, GSB) due to the excitation fields. As a result, it has less absorption of the third pulse, or more transmittance through the sample, with the net positive sign given $(-1)^2$ with two fields acting from the right (Rule 5 of the diagram). The second pathway, $R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)$ in rephasing and $R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)$ in nonrephasing corresponds to the stimulated emission (SE) from the first excited state, which also leads to more transmittance through the sample (net positively signed response). The last pathway shows the absorption from

the first excited state to the second excited state (excited state absorption, ESA), which has additional absorption or less transmittance through the sample. Note that it is negatively signed intrinsic in the response function. The assignment of these pathways are summarized in Table 2.1.

	GB	SE	ESA
Rephasing	$R_{IJKL}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3)$	$R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)$	$R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)$
Nonrephasing	$R_{IJKL}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3)$	$R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)$	$R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)$

Table 2.1: Assignments of the distinct pathways to the components of the third-order response functions

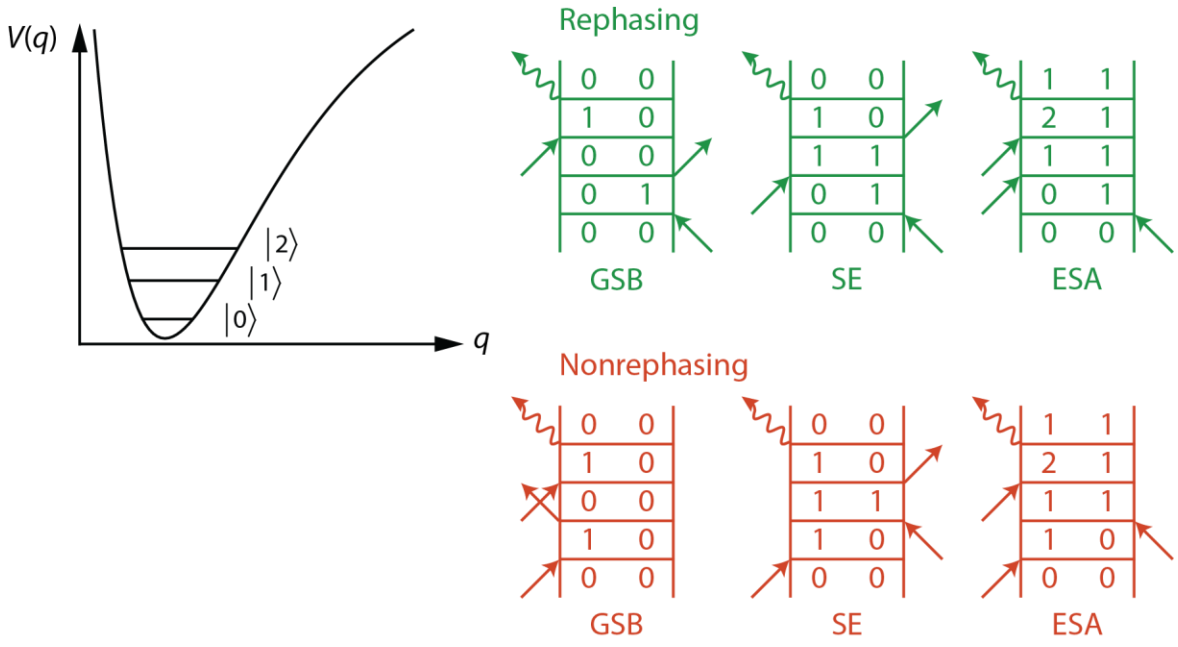

Figure 2.4: Diagrammatic representation for the third-order rephasing and nonrephasing response function of the three-level system.

2.4.5 Orientational Average

From the symmetry constraint in isotropic solution, we know that linear response has only one tensor element surviving, implying that it is insensitive to the underlying orientations of the system. Third-order spectroscopy, in contrast, carries the advantage of interrogating the relative orientation or anisotropy of molecules or vibrations using specifically a designed pulsed sequence with various polarizations.[24] In principle, a fully quantum mechanical treatment of the response function of the entire system in solution will inherently incorporate the orientational information. However, due to intractable number of degrees of freedoms, it is most common to treat low frequency motions such as translational motion and rotational motion classically, with more detail discussed later. The dipole moment operator under this classical treatment becomes a function of these classical motions, and we need to take rotational motion into account for describing the orientational effect onto the response functions.

Incorporating rotational and translational effects into the response functions requires an additional consideration of rotation between the lab frame and the molecular frame. Generally speaking, coordinate transform between two frames can be achieved by Euler angle transformation:

$$\boldsymbol{\mu}^{ab} = T(\alpha,\beta,\gamma)\boldsymbol{m}^{ab} \tag{2.120}$$

$$T(\alpha,\beta,\gamma) = \begin{bmatrix} \cos(\alpha)\cos(\gamma) - \sin(\alpha)\cos(\beta)\sin(\gamma) & \sin(\alpha)\cos(\gamma) + \cos(\alpha)\cos(\beta)\sin(\gamma) & \sin(\beta)\sin(\gamma) \\ -\cos(\alpha)\sin(\gamma) - \sin(\alpha)\cos(\beta)\cos(\gamma) & -\sin(\alpha)\sin(\gamma) + \cos(\alpha)\cos(\beta)\cos(\gamma) & \sin(\beta)\cos(\gamma) \\ \sin(\alpha)\sin(\beta) & -\cos(\alpha)\sin(\beta) & \cos(\beta) \end{bmatrix}$$

(2.121)

For isotropic orientation, the distribution of α and γ range uniformly from 0 to 2π whereas β is weighted by sinusoidal function from the angle of 0 to π. Because of the symmetry in isotropic solution, the orientationally averaged first-order response function in Eqn. (2.83) takes the form

$$R^{(1)}(\tau_1)=\Theta(\tau_1)\left(\frac{2}{\hbar}\right)\frac{1}{8\pi} \times \mathrm{Im}\left\{\int_0^{2\pi}d\alpha\int_0^{\pi}d\beta\sin(\beta)\int_0^{2\pi}d\gamma\left\langle\sum_{i=1}^{3}T_i^I(\alpha,\beta,\gamma)\hat{\mu}_i(\tau_1)T_i^I(\alpha,\beta,\gamma)\hat{\mu}_i(0)\right\rangle_0\right\} \tag{2.122}$$

Note that the summation goes through $\underline{i}$=*x*, *y*, and *z*, and each direction should be identical to the other directions, meaning that we can just choose z direction that leads to

$$\frac{1}{8\pi}\int_0^{2\pi}d\alpha\int_0^{\pi}d\beta\sin(\beta)\int_0^{2\pi}d\gamma T_z^2=\frac{1}{8\pi}\int_0^{2\pi}d\alpha\int_0^{\pi}d\beta\sin(\beta)\int_0^{2\pi}d\gamma\cos^2(\beta)=\frac{\pi}{3} \tag{2.123}$$

which is simply a constant, and the linear response function becomes

$$R^{(1)}(\tau_1)=\Theta(\tau_1)\frac{2\pi}{3\hbar}\sum_{i=1}^{3}\mathrm{Im}\left\langle\hat{\mu}_i(\tau_1)\hat{\mu}_i(0)\right\rangle_0 \tag{2.124}$$

Similarly, the orientationally-averaged third-order response functions require the integration over all dipole moments so that

$$\hat{\mu}_I\hat{\mu}_J\hat{\mu}_K\hat{\mu}_I\rightarrow T_{ijkl}^{IJKL}\hat{\mu}_i\hat{\mu}_j\hat{\mu}_k\hat{\mu}_l \tag{2.125}$$

And eventually we can write the expressions as follows

$$R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T_{ijkl}^{IJKL}\,\mathrm{Im}\{ R_{ijkl}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3)+R_{ijkl}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)-R_{ijkl}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)\} \tag{2.126}$$

$$R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T_{ijkl}^{IJKL}\,\mathrm{Im}\{ R_{ijkl}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3)+R_{ijkl}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)-R_{ijkl}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)\} \tag{2.127}$$

With each pathway mathematically same as before except for enumerating the orientations in the molecular frame:

$$R_{IJKL}^{(\mathrm{I})}(\tau_1,\tau_2,\tau_3)=\left\langle \hat{\mu}_j(\tau_1)\hat{\mu}_k(\tau_1+\tau_2)\hat{\mu}_l(\tau_1+\tau_2+\tau_3)\hat{\mu}_i(0)\right\rangle_0 \tag{2.128}$$

$$R_{IJKL}^{(\mathrm{II})}(\tau_1,\tau_2,\tau_3)=\left\langle \hat{\mu}_i(0)\hat{\mu}_k(\tau_1+\tau_2)\hat{\mu}_l(\tau_1+\tau_2+\tau_3)\hat{\mu}_j(\tau_1)\right\rangle_0 \tag{2.129}$$

$$R_{IJKL}^{(III)}(\tau_1,\tau_2,\tau_3)=\left\langle \hat{\mu}_i(0)\hat{\mu}_j(\tau_1)\hat{\mu}_l(\tau_1+\tau_2+\tau_3)\hat{\mu}_k(\tau_1+\tau_2)\right\rangle_0 \tag{2.130}$$

$$R_{IJKL}^{(IV)}(\tau_1,\tau_2,\tau_3)=\left\langle \hat{\mu}_l(\tau_1+\tau_2+\tau_3)\hat{\mu}_k(\tau_1+\tau_2)\hat{\mu}_j(\tau_1)\hat{\mu}_i(0)\right\rangle_0 \tag{2.131}$$

Here, T_{ijkl}^{IJKL} is the orientaional tensor that describes the weight depending on the microscopic orientation of the molecules labeled *i*, *j*, *k*, and *l* in molecular frame and the sequence of macroscopic linear polarization of electric field pulses and the signal, labeled *I*, *J*, *K*, and *L* in the lab frame. The values of T_{ijkl}^{IJKL} is summarized in Table 2.2.

Orientational Tensor	Value
T_{iiii}^{IIII}	1/5
$T_{iijj}^{IIII}=T_{ijji}^{IIII}=T_{ijij}^{IIII}$	1/15
$T_{iiii}^{IIJJ}=T_{iiii}^{IIJI}=T_{iiii}^{IIJJ}$	1/15
$T_{iijj}^{IIJJ}=T_{ijji}^{IIJI}=T_{ijij}^{IIJJ}$	2/15
$T_{ijji}^{IIJJ}=T_{ijij}^{IIJJ}=T_{iijj}^{IIJI}=T_{ijij}^{IIJI}=T_{iijj}^{IIJJ}=T_{ijji}^{IIJJ}$	-1/30

Table 2.2: Non-zero elements of the orientational tensors for third-order spectroscopy in the isotropic solution

2.4.6 Single Oscillator Model

It is useful to study simple model systems to illustrate the connection between molecular Hamiltonian and the 2D IR spectrum, and to identify the information that can be obtained from 2D IR spectroscopy. Two model systems will be studied here. One is a three-level single oscillator consisting of a ground state, first excited and second excited states to see the effect of anharmonicity. The other is two vibrationally coupled oscillators, resulting in six levels to investigate the effect of vibrational coupling and the orientational effects discussed above.

As the first example, we use a single weakly-anharmonic oscillator model that has only three states (see also Eqn. (2.119)):

$$\hat{H}_0 = \hbar \sum_{a=0}^{2} \omega_a |a\rangle\langle a|, \quad \omega_a = \begin{cases} 0, & a = 0 \\ \omega, & a = 1 \\ 2\omega - \Delta, & a = 2 \end{cases} \tag{2.132}$$

with the harmonic frequency ω, and the diagonal anharmonicity Δ. The diagonal anharmonicity of the amide I vibrations is typically 16 cm^{-1},[25] whereas the frequency of amide I is around 1650 cm^{-1}. Based on the argument made in Eqn. (2.86), we can safely assume the initial population is entirely at the vibrational ground state $|0\rangle$. The relationship of the transition dipole moments can be approximated by that of a simple harmonic oscillator:

$$\left|\mu^{10}\right| \equiv \left|\langle 1|\hat{\mu}|0\rangle\right| \tag{2.133}$$

$$\left|\mu^{21}\right| \equiv \left|\langle 2|\hat{\mu}|1\rangle\right| = \sqrt{2}\left|\mu^{10}\right| \tag{2.134}$$

The corresponding Liouville pathways are shown in Fig. 2.4. For simplicity, the transition dipole moment is directed at +*z*, meaning that we only need to evaluate *i*=*j*=*k*=*l*=*z*. The rephasing and nonrephasing response functions can be written as

$$
\begin{aligned}
R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl} T_{zzzz}^{IJKL}\,\mathrm{Im}\{ \\
+\left|\mu^{10}\right|^4 e^{+i\omega\tau_1}e^{-i\omega\tau_3} \\
+\left|\mu^{10}\right|^4 e^{+i\omega\tau_1}e^{-i\omega\tau_3} \\
-2\left|\mu^{10}\right|^4 e^{+i\omega\tau_1}e^{-i(\omega-\Delta)\tau_3}\}
\end{aligned}
\tag{2.135}
$$

$$
\begin{aligned}
R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3} T_{zzzz}^{IJKL}\,\mathrm{Im}\{ \\
+\left|\mu^{10}\right|^4 e^{-i\omega\tau_1}e^{-i\omega\tau_3} \\
+\left|\mu^{10}\right|^4 e^{-i\omega\tau_1}e^{-i\omega\tau_3} \\
-2\left|\mu^{10}\right|^4 e^{-i\omega\tau_1}e^{-i(\omega-\Delta)\tau_3}\}
\end{aligned}
\tag{2.136}
$$

One can see that the rephasing pathway has the oscillatory terms of $e^{+i\omega\tau_1}$ along τ_1 and $e^{-i\omega\tau_3}$ along τ_3 and the nonrephasing pathway has the oscillatory terms of $e^{-i\omega\tau_1}$ in τ_1 and $e^{-i\omega\tau_3}$ in τ_3, which are consistent with the Eqns. (2.104)–(2.105). For obtaining a 2D spectrum from one pathway, both τ_1 and τ_3 time interval are Fourier-transformed, so that

$$
\tilde{S}_{IJKL}^{(3)}(\omega_1,\tau_2,\omega_3) = \int_{-\infty}^{\infty} d\tau_1 e^{i\omega\tau_1}\int_{-\infty}^{\infty} d\tau_3 e^{i\omega\tau_3} R_{IJKL}^{(3)}(\tau_1,\tau_2,\tau_3) \tag{2.137}
$$

The observable $\tilde{S}_{IJKL}^{(3)}(\omega_1,\tau_2,\omega_3)$ in Eqn. (2.137) is essentially the 2D signal we observe, which may be convolved with additional pulse envelope as in Eqn. (2.98). This Eqn. also means that the rephasing pathway will have the oscillatory frequency of -ω in ω_1 (excitation frequency axis), and ω in ω_3 (detection frequency axis) whereas the non-rephasing pathway will have the oscillatory frequency of ω in both frequency axes. In practice, we need to flip the excitation frequency axis of the rephasing response in order to construct the correlation (C) spectral lineshape

$$
\tilde{S}_{IJKL}^{(\mathrm{C})}(\omega_1,\tau_2,\omega_3) = \tilde{S}_{IJKL}^{(\mathrm{R})}(-\omega_1,\tau_2,\omega_3) + \tilde{S}_{IJKL}^{(\mathrm{NR})}(\omega_1,\tau_2,\omega_3) \tag{2.138}
$$

As a side note, the transient absorption spectrum that measures the detection frequency as a function of τ_2 (waiting time) can be obtained from the projection-slice theorem:[26]

$$S^{\mathrm{TA}}\left(\tau_2,\omega_3\right)=S^{\mathrm{2D}}\left(\tau_1=0,\tau_2,\omega_3\right)=\int_{-\infty}^{\infty}d\omega_1 S^{\mathrm{2D}}\left(\omega_1,\tau_2,\omega_3\right) \tag{2.139}$$

One can also see from the Eqns. (2.135)–(2.136) that the ESA can cancel with both GSB and SE when $\Delta = 0$, meaning that a perfectly harmonic oscillator does not have any 2D IR signal. In reality, nearly all of the vibrational potentials are anharmonic.

Vibrational relaxation, which will be discussed later, is usually accounted for empirically by an exponential decay empirically with the rate constant Γ for convenience, although there are many dephasing mechanisms, including pure dephasing from random frequency modulation of the individual oscillator, and ensemble dephasing induced by oscillators with distribution of different frequencies. For amide I vibration, it is typically around 1.0 ps to 1.3 ps.[25, 27] For modeling the 2D spectra of both pathways, we multiply the response function with the exponential decay so that

$$\begin{aligned} R_{IJKL}^{(\mathrm{R})}\left(\tau_1,\tau_2,\tau_3\right)=\Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\Theta\left(\tau_3\right)\frac{2}{\hbar^3}\sum_{ijkl}T_{zzzz}^{IJKL}\,\mathrm{Im}\Big\{ \\ +2\left|\mu^{10}\right|^4 e^{+i\omega\tau_1+\Gamma\tau_1}e^{-\Gamma\tau_2}e^{-i\omega\tau_3+\Gamma\tau_3} \\ -2\left|\mu^{10}\right|^4 e^{+i\omega\tau_1+\Gamma\tau}e^{-\Gamma\tau_2}e^{-i(\omega-\Delta)\tau_3+\Gamma\tau_3}\Big\} \end{aligned} \tag{2.140}$$

$$\begin{aligned} R_{IJKL}^{(\mathrm{NR})}\left(\tau_1,\tau_2,\tau_3\right)=\Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\Theta\left(\tau_3\right)\frac{2}{\hbar^3}T_{zzzz}^{IJKL}\,\mathrm{Im}\Big\{ \\ +2\left|\mu^{10}\right|^4 e^{-i\omega\tau_1+\Gamma\tau_1}e^{-\Gamma\tau_2}e^{-i\omega\tau_3+\Gamma\tau_3} \\ -2\left|\mu^{10}\right|^4 e^{-i\omega\tau_1+\Gamma\tau_1}e^{-\Gamma\tau_2}e^{-i(\omega-\Delta)\tau_3+\Gamma\tau_3}\Big\} \end{aligned} \tag{2.141}$$

As a result, the 2D signal will take the form:

$$\tilde{S}_{IJKL}^{(\mathrm{R})}(\omega_1,\tau_2,\omega_3)=\Theta(\tau_2)e^{-\Gamma\tau_2}\frac{2i}{\hbar^3}\left(T_{zzzz}^{IJKL}\left|\mu^{10}\right|^4\right)\Bigg\{ +\frac{1}{i(\omega_1+\omega)-\Gamma}\frac{1}{i(\omega_3-\omega)-\Gamma} -\frac{1}{i(\omega_1+\omega)-\Gamma}\frac{1}{i(\omega_3-(\omega-\Delta))-\Gamma}\Bigg\} \tag{2.142}$$

$$\tilde{S}_{IJKL}^{(\mathrm{NR})}(\omega_1,\tau_2,\omega_3)=\Theta(\tau_2)e^{-\Gamma\tau_2}\frac{2i}{\hbar^3}\left(T_{zzzz}^{IJKL}\left|\mu^{10}\right|^4\right)\Bigg\{ +\frac{1}{i(\omega_1-\omega)-\Gamma}\frac{1}{i(\omega_3-\omega)-\Gamma} -\frac{1}{i(\omega_1-\omega)-\Gamma}\frac{1}{i(\omega_3-(\omega-\Delta))-\Gamma}\Bigg\} \tag{2.143}$$

Also the most common polarizations used are ZZZZ and ZZYY, meaning that

$$\begin{aligned} T_{zzzz}^{ZZZZ} &= 1/5 \\ T_{zzzz}^{ZZYY} &= 1/15 \end{aligned} \tag{2.144}$$

implying that the diagonal intensity in the ZZYY spectrum will be 1/3 of the intensity in the ZZZZ spectrum. The resulting ZZZZ-polarized spectra of this single oscillator are shown in Fig. 2.5. One can see phase twist in both rephasing and nonrephasing real parts, which is due to a mixing of absorptive and dispersive components of the signal. The correlation surface, which is the sum of both rephasing surface and nonrephasing surface gives a perfect two-dimensional Lorentzian lineshape. In practice, inaccurate phasing or determination of the timing can lead to a mixing of real part and imaginary part such as the 2D spectrum has some phase twist.

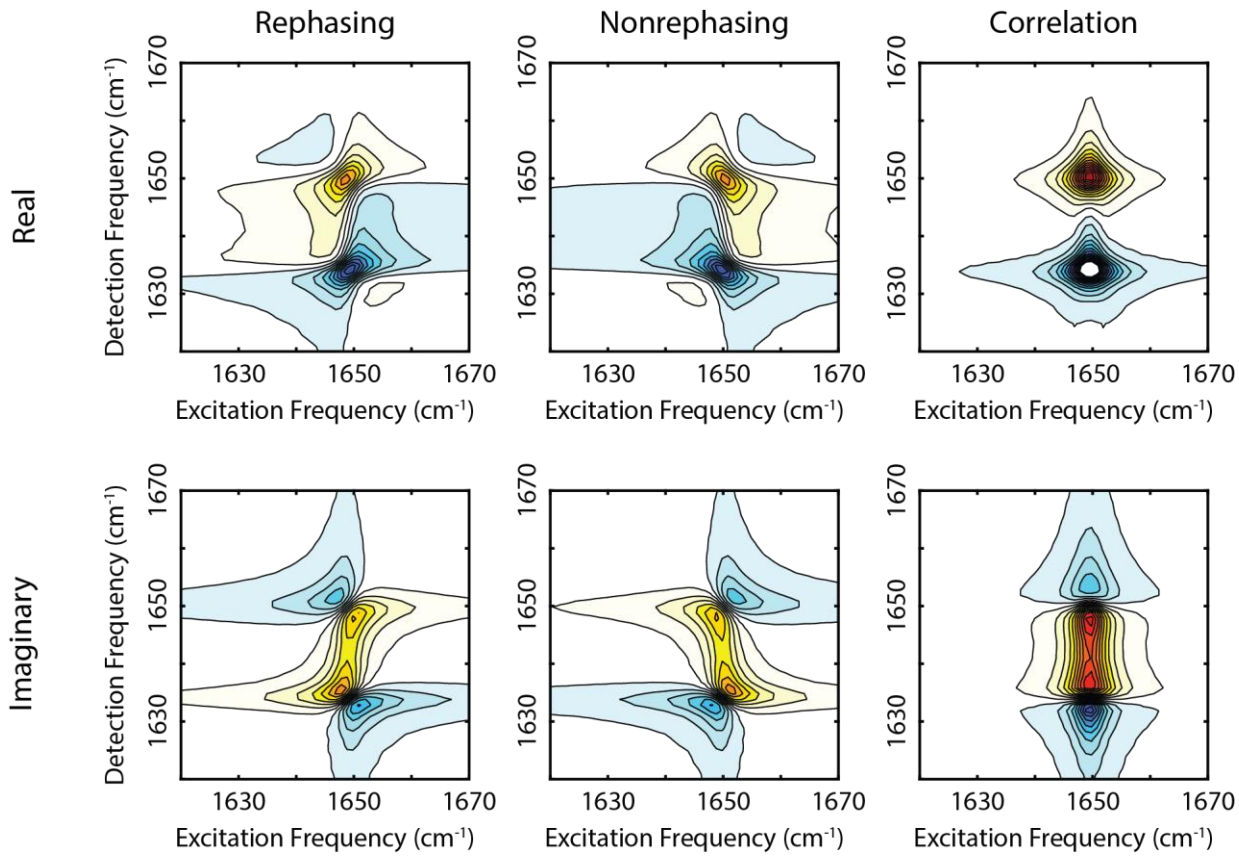

Figure 2.5: Comparison of the rephasing, nonrephasing, and correlation 2D spectra for a three-level single oscillator with both real part and imaginary part. The parameter used here: ω = 1650 cm^{-1}, Δ = 16 cm^{-1}, $1/\Gamma$ = 1 ps^{-1}, τ_2 = 0. Red color indicates positive contours while blue indicates negative. All of the spectra are normalized against the maximum magnitude of the real part correlation 2D spectrum.

2.4.7 Two Coupled Oscillator Model

For two coupled oscillators, the corresponding weakly-anharmonic Hamiltonian in the eigenbasis can be written as

$$\hat{H}_0 = \sum_{a,b} E_{ab} |ab\rangle\langle ab|, \quad E_{ab} = \begin{cases} 0 & a=0, b=0 \\ \omega_a & a=1, b=0 \\ \omega_b & a=0, b=1 \\ 2\omega_a - \Delta_a & a=2, b=0 \\ 2\omega_b - \Delta_b & a=0, b=2 \\ \omega_a + \omega_b + \Delta_{ab} & a=1, b=1 \end{cases} \tag{2.145}$$

For amide I spectroscopy, typically only diagonal anharmonicity of the overtone is taken into account, meaning that $\Delta_{ab} = 0$. The strength of transition dipole moment is the same for each oscillator as in Eqn. (2.133)–(2.134). We can follow the exact same argument as in the single oscillator. However, since enumerating all of the response functions in equation form will be messy, we instead present Liouville pathways that can contribute to the rephasing response and nonrephasing response in Fig. 2.6, and example 2D spectra shown in Fig. 2.7.

Clearly, one can see additional crosspeak signals in the 2D spectrum, which originates from the pathways oscillating with the frequency ω_a or ω_b along τ_1, but oscillating with ω_b or ω_a along τ_3. Also, the crosspeak is sensitive to the underlying relative orientation of the coupled transition dipoles. To illustrate this point, we compare 2D spectra for otherwise identical coupled oscillator with different relative orientation of the transition dipoles. The first case has the two transition dipoles are parallel to each other, while in the second they are perpendicular. The resulting ZZZZ-polarized and ZZYY-polarized 2D spectra (Fig. 2.7) show that ZZYY-polarized 2D spectrum preferentially enhances the crosspeak intensity of the perpendicular dipoles (bottom right), implying that the polarization control can inform us about the relative orientation of the coupled vibrations in general. Indeed, such behavior comes from the weighting due to orientational average discussed above. Also, one can use depolarization ratio of the cross-peak to characterize the relative orientation:

$$D_{\mathrm{CP}}\left(\Theta;\tau_2\right)=\frac{\tilde{S}_{ZZYY}\left(\omega_1,\tau_2,\omega_3\right)}{\tilde{S}_{ZZZZ}\left(\omega_1,\tau_2,\omega_3\right)}=\frac{1}{6}\frac{7-\cos^2\Theta}{1+2\cos^2\Theta} \tag{2.146}$$

with Θ the angle between the two transition dipoles. Or alternatively, one can use the anisotropy to investigate the orientational time correlation function:

$$\gamma(\tau_2) = \frac{\tilde{S}_{ZZZZ}(\tau_2) - \tilde{S}_{ZZYY}(\tau_2)}{\tilde{S}_{ZZZZ}(\tau_2) + 2\tilde{S}_{ZZYY}(\tau_2)} \equiv \frac{\tilde{S}_{\text{aniso}}}{\tilde{S}_{\text{iso}}} \tag{2.147}$$

where τ_{or} is the orientational correlation time of the species studied.

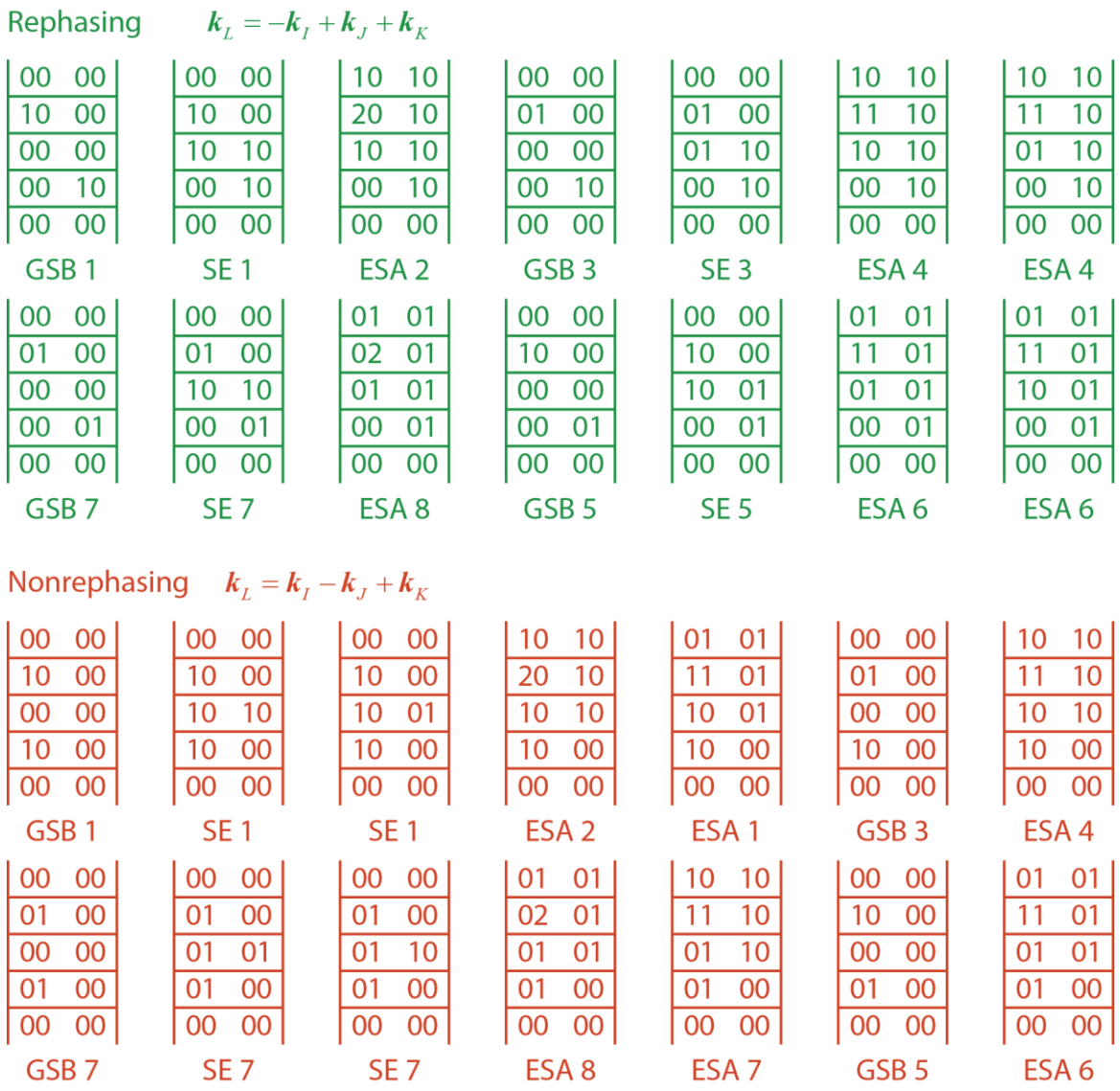

Figure 2.6: All possible Liouville pathways for the six-level coupled oscillator. For compactness, the arrows are dropped, and the notation being $|ab\rangle\langle ab| = |00\rangle\langle 00|$ for the starting vibrational ground state. GSB: Ground State Bleach. SE: Stimulated Emission. ESA: Excited State Absorption. Numbers are used to assign the peaks in Fig. 2.7.

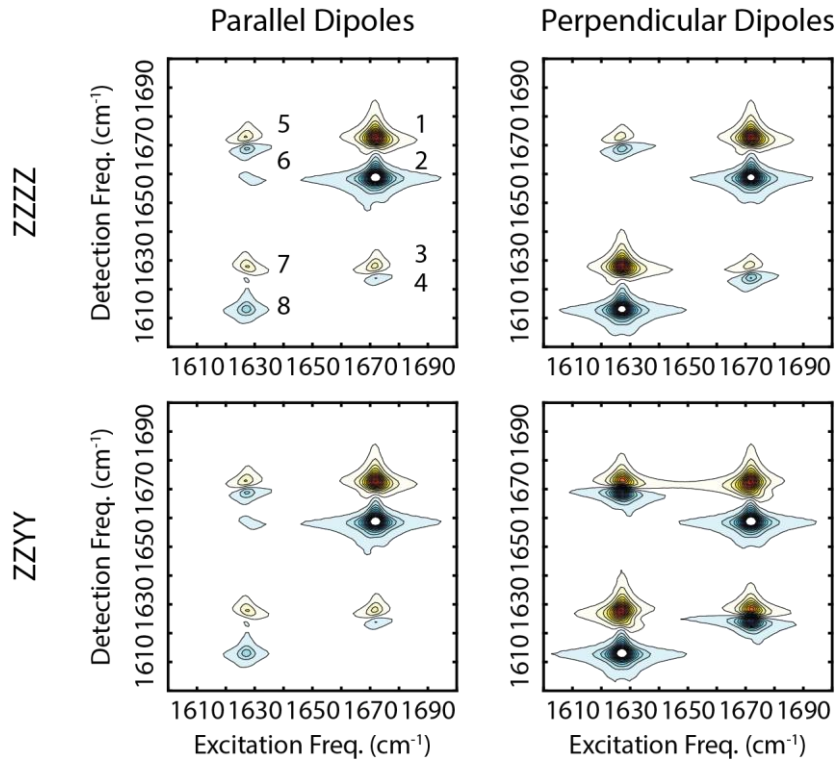

Figure 2.7: Polarization-dependent 2D IR spectra of the two coupled oscillator. Left: 2D IR spectra of the coupled oscillators with the parallel transition dipoles. Right: 2D IR spectra of the coupled oscillator with the perpendicular transition dipoles. Numbers in the top left correspond to the Liouville pathways shown in Fig. 2.6.

2.5 Response Functions Coupled to a Bath

2.5.1 System-Bath Hamiltonian

So far, we have treated isolated systems whose eigenstates exactly known. In reality, for condensed phase systems what we are really interested in is usually only a subset of the whole, such as protein amide I vibrations in bulk aqueous solution, meaning that we focus primarily on only a limited number of degrees of freedom (DOFs) instead. In practice, it is impossible to completely characterize all sort of the DOFs, and one has, to some degree, to employ stochastic descriptions of the remaining DOFs onto the physical observables we are interested in such as

amide I frequency fluctuations.[28] This is accomplished by partitioning the entire system into a system of interest comprising tractable DOFs, a bath consisting of the rest, and the interactions between the system and the bath illustrated in Fig. 2.8.

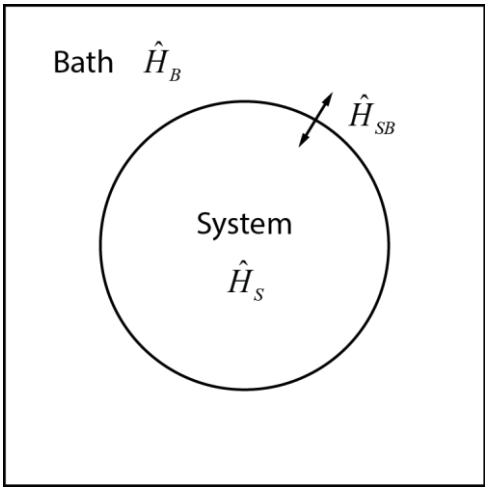

Figure 2.8: Schematic representation of the system-bath Hamiltonian.

The Hamiltonian in the Eqns. (2.84) and (2.110) can then be written as

$$\hat{H}_0(\boldsymbol{q},\boldsymbol{Q}) = \hat{H}_S(\boldsymbol{q}) + \hat{H}_B(\boldsymbol{Q}) + \hat{H}_{SB}(\boldsymbol{q},\boldsymbol{Q}) \tag{2.148}$$

in which the system Hamiltonian $\hat{H}_S(\boldsymbol{q})$ depends on its own nuclear coordinates $\boldsymbol{q}$. Given the environmental bath coordinates $\boldsymbol{Q}$, the bath Hamiltonian $\hat{H}_B(\boldsymbol{Q})$ is coupled to the system through the system-bath Hamiltonian $\hat{H}_{SB}(\boldsymbol{q},\boldsymbol{Q})$ that describes the interactions between the system and the bath. It is the system-bath Hamiltonian that induces the excess energy of the system exchanges with its surrounding bath, leading to the energy relaxation. One basis description for this Hamiltonian can be given as

$$\begin{aligned} \hat{H}_S(\boldsymbol{q}) &= \sum_a E_a |a\rangle\langle a| \\ \hat{H}_B(\boldsymbol{Q}) &= \sum_\alpha E_\alpha |\alpha\rangle\langle\alpha| \end{aligned} \tag{2.149}$$

where $|a\rangle$ and $|\alpha\rangle$ are referred to as the eigenbasis of the system Hamiltonian and the bath Hamiltonian without interactions between the two, meaning that

$$\left[\hat{H}_S(\boldsymbol{q})+\hat{H}_B(\boldsymbol{Q})\right]|a\alpha\rangle=\left(E_a+E_\alpha\right)|a\alpha\rangle \tag{2.150}$$

And we introduce the partial trace that performs the average over the bath coordinates:

$$\mathrm{Tr}_B\hat{\rho}=\sum_\alpha\langle\alpha|\hat{\rho}|\alpha\rangle=\hat{\sigma}_S=\sum_{a,b}\rho_{ab}|a\rangle\langle b| \tag{2.151}$$

with σ_S the reduced density matrix that describes the system. Or alternatively,

$$\mathrm{Tr}_S\hat{\rho}=\sum_a\langle a|\hat{\rho}|a\rangle=\hat{\sigma}_B=\sum_{\alpha,\beta}\rho_{\alpha\beta}|\alpha\rangle\langle\beta| \tag{2.152}$$

The expectation value of an observable $\hat{A}$ can be computed:

$$\left\langle\hat{A}\right\rangle=\mathrm{Tr}\left(\hat{A}\hat{\rho}\right)=\mathrm{Tr}\left(\hat{A}\mathrm{Tr}_B(\rho)\right)=\mathrm{Tr}\left(\hat{A}\hat{\sigma}_S\right) \tag{2.153}$$

2.5.2 Classical Approximation on the Bath

The linear response function with the system coupled to the bath can be written from Eqn. (2.83):

$$\begin{aligned}R^{(1)}(\tau_1)&=\Theta(\tau_1)\left(\frac{2\pi}{3\hbar}\right)\mathrm{Im}\sum_{i=1}^{3}\sum_{a,\alpha}\langle a\alpha|\hat{\mu}_i(\tau_1)\hat{\mu}_i(0)\hat{\rho}_0|a\alpha\rangle\\&=\Theta(\tau_1)\left(\frac{2}{\hbar}\right)\mathrm{Im}\sum_{i=1}^{3}\sum_{a}\langle a|\left\langle\hat{\mu}_i(\tau_1)\hat{\mu}_i(0)\right\rangle_B|a\rangle\end{aligned} \tag{2.154}$$

where

$$\left\langle\hat{\mu}_I(\tau_1)\hat{\mu}_I(0)\right\rangle_B=\left\langle e^{\frac{i}{\hbar}\hat{H}_0\tau_1}\hat{\mu}_I e^{-\frac{i}{\hbar}\hat{H}_0\tau_1}\hat{\mu}_I\right\rangle_B \tag{2.155}$$

is the dipole time correlation function averaged over the bath coordinates, but the system coordinates still remain active. Various approaches have been developed to treat this problem, which is of central importance in open quantum systems.[29-30] A very common approach is to assume a harmonic bath. This bath is described by an effective collective coordinate that models the statistical behavior of a collection of the bath coordinates, and the fluctuating behavior follows Gaussian statistics (through the Central limit Theorem).[4] The system-bath coupling is described by a bilinear coupling term in the ladder operator form, which is essentially displacement of the bath collective coordinate. This approach is usually assumed to have diagonal fluctuations on the system eigenbasis such that the resulting dipole time correlation function can be computed exactly with second-order cumulant expansion and spectral density that describes density of states of the bath collective modes.

For our purpose of connecting the response functions to computational spectroscopy with molecular simulations, here we instead introduce a mixed quantum-classical approach where the bath coordinates are treated entirely classically.[31-32] In other words, it is inherently imposing a high temperature limit onto the bath. This approximation is valid when either the temperature goes to the infinity, which is not a viable simulation condition, or the frequency of the bath modes is much smaller than the thermal energy (~200 cm^{-1}), such as collective protein conformational fluctuations, water libration, *etc*.

Here, we take adiabatic separation for granted to separate the system as a collection of high frequency amide I vibrations and the bath consisting of slow frequency modes. Many other examples are Born-Oppenheimer approximation that separates the fast electronic degrees of freedom and the slow nuclear degrees of freedom, and separation of different frequency ranges

such as UV/Vis, IR, microwave, *etc*. The goal of the mix quantum-classical description for the dipole operator is to find:

$$\hat{\mu}_I(\tau_1) \to \hat{U}^\dagger\left(\boldsymbol{Q}(\tau_1)\right)\hat{\mu}_I\left(\boldsymbol{Q}(\tau_1)\right)\hat{U}\left(\boldsymbol{Q}(\tau_1)\right) \tag{2.156}$$

where the operator $\hat{\mu}_I\left(\boldsymbol{Q}(\tau_1)\right)$ only depends on the classical bath coordinates $\boldsymbol{Q}(\tau_1)$, and the time evolution can still be evolved under some quantum mechanical Hamiltonian given the classical bath coordinates. And we choose to allow the classical coordinates $\boldsymbol{Q}(\tau_1)$ to evolve under the classical equations of motion for the ground state Hamiltonian

$$\hat{H}_g = \hat{\varepsilon}_0\left(\boldsymbol{Q}(\tau_1)\right) + \hat{T}_B \tag{2.157}$$

with the bath kinetic energy operator $\hat{T}_B$, and the potential energy of the bath plus the ground state energy of the system. Physically, similar to Born-Oppenheimer approximation, the bath degrees of freedom evolve with the potential as if the system stays in the ground state. Then the correspondence principle gives

$$\hat{\mu}_I\left(\boldsymbol{Q}(\tau_1)\right) \to e^{\frac{i}{\hbar}\hat{H}_g\tau_1}\hat{\mu}_I e^{-\frac{i}{\hbar}\hat{H}_g\tau_1} \equiv \hat{U}_g^\dagger(\tau_1)\hat{\mu}_I\hat{U}_g(\tau_1) \tag{2.158}$$

The total Hamiltonian can then be written as

$$\hat{H}_0 = \hat{H}_e + \hat{H}_g \tag{2.159}$$

with the system Hamiltonian being the energy gap Hamiltonian

$$\hat{H}_e = \sum_{a\neq 0}\left[\hat{\varepsilon}_a(\boldsymbol{Q}) - \hat{\varepsilon}_0(\boldsymbol{Q})\right]|a\rangle\langle a| \tag{2.160}$$

Then the corresponding $\hat{U}(\tau_1)$ is found to be

$$\hat{U}(\tau_1) = \exp_+\left(\frac{-i}{\hbar}\int_0^{\tau_1} dt\hat{H}_e(t)\right) \tag{2.161}$$

which is a time-ordered exponential given the time-dependent Hamiltonian.

With the classical equivalence, one can write

$$\hat{H}_e(\boldsymbol{Q}(t)) \to \hat{H}_e(t) = \hat{U}_g^\dagger(t)\hat{H}_e\hat{U}_g(t) \tag{2.162}$$

$$\hat{U}(\boldsymbol{Q}(\tau_1)) \to \hat{U}_g^\dagger(t)\hat{U}(t)\hat{U}_g(t) \tag{2.163}$$

Finally, the transition dipole operator has the desired correspondence

$$\hat{\mu}_I(\tau_1) \to \hat{U}^\dagger(\boldsymbol{Q}(\tau_1))\hat{\mu}_I(\boldsymbol{Q}(\tau_1))\hat{U}(\boldsymbol{Q}(\tau_1)) \tag{2.164}$$

Now, the classical approximation of the dipole time correlation function can be written as

$$\langle\hat{\mu}_I(\tau_1)\hat{\mu}_I(0)\rangle_B = \langle\hat{U}^\dagger(\boldsymbol{Q}(\tau_1))\hat{\mu}_I(\boldsymbol{Q}(\tau_1))\hat{U}(\boldsymbol{Q}(\tau_1))\hat{\mu}_I(\boldsymbol{Q}(0))\rangle_B \tag{2.165}$$

This setup that utilizes the classical bath coordinates under the ground state Hamiltonian is the basic assumption for well-known molecular dynamics (MD) simulations. Under the ergodic hypothesis, we can obtain

$$\langle\hat{\mu}_I(\tau_1)\hat{\mu}_I(0)\rangle_B = \langle\hat{m}_I(\tau_1)\hat{U}^\dagger(\tau_1)\hat{m}_I(0)\hat{U}(\tau_1)\rangle_{\mathrm{MD}} \tag{2.166}$$

with the notation

$$\langle A\rangle_{\mathrm{MD}} \equiv \lim_{T\to\infty}\frac{1}{T}\int_0^T dtA(t) \tag{2.167}$$

$$\hat{m}_I(\tau) \equiv \hat{\mu}_I(\boldsymbol{Q}(t_0+\tau)) \tag{2.168}$$

$$\hat{U}(t_0+\tau, t_0) \equiv \hat{U}(\boldsymbol{Q}(t_0+\tau), t_0) \tag{2.169}$$

Now, the linear response function can be written in the mix quantum-classical manner:

$$R^{(1)}(\tau_1) \approx \Theta(\tau_1)\left(\frac{2\pi}{3\hbar}\right)\mathrm{Im}\langle 0|\left\langle \hat{m}_I(\tau_1)\hat{U}^{\dagger}(\tau_1)\hat{m}_I(0)\hat{U}(\tau_1)\right\rangle_{\mathrm{MD}}|0\rangle \tag{2.170}$$

The system propagator $\hat{U}(\tau_1)$ is still quite general so that it can induce transition between system states. In practice, the block diagonal approximation is usually made in computational IR spectroscopy, which prohibits transitions between different vibrational quanta or manifolds. The singly excited states will stay in the singly excitation manifold. Also, the ground state remains unchanged. Therefore,

$$R^{(1)}(\tau_1) \approx \Theta(\tau_1)\left(\frac{2\pi}{3\hbar}\right)\sum_{i=1}^{3}\sum_{a,b}\mathrm{Im}\left\langle \hat{m}_i^{0b}(\tau_1)\hat{U}_{ba}(\tau_1,0)\hat{m}_i^{a0}(0)\right\rangle_{\mathrm{MD}} \tag{2.171}$$

The third-order response functions can be written in the same mixed quantum-classical manner:

$$\begin{aligned} R^{(\mathrm{R})}_{IJKL}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T^{IJKL}_{ijkl}\,\mathrm{Im}\{ \\ R^{(\mathrm{III})}_{ijkl}(\tau_1,\tau_2,\tau_3)+R^{(\mathrm{II})}_{ijkl}(\tau_1,\tau_2,\tau_3)-R^{(\mathrm{I})}_{ijkl}(\tau_1,\tau_2,\tau_3)\} \end{aligned} \tag{2.172}$$

$$\begin{aligned} R^{(\mathrm{NR})}_{IJKL}(\tau_1,\tau_2,\tau_3) = \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T^{IJKL}_{ijkl}\,\mathrm{Im}\{ \\ R^{(\mathrm{IV})}_{ijkl}(\tau_1,\tau_2,\tau_3)+R^{(\mathrm{I})}_{ijkl}(\tau_1,\tau_2,\tau_3)-R^{(\mathrm{II})}_{ijkl}(\tau_1,\tau_2,\tau_3)\} \end{aligned} \tag{2.173}$$

$$\begin{aligned} R^{(\mathrm{I})}_{ijkl}(\tau_1,\tau_2,\tau_3) = \sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}_i^{0a}(0)\hat{U}^{\dagger}_{aa'}(\tau_1+\tau_2+\tau_3,0)\hat{m}_l^{a'b}(\tau_1+\tau_2+\tau_3) \\ \times\hat{U}_{bb'}(\tau_1+\tau_2+\tau_3,\tau_1+\tau_2)\hat{m}_k^{b'c}(\tau_1+\tau_2)\hat{U}_{cc'}(\tau_1+\tau_2,\tau_1)\hat{m}_j^{c'0}(\tau_1)\Big\rangle_{\mathrm{MD}} \end{aligned} \tag{2.174}$$

$$\begin{aligned} R^{(\mathrm{II})}_{ijkl}(\tau_1,\tau_2,\tau_3) = \sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}_i^{0a}(0)\hat{U}^{\dagger}_{aa'}(\tau_1+\tau_2,0)\hat{m}_k^{a'b}(\tau_1+\tau_2)\hat{U}^{\dagger}_{bb'}(\tau_1+\tau_2,\tau_1+\tau_2+\tau_3) \\ \times\hat{m}_l^{b'c}(\tau_1+\tau_2+\tau_3)\hat{U}_{cc'}(\tau_1+\tau_2+\tau_3,\tau_1)\hat{m}_j^{c'0}(\tau_1)\Big\rangle_{\mathrm{MD}} \end{aligned} \tag{2.175}$$

$$R_{ijkl}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3) = \sum_{aa'bb'cc'}^{\mathrm{RWA}} \left\langle \hat{m}_i^{0a}(0)\hat{U}_{aa'}^{\dagger}(\tau_1,0)\hat{m}_j^{a'b}(\tau_1)\hat{U}_{bb'}^{\dagger}(\tau_1,\tau_1+\tau_2+\tau_3) \right. \\ \left. \times \hat{m}_l^{b'c}(\tau_1+\tau_2+\tau_3)\hat{U}_{cc'}(\tau_1+\tau_2+\tau_3,\tau_1+\tau_2)\hat{m}_k^{c'0}(\tau_1+\tau_2) \right\rangle_{\mathrm{MD}} \tag{2.176}$$

$$R_{ijkl}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3) = \sum_{aa'bb'cc'}^{\mathrm{RWA}} \left\langle \hat{m}_i^{0a}(0)\hat{U}_{aa'}^{\dagger}(\tau_1,0)\hat{m}_j^{a'b}(\tau_1)\hat{U}_{bb'}^{\dagger}(\tau_1,\tau_1+\tau_2) \right. \\ \left. \times \hat{m}_k^{b'c}(\tau_1+\tau_2)\hat{U}_{cc'}^{\dagger}(\tau_1+\tau_2,\tau_1+\tau_2+\tau_3)\hat{m}_l^{c'0}(\tau_1+\tau_2+\tau_3) \right\rangle_{\mathrm{MD}} \tag{2.177}$$

The Eqns. (2.171)–(2.177) constitute the set of fundamental equations to numerically simulate IR spectra throughout the thesis. Even the simpler sum-over-state method that neglects the dynamical behavior can be reduced from these equations and Fourier transformed into the eigenstate basis in frequency, and averaged by instantaneous MD snapshots.

Some comments need to be made before moving on to the next section. Firstly, one can see that both rephasing and non-rephasing pathways only require the Eqns. (2.174)–(2.177), and these equations are sharing similar propagator and variables albeit at different times. It is possible to devise wavefunction methods that propagates wavefunctions first and then calculating each Liouville pathway at the end, which will be discussed in detail in Chapter 3. Also, the bottleneck of the numerical simulations of the third-order spectroscopy comes down to two steps: generation of the propagator, and the propagation. Since the block diagonal approximation is usually made, the size of the two-quantum Hamiltonian of n oscillators is $n(n+1)/2$, meaning that generating the two-quantum propagator scales with $O(n^6)$ due to diagonalization of the Hamiltonian, which is actually really demanding. Numerical propagation of the two-quantum states scales with $O(n^4)$. Full simulations of the third-order response functions are prohibited once n reaches 40–50.[18] However, approximate schemes can be made to achieve numerical simulations feasible for n more than 100, which will also be discussed at Chapter 3.

2.5.3 Kubo Model for Spectral Diffusion and Lineshape

Before closing this chapter, it is still important to discuss the physical phenomena due to the system-bath coupling, and how such effects appears in the lineshape of the IR spectra. In particular, we will utilize one simple mixed quantum-classical model, the Kubo model, to discuss spectral diffusion, changes of spectral lineshape, and some connections to vibrational relaxation.

For the system coupled to the bath illustrated in Fig. 2.8, the Hamiltonian based on Eqn. (2.148) can be written as

$$\begin{aligned} \hat{H}_0 &= \hat{H}_S + \hat{H}_B + \hat{H}_{\mathrm{SB}} \\ &= \hat{H}_S + \delta\hat{H}(t) \end{aligned} \tag{2.178}$$

$$\begin{aligned} \hat{H}_S &= \sum_a E_a |a\rangle\langle a| \\ \hat{H}_B &= \sum_\alpha E_\alpha |\alpha\rangle\langle\alpha| \end{aligned} \tag{2.179}$$

where the bath degrees of freedom are inherently average by performing the partial trace. As a result, the system-bath coupling is instead described by a stochastic Hamiltonian $\delta\hat{H}(t)$. The equation is still completely general since we only average over the bath degrees of freedom. Physically, the stochastic Hamiltonian describes the fluctuations of energy acting on the system Hamiltonian including diagonal frequency and off-diagonal coupling. For instance, strength of hydrogen bonds between water and amide groups can induce frequency fluctuations, and dihedral rotation of the protein backbones can induce changes of through-bond coupling between adjacent amide I vibrations. The stochastic nature can also be physically interpreted as instantaneous reshaping of the vibrational potential to induce frequency shift, or vibrational Stark effect that changes the frequency of the chromophore due to local electrostatics.

For simplicity, we assume the stochastic Hamiltonian to be diagonal on the system eigenbasis. In other words,

$$\delta\hat{H}(t)=\sum_{n}\delta E_{n}(t)|n\rangle\langle n|,\quad \langle\delta E_{n}(t)\rangle=0 \tag{2.180}$$

meaning that

the coupling to the bath does not induce transitions between system eigenstates. In general, the stochastic Hamiltonian does not commute with itself at different times, but in this case, it will always commute with itself since it is diagonal. Such time-dependent evolution are known as spectral diffusion, and this spectral diffusion causes the randomization of the phase such that dephasing of the correlation function occurs (pure dephasing). The key question here is how the spectral diffusion dynamics influence the lineshape of absorption spectra.

Let us examine the dipole time response function first. Using the mixed quantum-classical treatment from Eqn. (2.171),

$$\langle\hat{\mu}_{I}(\tau_{1})\hat{\mu}_{I}(0)\rangle_{0}=\sum_{a,b}P_{a}|\mu_{ab}|^{2}e^{-i\omega_{ba}\tau_{1}}F(\tau_{1}) \tag{2.181}$$

$$F(\tau_{1})=\left\langle\exp\left\{-i\int_{0}^{t}d\tau\delta\omega_{ba}(\tau)\right\}\right\rangle_{B} \tag{2.182}$$

with $F(\tau_1)$ describing the irreversible decays of the correlation in time, and giving rise to the spectral lineshape of the linear absorption spectrum (dephasing). Here, we already made two approximations. First, the dipole operator does not depend on the nuclear coordinate (the Condon approximation). Second, the bath DOFs are assumed to behave classically. Assuming the bath collective coordinate leads to Gaussian statistics on the frequency fluctuation, second-order cumulant expansion will give

$$F(\tau_1) \approx e^{-g(\tau_1)} \tag{2.183}$$

$$g(\tau_1) = \int_0^{\tau_1} dt_2 \int_0^{t_2} dt_1 \left\langle \delta\omega_{ba}(t_2 - t_1)\delta\omega_{ba}(0)\right\rangle_B \tag{2.184}$$

which connects the irreversible dephasing to the equilibrium frequency time correlation function. The function $g(t)$ is sometimes referred to as the lineshape function.[4] The fluctuations due to the system-bath coupling induce dephasing of the dipole time correlation function. Under the assumption of Gaussian statistics of the bath, dephasing can be fully characterized by the frequency-frequency correlation function (FFCF) $C_{\delta\omega\delta\omega}(t) \equiv \left\langle \delta\omega_{ba}(t)\delta\omega_{ba}(0)\right\rangle_B$, and we can alternatively represent it by a spectral density that describes the system-bath coupling strength

$$J_{ba}(\omega) = \int_{-\infty}^{\infty} dt \left\langle \delta\omega_{ba}(t)\delta\omega_{ba}(0)\right\rangle_B e^{i\omega t} \tag{2.185}$$

The spectral density is a key quantity for describing spectral diffusion, relaxation, population transfer and so on for describing system-bath interactions,[4] and it can be obtained or estimated from experiments or simulations. The assumption here is that the system does not influence the bath state as suggested from the previous treatment of classical approximation, and the system-bath coupling is weak enough that it is separable in the system-bath Hamiltonian. As an additional note, the FFCF can be treated fully quantum mechanically to make this framework semi-classical, which can be obtained by performing ab initio MD of the full system. Or alternatively, a purely real classical FFCF can be calculated from a classical simulation, then a quantum correction factor to ensure the fluctuation-dissipation theorem holds applied to reconstruct the estimated quantum correlation function.[33]

As a simple model that can still capture useful physical insights, we assume the FFCF takes the simple exponential form without the imaginary part:

$$\left\langle \delta\omega_{ba}(t)\,\delta\omega_{ba}(0)\right\rangle_B = \Delta^2 e^{-t/\tau_c} \tag{2.186}$$

with Δ the standard deviation of the frequency fluctuation, and τ_c the correlation time. Plugging Eqn. (2.186) into Eqn. (2.184) gives

$$g(\tau_1) = \Delta^2\tau_c^2\left(e^{-\tau_1/\tau_c} + \tau_1/\tau_c - 1\right) \tag{2.187}$$

We can investigate the effect of the correlation time on the spectral lineshape. When $\tau_1 \ll \tau_c$, the FFCF is essentially static, and one can see

$$\begin{aligned} g(\tau_1) &= \Delta^2\tau_c^2\left[\left(1 - \tau_1/\tau_c + \frac{1}{2}\frac{\tau_1^2}{\tau_c^2} + \cdots\right) + \tau_1/\tau_c - 1\right] \\ &\approx \frac{1}{2}\Delta^2\tau_1^2 \end{aligned} \tag{2.188}$$

$$F(\tau_1) \approx e^{-\frac{1}{2}\Delta^2\tau_1^2} \tag{2.189}$$

meaning that the dipole correlation function has a Gaussian decay in time and also Gaussian spectral lineshape in frequency,

$$\left\langle \hat{\mu}_I(\tau_1)\,\hat{\mu}_I(0)\right\rangle_0 = \sum_{a,b} P_a \left|\mu_{ab}\right|^2 e^{-i\omega_{ba}\tau_1} e^{-\frac{1}{2}\Delta^2\tau_1^2} \tag{2.190}$$

$$\tilde{C}_{\mu\mu}(\omega) \equiv \int_{-\infty}^{\infty} dt e^{i\omega t}\left\langle \hat{\mu}_I(\tau_1)\,\hat{\mu}_I(0)\right\rangle_0 = \sqrt{\frac{2\pi}{\Delta^2}}\sum_a P_a\left|\mu_{ab}\right|^2 \exp\left(-\frac{(\omega-\omega_{ba})^2}{2\Delta^2}\right) \tag{2.191}$$

which is called the inhomogeneous limit.

When $\tau_c \ll \tau_1$, the FFCF de-correlates rapidly, resulting in the homogeneous limit.

$$\begin{aligned} g(\tau_1) &\approx \Delta^2\tau_c^2\left[1 + \tau_1/\tau_c - 1\right] \\ &\approx \Delta^2\tau_c\tau_1 \end{aligned} \tag{2.192}$$

$$F(\tau_1) = e^{-\Gamma\tau_1}, \quad \Gamma \equiv \Delta^2 \tau_c \tag{2.193}$$

The dipole time correlation function has a simple exponential decay in time, which in turn has a Lorentzian lineshape in frequency.

$$\left\langle \hat{\mu}_I(\tau_1)\hat{\mu}_I(0)\right\rangle_0 = \sum_{a,b} P_a \left|\mu_{ab}\right|^2 e^{-i\omega_{ba}\tau_1} e^{-\Gamma\tau_1} \tag{2.194}$$

$$\mathrm{Re}\,\tilde{C}_{\mu\mu}(\omega) = \sum_{a,b} P_a \left|\mu_{ab}\right|^2 \frac{\Gamma}{(\omega-\omega_{ba})^2 + \Gamma^2} \tag{2.195}$$

This simple model captures the two limiting cases that gives rise to distinct spectral lineshape. Experimentally and computationally, the spectral diffusion kinetics can be extracted by performing waiting time-dependent 2D IR spectroscopy that collects 2D spectra as a function of τ_2, and investigating the Center Line Slope (CLS) decay.[34-35] The CLS decay is a proxy to the underlying FFCF and can be used to extract the correlation time. One example of application to Ala–Ala is shown in Chapter 6.

2.5.4 Vibrational Relaxation

Here, we want to discuss how the excess energy of a chromophore is distributed to the surrounding bath after an optical excitation, which is a fundamental relaxation process for spectroscopy. We use the system-bath Hamiltonian in the Eqns. (2.178)–(2.179). Without the explicit separation of the system and the bath, let us say the initial state and the final state are given as

$$\begin{aligned} |i\rangle &= |a\alpha\rangle \\ |f\rangle &= |b\beta\rangle \end{aligned} \tag{2.196}$$

Based on the Golden rule, the rate constant of state transition from *i* to *f*, k_{fi}, can be written as[36-37]

$$\begin{aligned} k_{fi}\left(\omega_{fi}\right) &= \frac{2\pi}{\hbar}\sum_{i,f} P_i \left|\langle i|\delta\hat{H}|f\rangle\right|^2 \delta\left(\omega_f-\omega_i\right) \\ &= \frac{2\pi}{\hbar}\sum_{a,b,\alpha,\beta} P_{a,\alpha}\left|\langle a\alpha|\delta\hat{H}|b\beta\rangle\right|^2 \delta\left(\left(\omega_b+\omega_\beta\right)-\left(\omega_a+\omega_\alpha\right)\right) \\ &= \frac{1}{\hbar^2}\int_{-\infty}^{\infty} dt \sum_{a,b,\alpha,\beta} P_{a,\alpha}\left|\langle a\alpha|\delta\hat{H}|b\beta\rangle\right|^2 e^{-i\left[\left(\omega_b+\omega_\beta\right)-\left(\omega_a+\omega_\alpha\right)\right]t} \\ &= \frac{1}{\hbar^2}\int_{-\infty}^{\infty} dt \left\langle \hat{V}_I(t)\hat{V}_I(0)\right\rangle, \quad \hat{V}_I(t) = e^{\frac{i}{\hbar}\left(\hat{H}_S+\hat{H}_B\right)t}\hat{H}_{\mathrm{SB}}e^{-\frac{i}{\hbar}\left(\hat{H}_S+\hat{H}_B\right)t} \end{aligned} \tag{2.197}$$

Even though the system-bath coupling depends on the coordinates of both the system and the bath, we recognize

$$\langle a\alpha|\delta\hat{H}|b\beta\rangle = \langle\alpha|\delta\omega_{ab}|\beta\rangle \tag{2.198}$$

Then transition rate can eventually be obtained as

$$\begin{aligned} k_{ba}\left(\omega_{ba}\right) &= \frac{1}{\hbar^2}\int_{-\infty}^{\infty} dt \left\langle \delta\omega_{ba}(t)\delta\omega_{ba}\right\rangle_B e^{-i\omega_{ba}t} \\ &= \frac{1}{\hbar^2} J_{ba}\left(\omega_{ba}\right) \end{aligned} \tag{2.199}$$

Eqn. (2.199) is quite informative. First, the transition rate can be intimately related to the spectral density introduced in Eqn. (2.185), which described the coupling strength of the bath mode at the resonant frequency to the system. Also, the irreversible vibrational relaxation is connected to the equilibrium property: spectral density, or frequency time correlation function, which is another example of FDT. One experimentally consistent example is also illustrated in Chapter 6. Please note that in practice for computational amide I spectroscopy, since the block diagonal approximation is almost always applied, vibrational relaxation is not accounted through the coupling between different vibrational quanta. Instead, population relaxation is determined by a simple rate constant *ad hoc*.

2.6 Acknowlegments

I thank Lukas Whaley-Mayda and Luis Busto de Moner for carefully reading through the chapter and giving valuable and constructive comments.

2.7 References

1. Dirac, P. A. M., The quantum theory of the emission and absorption of radiation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1927,** *114* (767), 243-265.
2. Fermi, E., *Nuclear physics: a course given by Enrico Fermi at the University of Chicago*. University of Chicago Press: 1950.
3. Leroy, B., How to convert the equations of electromagnetism from Gaussian to SI units in less than no time. *American Journal of Physics* **1985,** *53* (6), 589-590.
4. Mukamel, S. In *Principles of Nonlinear Optical Spectroscopy*, 1995.
5. Cho, M., *Coherent multidimensional spectroscopy*. Springer: 2019; Vol. 226.
6. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press: 2011.
7. Jackson, J. D., Classical electrodynamics. American Association of Physics Teachers: 1999.
8. Shen, Y.-R., The principles of nonlinear optics. *wi* **1984**.
9. Cho, M., Coherent two-dimensional optical spectroscopy. *Chem Rev* **2008,** *108* (4), 1331-418.
10. Onsager, L., Reciprocal Relations in Irreversible Processes. I. *Physical Review* **1931,** *37* (4), 405-426.
11. Callen, H. B.; Welton, T. A., Irreversibility and Generalized Noise. *Physical Review* **1951,** *83* (1), 34-40.
12. Kubo, R., Statistical-Mechanical Theory of Irreversible Processes. I. General Theory and Simple Applications to Magnetic and Conduction Problems. *Journal of the Physical Society of Japan* **1957,** *12* (6), 570-586.
13. Mukamel, S., Nonimpact unified theory of four-wave mixing and two-photon processes. *Physical Review A* **1983,** *28* (6), 3480-3492.
14. Mukamel, S.; Loring, R. F., Nonlinear response function for time-domain and frequency-domain four-wave mixing. *Journal of the Optical Society of America B* **1986,** *3* (4), 595.
15. Mukamel, S., Femtosecond Optical Spectroscopy: A Direct Look at Elementary Chemical Events. *Annual Review of Physical Chemistry* **1990,** *41* (1), 647-681.
16. Hu, B. Y. K., Kramers–Kronig in two lines. *American Journal of Physics* **1989,** *57* (9), 821-821.
17. Boyd, R. W., *Nonlinear optics*. Academic press: 2008.
18. Schweigert, I. V.; Mukamel, S., Simulating multidimensional optical wave-mixing signals with finite-pulse envelopes. *Physical Review A* **2008,** *77* (3).

19. Do, T. N.; Gelin, M. F.; Tan, H. S., Simplified expressions that incorporate finite pulse effects into coherent two-dimensional optical spectra. *J Chem Phys* **2017,** *147* (14), 144103.
20. Cheng, Y. C.; Lee, H.; Fleming, G. R., Efficient simulation of three-pulse photon-echo signals with application to the determination of electronic coupling in a bacterial photosynthetic reaction center. *J Phys Chem A* **2007,** *111* (38), 9499-508.
21. Gelin, M. F.; Egorova, D.; Domcke, W., Efficient calculation of time- and frequency-resolved four-wave-mixing signals. *Acc Chem Res* **2009,** *42* (9), 1290-8.
22. Li, H.; Spencer, A. P.; Kortyna, A.; Moody, G.; Jonas, D. M.; Cundiff, S. T., Pulse propagation effects in optical 2D Fourier-transform spectroscopy: experiment. *J Phys Chem A* **2013,** *117* (29), 6279-87.
23. Perlík, V.; Hauer, J.; Šanda, F., Finite pulse effects in single and double quantum spectroscopies. *Journal of the Optical Society of America B* **2017,** *34* (2), 430.
24. Khalil, M.; Demirdöven, N.; Tokmakoff, A., Coherent 2D IR Spectroscopy: Molecular Structure and Dynamics in Solution. *The Journal of Physical Chemistry A* **2003,** *107* (27), 5258-5279.
25. Hamm, P.; Lim, M.; Hochstrasser, R. M., Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *The Journal of Physical Chemistry B* **1998,** *102* (31), 6123-6138.
26. Gallagher Faeder, S. M.; Jonas, D. M., Two-Dimensional Electronic Correlation and Relaxation Spectra: Theory and Model Calculations. *The Journal of Physical Chemistry A* **1999,** *103* (49), 10489-10505.
27. Feng, C. J.; Tokmakoff, A., The dynamics of peptide-water interactions in dialanine: An ultrafast amide I 2D IR and computational spectroscopy study. *J Chem Phys* **2017,** *147* (8), 085101.
28. Nitzan, A., *Chemical dynamics in condensed phases: relaxation, transfer and reactions in condensed molecular systems*. Oxford university press: 2006.
29. Leggett, A. J.; Chakravarty, S.; Dorsey, A. T.; Fisher, M. P. A.; Garg, A.; Zwerger, W., Dynamics of the dissipative two-state system. *Reviews of Modern Physics* **1987,** *59* (1), 1-85.
30. Breuer, H.-P.; Petruccione, F., *The theory of open quantum systems*. Oxford University Press on Demand: 2002.
31. Torii, H., Effects of intermolecular vibrational coupling and liquid dynamics on the polarized Raman and two-dimensional infrared spectral profiles of liquid N,N-dimethylformamide analyzed with a time-domain computational method. *J Phys Chem A* **2006,** *110* (14), 4822-32.
32. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
33. Egorov, S. A.; Skinner, J. L., Semiclassical approximations to quantum time correlation functions. *Chemical Physics Letters* **1998,** *293* (5-6), 469-476.
34. Kwak, K.; Park, S.; Finkelstein, I. J.; Fayer, M. D., Frequency-frequency correlation functions and apodization in two-dimensional infrared vibrational echo spectroscopy: a new approach. *J Chem Phys* **2007,** *127* (12), 124503.
35. Fenn, E. E.; Fayer, M. D., Extracting 2D IR frequency-frequency correlation functions from two component systems. *J Chem Phys* **2011,** *135* (7), 074502.
36. Oxtoby, D. W., *Vibrational Population Relaxation in Liquids*. 1981; p 487-519.
37. Egorov, S. A.; Skinner, J. L., A theory of vibrational energy relaxation in liquids. *The Journal of Chemical Physics* **1996,** *105* (16), 7047-7058.

Chapter 3

Computational Amide I Spectroscopy

3.1 Introduction

In the previous chapter, we discussed the theory of nonlinear spectroscopy that eventually relates the observed spectrum with the molecular response function. We also developed a mixed quantum classical (MQC) model for evaluating the orientationally averaged response tensors in a classical bath, which can in turn be simulated with classical molecular dynamics. As a continuing effort, this chapter will focus on applying this MQC model to calculate protein amide I IR spectra, which allows a direct comparison with experiments, and quantitative determination of conformational ensemble against experimental data (See Chapter 5 and Chapter 8).

There are numerous approaches to calculate amide I IR spectra, including *ab initio* molecular dynamics (AIMD), quantum mechanics/molecular mechanics (QM/MM), classical MD, *etc*. In particular, the combination of the MQC model with spectroscopic maps will be used throughout this thesis, which translates protein instantaneous structure with its surrounding solvation environment into amide I vibrational frequencies and couplings between amide I vibrations. This mapping approach on amide I computational spectroscopy has been widely applied on various peptides and proteins,[1, 5-18] and summarized in detail in quite many reviews.[19-23] It indicates that amide I spectroscopic modeling is quite a well-developed field, and actually close to the points where the frequency accuracy reaches a few wavenumber difference from

experiments,[19] and where many variants of spectroscopy maps have been available.[20] Although there is still room for improving the quality of spectroscopic maps, including detailed (experimental) parametrization and rigorous theoretical formulation that accounts for various intermolecular effects, what this chapter does is to present the basics of amide I spectroscopic maps, a brief summary of map parameterizations, and how one can computationally obtain linear IR and 2D IR spectra from molecular dynamics simulations.

The content of this chapter is summarized as follows. Section 3.2 will briefly discuss various approaches of computationally predicting amide I spectra, and why the MQC approach is preferred. Section 3.3 describes the exciton Hamiltonian and the approximations made to allow relatively straightforward parameterizations of spectroscopic maps and spectral simulations introduced in Section 3.4. Section 3.5 presents the approach of numerically simulating IR and 2D IR spectra, with split-operator method to allow amide I spectral simulation of a large protein.

3.2 Short Summary of Various Computational Approaches

Before going into the detail of the spectroscopic maps, it is worth discussing computational approaches of predicting amide I spectroscopy and why this MQC approach is preferred on computationally predicting amide I spectra. All computational approaches are inevitably required to define what constitutes the system and the bath. The differences between these methods in essence originate from such definition, and which degrees of freedom (DOFs) are treated quantum mechanically or classically.

In theory, treating the whole protein and the surrounding bath fully quantum mechanically should be the most accurate representation and apparently the most computationally intensive. One example is to use *ab initio* MD simulations.[26] In this situation, the whole simulation box is treated

quantum mechanically, and all of the electronic DOFs and nuclear DOFs constitutes the system. In that sense, there is no clear cutoff between the system and the bath. Linear IR spectrum can be computed from computing the two-point transition dipole time correlation function presented in Chapter 2. However, nonlinear response function remains too computationally expensive. A perhaps feasible approach is semi-empirical quantum chemistry calculations such as ZINDO/S,[27] which stands for Zerner's Intermediate Neglect of Differential Overlap/Spectroscopy such that it can achieve somewhat efficient calculations while at the same able to reproduce experimental peak frequency. However, this approach cannot properly capture the accurate geometry of the molecules, which is essential for reasonable AIMD simulations and structural interpretations.

In quantum mechanics/molecular mechanics methods, the peptide or the protein is treated quantum mechanically, whereas the surrounding solvent and ions are treated classically.[28] This approach allows a more efficient simulation than the full AIMD while still having full interactions between the quantum systems and the classical bath. Also, because the vibrational motions of the quantum system are still anharmonic, which are not described by classical harmonic potential, QM/MM simulations can still be used to directly compute 2D IR spectra. However, it does require to have well-parameterized force field compatible with QM/MM, and the computational expense is still limited to simple peptides, and prohibitive for longer peptide chains or a large protein.

Classical MD treats the entirety of the degrees of freedom classically, and allows efficient protein simulations solvated with water. The corresponding linear IR spectrum can be computed using normal mode analysis from classical Hessian matrix. However, due to the nature of harmonic potential to describe chemical bond, nonlinear response is disappeared,[28] as demonstrated in Subsection 2.4.6. For our application to predict amide I 2D IR spectroscopy, classical MD is not

the most suitable approach, even though it has been demonstrated that nonperturbative approach to perform classical simulations can be used to compute nonlinear spectroscopy.[29]

Mixed Quantum-Classical (MQC) Model for the application to amide I spectroscopy treats the amide I vibrational excitation quantum mechanically whereas all the other DOFs are treated classically. This type of floating oscillator approach was firstly developed by Miyazawa *et al*.[30-31] and much later by Torii and Tasumi,[32] which treats the amide I vibrations adiabatically separated from the other DOFs and behaves like delocalized exciton across amide I vibrations. The distinct difference between the MQC model and the previous QM/MM method is that the amide I vibrations during the molecular simulations are not directly modeled quantum mechanically. Instead, mapping approach, which will be introduced later, is used to construct the quantum Hamiltonian from the classical simulation. Such approach allows a relatively straightforward parametrization of frequencies, coupling, and transition dipole moments. Spectral simulation using this approach utilizes the MQC formalism introduced in Section 2.5, which is computationally feasible for large proteins and still able to predict amide I spectra with reasonable accuracy to a few cm^{-1}.

3.3 Exciton Hamiltonian

In computational amide I spectroscopy, the amide I Hamiltonian is described through an exciton Hamiltonian accounting for the 0–1 transition:

$$\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}} = \hbar\left[\sum_{a}\omega_{a}\left(\boldsymbol{q};\boldsymbol{Q}\right)\left|a\right\rangle\left\langle a\right| + \sum_{a,b\neq a}J_{ab}\left(\boldsymbol{q};\boldsymbol{Q}\right)\left|a\right\rangle\left\langle b\right|\right] \tag{3.1}$$

In this Hamiltonian, a basis vector $|a\rangle$ is in essence in the diabatic basis (local basis) that describes a single excitation on a^{th} amide I vibration or site a, meaning that only the a^{th} amide I vibration is excited to the first excited state while all the other amide I vibrations or sites remain in the ground state. In the wavefunction picture, it means that

$$|a\rangle = \psi_1(\boldsymbol{r}_a)\prod_{b\neq a}\psi_0(\boldsymbol{r}_b) \tag{3.2}$$

where $\psi_1(\boldsymbol{r}_a)$ is the first excited state vibrational wavefunction of the site a and the rest of the sites are described by the ground state vibrational wavefunctions $\prod_{b\neq a}\psi_0(\boldsymbol{r}_b)$. The diagonal elements $\hbar\omega_a$ is referred to as the site energy with the corresponding site frequency ω_a (in cm^{-1}). In this diabatic basis, the site frequency of the site a corresponds to the amide I frequency of the site a without any coupling to all the other amide I vibrations. The coupling between amide I vibrations is then described by the off-diagonal element J_{ab}. The dependence to the nuclear coordinates are described by the system coordinate $\boldsymbol{q}$ based on a bath configuration described by the bath coordinates $\boldsymbol{Q}$. Please note that we already implicitly used the system-bath Hamiltonian, and the system now is the amide I Hamiltonian that consists of amide I vibrations.

In 2D IR spectroscopy, the doubly excited states are necessary to describe excited state absorption, and the corresponding diabatic state representation is:

$$|ab\rangle = \begin{cases} \psi_1(\boldsymbol{r}_a)\psi_1(\boldsymbol{r}_b)\prod_{c\neq a,b}\psi_0(\boldsymbol{r}_c) & a\neq b \\ \psi_2(\boldsymbol{r}_a)\prod_{c\neq a}\psi_0(\boldsymbol{r}_c) & a=b \end{cases} \tag{3.3}$$

The corresponding amide I Hamiltonian with zero, one, and two quantum states can be written as

$$\hat{H}_{\mathrm{AmI}} = \begin{bmatrix} 0 & \hat{C}_{10} & \hat{C}_{20} \\ \hat{C}_{01} & \hat{H}^{1\mathrm{Q}}_{\mathrm{AmI}} & \hat{C}_{21} \\ \hat{C}_{02} & \hat{C}_{12} & \hat{H}^{2\mathrm{Q}}_{\mathrm{AmI}} \end{bmatrix} \tag{3.4}$$

$$\hat{H}^{2\mathrm{Q}}_{\mathrm{AmI}} = \hbar \left[\sum_{ab} \omega_{a+b}(\boldsymbol{Q}) |ab\rangle\langle ab| + \sum_{a,b,c,d} J_{ab,cd}(\boldsymbol{Q}) |ab\rangle\langle cd| \right] \tag{3.5}$$

In the Eqn. (3.4), $\hat{C}$ describes the inter-manifold coupling including 0–1, 0–2, and 1–2 transitions, which is essential for describing the vibrational relaxation discussed in Section 2.5. ω_{a+b} and $J_{ab,cd}$ are the site frequency of the two-quantum state $|ab\rangle$ and the vibrational coupling between two-quantum states $|ab\rangle$ and $|cd\rangle$, respectively. This form of exciton Hamiltonian using the diabatic basis has two main advantages. First, the diabatic basis allows direct parameterization of the site frequencies and couplings on isolated amide I oscillators such as *N*-methylacetamide (NMA) or glycine dipeptide (GLDP) without explicitly considering the coupling to the bath, and the parameters can be transferrable to larger proteins in the solvent.[25, 33-35] Second, off-diagonal coupling is physically intuitive and easier to compute since the local amide I vibrational mode is well-defined in this situation and the kinetic coupling from the kinetic energy operator of the bath does not influence the coupling of amide I vibrations in the diabatic basis.

For the application to amide I 2D IR spectroscopy , the form of (3.4)–(3.5) can be further simplified. The first approximation is the block diagonal approximation, which neglects the inter-manifold coupling operator $\hat{C}$ such that

$$\hat{H}_{\mathrm{AmI}} \approx \begin{bmatrix} 0 & 0 & 0 \\ 0 & \hat{H}^{1\mathrm{Q}}_{\mathrm{AmI}} & 0 \\ 0 & 0 & \hat{H}^{2\mathrm{Q}}_{\mathrm{AmI}} \end{bmatrix} \tag{3.6}$$

It allows us to characterize the coupling only within the same quantum states, either in the one-quantum states or two-quantum states. Physically, it means that the (stochastic) motion of the bath and the motion of the system cannot induce vibrational transitions. This assumption is usually reasonable when the inter-manifold coupling (on the order of less than 100 cm^{-1}) is much smaller than the diagonal site frequency (~1700 cm^{-1}). Please see Mike Reppert's thesis for more detail.[36]

The second approximation made for amide I 2D IR spectroscopy is regarding the anharmonicity of the coupled vibrations. Physically, two coupled vibrations can influence each other such that the combination band will exhibit anharmonicity due to reshaping of the vibrational potential. For instance, a simple water molecule has symmetric stretch, anti-symmetric stretch and a bending vibration. Exciting bending vibration will induce time-dependent geometrical change of the water molecule including subtle change of bond length, and subsequent excitation of either stretch will have a different frequency. For weakly anharmonic oscillators, this off-diagonal anharmonicity can be neglected. Only the diagonal anharmonicity Δ is taken into account, such that the frequency in the two-quantum Hamiltonian can be written as:

$$\omega_{a+b} \approx \omega_a + \omega_b + \Delta\delta_{ab} \tag{3.7}$$

In Eqn. (3.7), δ_{ab} is the Kronecker delta, which is equal to 1 only when $a = b$, and zero otherwise. Also, the coupling between two-quantum states can be related to one-quantum vibrational coupling using ladder operators:

$$\begin{aligned} J_{ca,cb} &= \sqrt{1+\delta_{ca}}\, J_{ab} \\ J_{ca,cb} &= J_{cb,ca} \end{aligned} \tag{3.8}$$

This weakly anharmonic approximation allows us to directly obtain the two-quantum Hamiltonian from the one-quantum Hamiltonian, without explicit anharmonic quantum calculations or detailed experimental parameterization from the excited state absorption. The diagonal anharmonicity can

be determined in a straightforward manner in nonlinear spectroscopy, which has been determined to be 16 cm^{-1}.[37] Parameterization of the one-quantum Hamiltonian including site frequencies and couplings will be sufficient for computing amide I IR and 2D IR spectra, and the set of parameters that correlate coordinates to amide I frequencies and coupling constitute the spectroscopic maps. Please note that such approximation does not always hold. For instance, in water and excess proton problems, anharmonicity is so strong that weakly anharmonic approximation breaks down and one has to perform anharmonic calculations in order to investigate the ESA.[20, 38-39] We will discuss amide I spectroscopic maps in detail in the following section.

3.4 Spectroscopic Maps

We briefly closed the previous section with a few points. First, the exciton Hamiltonian is suitable for parameterization of diagonal site frequencies and off-diagonal coupling, which can be transferrable. Second, weakly anharmonic approximation and block diagonal approximation leads to simpler parameterization of spectroscopic maps that only requires characterizing the one-quantum site frequencies and couplings. The workflow of using spectroscopic maps is summarized in Fig. 3.1. In short, spectroscopic maps translate the time-dependent protein structure with its surrounding bath into amide I parameters including site frequencies, coupling, and transition dipole moments of the local (amide I) mode to generate time-dependent amide I Hamiltonian with associated transition dipole moments. The Hamiltonian trajectory and the dipole moment trajectory can then be used to compute the response tensors for linear IR and 2D IR spectra. In this section, the details of spectroscopic maps will be discussed including frequency maps, coupling maps, and transition dipole moments.

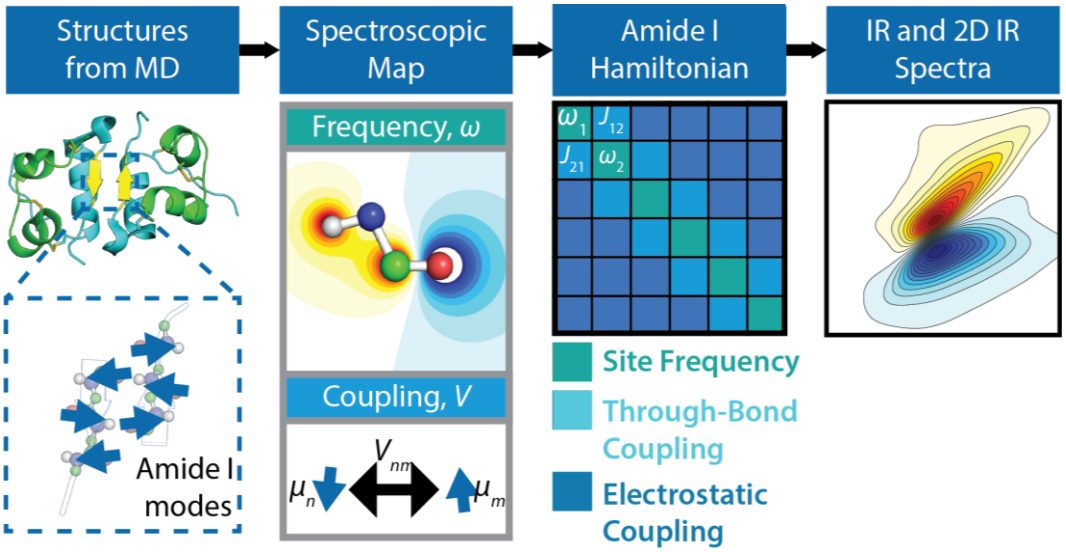

Figure 3.1: Workflow of the MQC model to predict amide I spectra.

3.4.1 Frequency Map

In the Eqn. (3.1), ω_a corresponds to the site frequency of the local amide I mode on site a. Now, taking the system-bath interactions into account, the amide I Hamiltonian can be written as

$$\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}} = \hat{H}_0 + \hat{V}(\boldsymbol{r};\boldsymbol{R}) \tag{3.9}$$

where $\hat{V}$ acts as the perturbing potential that can change the shape of the vibrational potential. As demonstrated in the previous chapter, $\hat{V}$ induces the site frequency variation. To first order, site frequency takes the form:

$$\omega_a \approx \omega_a^0 + \langle a|\hat{V}_a(r_a,\boldsymbol{R})|a\rangle \tag{3.10}$$

where r_a is the local mode coordinate of the site a. For small amplitude oscillations, the interaction term $\hat{V}_a$ can be expanded:

$$\hat{V}_a\left(r_a,\boldsymbol{R}\right)\approx V_a\left(0,\boldsymbol{R}\right)-F_a\left(\boldsymbol{R}\right)\cdot r_a \tag{3.11}$$

$$F_a\left(\boldsymbol{R}\right)=-\left(\frac{\partial V_a\left(r_a;\boldsymbol{R}\right)}{\partial r_a}\right)_{r_a^{\mathrm{eq}}} \tag{3.12}$$

In the Eqn. (3.12), F_a is the force exerted on the amide I from the surrounding bath in the MQC framework. The origin of the force includes all sorts of non-bonded interactions, such as electrostatic interactions, van der Waals and dispersion forces, *etc*.[19-20] However, it has been shown that electrostatic interactions dominate the frequency shift of amide I vibrations.[40-42]

The electrostatic contribution to the perturbing potential takes the form of Coulomb interactions between the amide bond charge density and the charge density from the surrounding bath. Approximately, these charge densities can be coarse-grained by atomic partial charges of the amide I system and the bath. Given a set of atomic partial charges Q_i with the associated coordinate R_i on the amide group, the electrostatic interaction takes the form:

$$\left\langle a\middle|\hat{V}_a\left(r_a;\boldsymbol{R}\right)\middle|a\right\rangle=\sum_i Q_i\left(r_a\right)\Phi\left(\boldsymbol{R}_i\left(r_a\right)\right) \tag{3.13}$$

In the Eqn. (3.13), the atomic coordinate depends on the local mode displacement r_a away from the equilibrium position $\boldsymbol{R}_i^{\mathrm{eq}}$:

$$\boldsymbol{R}_i\left(\boldsymbol{r}_a\right)=\boldsymbol{R}_i^{\mathrm{eq}}+\hat{\boldsymbol{r}}_i r_a \tag{3.14}$$

And the electrostatic potential, $\Phi\left(\boldsymbol{R}_i\left(r_a\right)\right)$, experienced by the amide atom i is

$$\Phi\left(\boldsymbol{R}_i\left(r_a\right)\right)=\sum_\alpha\frac{Q_\alpha}{\left|\boldsymbol{R}_i\left(r_a\right)-\boldsymbol{R}_\alpha\right|} \tag{3.15}$$

in which $\boldsymbol{R}_\alpha$ corresponds to the atomic position of the α^{th} atom in the bath. Please note that in the Eqn. (3.13), the atomic partial charges depend on the amide I local mode coordinate to account for the polarizability of the amide charge density. To express the dependence, we can further expand the partial charges:

$$Q_i\left(r_a\right) = Q_i^{\text{eq}} + \delta Q_i r_a + \cdots \tag{3.16}$$

where $\delta Q_i = \partial Q_i / \partial r_a$ is the charge flux along the local mode displacement. Also, the electrostatic potential can be perturbed from the local mode displacement so that

$$\Phi\left(\boldsymbol{R}_i\left(r_a\right)\right) = \Phi\left(\boldsymbol{R}_i^{\text{eq}}\right) + \left[\boldsymbol{E}\left(\boldsymbol{R}_i\right)\cdot\hat{\boldsymbol{r}}_i\right] r_a + \cdots \tag{3.17}$$

Plugging the Eqns. (3.16)–(3.17) into the Eqn. (3.13) leads to

$$\left\langle a\middle|\hat{V}_a\left(r_a;\boldsymbol{R}\right)\middle|a\right\rangle \approx \sum_i \left[Q_i^{\text{eq}}\Phi\left(\boldsymbol{R}_i\right) - \left(Q_i^{\text{eq}}\boldsymbol{E}\left(\boldsymbol{R}_i\right)\cdot\hat{\boldsymbol{r}}_i - \delta Q_i\Phi\left(\boldsymbol{R}_i\right)\right) r_a \right] \tag{3.18}$$

Physically, the zero-order contribution, $Q_i^{\text{eq}}\Phi\left(\boldsymbol{R}_i\right)$, describes the static frequency offset of the amide I local mode solvated in the environment. The first-order contributions have two terms, $Q_i^{\text{eq}}\boldsymbol{E}\left(\boldsymbol{R}_i\right)\cdot\hat{\boldsymbol{r}}$ and $\delta Q_i\Phi\left(\boldsymbol{R}_i\right)$. The first term corresponds to the motion of the equilibrium atomic partial charges along the local mode coordinate whereas the second term corresponds to the charge flux through the local mode coordinate, as illustrated in Fig. 3.2.

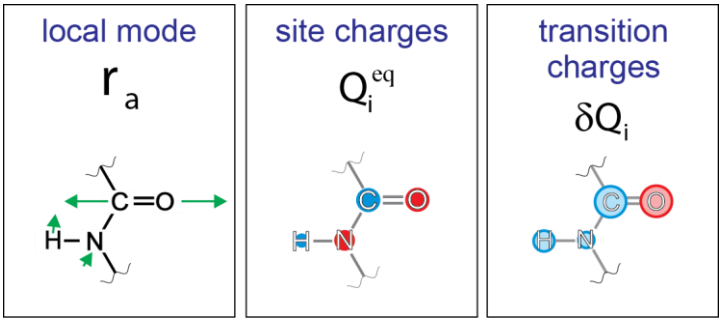

Figure 3.2: First-order contributions within the electrostatic interaction potential as the primary determinants of amide I frequencies in solution. Taken together, the first and second terms from Equation 11 provide a concise, but approximate, physical picture of the electrostatic interaction of an oscillating amide I bond with its environment: (a) local mode, (b) site charges, and (c) transition charges. In panels b and c, red circles indicate negative charges or charge fluxes; blue circles indicate positive charges or charge fluxes; and circle size indicates relative magnitude of atomic charge and charge flux values. The values shown here are meant to be illustrative and are not based on quantitative calculations. Reproduced from Figure 6 of Ref. 19. Copyright 2016 Annual Reviews

In practice, it is more computationally efficient to regard the Eqn. (3.18) as an semi-empirical equation, and treat the variables in the equation as the fit parameters, meaning that

$$\begin{aligned} \left\langle a\left|\hat{V}_a\left(r_a;\boldsymbol{R}\right)\right|a\right\rangle &= \sum_i \sum_{\beta=1}^{3} C_{i,\beta}^{(\mathrm{Field})} E_\beta^i + \sum_i C_i^{(\mathrm{pot})} \Phi_i \\ C_{i,\beta}^{\mathrm{Field}} &= Q_i^{\mathrm{eq}} r_i^\beta r_a \\ C_i^{\mathrm{pot}} &= \delta Q_i r_a \end{aligned} \tag{3.19}$$

Evaluating the site frequency leads to

$$\begin{aligned} \omega_a &\approx \omega_a^0 + \left\langle a\left|\hat{V}_a\left(r_a,\boldsymbol{R}\right)\right|a\right\rangle \\ &= \omega_0 + \sum_i \sum_{\beta=1}^{3} C_{i,\beta}^{(\mathrm{Field})} E_\beta^i + \sum_i C_i^{(\mathrm{pot})} \Phi_i \end{aligned} \tag{3.20}$$

The reference frequency of the site a, ω_a^0, is usually set to a constant across different sites such that $\omega_a^0 \equiv \omega_0$. In this form, the reference frequency ω_0, coefficients $C_i^{(\mathrm{pot})}$ and $C_{i,\beta}^{(\mathrm{Field})}$ constitute an electrostatic frequency map for amide I site energy calculations. Nearly all of the electrostatic

frequency maps available in this field follow the form in Eqn. (3.20) with a few exceptions such as Jansen's NMA frequency map, which accounts for the gradient of the field:[35]

$$\omega_a = \omega_0 + \sum_i \left[C_i^{(\mathrm{pot})} \Phi_i + \sum_{\beta=1}^{3} C_{i,\beta}^{(\mathrm{Field})} E_\beta^i + \sum_{\beta=1}^{3} \sum_{\gamma=\beta}^{3} C_{i,\beta\gamma}^{(\mathrm{Grad})} G_{\beta\gamma}^i \right] \tag{3.21}$$

3.4.2 Coupling Map

Coupling between two amide I vibrations can have many physical origins, including through-space electrostatic coupling, polarizability, steric interactions from nearby amide I vibrations, mechanical coupling between two adjacent amide groups, and spatial overlap of vibrational wavefunctions. In this section, we briefly discuss how the coupling is formulated and what are the forms for parameterizations.

The electrostatic, through-space coupling is similar to electrostatic site energy, which takes the form:[25, 33, 43]

$$\left\langle a \middle| \hat{V}\left(r_a, r_b; \boldsymbol{R}\right) \middle| b \right\rangle \equiv V_{ab} = \sum_{i,j} \frac{Q_i\left(r_a\right) Q_j\left(r_b\right)}{\left| \boldsymbol{R}_i\left(r_a\right) - \boldsymbol{R}_j\left(r_b\right) \right|} \tag{3.22}$$

Similarly, expanding the atomic partial charges as a function of local mode coordinates to first order results in the transition charge coupling (TCC) model:

$$J_{ab}^{\mathrm{TCC}} \approx \sum_{i,j} \left[\frac{Q_i^{\mathrm{eq}} Q_j^{\mathrm{eq}}}{\left| \boldsymbol{r}_{i,j}^{a,b} \right|^3} T_{i,j}^{a,b} + \frac{\left[\delta Q_i Q_j^{\mathrm{eq}} \left(\hat{\boldsymbol{r}}_j^b \cdot \hat{\boldsymbol{r}}_{i,j}^{a,b} \right) - Q_i^{\mathrm{eq}} \delta Q_j \left(\hat{\boldsymbol{r}}_i^a \cdot \hat{\boldsymbol{r}}_{i,j}^{a,b} \right) \right]}{\left| \boldsymbol{r}_{i,j}^{a,b} \right|^3} + \frac{\delta Q_i \delta Q_j}{\left| \boldsymbol{r}_{i,j}^{a,b} \right|} \right] \tag{3.23}$$

$$T_{i,j}^{a,b} = \delta_{ij} - 3\left(\hat{\boldsymbol{r}}_i^a \cdot \hat{\boldsymbol{r}}_{i,j}^{a,b} \right)\left(\hat{\boldsymbol{r}}_j^b \cdot \hat{\boldsymbol{r}}_{i,j}^{a,b} \right) \tag{3.24}$$

which $\boldsymbol{r}_{i,j}^{a,b}$ is the displacement vector between the amide atom *i* on the site *a* and the amide atom *j* on the site *b*. In this model, the first term in the Eqn. (3.23) describes the coupling arising from the motion of equilibrium charges relative to each other along the local mode coordinates. The second term describes the coupling between the motion of equilibrium charges along their nuclear motion and the charge flux along the corresponding local mode coordinate. The third term describes the coupling between charge fluxes of the two amide I vibrations on sites *a* and *b*. Neglecting the polarizability of the charge density leads to the transition dipole coupling (TDC) model:

$$J_{ab}^{\mathrm{TDC}} \approx \frac{\boldsymbol{m}^{0a} \cdot \boldsymbol{m}^{0b}}{R_{ab}^{3}} - 3\frac{\left(\boldsymbol{R}_{ab} \cdot \boldsymbol{m}^{0a}\right)\left(\boldsymbol{R}_{ab} \cdot \boldsymbol{m}^{0b}\right)}{R_{ab}^{5}} \tag{3.25}$$

$\boldsymbol{m}^{0a}$ describes the transition dipole moment of the site *a* in the MQC model, and R_{ab} denotes the distance between the two transition dipoles.

At beginning, we also mentioned there are additional contributions such as steric interactions, mechanical coupling between two adjacent amide groups, and spatial overlap of vibrational wavefunctions. These interactions are usually lumped into an empirical nearest neighbor coupling (NNC) map that depends on the backbone dihedral angles:

$$J_{a-1,a}^{\mathrm{NNC}} = J\left(\phi_n, \psi_n\right) \tag{3.26}$$

This map assigns coupling constants due to bonded nearest neighbor interactions using an empirical function $J\left(\phi_n, \psi_n\right)$, which is typically constructed as a grid-based map using quantum chemistry calculations on model peptides such as GLDP, or empirically determined using

experimentally measured protein spectra.[25, 33, 44-46] An example of this grid-based nearest-neighbor coupling map from Jansen *et al.* is shown in Fig. 3.3.

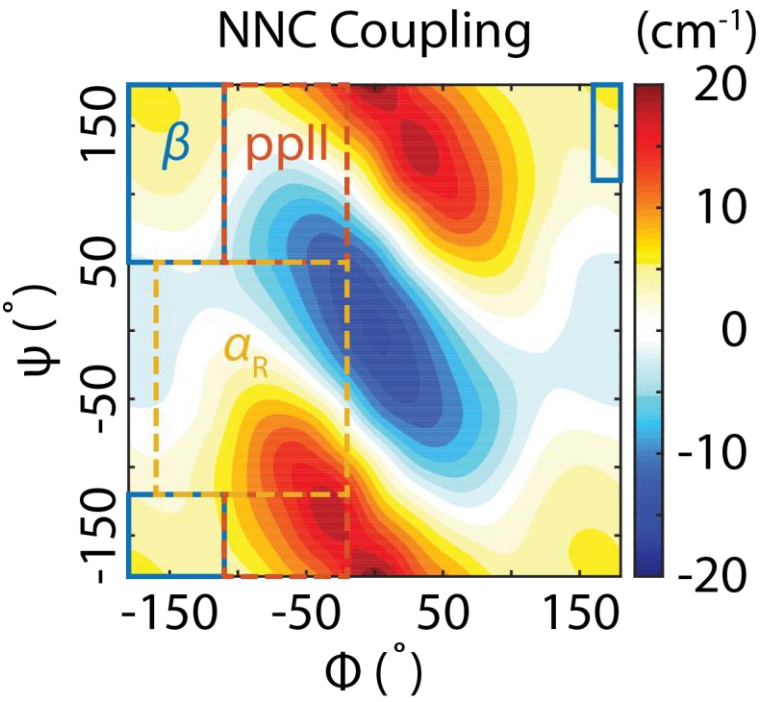

Figure 3.3: Nearest Neighbor Coupling Map as a function of backbone dihedral angles ϕ and ψ derived from DFT calculations of GLDP. Colored boxes show one definition of the conformers present in the literature.[1-3] The data was extracted from Ref. 25.

3.4.3 Transition Dipole Moments

Electrostatic descriptions of site energies and couplings can also be readily applied on the transition dipole moments for IR spectroscopy. The dipole moment operator of the local mode is given by

$$\hat{\boldsymbol{m}}(r_a) = \sum_i Q_i(r_a)\,\boldsymbol{R}_i(r_a) \tag{3.27}$$

The same treatment of expansion along the local mode coordinate based on the Eqn. (3.14) gives

$$\hat{\boldsymbol{m}}(r_a) \approx \sum_i \left[Q_i^{\mathrm{eq}} \boldsymbol{R}_i^{\mathrm{eq}} + \left(Q_i^{\mathrm{eq}} \hat{\boldsymbol{r}}_i + \delta Q_i \boldsymbol{R}_i^{\mathrm{eq}} \right) r_a \right] \tag{3.28}$$

Physically similar to the interpretation of the site energy, the first term indicates the static dipole moment that does not contribute to the amide I spectroscopy, and the second term describes the variation along the amide I coordinate, which gives rise to the intensity of the IR spectrum (Selection rule of infrared spectrscopy). Evaluating the transition dipole moment gives

$$\begin{aligned} \boldsymbol{m}^{0a} &= \langle 0 | \hat{\boldsymbol{m}}(r_a) | a \rangle \approx \frac{\partial \boldsymbol{m}}{\partial r_a} \langle 0 | r_a | a \rangle = \sqrt{\frac{\hbar}{2m\omega_a}} \sum_{i \in a} \left(Q_i^{\mathrm{eq}} \hat{\boldsymbol{r}}_i + \delta Q_i \boldsymbol{R}_i^{\mathrm{eq}} \right) \\ &\approx T^{-1}(\boldsymbol{R}) \mu_{\mathrm{ref}} \end{aligned} \tag{3.29}$$

where m is the reduced mass of the site a, and T is a rotation matrix to perform coordinate transform between the lab frame and the molecular frame, introduced in the Eqn. (2.122). Eqn. (3.29) indicates the magnitude of the transition dipole depends on both the equilibrium charge displacement and the charge flux. Amide I vibration does have significant charge flux component such that it turns out to have a large oscillator strength of ~0.12 D^2.[32, 47-48] More commonly, the transition dipole moment are parameterized using an effective transition dipole moment.[4, 49] Within the harmonic approximation, the transition dipole moment for 1–2 transition can be computed from

$$\boldsymbol{m}^{a,ab} = \sqrt{1+\delta_{ab}}\, \boldsymbol{m}^{0a} \tag{3.30}$$

3.4.4 Model Summary

We have introduced basic formalism of the frequency map, the coupling map, and transition dipole moments, which is readily applicable for parameterizations. Now, it is a good place to summarize these maps and to identify fitting parameters involved in these maps.

1. Frequency Map: Eqn. (3.21) is the key equation for computing frequency shifts due to electrostatic environments. The fitting parameters can be ω_0, $\left\{C_i^{(\mathrm{pot})}\right\}$, $\left\{C_{i,\beta}^{(\mathrm{Field})}\right\}$, and

$\left\{C_{i,\beta\gamma}^{(\mathrm{Grad})}\right\}$ based on the selection of atoms in "the amide system", which can be simply C=O or all the amide atoms C, O, N, H, and so on.

2. Through-Space Coupling Map: The electrostatic through-space coupling follows either the TDC model in the Eqn. (3.25), or the TCC model in the Eqn. (3.23). The fitting parameters are the charge fluxes $\{\delta Q_i\}$, and the local mode displacement vectors, which may be computed from normal mode calculations. Please note that the equilibrium charges $\left\{Q_i^{\mathrm{eq}}\right\}$ can be used directly from the protein force fields.
3. Through-Bond Coupling Map: Eqn. (3.26) is an effective model for describing such coupling map, which is typically parameterized by quantum chemistry calculations of a simple peptide as a function of backbone dihedral angles and Hessian matrix reconstruction method introduced by Cho and co-workers.[33]
4. Transition Dipole Moments: Transition dipole moments are obtained from the Eqn. (3.29), which can be computed either from equilibrium charges and transition charges, the same as the previous maps, or using an effective transition dipole moment μ_{ref}.

Note that in principle it is possible to use equilibrium charges, transition charges to determine all of these maps. However, in practice the parameterizations of each map can also be done separately. Also note that the construction of the two quantum Hamiltonian can be achieved using the Eqns. (3.7), (3.8), and (3.30) assuming a weakly anharmonic model.

3.4.5 Map Parameterizations

This part will briefly summarize parameterizations of spectroscopic maps and spectroscopic maps that are used throughout the thesis. For detailed reviews of the parameterization, please refer to the previous review papers.[19-20, 23]

Historically, the empirical correlation between amide I frequency shift $\Delta\omega$ in cm^{-1} and the distance to the hydrogen bond donor in Å was used:[37]

$$\Delta\omega = 30\left(r_{OH} - 2.6\right) \tag{3.31}$$

Later, Cho and co-workers developed an electrostatic-based frequency map by performing *ab initio* quantum chemistry calculations on NMA in *n*-D_2O complexes.[34] Similar approaches of *ab initio* parameterization have been adopted for many groups using the Eqn. (3.21) with different selections of atomic sites to evaluate the electrostatic variables such as field, potential, or even gradient.[35, 50-54]

This approach has been proven useful for qualitative interpretations of the experimental amide I vibrational spectra, but quantitative prediction has been challenging.[24, 49, 55-56] One of many challenging issues is rooted in relying heavily on model accuracy including but not limited to quantum chemistry calculations that are popular for map developments, and electrostatics and sampling of conformational distribution in MD simulation. One potential issue of incorporating *ab initio* maps from quantum chemistry calculations includes mismatch of the ground state geometry from ab initio optimized structures and structures from molecular-mechanic based structures from MD trajectories. Due to this mismatch, the interaction potential experienced by the amide group and resulting frequency shift can vary drastically, causing difficulty in quantitative predictions.[20] Another challenge is a lack of reliable experimental standards that provide a direct and systematic map evaluation or empirical map parameterization. Isotope labeling provides a unique opportunity to examine the frequency map prediction isolated from coupling to other unlabeled amide I vibrations. However, preparing isotope-edited protein standards falls into a dilemma between synthetic difficulty and reliability of underlying protein conformation to

decouple uncertainty from structural variations.[24] While isotope-labeled short peptides can be obtained using standard peptide synthesis approach, these peptides are usually conformationally disordered. For example in Chapter 7, even one of the well-studied peptides in IR spectroscopy, Ala–Ala–Ala, has been shown its conformational heterogeneity, and current protein force fields cannot predict the conformational distribution quantitatively.[57-58] On the other hand, larger proteins usually have well-folded conformations but synthesis with site-specific isotope-labeling is almost prohibitive. Early empirical map developed by Skinner and co-workers can achieve qualitative agreement, but it is still limited to reproduce the qualitative spectral lineshape.[49]

To develop experimental standards, the first step in our group was to perform an extensive IR measurements of dipeptide variants at various pH as a library for frequency predictions.[4] The advantage of this approach is that each dipeptide has only a single amide I vibration, without the need to account for coupling effects. Empirically, the map parameterization was done by extensively computing the electrostatic variables and finding the correlation between these variables and the experimental peak frequency shown in Fig. 3.4. It was found the electric field evaluated on the carbonyl oxygen atom along the C=O bond axis has the strongest correlation to the experimental peak frequency, and the corresponding one-site field map (1F) can achieve ~3 cm^{-1} frequency error.

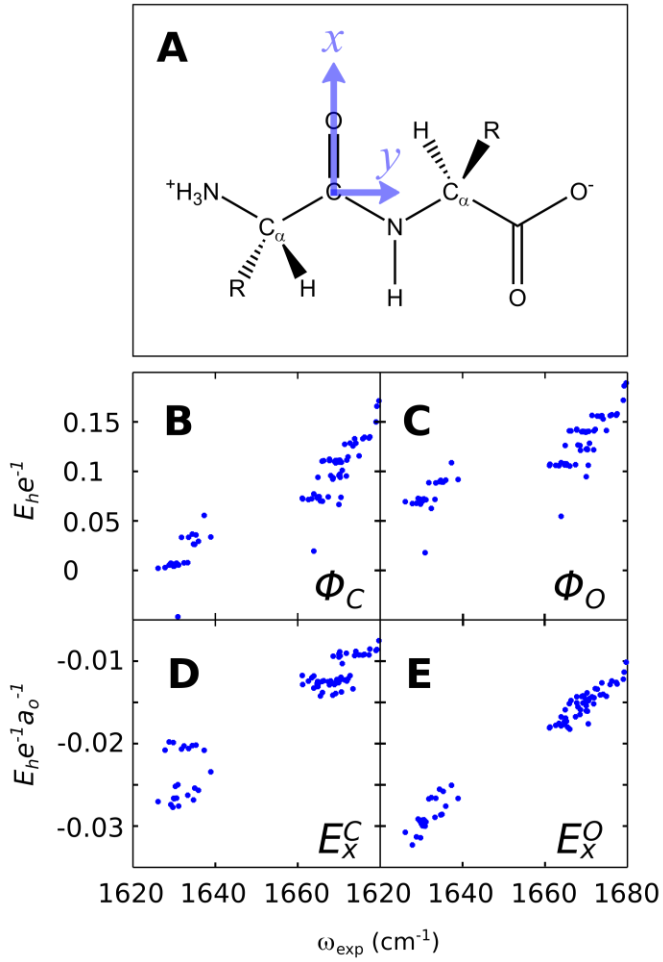

Figure 3.4: (A) Structure of a generic dipeptide; our coordinate system is defined so that the *x*-axis points along the amide C=O bond and the *y*-axis is in the plane of the amide unit. (B – E) Scatter plots of experimental peak frequencies for 23 standard dipeptides with individual electrostatic variables evaluated from 5 ns CHARMM27 MD simulations (see labels in figure). Taken from Figure 2 of Ref. 4. Copyright 2013 AIP Publishing

Additionally, our group used isotope-enriched protein expression to produce residue-specific isotope-edited NuG2b protein, which is extremely structurally stable to mitigate the issue of conformational disorder present in short peptides and prediction of conformational distribution from force fields.[24] From the direct evaluation of maps against experiments, the empirical one-site field (1F) map parameterized against dipeptide set fails to qualitatively describe some of the

isotope labeled spectra even though the map has a strong correlation between field along C=O axis and experimental peak frequency within the dipeptide set. It is found to require representative sampling of N–H electrostatics to obtain qualitative agreement since hydrogen bonding around N–H can also affect amide I frequency.[34] In addition, the labeled peak frequencies depend highly on the force field charges employed in the spectroscopic simulations since electrostatics from the force field charges are optimized for MD simulations but not for spectroscopic simulations,[4, 10] which creates issues of map transferability across force fields and quantitative prediction on amide I spectroscopy. Our group later developed an empirically optimized four-site potential (4P) map against isotope-edited NuG2b spectra, which calculates frequency using potential value evaluated at the C, O, N, and D atoms.[59] The map quantitatively described the isotope-labeled spectral features (Fig. 3.5) and achieved ~2 cm^{-1} frequency error against dipeptide spectra.

Coupling maps used throughout the thesis are based on the transition charge coupling map and the nearest neighbor coupling map (Fig. 3.3) developed by Jansen *et al*,[25] which are constructed based on quantum calculations of GLDP at the DFT level. Transition dipole moments are taken from the zero-field component from Jansen *et al*.[25] A summary of the maps used in this thesis is shown in Table 3.1.

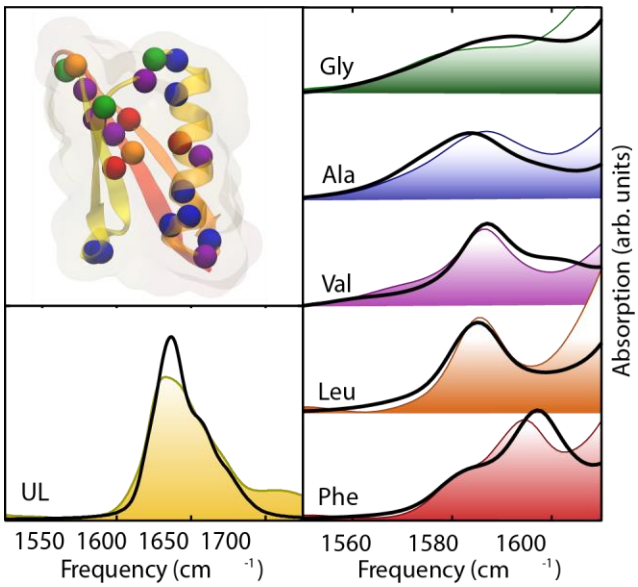

Figure 3.5: Comparison of the frequencies, line widths, and intensities of experimental spectra with simulated spectra starting from MD simulations of the crystal structure of the protein G mutant NuG2b.[24] Taken from Figure 8 of Ref. 19. Copyright 2016 Annual Reviews

Map	**Site Energy ω_a**		**Coupling J_{ab}**		**Transition Dipole Model**
	Electrostatic Sites	Electrostatic Components	Through-Space Coupling	NNC Dihedral Map	
DC15[4, 59]	O	E_x	Jansen TCC[25]	Jansen	Jansen Zero-Field[25]
4PN-150[59]	C, O, N, H	$\Phi_C, \Phi_O, \Phi_N, \Phi_H$	Jansen TCC[25]	Jansen	Jansen Zero-Field[25]

Table 3.1: Spectroscopic Maps used in this thesis.

3.5 Calculating IR Spectrum

So far, we have introduced spectroscopic maps useful for computing the quantities in amide I Hamiltonian from MD simulations, including site frequency, coupling, and transition dipole moments. In this section, numerical methods of simulating an IR spectrum will be presented given the amide I Hamiltonian and response function formalism in Chapter 2.

3.5.1 Numerical Wavefunction Propagation

In Subsection 2.5.2, we applied the MQC model to obtain the response functions in a classical bath that can be readily used in spectral simulations from the Eqns. (2.173)–(2.179). The simulation approach here is to numerically construct discrete time evolution operators based on the time-dependent amide I Hamiltonian, and numerically propagate an initial amide I excitation as a function of time. Once the response functions are calculated, the spectra can be obtained through Fourier transform along the appropriate time interval. In the literature, because this method is formally equivalent to solving a time-dependent Schrödinger Equation using the time-dependent Hamiltonian, this approach is frequency referred to as numerical integration of Schrödinger Equation (NISE). However, the term numerical wavefunction propagation is used since the response function can be formally interpreted as propagating dipole-moment weighted wavefunction and eventually calculating the overlap between vibrational wavepackets.

To demonstrate the physical meaning of this, we start with the linear response:

$$R^{(1)}(\tau_1) = \Theta(\tau_1)\left(\frac{2\pi}{3\hbar}\right)\sum_{i=1}^{3}\sum_{a,b}\mathrm{Im}\left\langle \hat{m}_i^{0b}(\tau_1)\hat{U}_{ba}(\tau_1,0)\hat{m}_i^{a0}(0)\right\rangle_{\mathrm{MD}} \tag{3.32}$$

We can define an effective dipole-moment weighted one-quantum wavefunction

$$\left|\psi_i(t_0,\tau)\right\rangle = \hat{U}(t_0+\tau,t_0)\hat{m}_i(t_0)\left|0\right\rangle \tag{3.33}$$

with its associated vector in the dual space

$$\left\langle\psi_i(t_0,\tau)\right| = \left\langle 0\right|\hat{m}_i(t_0)\hat{U}^{\dagger}(t_0+\tau,t_0) \tag{3.34}$$

An effective two-quantum wavefunction can be defined in a similar way

$$\begin{aligned}\left|\psi_{ij}(t_0,\tau,\Delta t)\right\rangle &= \hat{U}(t_0+\tau+\Delta t,t_0+\tau)\hat{m}_j(t_0+\tau)\left|\psi_i(t_0,\tau)\right\rangle \\ &= \hat{U}(t_0+\tau+\Delta t,t_0+\tau)\hat{m}_j(t_0+\tau)\hat{U}(t_0+\tau,t_0)\hat{m}_i(t_0)\left|0\right\rangle\end{aligned} \tag{3.35}$$

with its associated vector in the dual space

$$\begin{aligned}\left\langle \psi_{ij}\left(t_0,\tau,\Delta t\right)\right| &= \left\langle \psi_i\left(t_0,\tau\right)\right|\hat{m}_j\left(t_0+\tau\right)\hat{U}^{\dagger}\left(t_0+\tau+\Delta t,t_0+\tau\right) \\ &= \left\langle 0\right|\hat{m}_i\left(t_0\right)\hat{U}^{\dagger}\left(t_0+\tau,t_0\right)\hat{m}_j\left(t_0+\tau\right)\hat{U}^{\dagger}\left(t_0+\tau+\Delta t,t_0+\tau\right)\end{aligned} \tag{3.36}$$

The effective one-quantum wavefunction in the Eqns. (3.33)–(3.34) represents the time-dependent wavepacket of the system, which is a superposition of the exciton states, evolves for the duration of τ after the initial excitation from the ground state $\left|0\right\rangle$ at t_0. Then, the linear response function in the Eqn. (3.32) can be simply written as

$$R^{(1)}\left(\tau_1\right) = \Theta\left(\tau_1\right)\left(\frac{2\pi}{3\hbar}\right)\mathrm{Im}\sum_{i=1}^{3}\left\langle\left\langle \psi_i\left(\tau_1,0\right)\middle|\psi_i\left(0,\tau_1\right)\right\rangle\right\rangle_{\mathrm{MD}} \tag{3.37}$$

This equation can be interpreted as the overlap between effective wavefunctions. Similarly, for third-order response, the rephasing (R) and nonrephasing (NR) pathways can be computed via

$$\begin{aligned}R^{(\mathrm{R})}_{IJKL}\left(\tau_1,\tau_2,\tau_3\right) = \Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\Theta\left(\tau_3\right)\frac{2}{\hbar^3}\sum_{ijkl}T^{IJKL}_{ijkl}\,\mathrm{Im}\Big\{ \\ R^{(\mathrm{III})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)+R^{(\mathrm{II})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)-R^{(\mathrm{I})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)\Big\}\end{aligned} \tag{3.38}$$

$$\begin{aligned}R^{(\mathrm{NR})}_{IJKL}\left(\tau_1,\tau_2,\tau_3\right) = \Theta\left(\tau_1\right)\Theta\left(\tau_2\right)\Theta\left(\tau_3\right)\frac{2}{\hbar^3}\sum_{ijkl}T^{IJKL}_{ijkl}\,\mathrm{Im}\Big\{ \\ R^{(\mathrm{IV})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)+R^{(\mathrm{I})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)-R^{(\mathrm{II})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right)\Big\}\end{aligned} \tag{3.39}$$

$$\begin{aligned}R^{(\mathrm{I})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right) = \sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}^{0a}_i\left(0\right)\hat{U}^{\dagger}_{aa'}\left(\tau_1+\tau_2+\tau_3,0\right)\hat{m}^{a'b}_l\left(\tau_1+\tau_2+\tau_3\right) \\ \times\hat{U}_{bb'}\left(\tau_1+\tau_2+\tau_3,\tau_1+\tau_2\right)\hat{m}^{b'c}_k\left(\tau_1+\tau_2\right)\hat{U}_{cc'}\left(\tau_1+\tau_2,\tau_1\right)\hat{m}^{c'0}_j\left(\tau_1\right)\Big\rangle_{\mathrm{MD}}\end{aligned} \tag{3.40}$$

$$\begin{aligned}R^{(\mathrm{II})}_{ijkl}\left(\tau_1,\tau_2,\tau_3\right) = \sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}^{0a}_i\left(0\right)\hat{U}^{\dagger}_{aa'}\left(\tau_1+\tau_2,0\right)\hat{m}^{a'b}_k\left(\tau_1+\tau_2\right)\hat{U}^{\dagger}_{bb'}\left(\tau_1+\tau_2,\tau_1+\tau_2+\tau_3\right) \\ \times\hat{m}^{b'c}_l\left(\tau_1+\tau_2+\tau_3\right)\hat{U}_{cc'}\left(\tau_1+\tau_2+\tau_3,\tau_1\right)\hat{m}^{c'0}_j\left(\tau_1\right)\Big\rangle_{\mathrm{MD}}\end{aligned} \tag{3.41}$$

$$R_{ijkl}^{(\mathrm{III})}(\tau_1,\tau_2,\tau_3)=\sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}_i^{0a}(0)\hat{U}_{aa'}^{\dagger}(\tau_1,0)\hat{m}_j^{a'b}(\tau_1)\hat{U}_{bb'}^{\dagger}(\tau_1,\tau_1+\tau_2+\tau_3) \times\hat{m}_l^{b'c}(\tau_1+\tau_2+\tau_3)\hat{U}_{cc'}(\tau_1+\tau_2+\tau_3,\tau_1+\tau_2)\hat{m}_k^{c'0}(\tau_1+\tau_2)\Big\rangle_{\mathrm{MD}} \tag{3.42}$$

$$R_{ijkl}^{(\mathrm{IV})}(\tau_1,\tau_2,\tau_3)=\sum_{aa'bb'cc'}^{\mathrm{RWA}}\Big\langle \hat{m}_i^{0a}(0)\hat{U}_{aa'}^{\dagger}(\tau_1,0)\hat{m}_j^{a'b}(\tau_1)\hat{U}_{bb'}^{\dagger}(\tau_1,\tau_1+\tau_2) \times\hat{m}_k^{b'c}(\tau_1+\tau_2)\hat{U}_{cc'}^{\dagger}(\tau_1+\tau_2,\tau_1+\tau_2+\tau_3)\hat{m}_l^{c'0}(\tau_1+\tau_2+\tau_3)\Big\rangle_{\mathrm{MD}} \tag{3.43}$$

In the wavefunction representation with the Eqns. (3.33)–(3.36), the rephasing response function can be formulated as

$$\begin{aligned}R_{IJKL}^{(\mathrm{R})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T_{ijkl}^{IJKL}\,\mathrm{Im}\{ &\\ +\Big\langle\big\langle\psi_i(0,\tau_1)\big|\psi_j(\tau_1,0)\big\rangle\big\langle\psi_l(\tau_1+\tau_2+\tau_3,0)\big|\psi_k(\tau_1+\tau_2,\tau_3)\big\rangle\Big\rangle_{\mathrm{MD}} &\\ +\Big\langle\big\langle\psi_i(0,\tau_1+\tau_2)\big|\psi_k(\tau_1+\tau_2,0)\big\rangle\big\langle\psi_l(\tau_1+\tau_2+\tau_3,0)\big|\psi_j(\tau_1,\tau_2+\tau_3)\big\rangle\Big\rangle_{\mathrm{MD}} &\\ -\Big\langle\big\langle\psi_{il}(0,\tau_1+\tau_2+\tau_3,0)\big|\psi_{jk}(\tau_1,\tau_2,\tau_3)\big\rangle\Big\rangle_{\mathrm{MD}}\}&\end{aligned} \tag{3.44}$$

The nonrephasing response function can similarly be written as

$$\begin{aligned}R_{IJKL}^{(\mathrm{NR})}(\tau_1,\tau_2,\tau_3)=\Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3)\frac{2}{\hbar^3}\sum_{ijkl}T_{ijkl}^{IJKL}\,\mathrm{Im}\{ &\\ +\Big\langle\big\langle\psi_i(0,\tau_1)\big|\psi_j(\tau_1,0)\big\rangle\big\langle\psi_k(\tau_1+\tau_2,\tau_3)\big|\psi_l(\tau_1+\tau_2+\tau_3,0)\big\rangle\Big\rangle_{\mathrm{MD}} &\\ +\Big\langle\big\langle\psi_i(0,\tau_1+\tau_2+\tau_3)\big|\psi_l(\tau_1+\tau_2+\tau_3,0)\big\rangle\big\langle\psi_k(\tau_1+\tau_2,0)\big|\psi_j(\tau_1,\tau_2)\big\rangle\Big\rangle_{\mathrm{MD}} &\\ -\Big\langle\big\langle\psi_{ik}(0,\tau_1+\tau_2,\tau_3)\big|\psi_{jl}(\tau_1,\tau_2+\tau_3,0)\big\rangle\Big\rangle_{\mathrm{MD}}\}&\end{aligned} \tag{3.45}$$

Note that the time evolution of the ground state has been dropped since the ground sate energy/frequency is set to zero in the amide I Hamiltonian. The equations (3.37), (3.44), and (3.45) are the essential equations for computing the linear response and third-order response using

numerical wavefunction progration. The computation requires propagating the one-quantum wavefunction and the two quantum wavefuction through the appropriate time interval.

In practice, the key steps are generating the time evolution operator $\hat{U}$, and propagating the wavefunctions. Recall that the time evoluation operating under the time-dependent amide I Hamiltonian is

$$\hat{U}\left(t_0+\tau,t_0\right)=\exp_+\left(-\frac{i}{\hbar}\int_{t_0}^{t_0+\tau}\hat{H}_{\mathrm{AmI}}\left(t\right)dt\right) \tag{3.46}$$

Assuming the time step is small enough, one can obtain the discretized form:

$$\hat{U}\left(t_0+\tau,t_0\right)\approx\prod_{n=0}^{N_{\mathrm{frames}}}\exp\left(-\frac{i}{\hbar}\Delta t\hat{H}_{\mathrm{AmI}}\left(t_0+n\Delta t\right)\right) \tag{3.47}$$

with number of frames $N_{\mathrm{frames}}=\tau/\Delta t$. Note that number of frames usually can be pre-determined, and it is related to the scan time in g_spec, which is our home-build spectral simulation program.[60] Also note that the propagation in the spectroscopic simulations can only be done in the diabatic states (local mode basis) since the exciton states change in time as well. Formally, solving the short-time propagator requires the diagonalization of the amide I Hamiltonian:

$$\hat{H}_{\mathrm{AmI}}\left(\tau\right)=C\left(\tau\right)\Lambda\left(\tau\right)C^{-1}\left(\tau\right) \tag{3.48}$$

where $\Lambda(\tau)$ is the eigenvalue matrix at time τ and $C(\tau)$ is the eigenvector matrix in the local mode basis at time τ

$$C\left(\tau\right)=\left[\left|e_1\left(\tau\right)\right\rangle\quad\left|e_2\left(\tau\right)\right\rangle\quad\cdots\quad\left|e_n\left(\tau\right)\right\rangle\right] \tag{3.49}$$

with the eigenvector

$$\left|e_\alpha(\tau)\right\rangle = \sum_{a=1}^{n}\left\langle a\middle|e_\alpha(\tau)\right\rangle\left|a\right\rangle \equiv \sum_{a=1}^{n} C_a^\alpha(\tau)\left|a\right\rangle = \begin{bmatrix} C_1^\alpha(\tau) \\ C_2^\alpha(\tau) \\ \vdots \\ C_n^\alpha(\tau) \end{bmatrix} \tag{3.50}$$

The short-time propagator can then be calculated using the matrix exponential method:

$$e^{-\frac{i}{\hbar}\Delta t\hat{H}_{\mathrm{AmI}}(t_0+n\Delta t)} = C(t_0+n\Delta t)\,e^{-\frac{i}{\hbar}\Lambda(t_0+n\Delta t)\Delta t}\,C^{-1}(t_0+n\Delta t) \tag{3.51}$$

Generally speaking, this form in the Eqn. (3.51) propagates all states, including ground state, one-quantum states, and two-quantum states in the response functions. Because of the block diagonal approximation introduced in the Eqn. (3.6), one can separately propagate one-quantum wavefunctions and two-quantum wavefunctions in the Eqns. (3.44)–(3.45) with the one-quantum Amide I Hamiltonian and the two-quantum Hamiltonian, respectively. In other words,

$$\hat{U}(t_0+\tau,t_0)\left|0\right\rangle = e^{-\frac{i}{\hbar}\tau\cdot 0}\left|0\right\rangle = \left|0\right\rangle \tag{3.52}$$

$$\begin{aligned}\left|\psi_i(t_0,\tau)\right\rangle &= \hat{U}^{1\mathrm{Q}}(t_0+\tau,t_0)\,\hat{m}_i^{10}(t_0)\left|0\right\rangle \\ &= \prod_{n=0}^{N_{\mathrm{frames}}}\exp\left(-\frac{i}{\hbar}\Delta t\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}}(t_0+n\Delta t)\right)\hat{m}_i^{10}(t_0)\left|0\right\rangle\end{aligned} \tag{3.53}$$

$$\begin{aligned}\left|\psi_{ij}(t_0,\tau,\Delta t)\right\rangle &= \hat{U}^{2\mathrm{Q}}(t_0+\tau+\Delta t,t_0+\tau)\,\hat{m}_j^{21}(t_0+\tau)\left|\psi_i(t_0,\tau)\right\rangle \\ &= \prod_{n=0}^{N_{\mathrm{frames}}}\exp\left(-\frac{i}{\hbar}\Delta t\hat{H}_{\mathrm{AmI}}^{2Q}(t_0+\tau+n\Delta t)\right)\hat{m}_j^{21}(t_0+\tau)\left|\psi_i(t_0,\tau)\right\rangle\end{aligned} \tag{3.54}$$

with the superscript 10 and 21 indicating the 0–1 and 1–2 transition of the wavefunction.

3.5.2 Computational Procedure for the Numerical Wavefunction Propagation

The workflow of computing linear response functions given the amide I Hamiltonian trajectory, $\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}}(\tau)$, and transition dipole moment trajectory $\hat{\boldsymbol{m}}^{10}(\tau)$ can be summarized below:

1. Diagonalize the one-quantum amide I Hamiltonian and construct the matrix *C* using the equations (3.49)–(3.50).
2. Construct the propagator using the Eqn. (3.51).
3. Choose a starting frame in the trajectory, and create the initial excitation of the one-quantum wavefunction $\hat{m}_i^{10}(t_0)|0\rangle$, mediated by the transition dipole operator in the Eqn. (3.53).
4. Loop through τ_1 and *i*, and propagate the one-quantum wavefunction using the Eqn. (3.53) given $\tau=\tau_1$.
5. At each τ_1, create another one-quantum wavefunction $\langle\psi_i(\tau_1,0)| = \langle 0|\hat{m}_i^{01}(\tau_1)$, and compute the response function in the Eqn. (3.37).
6. Once τ_1 reaches the scan time or the number of frames have been reached, perform Fourier transform along τ_1 to obtain the linear spectrum.
7. Choose the next starting frame, and repeat the steps 1 to 6. Average the resulting IR spectra at different starting frames.

The workflow of computing third-order response functions is more complicated due to the additional time variable τ_3 and vector components *i*, *j*, *k*, and *l*. But it can be simplified by recognizing that *i*, *j*, *k*, and *l* are dummy vector component variables when computing the wavefunction overlaps in the Eqns. (3.44) and (3.45). In practice, only two independent variable needs to be specified: *a*, and *b*. Also, one additional variable *p* needs to be specified to enumerate

all possible situations of the orientational average tensor T_{ijkl}^{IJKL} summarized in Table 3.2. Nested loops are required with the time variables τ_1 and τ_3, and the vector component variables *a*, *b*, and *p*. Note that τ_2 is an input variable since a typical 2D spectrum will be acquired by fixing the waiting time (τ_2), as well as *I*, *J*, *K*, and *L*.

p	0	1	2	3
ijkl	*aaaa*	*aabb*	*abba*	*abab*

Table 3.2: All possible enumerations of the tensorial components

Given the amide I Hamiltonian trajectory, $\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}}(\tau)$, and transition dipole moment trajectory, $\hat{\boldsymbol{m}}^{10}(\tau)$, the computational procedure of the third-order response functions can be summarized as follows:

1. Construct the two-quantum Hamiltonian trajectory and the two-quantum transition dipole moments using the Eqns. (3.7), (3.8) and (3.30).
2. Diagonalize $\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}}(\tau)$ and $\hat{H}_{\mathrm{AmI}}^{2\mathrm{Q}}(\tau)$ to construct the matrices $C^{1\mathrm{Q}}$ and $C^{2\mathrm{Q}}$ using the equations (3.49)–(3.50).
3. Choosing a starting frame, create the initial one-quantum wavefunction $\left|\psi_a(t_0,0)\right\rangle = \hat{m}_a^{10}(t_0)\left|0\right\rangle$, adapted from the Eqn. (3.53).
4. First, loop through τ_1 to propagate $\left|\psi_a(t_0,\tau_1)\right\rangle$ and $\left|\psi_a(t_0,\tau_1+\tau_2)\right\rangle$ using the Eqn. (3.53).
5. At each τ_1, create the one-quantum wavefunctions $\left|\psi_a(t_0+\tau_1,0)\right\rangle$, $\left|\psi_a(t_0+\tau_1,\tau_2)\right\rangle$, and $\left|\psi_a(t_0+\tau_1+\tau_2,0)\right\rangle$ using the Eqn. (3.53).

6. At each τ_1, create the two-quantum wavefunctions $\left|\psi_{ab}\left(t_0,\tau_1+\tau_2,0\right)\right\rangle$ and $\left|\psi_{ab}\left(t_0+\tau_1,\tau_2,0\right)\right\rangle$ using the Eqn. (3.54).
7. Loop through τ_3. At each τ_3, propagate the one-quantum and two-quantum wavefunctions $\left|\psi_{a}\left(t_0+\tau_1,\tau_2+\tau_3\right)\right\rangle$, $\left|\psi_{a}\left(t_0+\tau_1+\tau_2,\tau_3\right)\right\rangle$, $\left|\psi_{ab}\left(t_0,\tau_1+\tau_2,\tau_3\right)\right\rangle$, and $\left|\psi_{ab}\left(t_0+\tau_1,\tau_2,\tau_3\right)\right\rangle$ using the Eqn. (3.53)–(3.54).
8. At each τ_3, create the wavefunctions $\left|\psi_{a}\left(t_0,\tau_1+\tau_2+\tau_3\right)\right\rangle$, $\left|\psi_{ab}\left(t_0,\tau_1+\tau_2+\tau_3,0\right)\right\rangle$, $\left|\psi_{ab}\left(t_0+\tau_1,\tau_2+\tau_3,0\right)\right\rangle$ using the Eqns. (3.53)–(3.54).
9. Loop through *a*, *b*, and *p* to compute the rephasing and nonrephasing response function at each combination of τ_1 and τ_3 using the Eqns. (3.44)–(3.45).
10. Once τ_1, τ_3 reaches the scan time, perform two-dimensional Fourier transform on both τ_1 and τ_3 to obtain the rephasing and nonrephasing 2D spectra at a given polarization *I*, *J*, *K*, and *L*.
11. Move to the next starting frame, and repeat the computation from steps 4 to 10. Average the resulting 2D IR spectra at different starting frames.

3.5.3 Suzuki-Trotter Expansion

The bottleneck of using numerical wavefunction propagation comes from the diagonalization of the amide I Hamiltonian, which is $O(n^3)$ in linear IR simulations, and $O(n^6)$ in 2D IR simulations where n is the number of amide I oscillators. Another tremendous contribution to the computation of response functions is propagation of the wavefunction in time. Computationally, it is equivalent to the matrix-vector multiplication, resulting in $O(n^2)$ in linear IR simulations and $O(n^4)$ in 2D IR simulations. Practically, it is prohibitive to simulate a

peptide/protein of $n \geq 50$,[61] which is effectively the size of insulin monomer ($n = 49$). It is necessary to find an approximate but accurate enough method to reduce the scaling and consequently the computational cost so that it becomes viable for simulating large proteins like insulin dimer ($n = 98$).

The split operator technique has been well-developed and applied on numerical simulations on quantum systems with the benefit of avoiding the expensive diagonalization to reduce the scaling and still preserving the unitary property of the operator.[62-64] For the application to numerical wavefunction propagations, we summarize the Suzuki-Trotter expansion scheme proposed by Jansen and co-workers.[65]

For a time-dependent Amide I Hamiltonian, it can be split into two contributions

$$\hat{H}_{\mathrm{AmI}}(\tau) = \hat{A}(\tau) + \hat{B}(t) \tag{3.55}$$

where $\hat{A}$ is a diagonal matrix in the local mode basis

$$\hat{A}(t) = \sum_{a} \omega_a(t) |a\rangle\langle a| \tag{3.56}$$

and $\hat{B}$ is the off-diagonal coupling matrix that is in essence much smaller than $\hat{A}$

$$\begin{aligned} B(t) &= \sum_{a \neq b} B_{ab}(t) \\ B_{ab}(t) &= J_{ab}(t) |a\rangle\langle b| + J_{ba}(t) |b\rangle\langle a| \end{aligned} \tag{3.57}$$

The corresponding discretized propagator in the Eqn. (3.53) can be expanded as

$$\begin{aligned} \hat{U}(t_0 + \Delta t, t_0) &= \exp\left(-\frac{i}{\hbar} \hat{H}_{\mathrm{AmI}}(t_0) \Delta t \right) \\ &= \exp\left(-\frac{1}{2} i\hat{A}(t_0) \Delta t \right) \exp\left(-i\hat{B}(t_0) \Delta t \right) \exp\left(-\frac{1}{2} i\hat{A}(t_0) \Delta t \right) + O(t^3) \end{aligned} \tag{3.58}$$

The truncation is valid when

$$\left\|B(t)\right\|\Delta t \ll \hbar \sim 5.3\times10^{3}\ \left(\mathrm{cm}^{-1}\cdot\mathrm{fs}\right) \tag{3.59}$$

Typically, the extreme coupling value in real simulations is around 10 cm^{-1}, and the time step is set to 20 fs such that the inequality in the Eqn. (3.59) is hold. This decomposition has the benefit of avoiding matrix diagonalization. Since $\hat{A}$ is diagonal in the local mode basis, the corresponding propagator is also diagonal:

$$\exp\left(-\frac{1}{2}iA(t)\Delta t\right)=\sum_{a}\exp\left(-\frac{1}{2}\omega_{a}(t)\Delta t\right)|a\rangle\langle a| \tag{3.60}$$

For the propagator under the off-diagonal coupling matrix $\hat{B}$, it has the analytical solution:

$$\exp\left(-iB(t)\Delta t\right)=\prod_{a\neq b}\exp\left(-iB_{ab}(t)\Delta t\right) \tag{3.61}$$

$$\begin{aligned}\exp\left(-iB_{ab}(t)\Delta t\right)=&\sum_{c\neq a,b}1|c\rangle\langle c|+\cos\left(J_{ab}(t)\Delta t\right)|a\rangle\langle a|+\cos\left(J_{ab}(t)\Delta t\right)|b\rangle\langle b|\\&-i\sin\left(J_{ab}(t)\Delta t\right)|a\rangle\langle b|-i\sin\left(J_{ab}(t)\Delta t\right)|b\rangle\langle a|\\=&\begin{bmatrix}1&0&\cdots&0&0&\cdots&0\\0&1&\cdots&0&0&\cdots&0\\\vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\cos\left(J_{ab}(t)\Delta t\right)&-i\sin\left(J_{ab}(t)\Delta t\right)&\cdots&0\\0&0&\cdots&-i\sin\left(J_{ab}(t)\Delta t\right)&\cos\left(J_{ab}(t)\Delta t\right)&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&0&0&\cdots&1\end{bmatrix}\end{aligned} \tag{3.62}$$

Using Suzuki-Trotter expansion can significantly speed up the calculation, which can be examined through scaling analysis. For linear spectroscopy, the bottleneck includes diagonalizing the one-quantum Hamiltonian, which is $O(n^3)$, and the propagation of wavefunction, which is $O(n^2)$. Applying Suzuki-Trotter expansion requires generating $\hat{A}$ and $B_{ij}^{\wedge}$, which have n diagonal

elements and $n(n\text{-}1)/2$ pairs, resulting in $O(n)+O(n^2)$ instead of $O(n^3)$ for matrix diagonalization. Propagation using Suzuki-Trotter expansion requires $O(n(n\text{-}1)/2) * O(1) \sim O(n^2)$, which is the same as in the original matrix exponential method. Applying Suzuki-Trotter Expansion to linear response function reduces the scaling from $O(n^3)$ to $O(n^2)$.

For computing third-order response functions, however, the reduction of the scaling becomes significant. Since there are $n(n+1)/2$ two-quantum modes and $n^2(n\text{-}1)/2$ non-zero coupling pairs, diagonalizing the two-quantum Hamiltonian has the scaling of $O(n^6)$, and the propagation of wavefunction scales with $O(n^4)$. Generating $\hat{A}$ and $\hat{B_{ij}}$ from the two-quantum Hamiltonian has the scaling of $O(n^2)+O(n^2(n\text{-}1)/2) \sim O(n^3)$ instead of $O(n^6)$. Propagation using Suzuki-Trotter expansion scales with $O(n^2(n\text{-}1)/2)) * O(1) \sim O(n^3)$ instead of $O(n^4)$. In total, Suzuki-Trotter Expansion to third-order response function calculations reduces the scaling from $O(n^6)$ to $O(n^3)$.

Investigation of the scaling on a single thread is shown in Fig. 3.6. Overall, using Suzuki-Trotter expansion significantly speeds up the calculation, and makes 2D simulation viable for larger systems up to ~100 oscillators within a day. However, the scaling of generating the two-quantum Hamiltonian is still above $O(n^4)$, which needs further investigation. To illustrate the feasibility to perform 2D spectral simulations, Fig. 3.7 shows 2D IR spectral simulations of insulin dimer using structures drawn from the Markov State Model (MSM) presented in Chapter 8, which has $n = 98$ oscillators, and a single trajectory with 1 ns in length can be finished in 2 days.

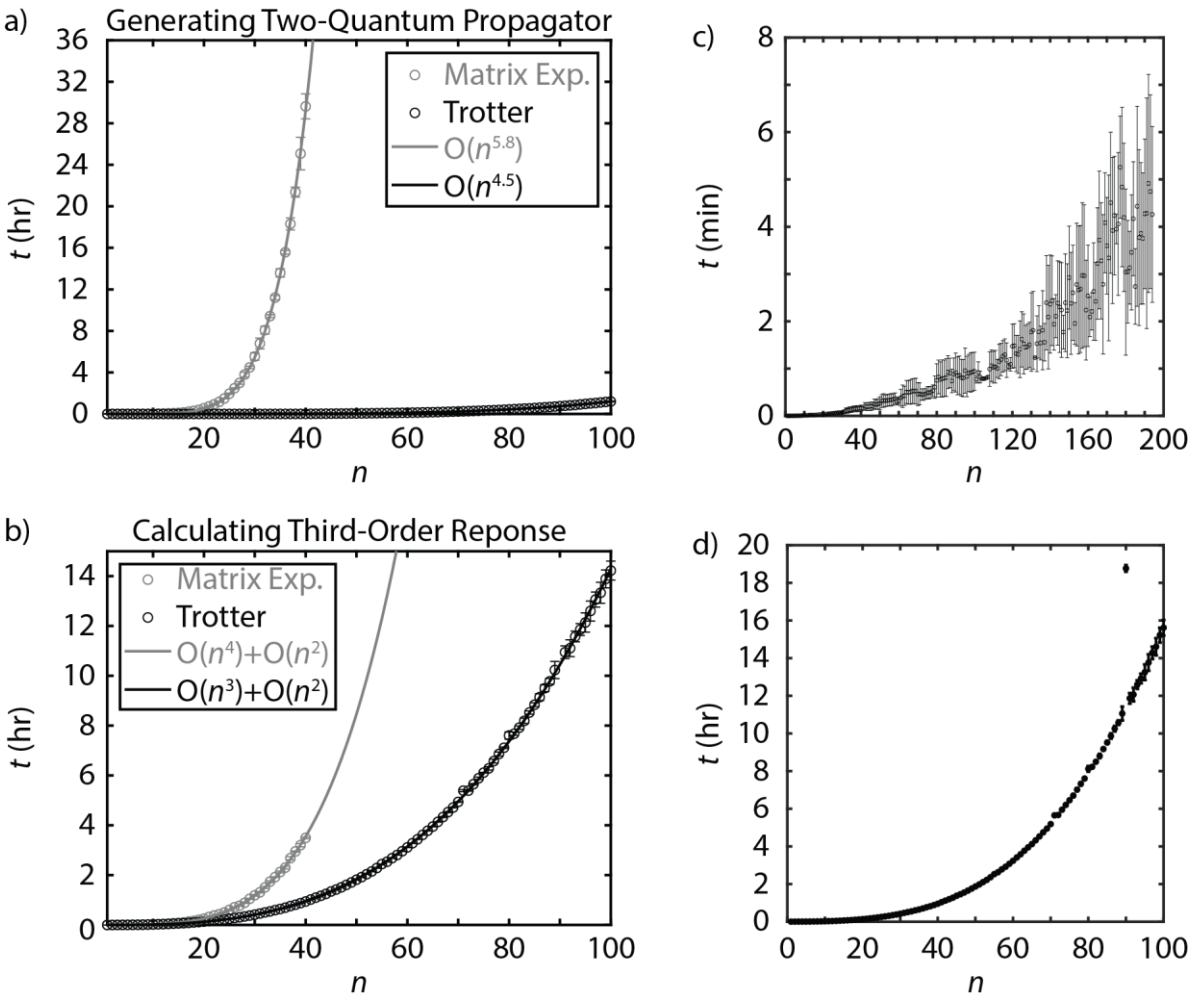

Figure 3.6: Scaling analysis of numerical wavefunction propagation method. (a) Generating the two-quantum propagator using matrix exponential method (gray) and Suzuki-Trotter expansion (black). (b) Calculating the third-order response functions (c) Total simulation time of the linear IR spectra as a function of number of amide I oscillators *n*. (d) Total simulation time of the 2D IR spectra as a function of *n*. The simulations were set up using the following parameters: Δt = 20 fs, scan time: 2.5 ps, 123 realizations for a 250 ps trajectory.

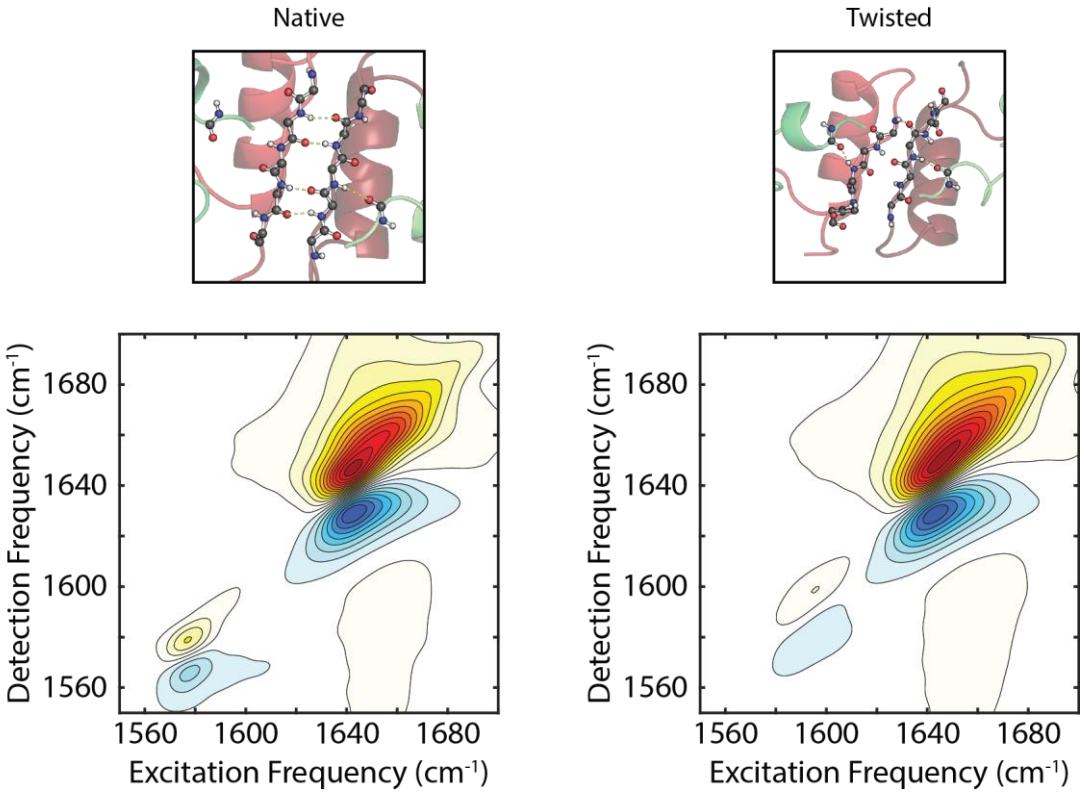

Figure 3.7: 2D IR spectral simulation of insulin dimer (*n* = 98). Left: Isotope-labeled ZZZZ-polarized 2D spectrum of the native dimer in the Markov State Model (MSM) presented in Chapter 8. (Right) Isotope-labeled ZZZZ-polarized 2D spectrum of the twisted dimer in the MSM.

3.6 Acknowledgments

I thank Mike Earl Reppert for helpful discussions on spectroscopic maps and potential improvements on g_amide and g_spec. I also thank Luis Busto de Moner for providing feedback to improve the chapter.

3.7 References

1. Feng, Y.; Huang, J.; Kim, S.; Shim, J. H.; MacKerell, A. D., Jr.; Ge, N. H., Structure of Penta-Alanine Investigated by Two-Dimensional Infrared Spectroscopy and Molecular Dynamics Simulation. *J Phys Chem B* **2016,** *120* (24), 5325-39.

2. Best, R. B.; Buchete, N. V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophys J* **2008,** *95* (1), L07-9.
3. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D., Jr., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput* **2012,** *8* (9), 3257-3273.
4. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
5. Woutersen, S.; Hamm, P., Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *Journal of Physical Chemistry B* **2000,** *104* (47), 11316-11320.
6. Smith, A. W.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A.; Roy, S.; Jansen, T. L.; Knoester, J., Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J Phys Chem B* **2010,** *114* (34), 10913-24.
7. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. 2011.
8. Woys, A. M.; Almeida, A. M.; Wang, L.; Chiu, C. C.; McGovern, M.; de Pablo, J. J.; Skinner, J. L.; Gellman, S. H.; Zanni, M. T., Parallel beta-sheet vibrational couplings revealed by 2D IR spectroscopy of an isotopically labeled macrocycle: quantitative benchmark for the interpretation of amyloid and protein infrared spectra. *J Am Chem Soc* **2012,** *134* (46), 19118-28.
9. Buchanan, L. E.; Dunkelberger, E. B.; Tran, H. Q.; Cheng, P. N.; Chiu, C. C.; Cao, P.; Raleigh, D. P.; de Pablo, J. J.; Nowick, J. S.; Zanni, M. T., Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. *Proc Natl Acad Sci U S A* **2013,** *110* (48), 19285-90.
10. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
11. Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D.; Skinner, J. L.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016,** *353* (6303), 1040-1044.
12. Lomont, J. P.; Ostrander, J. S.; Ho, J. J.; Petti, M. K.; Zanni, M. T., Not All beta-Sheets Are the Same: Amyloid Infrared Spectra, Transition Dipole Strengths, and Couplings Investigated by 2D IR Spectroscopy. *J Phys Chem B* **2017,** *121* (38), 8935-8945.
13. Cheatum, C. M.; Tokmakoff, A.; Knoester, J., Signatures of beta-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J Chem Phys* **2004,** *120* (17), 8201-15.
14. Sengupta, N.; Maekawa, H.; Zhuang, W.; Toniolo, C.; Mukamel, S.; Tobias, D. J.; Ge, N. H., Sensitivity of 2D IR spectra to peptide helicity: a concerted experimental and simulation study of an octapeptide. *J Phys Chem B* **2009,** *113* (35), 12037-49.
15. Wang, L.; Middleton, C. T.; Singh, S.; Reddy, A. S.; Woys, A. M.; Strasfeld, D. B.; Marek, P.; Raleigh, D. P.; de Pablo, J. J.; Zanni, M. T.; Skinner, J. L., 2DIR spectroscopy of human amylin fibrils reflects stable beta-sheet structure. *J Am Chem Soc* **2011,** *133* (40), 16062-71.
16. Mukherjee, P.; Kass, I.; Arkin, I. T.; Zanni, M. T., Structural disorder of the CD3zeta transmembrane domain studied with 2D IR spectroscopy and molecular dynamics simulations. *J Phys Chem B* **2006,** *110* (48), 24740-9.

17. Stevenson, P.; Gotz, C.; Baiz, C. R.; Akerboom, J.; Tokmakoff, A.; Vaziri, A., Visualizing KcsA conformational changes upon ion binding by infrared spectroscopy and atomistic modeling. *J Phys Chem B* **2015,** *119* (18), 5824-31.
18. Kratochvil, H. T.; Maj, M.; Matulef, K.; Annen, A. W.; Ostmeyer, J.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Probing the Effects of Gating on the Ion Occupancy of the K(+) Channel Selectivity Filter Using Two-Dimensional Infrared Spectroscopy. *J Am Chem Soc* **2017,** *139* (26), 8837-8845.
19. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
20. Baiz, C. R.; Blasiak, B.; Bredenbeck, J.; Cho, M.; Choi, J. H.; Corcelli, S. A.; Dijkstra, A. G.; Feng, C. J.; Garrett-Roe, S.; Ge, N. H.; Hanson-Heine, M. W. D.; Hirst, J. D.; Jansen, T. L. C.; Kwac, K.; Kubarych, K. J.; Londergan, C. H.; Maekawa, H.; Reppert, M.; Saito, S.; Roy, S.; Skinner, J. L.; Stock, G.; Straub, J. E.; Thielges, M. C.; Tominaga, K.; Tokmakoff, A.; Torii, H.; Wang, L.; Webb, L. J.; Zanni, M. T., Vibrational Spectroscopic Map, Vibrational Spectroscopy, and Intermolecular Interaction. *Chem Rev* **2020,** *120* (15), 7152-7218.
21. Ganim, Z.; Chung, H. S.; Smith, A. W.; Deflores, L. P.; Jones, K. C.; Tokmakoff, A., Amide I two-dimensional infrared spectroscopy of proteins. *Acc Chem Res* **2008,** *41* (3), 432-41.
22. Ghosh, A.; Ostrander, J. S.; Zanni, M. T., Watching Proteins Wiggle: Mapping Structures with Two-Dimensional Infrared Spectroscopy. *Chem Rev* **2017,** *117* (16), 10726-10759.
23. Baiz, C. R.; Reppert, M.; Tokmakoff, A., An Introduction to Protein 2D IR Spectroscopy. In *Ultrafast Infrared Vibrational Spectroscopy*, 2013; p 361.
24. Reppert, M.; Roy, A. R.; Tokmakoff, A., Isotope-enriched protein standards for computational amide I spectroscopy. *J Chem Phys* **2015,** *142* (12), 125104.
25. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
26. Gaigeot, M. P., Theoretical spectroscopy of floppy peptides at room temperature. A DFTMD perspective: gas and aqueous phase. *Phys Chem Chem Phys* **2010,** *12* (14), 3336-59.
27. Ridley, J.; Zerner, M., An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines. *Theoretica Chimica Acta* **1973,** *32* (2), 111-134.
28. Jeon, J.; Cho, M., Direct quantum mechanical/molecular mechanical simulations of two-dimensional vibrational responses:N-methylacetamide in water. *New Journal of Physics* **2010,** *12* (6), 065001.
29. Reppert, M.; Brumer, P., Classical coherent two-dimensional vibrational spectroscopy. *J Chem Phys* **2018,** *148* (6), 064101.
30. Miyazawa, T., Perturbation Treatment of the Characteristic Vibrations of Polypeptide Chains in Various Configurations. *The Journal of Chemical Physics* **1960,** *32* (6), 1647-1652.
31. Miyazawa, T.; Blout, E. R., The Infrared Spectra of Polypeptides in Various Conformations: Amide I and II Bands1. *Journal of the American Chemical Society* **1961,** *83* (3), 712-719.
32. Torii, H.; Tasumi, M., Model calculations on the amide-I infrared bands of globular proteins. *The Journal of Chemical Physics* **1992,** *96* (5), 3379-3387.
33. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
34. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.

35. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
36. Reppert, M. Computational Amide I Spectroscopy from the ground up : building and benchmarking new tools to study disordered peptide ensembles. Massachusetts Institute of Technology, Massachusetts Institute of Technology, 2016.
37. Hamm, P.; Lim, M. H.; Hochstrasser, R. M., Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *Journal of Physical Chemistry B* **1998,** *102* (31), 6123-6138.
38. Gruenbaum, S. M.; Tainter, C. J.; Shi, L.; Ni, Y.; Skinner, J. L., Robustness of Frequency, Transition Dipole, and Coupling Maps for Water Vibrational Spectroscopy. *J Chem Theory Comput* **2013,** *9* (7), 3109-17.
39. Yu, Q.; Carpenter, W. B.; Lewis, N. H. C.; Tokmakoff, A.; Bowman, J. M., High-Level VSCF/VCI Calculations Decode the Vibrational Spectrum of the Aqueous Proton. *J Phys Chem B* **2019,** *123* (33), 7214-7224.
40. Choi, J.-H.; Ham, S.; Cho, M., Local Amide I Mode Frequencies and Coupling Constants in Polypeptides. *The Journal of Physical Chemistry B* **2003,** *107* (34), 9132-9138.
41. Blasiak, B.; Cho, M., Vibrational solvatochromism. II. A first-principle theory of solvation-induced vibrational frequency shift based on effective fragment potential method. *J Chem Phys* **2014,** *140* (16), 164107.
42. Blasiak, B.; Cho, M., Vibrational solvatochromism. III. Rigorous treatment of the dispersion interaction contribution. *J Chem Phys* **2015,** *143* (16), 164111.
43. Gorbunov, R. D.; Kosov, D. S.; Stock, G., Ab initio-based exciton model of amide I vibrations in peptides: definition, conformational dependence, and transferability. *J Chem Phys* **2005,** *122* (22), 224904.
44. Hamm, P.; Woutersen, S., Coupling of the Amide I Modes of the Glycine Dipeptide. *Bulletin of the Chemical Society of Japan* **2002,** *75* (5), 985-988.
45. Maekawa, H.; De Poli, M.; Moretto, A.; Toniolo, C.; Ge, N. H., Toward detecting the formation of a single helical turn by 2D IR cross peaks between the amide-I and -II modes. *J Phys Chem B* **2009,** *113* (34), 11775-86.
46. Karjalainen, E. L.; Ersmark, T.; Barth, A., Optimization of model parameters for describing the amide I spectrum of a large set of proteins. *J Phys Chem B* **2012,** *116* (16), 4831-42.
47. Torii, H.; Tasumi, M., Infrared intensities of vibrational modes of an α-helical polypeptide: Calculations based on the equilibrium charge/charge flux (ECCF) model. *Journal of Molecular Structure* **1993,** *300*, 171-179.
48. Cheam, T. C.; Krimm, S., Infrared intensities of amide modes in N-methylacetamide and poly(glycine I) from ab initio calculations of dipole moment derivatives of N-methylacetamide. *The Journal of Chemical Physics* **1985,** *82* (4), 1631-1641.
49. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.
50. Ham, S.; Cho, M., Amide I modes in the N-methylacetamide dimer and glycine dipeptide analog: Diagonal force constants. *The Journal of Chemical Physics* **2003,** *118* (15), 6915-6922.
51. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.

52. Schmidt, J. R.; Corcelli, S. A.; Skinner, J. L., Ultrafast vibrational spectroscopy of water and aqueous N-methylacetamide: Comparison of different electronic structure/molecular dynamics approaches. *J Chem Phys* **2004,** *121* (18), 8887-96.
53. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
54. Watson, T. M.; Hirst *, J. D., Theoretical studies of the amide I vibrational frequencies of [Leu]-enkephalin. *Molecular Physics* **2005,** *103* (11-12), 1531-1546.
55. Bondarenko, A. S.; Jansen, T. L., Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins. *J Chem Phys* **2015,** *142* (21), 212437.
56. Cunha, A. V.; Bondarenko, A. S.; Jansen, T. L., Assessing Spectral Simulation Protocols for the Amide I Band of Proteins. *J Chem Theory Comput* **2016,** *12* (8), 3982-92.
57. Woutersen, S.; Pfister, R.; Hamm, P.; Mu, Y.; Kosov, D. S.; Stock, G., Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations. *The Journal of Chemical Physics* **2002,** *117* (14), 6833-6840.
58. Feng, C. J.; Dhayalan, B.; Tokmakoff, A., Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophys J* **2018,** *114* (12), 2820-2832.
59. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
60. Reppert, M.; Feng, C.-J., g_spec. **2017**.
61. Paarmann, A.; Hayashi, T.; Mukamel, S.; Miller, R. J., Nonlinear response of vibrational excitons: simulating the two-dimensional infrared spectrum of liquid water. *J Chem Phys* **2009,** *130* (20), 204110.
62. Suzuki, M., Generalized Trotter's formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems. *Communications in Mathematical Physics* **1976,** *51* (2), 183-190.
63. Kosloff, R., Time-dependent quantum-mechanical methods for molecular dynamics. *The Journal of Physical Chemistry* **1988,** *92* (8), 2087-2100.
64. Suzuki, M., General Decomposition Theory of Ordered Exponentials. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences* **1993,** *69* (7), 161-166.
65. Liang, C.; Jansen, T. L., An Efficient N(3)-Scaling Propagation Scheme for Simulating Two-Dimensional Infrared and Visible Spectra. *J Chem Theory Comput* **2012,** *8* (5), 1706-13.
66. Reppert, M., g_amide. **2017**.
67. Auer, B. M.; Skinner, J. L., Dynamical effects in line shapes for coupled chromophores: time-averaging approximation. *J Chem Phys* **2007,** *127* (10), 104105.
68. Torii, H., Effects of intermolecular vibrational coupling and liquid dynamics on the polarized Raman and two-dimensional infrared spectral profiles of liquid N,N-dimethylformamide analyzed with a time-domain computational method. *J Phys Chem A* **2006,** *110* (14), 4822-32.
69. Jansen, T.; Knoester, J., Nonadiabatic effects in the two-dimensional infrared spectra of peptides: application to alanine dipeptide. *J Phys Chem B* **2006,** *110* (45), 22910-6.

Appendix 3A: Mapping Structures into Amide I Parameters: g_amide

Given an MD trajectory generated from a MD engine like GROMACS, one can utilize amide I spectroscopic maps to translate a protein structural trajectory into a trajectory of time-dependent amide I Hamiltonian and transition dipole moments. One of the programs performing such translation is a home-built program, g_amide.[66] g_amide identifies the topology of a given protein including atomic identity, connectivity (chemical bond) between atoms, position of each atom, and the corresponding atomic partial charge. In addition, g_amide reads supplied spectroscopic maps that convert the topology and collective structural and electrostatic variables into resulting spectroscopic variables such as amide I vibrational frequency of each amide group, coupling between these vibrations, and transition dipole moments. These spectroscopic variables will be treated as input parameters for spectral simulations described in Appendix 3B.

Below we provided descriptions of g_amide including required input files, and corresponding command-line arguments. For simplicity, we denote a variable called *A* as $[*A*].

Example command for running g_amide:

***g_amide -s* $[*tpr*] *-f* $[*xtc*] *-mapfile* $[*map*]**

Required inputs:

-s* $[*tpr*]**: Binary topology file (tpr file) from a GROMACS MD simulation. This tpr file provides essential parameters such as identify of each atom, connectivity between atoms, and atomic partial charges. The default file name is ***topol.tpr if nothing is specified. Note that only tpr files generated from GROMACS 4.6.7 and below can be used in g_amide.

-f* $[*xtc*]**: Binary trajectory file (xtc file) from a GROMACS MD simulation. The xtc file only contains coordinate trajectories. The default file name is ***traj.xtc if nothing is specified.

Note that other structural trajectories are also available for g_amide, including trr file and pdb file. However, most of the time, xtc will be recommended due to its minimal storage requirement compared to other formats.

***-mapfile* $[*map*]**: Text-based map file specified as **$[*map*]**, containing all of the spectroscopic maps for translating a protein structure into amide I parameters, with details described later. No default option is available.

Descriptions about output files:

ham.txt: One-quantum Hamiltonian trajectory of the system. Given a trajectory of M frames of a single-chain protein with (N+1) residues, the resulting ham.txt consists of a M-by-N^2 matrix, with each row representing the Hamiltonian at that frame, and the corresponding Hamiltonian sized N-by-N has been reshaped into N^2 row vector. Order of the matrix elements in the Hamiltonian is column-major, meaning that the order will go through each column first and then move to the next column. For instance, given a system of two coupled oscillators (2-by-2 Hamiltonian at frame t) below in Eqn. (3.63),

$$H(t)=\begin{bmatrix} \omega_1(t) & J_{12}(t) \\ J_{21}(t) & \omega_2(t) \end{bmatrix} \tag{3.63}$$

the corresponding reshaped row vector will follow the order in Eqn. (3.64).

$$H'(t)=\begin{bmatrix} \omega_1(t) & J_{21}(t) & J_{12}(t) & \omega_2(t) \end{bmatrix} \tag{3.64}$$

dipx.txt, ***dipy.txt***, and ***dipz.txt***: These files store the x, y, and z component of transition dipole moments of each amide I vibration. Given a trajectory of M frames of a single-chain

protein with (N+1) residues, each file will consist of M-by-N matrix listing the corresponding x, y, and z component of the transition dipole moment of N amide I vibrations. For instance, the same coupled oscillator above will have the row vectors of each component at frame t as

$$m_i(t) = \begin{bmatrix} m_i^1(t) & m_i^2(t) \end{bmatrix}, i = x, y, z \tag{3.65}$$

sites.txt: This file only stores the amide I site frequency trajectory of each amide group in the system, equivalent with the diagonal frequency elements as in Eqn. (3.63) and in ***ham.txt***. This file will be useful when swapping frequencies while controlling coupling values the same.

info.txt: This file contains information about the composition of amide I vibrations, including the number of amide I vibrations, and the residue composition of each amide I vibration. For instance, the ***info.txt*** of Ala–Ala–Ala gives

BONDS: 2

ALA 1 ALA 2

ALA 2 ALA 3

The first line indicates the total number of amide I vibrations or amide bonds. As described above, the (N+1)-residue peptide/protein results in N amide bonds or amide I vibrations. The following lines indicate the composition of each amide group/vibration as identified in the (binary) topology file. This file is particular useful when performing isotope-labeled spectral simulations, in particular identifying the residue to be isotopically substituted.

Additional optional command line arguments:

***-chargefile* $[*charge*]**: A flag specifying customized atomic partial charge files in the simulation box. By default, g_amide utilized charges provided from tpr file. This flag allows user to perform studies with different electrostatic setup, including testing the effect of switching off electrostatic interactions from water/specific residue/*etc*, using optimized charge map that gives rise to better quantitative agreement such as Glycine-corrected CHARMM27 charge map,[59] and swapping atomic partial charges between two force fields to test the error coming from electrostatic effects.[24] By default, it is not specified.

***-promapfile* $[*promap*]**: A flag specifying proline spectroscopic map. By default, it is not specified.

-outname* $[*outname*]**: This flag specifies a file name prefix to be added to all the output files described above, including ***ham.txt, ***dipx.txt***, ***dipy.txt***, ***dipz.txt***, ***sites.txt***, and ***info.txt***. For example, the resulting ham.txt will be named as **$[*outname*]*_ham.txt***, or **$[*outname*]*/ham.txt*** when **$[*outname*]** is a directory specified as "$[dir]/". **$[*outname*]** can be used to specify both directory and the prefix. By default, it is not specified.

***-nt* $[*nthreads*]**: Number of threads used for mapping. By default, it is 1.

-[no]print_elec: Print electrostatic values or not. By default, it is –noprint_elec, which does not print electrostatics.

-[no]print_angles: Print dihedral angle values of each amide group or not. By default, it is –noprint_angles, which does not print ϕ and ψ dihedral angles.

-cutoff: Cutoff distance for electrostatic calculations. By default, it is 1000 nm.

***-osc* $[*osc_strength*]**: Oscillator strength in unit of Debye^2. By default, it is -1, which means using the values provided from **$[*map*]**. If provided with a positive, this value will overwrite the value in **$[*map*]** and used for normalizing transition dipole moments.

-chunk: Range for selecting atoms of the system to be included for electrostatic frequency calculations. The format is [a-b;…;c-d]. Atomic indexing begins at zero. This command can be useful when including some parts of protein/water molecules for electrostatic calculations such as nearby atoms of an amide group within 5 Å. For instance, to include the first 10 atoms for electrostatic calculations, the resulting flag will be

–chunk [0-9]

If additional atoms from atom 15 to 20 are included as well, the corresponding flag will look like

–chunk [0-9;15-20]

Appendix 3B: Calculating IR Spectra: g_spec

With trajectories of time-dependent amide I Hamiltonian and transition dipole moments, one can calculate the corresponding amide I spectra. The calculation can be achieved utilizing the home-build program g_spec.[60] g_spec calculates FTIR spectra and non-linear 2D IR spectra using one of the three schemes: Static averaging, time-averaging approximation (TAA),[67] and numerical integration of Schrödinger equation (NISE).[68-69]

Below we provided descriptions of g_spec including required input files, and corresponding command-line arguments. For simplicity, we denote a variable called *A* as $[*A*]. Example command for running g_spec:

***g_spec -deffnm* $[*deffnm*] *-outname* $[*outname*] *-tstep* $[*tstep*] - *tscan* $[*tscan*]**

Command line arguments:

-deffnm* $[*deffnm*]**: This flag sets a prefix to all of the input files. For instance, the input Hamiltonian file will be $[deffnm]_ham.txt or $[deffnm]/ham.txt if $[deffnm] is a directory specified as "$[dir]/". Two exceptions are the optional flags ***-sites and ***-shift*** described below. By default, it is not specified.

***-outname* $[*outname*]**: This flag set a prefix to all of the output files just as in g_amide. Please see -outname flag in g_amide for more detail.

***-tstep* $[*tstep*]**: This flag specifies time step of the spectral simulation in unit of fs. To properly determine the time step, one should take Nyquist frequency, undersampling, and effects of motional narrowing into account. In practice, 20 fs is reasonable for accounting for motional narrowing in amide I spectroscopy. By default, it is specified as 20 (fs).

-tscan* $[*tscan*]**: This flag specifies the scan time for spectral simulations in unit of fs. In TAA scheme, this sets total length of the averaging window, and time-dependent Hamiltonians within the window will be averaged equally or with the weight specified by ***–whann flag. The averaging accounts for motional narrowing in TAA schemes, and the optimal scan time in TAA scheme can be determined to be ~5/Γ by comparing with the exact NISE results, in which Γ is the full width at half maximum (FWHM) of the spectral lineshape. In practice, several hundreds of fs may be in a good range for amide I spectra.

In NISE scheme, it sets the scan time for each time period when calculating the response functions. For instance, ***-tscan*** 2500 with ***-tstep*** 20 in FTIR simulations will correspond to scan from starting frame of each realization to 2.5 ps, effectively 2500/20+1=126 frames. In 2D IR simulations, this will correspond to scan 126 frames both in τ_1 and τ_3. This flag also sets the true frequency resolution in NISE scheme, determined by $\Delta\omega = \frac{1}{c \cdot T_{\text{scan}}}$, in which T_{scan} is the scan time and $\Delta\omega$ is the frequency resolution. For example, scan time of 2.5 ps sets the resolution of 13.2 cm^{-1}.

T_{scan} is intimately related to the computational time. In FTIR simulations, the computation time scales with $O(T_{\text{scan}})$ whereas computational time in 2D IR simulations scales with $O(T_{\text{scan}}^2)$. In practice, one would like to reach a balance between computational time and accuracy of the simulations. A rule of thumb is to set the scan time to be 5 times longer than the population lifetime described below. However, sometimes 2.5 times longer is also applicable.

-taup **$[*taup*]**: ad hoc one-quantum population lifetime in unit of fs. Default value is 1300 fs measured for N-methylacetamide (NMA) in D_2O.[37] The population lifetime is used to construct a single exponential decay convolved with response functions, effectively Lorentzian convolution in frequency domain with the FWHM of $1/c \cdot \tau_p$.

-ham **$[*ham*]**: This optional flag sets a custom file for the input one-quantum Hamiltonian trajectory described previously in g_amide.

-dipx **$[*dipx*]**, ***-dipy*** **$[*dipy*]**, ***-dipz*** **$[*dipz*]**: These optional flags set custom files for the input x,y, and z components of transition dipole moment trajectory of each amide I vibration, described previously in g_amide.

-sites* $[*sites*]**: This optional flag specifies a file of site energy trajectory that will replace the diagonal frequency of the input Hamiltonian trajectory. Note that the ***-deffnm flag will not affect the supplied site energy file name in this flag.

***-shift* $[*shift*]**: This optional flag species a text file containing frequency shifts to be applied to individual sites. Index starts at 0. The first line should contain the keyword SHIFT, followed by the total number of sites to be shifted. The remaining lines should specify the index of the sites and the corresponding frequency shift in cm^{-1}. An example file is as follows.

SHIFT 4

43 -65

44 -65

92 -65

93 -65

This file describes a set of frequency shifts, including 4 amide I sites on site 43, 44, 92, and 93. Each site is frequency shifted by -65 cm^{-1}.

-nise: This flag specifies the use of NISE method to calculate spectra. By default, it is false.

-trotter: Trotter approximation on one-quantum and two-quantum propagators. All pairs of off-diagonal couplings are used as perturbations to factorize the amide I Hamiltonian into analytically solvable 2-by-2 Hamiltonians, and to avoid diagonalization. Given the number of amide I vibrations as N, it reduces the scaling of FTIR simulations from $O(N^3)$ to $O(N^2)$, and the scaling of 2D IR simulations from $O(N^6)$ to $O(N^3)$.[65] This flag is incompatible with

-pert described below, and Trotter expansion always performs more accurate and faster than perturbative correction.

-wstart **$[*wstart*]**: Start frequency for output spectra and frequency axis in cm^{-1}. By default, it is 1500 cm^{-1}.

-wstop **$[*wstop*]**: End frequency for output spectra and frequency axis in cm^{-1}. By default, it is 1800 cm^{-1}.

-winterp **$[*winterp*]**: Frequency spacing for output spectra and frequency axis in cm^{-1}. In static averaging and TAA schemes, this sets the spacing between frequency bins such that it is the true frequency resolution. In NISE scheme, it determines the length of zero padding, which is effectively interpolation frequency. Instead, the true frequency resolution is set by **$[*tscan*]** specified in the ***-tscan*** flag. By default, it is 1 cm^{-1}.

-whann: This flag specifies additional Hann window in time domain for windowing purpose. In NISE scheme, this windows the response functions before Fourier transform, while in TAA scheme, this reweights the averaged Hamiltonian.

-nt **$[*nthreads*]**: Number of threads used for parallel calculations of g_spec. Currently the parallelization is not ideal for 2D IR calculations. It is advisory to split long trajectories into many sub-trajectories to perform spectral simulations and to recombine the results afterwards. By default, it is 1.

-skip **$[*skip*]**: The flag specifies frame spacing between starting frames for response function calculations. By default, it is set to 1, meaning each frame will be used as the starting frame for spectral simulations. Larger **$[*skip*]** gives less averaged spectra while

speeding up the spectral simulations. Usually a value of 10 is a good balance between computational time and accuracy of the simulations.

-dump* $[*dump*]**: Output time interval for FTIR and 2D IR simulations in unit of ps. By default, it is -1, indicating that spectra will be written only at the end of the trajectory. Positive value will create additional time stamp file, and spectral trajectory files with the suffix _traj. For instance, FTIR spectral trajectory file and time stamp file will be named as **$[*deffnm*]**_ftir_traj.txt, and **$[*deffnm*]**_tstamp.txt, respectively. When turning on ***-2dir flag described below, additional 2D IR spectral trajectory files will be written.

Additional command line arguments specific to 2D IR simulations:

-2dir: This flag turns on 2D IR spectral simulations. By default, it is false.

***-ham2Q* $[*ham2Q*]**: This optional flag specifies the file name of two-quantum Hamiltonian trajectory, which will overwrite the file name specified by ***-deffnm* $[*deffnm*]**. This flag is particularly useful when the weakly anharmonic oscillator model breaks down such as strongly anharmonic systems or systems with inverted anharmonicity. Given the number of amide groups as *N*, the two-quantum states are numbered in order of 11, 21, 22, 31, 32, 33, …, *NN* such that the resulting two-quantum Hamiltonian H_{2Q} is constructed as

$$H_{2Q}(t)=\begin{bmatrix} \omega_{1,1}(t) & J_{11,21}(t) & \cdots & J_{11,N(N-1)}(t) & J_{11,NN}(t) \\ J_{21,11}(t) & \omega_{1,2}(t) & & & J_{21,NN}(t) \\ \vdots & & \omega_{2,2}(t) & & \vdots \\ J_{N(N-1),11}(t) & & & \ddots & J_{N(N-1),NN}(t) \\ J_{NN,11}(t) & J_{NN,21}(t) & \cdots & J_{NN,N(N-1)}(t) & \omega_{N,N}(t) \end{bmatrix} \tag{3.66}$$

The H_{2Q} is reshaped into a row vector in $[ham2Q] at each frame (row-major) as

$$H_{2Q}^{'}(t)=\begin{bmatrix}\omega_{1,1}(t) & J_{11,21}(t) & \cdots & \omega_{N,N}(t)\end{bmatrix} \tag{3.67}$$

-dipx2Q* $[*dipx2Q*]**, ***-dipy2Q* $[*dipy2Q*]**, ***-dipz2Q* $[*dipz2Q*]**: These optional flags specify the file names of two-quantum transition dipole moment trajectory, which will overwrite the file name specified by ***-deffnm* $[*deffnm*]**. Ordering of the 2Q states follows the same way as defined in the ***-ham2Q flag.

$$m_i(t)=\begin{bmatrix}m_i^{1,1}(t) & m_i^{2,1}(t) & m_i^{2,2}(t) & \cdots & m_i^{N,N}(t)\end{bmatrix}, i=x,y,z \tag{3.68}$$

-[no]reph, -[no]nreph: These flags specify rephrasing spectra and non-rephasing spectra will be computed or not. By default, they are specified as -reph and -nreph, indicating that both rephrasing and non-rephasing spectra will be computed.

-[no]zzzz, -[no]zzyy, -[no]zyyz, -[no]zyzy: These flags indicate that ZZZZ-, ZZYY-, ZYYZ-, and ZYZY-polarized 2D IR spectra will be computed or not. By default, they are specified as ***-zzzz***, ***-zzyy***, ***-nozyyz***, ***-nozyzy***, meaning that ZZZZ- and ZZYY-polarized spectra will be computed.

***-taup2Q* $[*taup2Q*]**: ad hoc two-quantum population lifetime in unit of fs. Default value is 1300 fs. The population lifetime is used to construct a single exponential decay convolved with third-order response functions, specifically responses related to the two-quantum manifold.

***-tau2* $[*tau2*]**: Waiting time/Population time for NISE 2D IR calculations, in unit of fs. By default, it is 0 fs.

***-delta* $[*delta*]**: Diagonal anharmonicity in cm^{-1}, assuming a weakly anharmonic oscillator model.[19, 37] By default, it is 16 cm^{-1} extracted from N-methylacetamide (NMA) in D_2O.

-pert: This flag specifies the use of weakly anharmonic perturbative approximation to calculate 2D IR spectra. In static averaging and TAA schemes, this corresponds to a first-ordered correction on the harmonic two-quantum Hamiltonian with diagonal anharmonicity **$[*delta*]** specified in ***-delta*** flag. In NISE scheme, it corresponds to a first-ordered correction on the harmonic two-quantum propagator with diagonal anharmonicity, which reduces the computational scaling from $O(N^6)$ to $O(N^4)$. By default, it is false. Note that this flag is incompatible with ***-trotter*** flag, and Trotter expansion always performs more accurate and faster than perturbative correction.

Chapter 4

Experimental Methods of IR Spectroscopy and Sample Characterization of Insulin

4.1 Introduction

Insulin is notorious for irreversible fibril formation or protein aggregation, which can be activated and facilitated under destabilizing conditions such as low pH, high temperature, high ionic strength, character of ions, and agitation.[1-4] The instability of insulin hinders widely accessible pharmaceutical applications of insulin. Or perhaps insulin is one of the best protein models to study behaviors of fibril formations, in which it is required to have an early intermediate of the partially unfolded monomer exhibiting exposure of the hydrophobic residues close to the A-chain N-terminus.[5-7] However, the same aggregation issue persists for IR spectroscopy of insulin, in particular studying the dissociation dynamics of the homodimer, which usually requires thermal dissociation at high temperature and low pH to ensure workable solubility for spectroscopy as well as the spectroscopic benefit of removing COO^- vibration from $^{13}C^{18}O$-labeled amide I vibrations. Some preliminary efforts have been made to find a condition workable for IR spectroscopy of insulin.[8] Still, it is especially challenging but critically important to find an optimal condition that maximizes the stability of insulin while still being able to perform equilibrium characterization

and dimer dissociation dynamics using IR spectroscopy on precious samples such as isotope-edited human insulin.

This chapter summarizes recent progress on experimental characterization of human insulin that makes equilibrium and transient IR experiments feasible, in particular for isotope-edited insulin across the entire dissociation transition and characterizing monomeric state at high temperature. These are written with the aim of making it a good reference for people in the future to reproduce the experimental results of human insulin and to potentially save time for characterizing conditions for insulin, insulin mutant, or some other proteins/peptides. With this specific goal in mind, section 3.2 provides detailed descriptions of up-to-date practice of sample preparation in order to get quality spectra with minimal artifacts and the synthesis of isotope-edited human insulin. Section 3.3 describes dependence of human insulin across multiple experimental variables such as pH, ionic strength, and additional treatments to maximize the stability of human insulin. Section 3.4 contains a brief summary of experimental IR techniques used for human insulin and tricks such as fast T-ramp FTIR spectroscopy and undersampling technique, that speed up the data collection to minimize the time spent at destabilizing conditions.

4.2 Sample Preparation of Human Insulin for IR Spectroscopy

This section provides detailed descriptions of preparing human insulin samples for IR spectroscopy, including synthesis of $^{13}C^{18}O$ isotope-edited insulin adopted from Balamurugan Dhayalan's protocal, hydrogen-deuterium (H–D) exchange, TFA removal for isotope-edited samples, lyophilization, additional descriptions about preparation for the IR spectroscopy, and cleaning the window with protein aggregations. For more detailed synthesis of isotope-edited insulin, please refer to Ref. 9.

4.2.1 Synthesis of Site-Specific Isotope Labeled Samples

Unlabeled (UL) zinc-free human insulin can be purchased from Sigma-Aldrich (91077C) The potency of the sample is ≥ 27.5 insulin unit (IU)/mg. Each IU corresponds to 0.0347 mg of human insulin based on WHO, which translates the potency to ≥ 95.4% weight percentage purity. The impurity has ≤ 1.0 endotoxin unit/mL endotoxin. Site-specific $^{13}C^{18}O$ labeled insulin samples used throughout this thesis were synthesized using solid phase peptide synthesis (SPPS) with one-pot native sequential chemical ligation.[9] The strategy to make the insulin involves one-pot sequential native chemical ligation of three unprotected peptides to get the full-length ester-insulin polypeptide,[10] folding and saponification as the key steps. Ester insulin - a Glu^{A4}-Thr^{B30} side-chain linked single chain insulin molecule was used as the key intermediate in generating native insulin.[11]

Isotope labeled amino acid: 1-^{13}C-Phenylalanine was isotopically exchanged with ^{18}O-water in the presence of dry hydrochloric acid at 100 ^{o}C. The dried 1-$^{13}C^{18}O$-phenylalanine was then N^{α}-protected as tert-Butoxy carbonyl for the use in the peptide synthesis. Exact mass and isotopic enrichment (up to 97%-^{18}O) was assessed by ESI-mass spectrometry.

Peptide synthesis and native chemical ligation: All the peptides were synthesized by manual Boc chemistry SPPS using "in situ neutralization" protocols.[12] Three segments Phe^{B1}-Val^{B18}-$^{\alpha}$thioester, ester linked Gly^{A1}-Glu^{A4}[O$\beta$$Thr^{B30}$-$Cys(Thz)^{B19}$]-$Cys^{A6}$-$^{\alpha}$thioester and Cys^{A7}-Asn^{A21} were chosen as previously described.[13] The $^{13}C^{18}O$-labelled amino acids were incorporated at appropriate positions 1) Phe^{B24} and 2) Phe^{B24}/Phe^{B25} during the solid phase peptide synthesis. Peptides were then assembled sequentially from C-to-N direction by one-pot native chemical ligation strategy as shown in Fig. 4.1.

Folding/saponification: The full-length 51-mer ester linked polypeptide was subjected to folding reaction using a cysteine/cystine redox couple at pH 7.6 at a polypeptide concentration of 0.05 mg/mL. After the folding was complete as indicated by the early eluting peak in the HPLC, the ester insulin single chain precursor was purified and subjected to lithium hydroxide mediated saponification to form the native insulin that had an 18 Da mass increase indicative of addition of the elements of water. The homogeneity of the insulin molecule was confirmed by LC-MS analysis.

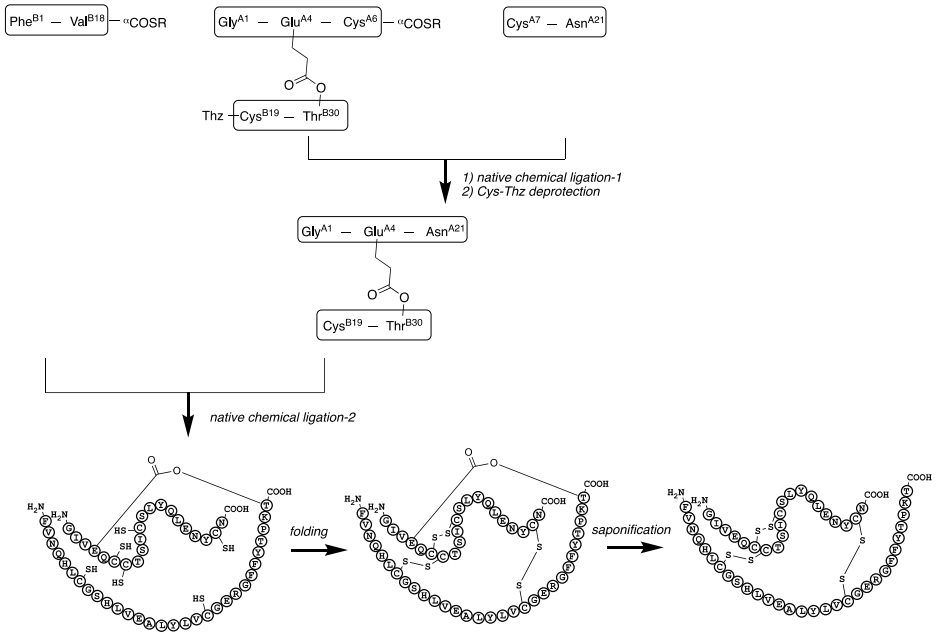

Figure 4.1: Outline of the synthesis of insulin through ester insulin.

4.2.2 Hydrogen-Deuterium Exchange and Lyophilization

To avoid spectral overlap between amide I vibration and water bend vibration around 1650 cm^{-1}, hydrogen-deuterium (H–D) exchange on the amide hydrogen atoms was performed by dissolving human insulin in 10 mM DCl/D_2O solution (pH* = 1.7). pH* is the pH reading of a

deuterated aqueous solution using a pH meter with the glass electrode. As a side note, the conversion between pH* of a D_2O solution measured from a pH meter and pD can be empirically determined as[14]

$$\mathrm{pD} = \mathrm{pH}^{*} + 0.41 \tag{4.1}$$

However, several different conversion formulae have also been proposed.[15] The pH* values are reported to avoid slight differences due to the conversion. Also note that the isoelectric point (pI) of human insulin is 5.3–5.35 based on Sigma-Aldrich's data sheet, so insulin is insoluble in pure D_2O or D_2O solution with the pH value ranging from 4.5 to 7 (Fig. 4.6). The concentration of insulin for H–D exchange was prepared at ~1 mg/mL and checked on a Nanodrop UV/Vis spectrometer (Thermo Scientific). One can monitor the concentration of insulin by looking at the absorption at either 276 nm or 280 nm, with the A280 mode that is a built-in feature in Nanodrop. The absorption peak at 276 nm is attributed to the tyrosine absorption with the corresponding extinction coefficient of 6.2×10^3 M^{-1} cm^{-1}.[16-17] Since the molecular weight of human insulin is 5808 g/mol, and the absorbance of the Nanodrop UV/Vis spectrometer is normalized to the path length of 1 cm (in A280 mode). The conversion from the optical density (OD) at 276 nm to mg/mL can be found as

$$C \text{ mg/mL} = \frac{5808}{6200} \times \mathrm{OD}(276 \text{ nm}) = 0.937 \times \mathrm{OD}(276 \text{ nm}) \tag{4.2}$$

The absorbance at 280 nm (A280) based on a general reference in Nanodrop is assumed to be

$$C \text{ mg/mL} = \mathrm{OD}(280 \text{ nm}) \tag{4.3}$$

It was found that concentration values determined based on absorption at 276 nm and at 280 nm are consistent with each other within our experimental precision.

Effective H–D exchange requires heating the sample due to buried amide hydrogen atoms not accessible in the dimeric state, and the dissociation temperature of 10 mg/mL human insulin in 10 mM DCl/D_2O solution is around 68 °C (Please see Table 5.2 in subsection 5.3.3 or subsection 9.4.1). To estimate how long H–D exchange has to take to be completed, an un-exchanged sample of 1 mg/mL human insulin in 10 mM DCl/D_2O solution was placed into the FTIR spectrometer at high temperature. The recirculating chiller and the brass sample holder were firstly equilibrated at the chiller temperature of 63.6 °C, which translates into the sample temperature of 60 °C. The sample cell at room temperature was placed into the brass sample holder and monitored the progress of H–D exchange using the FTIR spectrometer as soon as possible. The FTIR spectra are shown in Fig. 4.2(a), in which there are an absorbance decay at 1544 cm^{-1} and a slight red-shift of the amide I vibration around 1650 cm^{-1} on top of huge D_2O solvent background. To highlight these changes, the second derivatives were computed and shown in Fig. 4.2(b). The decay at 1544 cm^{-1} corresponds to the un-exchanged amide II vibrations,[18] which happens on the minute time scale and goes to completion on the order of 20 min, and the corresponding kinetic trace is plotted in Fig. 4.2(c). The kinetic trace appears to be a single exponential decay, and the fit yields a time constant of 3.4 min, meaning that the H–D exchange is completed after 20 min. Note that this condition is at low pH, H–D exchange at higher pH will slow down due to the lower concentration of $[H^+]$.

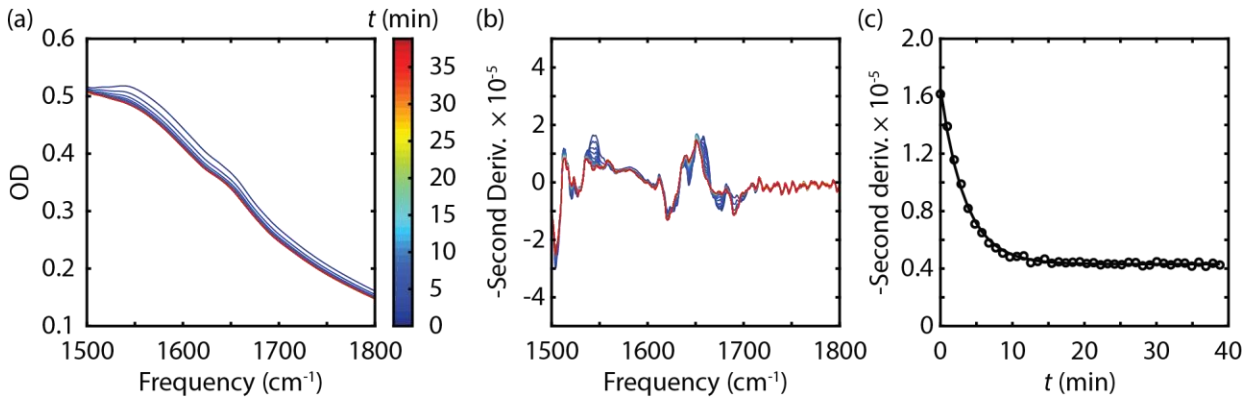

Figure 4.2: Monitoring the H–D exchange process from FTIR spectra. (a) Time-dependent FTIR spectra of 1 mg/mL insulin in 10 mM DCl. (b) Second derivative of the FTIR spectra to highlight the un-exchanged amide II vibrations around 1540 cm^{-1}. (c) Kinetic trace of the second derivative slice at 1544 cm^{-1}.

To sufficiently exchange amide hydrogen into deuterium, the sample was heated at the incubator temperature of 80 °C for 30 minutes in the incubator (Denville scientific Incublock), which will roughly translate into the sample temperature of 70 °C in a 1.5 mL plastic falcon tube. The heated solution was then lyophilized to remove excess DCl and D_2O. Usually the concentration of insulin for IR spectroscopy is ~1.7 mM or 10 mg/mL, meaning that the concentration for H–D exchange is the 10-fold dilution. Therefore, the heated solution was split into aliquots of 350 μL in falcon tubes, and then lyophilized overnight. Each tube would eventually be rehydrated into the total volume of 35 μL so that the final concentration would be around 1.7 mM.

4.2.3 Removal of Trifluoroacetid Acid

During chemical synthesis of isotope-edited human insulin, trifluoroacetic acid (TFA) is a common reactant to cleave sidechain protecting groups. Unfortunately, TFA is extremely bright in the 6 μm range, having a peak at 1672 cm^{-1} as the COO^- stretch, and (potentially) an additional

peak at ~1760 cm^{-1}, which is possibly associated with the C=O stretch from protonated COOH group at low pH (Fig. 4.3). The presence of TFA strongly interferes with amide I vibrational spectra of any protein and peptide so that the removal of TFA is critical for quality data and potentially interpretations of the data. The removal of TFA can be achieved by iteratively lyophilizing the protein sample three times or more at ~10 mM DCl.[19] When possible, isotope-edited human insulin samples were iteratively lyophilized with 10 mM DCl/D_2O solution. A quick check of TFA removal can be done by firstly dissolving insulin in pure D_2O, where residual TFA still makes human soluble.

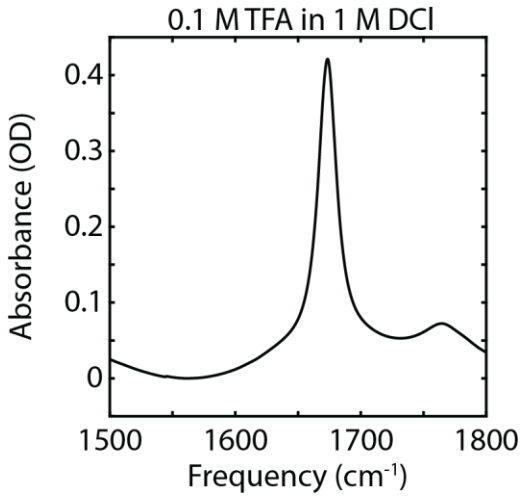

Figure 4.3: FTIR spectrum of 0.1 M in 1M DCl at room temperature.

As an unfortunate example of unsuccessful TFA removal, Fig. 4.4 shows a comparison between isotope-edited human insulin with and without residual TFA, where there is a subtle shoulder around 1670 cm^{-1} with residual TFA, and it becomes more apparent by looking at the difference spectrum between labeled insulin and unlabeled insulin. In particular, unlabeled insulin does not have any TFA so that it can be a good reference spectrum to be subtracted off. One can

clear see in the difference spectrum that has a clear peak at 1672 cm^{-1}, and a little peak around 1760 cm^{-1}, which is consistent with TFA at low pH shown in Fig. 4.3. As a side note, the additional peak at 1620 cm^{-1} in the high temperature spectra indicates irreversible insulin aggregation.

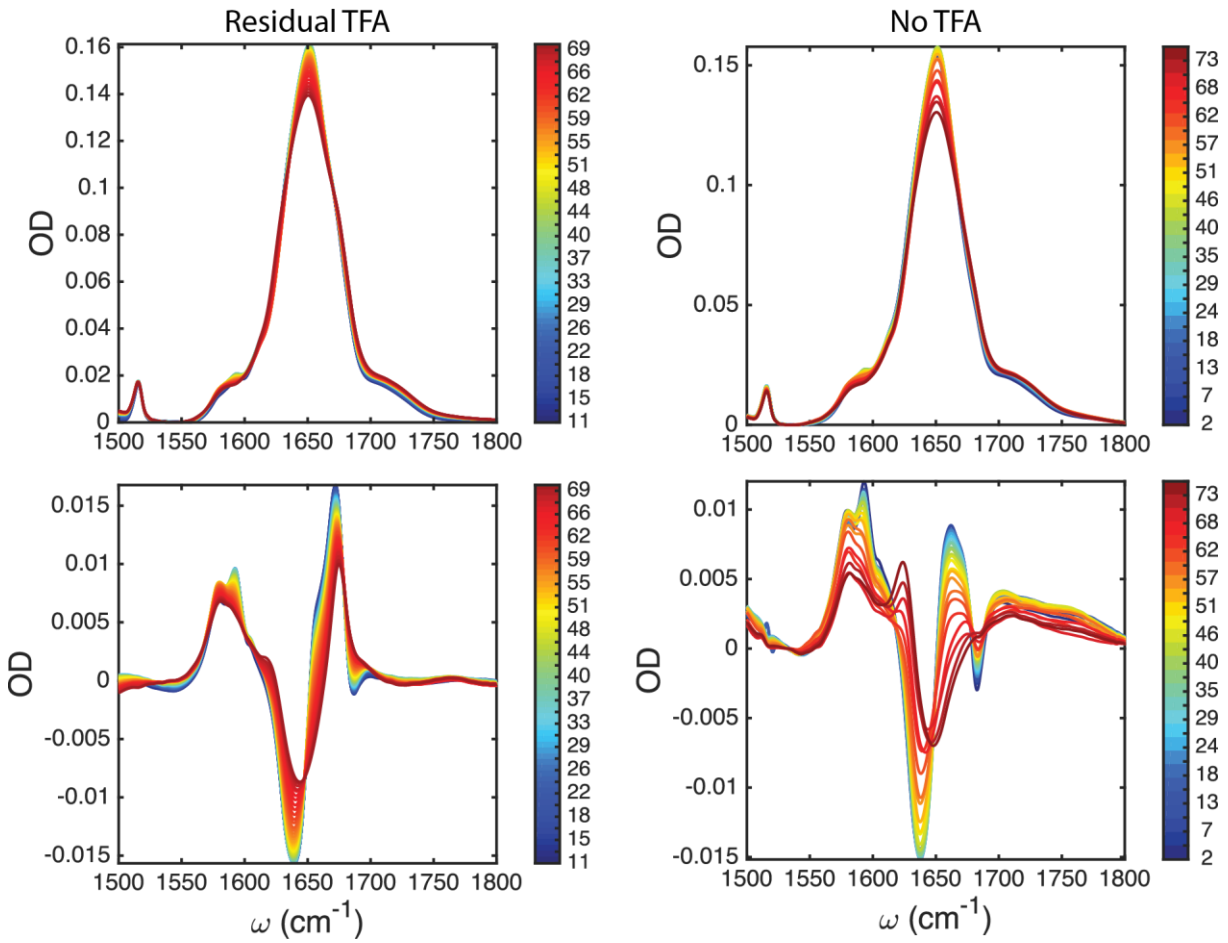

Figure 4.4: Top: Temperature-dependent FTIR spectrum of isotope-edited human insulin with residual TFA (Top left), and without residual TFA (Top right). Bottom: Temperature-dependent difference spectra between isotope-edited insulin and UL insulin with residual TFA (Bottom left), and without residual TFA (Bottom right).

4.2.4 Preparing Insulin Sample for IR Spectroscopic Measurements

For IR measurements, low pH solution was required to protonate the COO^- group to shift its carbonyl vibration to ~1720 cm^{-1}. The lyophilized insulin samples were then dissolved in either 10 mM DCl/D_2O solution (pH* = 1.7, low salt condition) or 270 mM DCl/100 mM NaCl/D_2O solution (pH* = 0.5, high salt condition) to the final concentration of 10 ± 0.3 mg/mL, or equivalently ~1.7 ± 0.1 mM. The condition of 10 mM DCl/D_2O was chosen based on the objective of mitigating irreversible fibril formation due to high ionic strength[20] while maintaining at low pH to avoiding spectral overlap between $^{13}C^{18}O$ amide I vibrations and COO^- asymmetric stretch from sidechains.[21] The other high salt condition was chosen to compare with previous studies on bovine insulin.[8, 22-23] To study ionic strength dependence on thermodynamics below and in Chapter 9, the H–D exchanged insulin samples were dissolved in series of 0, 10, 20, 50, 100, 200, 360 mM NaCl in 10 mM DCl/D_2O solution. Other conditions that are not included here will be mentioned when necessary.

4.2.5 Window Cleaning

Insulin can get stuck on the CaF_2 windows after IR spectroscopic measurements at high temperature due to irreversible aggregation. Usually trying to minimize the amount of time between the window cleaning and the onset of the aggregation will make cleaning easier, but it can still be difficult to clean the aggregation by typical cleaning procedure with methanol and water. To clean residual proteins on the window surface, NOCHROMIX™ (Cabin John, MD, purchased from Sigma Aldrich) in concentrated sulfuric acid solution is used. The cleaning procedure is described below, and example photos are shown in Fig. 4.5.

1. Preparing the NOCHROMIX™ solution with 1 package of NOCHROMIX™ (3.1 oz, approximately 88 gram) per gallon of concentrated sulfuric acid (~3.7 L). Use the same ratio for scaling down the amount of the solution. Usually, the solution can be prepared with a few grams per 100 mL concentrated sulfuric acid.
2. Wait the solution to dissolve NOCHROMIX™ crystals. The solution will be the most effective for the first few days.
3. Cleaning the CaF_2 window first by alternating water and MeOH rinses. Use lab air to blow CaF_2 window completely dry and out of dust.
4. Immersing the window into the NOCHROMIX™ solution for 1–2 minutes. The amount of time can vary depending on how clean the window is before immersion.
5. Rinsing the window with water completely multiple times.
6. Use lens paper and lab air to dry the CaF_2 window.
7. The waste should be neutralized first before going into the waste bottle.

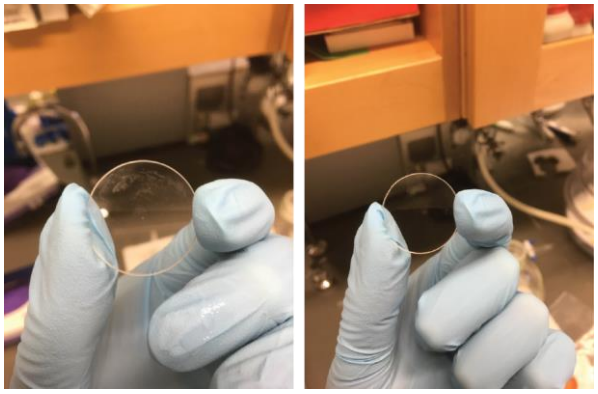

Figure 4.5: Left: CaF_2 window with insulin aggregation after a run of temperature ramp FTIR from 2 °C to 98 °C. Right: CaF_2 window after cleaning with NOCHROMIX™.

4.3 Sample Characterization of Insulin Solution Conditions

Since insulin can easily undergo irreversible aggregation under destabilizing conditions such as low pH and high ionic strength, it is important to explore the dependence to these experimental variables and potentially finding an optimal condition that is workable for thermodynamic characterization and kinetic experiments.

4.3.1 pH Dependence of Human Insulin

Acid titration can help us identify the pK_a of insulin's titratable groups so that one can effectively remove shift the COO^- vibration from the isotope-labeled amide I vibration window, and also characterize the values of pK_a of insulin.[24] pH titration has been performed on bovine insulin with zinc (primarily hexameric), and the results showed that all of the titratable groups were freely accessible to the solvent even though some precipitates can be formed around pH 4–7, and that the lowest pK_a determined at that time was 3.6.[24]

To make sure the pH dependence is consistent with the previous study, titration of zinc-free human insulin was performed by adding NaOH into the insulin solution at pH = 1 prepared with excess HCl. Two independent repeats were performed to make sure the reproducibility of the data. For each repeat, a stock solution of 0.1 N HCl was prepared for dissolving insulin and the corresponding pH was checked by the pH meter (Accument™ AB150 pH benchtop meter with METTKER TOLEDO pH electrode InLab Micro). The basic solution of 1 N NaOH was prepared by dissolving NaOH into water, and checked by the pH meter. For each titration, insulin was dissolved into 5 mL of 1 N HCl solution, and the concentration was checked by Nanodrop UV/Vis spectrometer (Thermo Scientific). The concentration of insulin and pH values of the acid/base solution are summarized in Table 4.1. The setup of the pH titration is shown in Fig. 4.6. For each

pH point, we recorded the volume of added NaOH and pH reading when the stirred solution reaches stable reading. The corresponding pH titration curves are presented in Fig. 4.7. From the titration curves of the two repeats, the consistent lowest pK_a lies around pH=2.6 instead of 3.6 reported previously.[24]

One can also computationally estimate pKa values of the human insulin as another way to examine the pH dependence. Here, we use PROPKA[25-26] on the insulin dimer structure (PDB: 3W7Y). The estimated pK_a values are listed in Table 4.2. From the calculation, titratable groups with the pK_a values below 3 are A4E, and A21 C-terminus, which are consistent with experimentally observed pK_a of 2.6. Meaning that at low pH around and below 2, there will be minimal contribution from the asymmetric COO- vibrations appearing at 1590 cm^{-1} that can interfere with $^{13}C^{18}O$ isotope-edited amide I vibrations.[21]

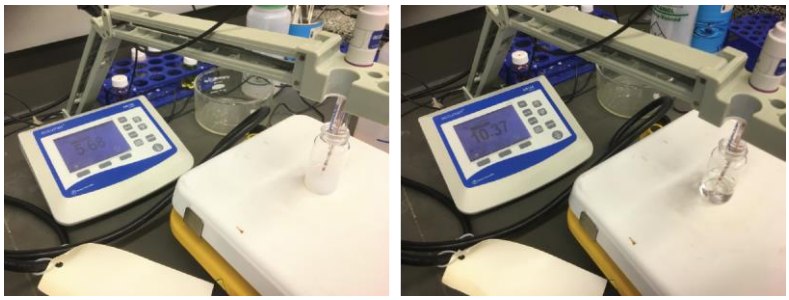

Figure 4.6: pH titration of human insulin. Left: Cloudy solution due to insoluble insulin at pH = 5.68 close to pI. Right: Clear solution of insulin at high pH (10.37).

	[Ins] (mM)	pH(HCl)	pH(NaOH)
Repeat 1	0.63	1.00	13.98
Repeat 2	0.34	0.98	14.00

Table 4.1: Concentration of insulin and pH values of the stock HCl solution and NaOH solution in each repeat for pH titrations.

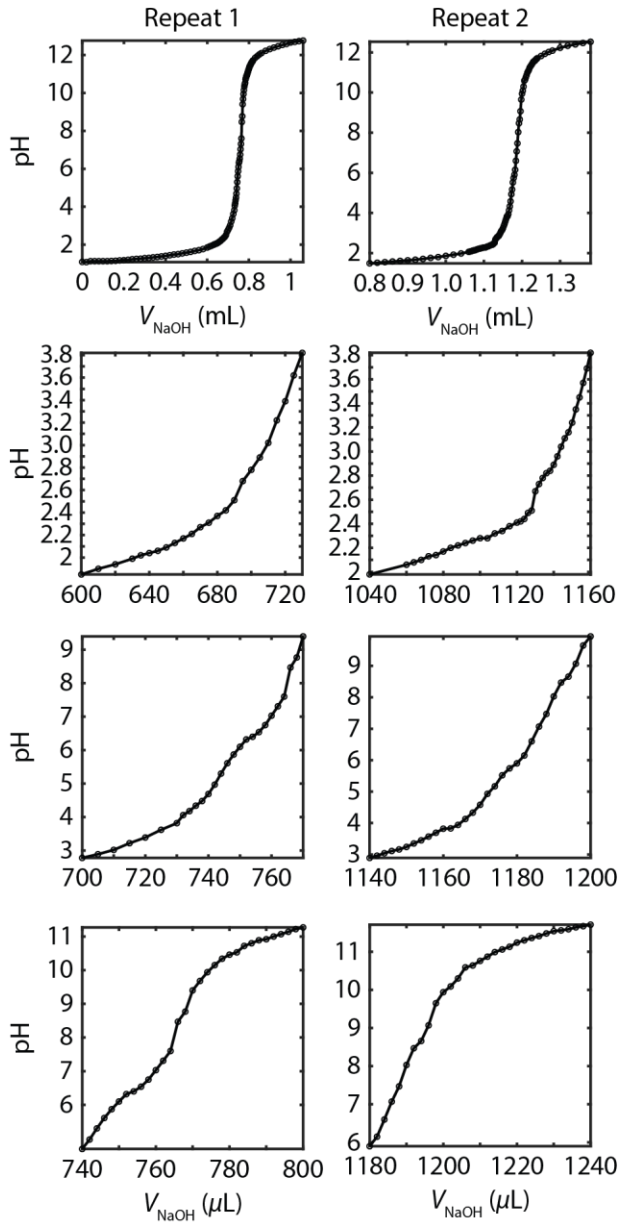

Figure 4.7: pH value of human insulin as a function of added volume of NaOH. The left column represents the first repeat of the pH titration, and the right column represents the second repeat of the pH titration. Titration curves in the second to the fourth row show different pH ranges of the same titration curve in the first row.

	$pK_{a,1}$	$pK_{a,2}$	$pK_{a,avg}$
A1 N-ter ($–NH_2$)	8.67	9.06	8.87
A4 GLU	2.37	3.03	2.70
A14 TYR	10.37	10.21	10.29
A17 GLU	4.57	4.64	4.61
A19 TYR	10.45	10.90	10.68
A21 C-ter ($–COO^-$)	2.65	3.05	2.85
B1 N-ter ($–NH_2$)	7.86	7.92	7.89
B5 HIS	6.85	7.07	6.96
B10 HIS	5.54	5.51	5.53
B13 GLU	6.23	3.95	5.09
B16 TYR	10.54	10.46	10.50
B21 GLU	4.42	4.60	4.51
B22 ARG	13.27	12.84	13.06
B26 TYR	12.01	11.91	11.96
B29 LYS	10.53	10.72	10.63
B30 C-ter ($–COO^-$)	3.38	3.32	3.35

Table 4.2: Estimated pKa values in the insulin dimer, with the subscript 1 and 2 indicating the monomer 1 and 2 in the dimer structure (PDB: 3W7Y).

4.3.2 Ionic Strength Dependence

To investigate the ionic strength dependence on the tendency of irreversible insulin aggregation, and to evaluate if ionic strength can perturb the apparent equilibrium between thermodynamic dimer state and monomer state, we performed a series of temperature-dependent

FTIR experiments with increasing ionic strength in 10 mM DCl/D_2O solution shown in Figs. 4.8–4.9. The solution condition of 10 mM DCl/D_2O is chosen because insulin is soluble at this condition and the ionic strength is minimal. Fig. 4.8 shows that high temperature spectra have an increasing peak intensity centered at 1620 cm^{-1} with increasing ionic strength, which is indicative of the irreversible protein aggregation, and consistent with the previous study.[2] These experiments were performed and heated up step by step over the course of 3 hours. Therefore, the peak intensity at 1620 cm^{-1} reports the tendency of aggregation within 3 hours, and kinetically high ionic strength gives faster growth of aggregation. The corresponding SVD curves as proxies of thermal dissociation curves also show that the shape of the dissociation curves does not vary across a wide range of ionic strength other than the conditions of more than 200 mM NaCl in 10 mM DCl, meaning that the thermodynamic behavior is perturbed at high enough ionic strength. However, the curve taken from 100 mM NaCl/270 mM DCl has the same shape as the curves at low ionic strengths in 10 mM DCl, suggesting that the details of two-state dimer dissociation involves a complex interplay between ionic strength and identity of cations instead of purely ionic strength or electrostatic screening.

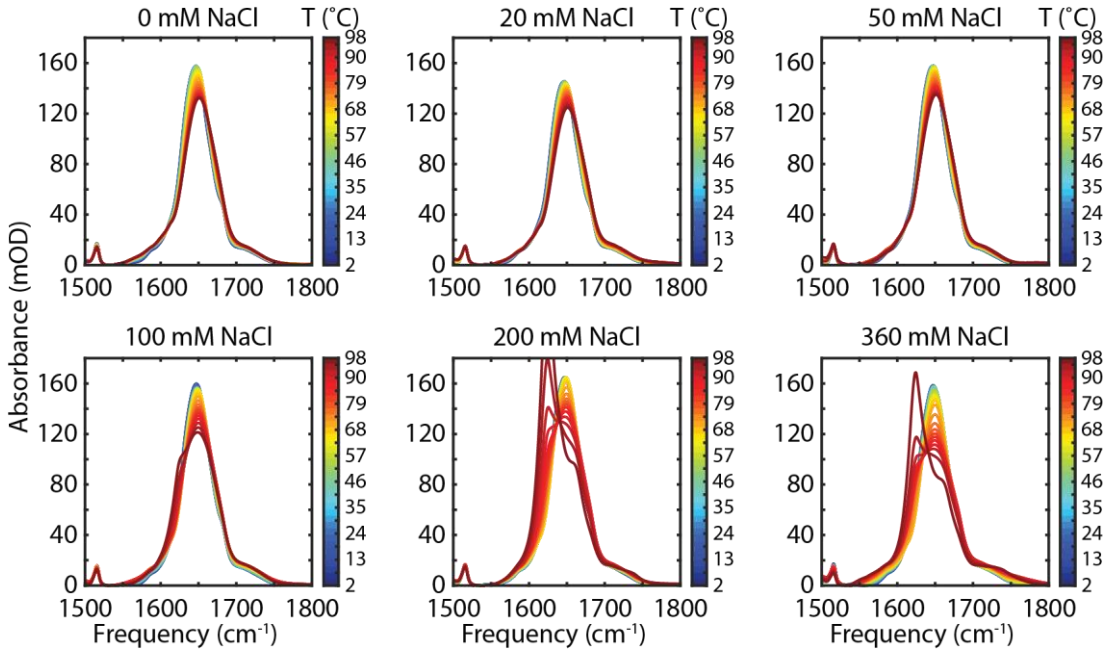

Figure 4.8: Temperature-dependent FTIR spectra of UL insulin in 10 mL DCl at various ionic strength. The concentrations of additional NaCl from the top left to top right are 0, 20, 50 mM, and those from the bottom left to bottom right are 100 mM, 200 mM, 360 mM.

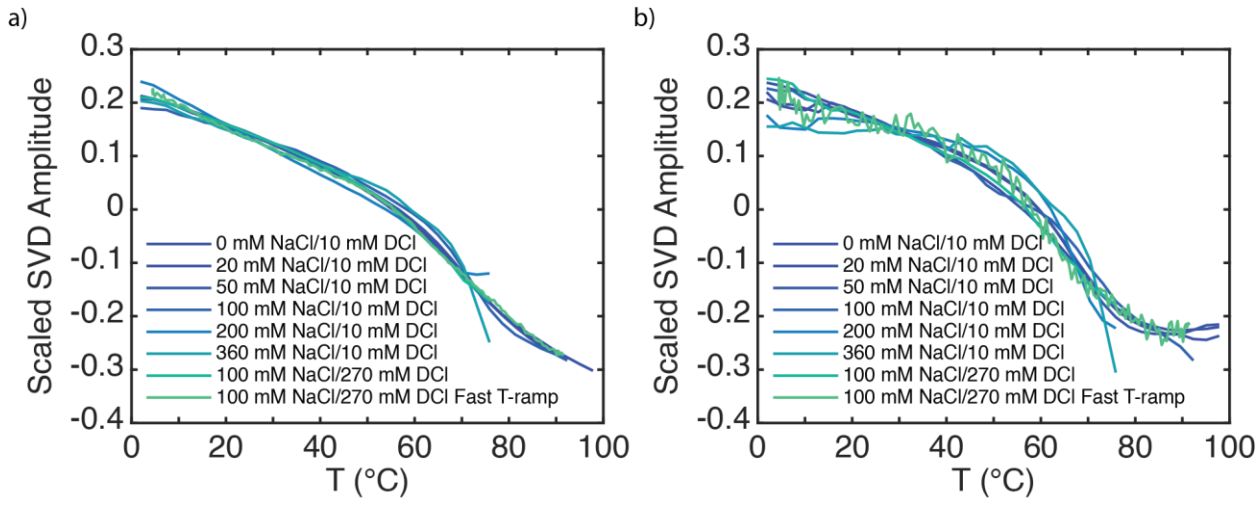

Figure 4.9: Ionic strength dependence on the thermodynamics of dimer-monomer equilibrium. (a–b) Scaled SVD amplitude as a function of temperature, with the SVD performed using the frequency range of (a) 1560–1700 cm^{-1}, and (b) 1670–1700 cm^{-1}.

4.3.3 Searching Conditions to Minimize Irreversible Aggregation at High Temperature

From the previous characterizations, we know that low ionic strength is one key to minimizing irreversible aggregations, and that pH 2 is a soluble condition for insulin and still has the spectroscopic advantages of shifting the 1590 cm^{-1} COO^- asymmetric stretch to 1720 cm^{-1} C=O stretch of the COOH group. Previous experiences also indicate that at the sample temperature of 65 °C with the presence of 6 μm lasers, insulin starts aggregation in 10 min at the high salt condition (100 mM NaCl/270 mM DCl, pH* = 0.5). In contrast, a significant improvement can be made by tuning the condition to 10 mM DCl without any additional salt. In this condition, human insulin has a half-lifetime of 75 min at the sample temperature of 83° C. However, 1-hour time window for measuring IR spectra is still not ideal for long-lasting experimental characterizations such as temperature-dependent 2D IR spectroscopy and transient 2D IR spectroscopy that can easily go beyond 6 hours. Further optimizations for the sample condition are needed in order to perform such time-consuming experiments that are crucial to understand the structural dynamics of insulin. This section summarizes the attempts made by Paul Sanstead and me to seek a more robust and feasible condition.

As the first attempt, we tried to see if further minimizing the ionic strength is possible. We observed that there is residual DCl in the lyophilized insulin sample so that is can be dissolved in neat D_2O solution. The measured pH* was actually around 3.0, which may vary since the solution is not buffered. The resulting T-ramp FTIR spectra show little difference compared to the spectra taken at pH* 1.7. (Figs. 4.10a–b), and the corresponding second SVD components do not qualitatively differ from each other than slight baseline differences (Fig. 4.10c). However, the stability of insulin at pH* = 3.0 can go beyond >7 hrs at 84 °C, which is definitely encouraging and already feasible to do T-jump 2D IR experiments.

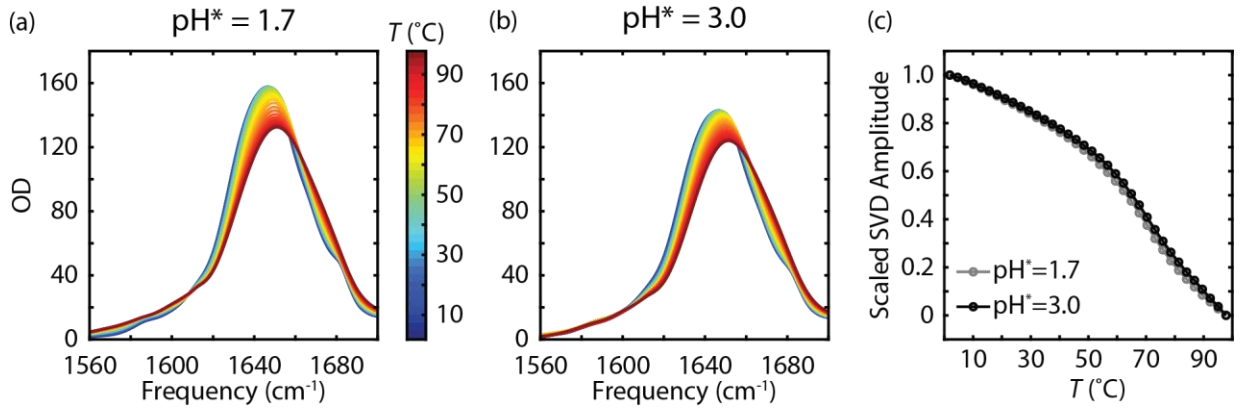

Figure 4.10: T-ramp FTIR spectra of human insulin at (a) pH* = 1.7 and (b) pH* = 3.0. (c) Scaled second SVD temperature components at these two conditions.

Different ions in the solution also affects the stability of insulin. It was reported that the fibrillation rates follow this order: H_2SO_4 > HCl > H_3PO_4 ≈ citric acid ≫ acetic acid.[3] Additionally, simulation study suggests that Na^+ induce the most aggregated structures on charged peptides in water whereas this behavior is reduced by swapping the cation to K^+.[27] FTIR Aggregation tests at T = 84 °C showed that insulin in 100 mM NaD_2PO_4/D_3PO_4 started aggregation in ~1 hr comparable to 10 mM DCl/D_2O, shown in Fig. 4.11. Unexpectedly, no improvements were found when lowering the concentration of the buffer from 100 mM to 10 mM. Also, swapping the cations from Na^+ to K^+ show little improvement on the aggregation onset. As a side note, none of these changes affects the thermodynamic behaviors of insulin. The aggregation onsets across some conditions are summarized in Table 4.3.

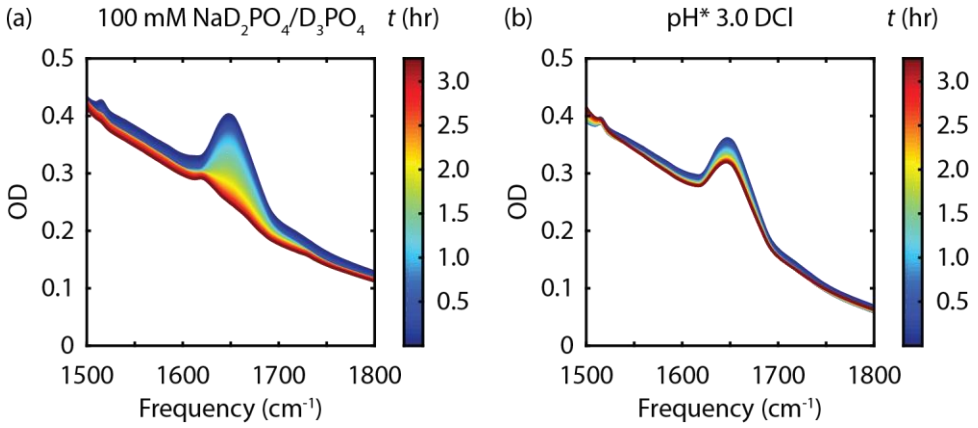

Figure 4.11: T-ramp FTIR spectra of human insulin in (a) 10 mM OGP in 100 mM NaD_2PO_4/D_3PO_4 (pH* ≈ 1.6) and (b) 10 mM OGP in DCl/D_2O solution (pH* = 3.0).

Mechanical agitation is another key factor to facilitating the insulin aggregation. It has been shown that hydrophobic surfaces such as CaF_2 and Teflon induce monomer denaturation followed by formation of intermediate species and aggregation.[1] This suggests that preventing the contact between insulin and the hydrophobic interface can mitigate the aggregation, which can be achieved by adding surfactants that occupy the interfacial sites of CaF_2 window. In their study, *n*-octyl-*β*-D-glucopyranoside (OGP) and *n*-dodecyl-*β*-D-maltoside were used to demonstrate the enhancement of insulin stability. It was cited to have completion of insulin aggregation within 24 hours without the surfactant whereas no detectable aggregation was found for more than 6 weeks at 37 °C, which seems to be a promising approach for further improving the stability.

Paul and I tested the effect of adding the OGP surfactant into the solution. The H–D exchanged stock solution of 10 mM OGP in D_2O was prepared, and the OGP solution was added during the rehydration step of the lyophilized insulin powders to the final OGP concentration of 10 mM. The aggregation test indicates that OGP surfactant does not appear to influence the IR spectral lineshape and thermodynamics, but it increases the stability of insulin in phosphate buffer solution (Table 4.3). Additionally, it helps reduce the evaporation at high temperature (T = 84 °C),

as can be seen by the OD stretch at 2500 cm^{-1} (Fig. 4.12). Note that the critical micelle concentration (CMC) of the OGP is around 23 mM.[28] Thus, having the OGP concentration above the CMC cannot help prevent aggregation since the OGP molecules do not cover the hydrophobic surfaces.

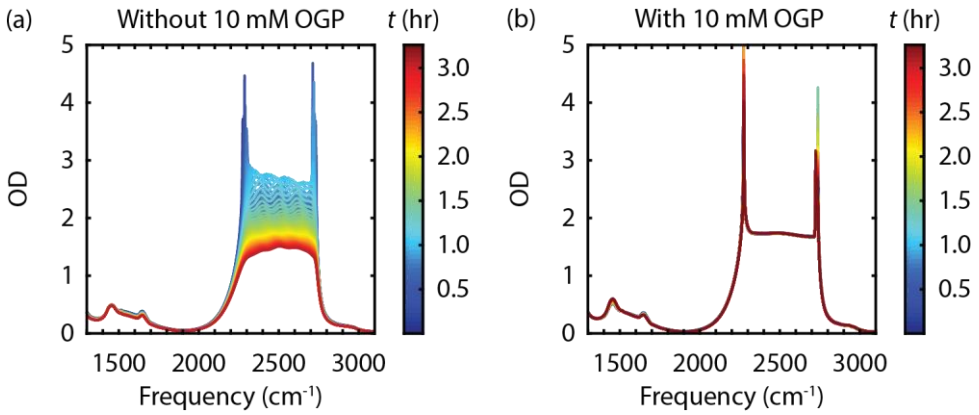

Figure 4.12: T-ramp FTIR spectra of human insulin in (a) DCl/D_2O solution without 10 mM OGP (pH* = 3.0), and in (b) DCl/D_2O solution with 10 mM OGP (pH* = 3.0).

As the current summary, the most robust and feasible condition is 10 mM OGP surfactant in D_2O without additional DCl added during the rehydration step of the H–D exchanged and lyophilized insulin sample, which gives roughly pH* = 3.0. This condition provides more than 7 hours of insulin stability at T = 84 °C, meaning that this condition should be feasible for equilibrium characterizations of wild-type insulin monomer at high temperature and T-jump experiment on precious isotope-edited samples.

Condition	Aggregation Onset
100 mM D_3PO_4	~1 hr
100 mM D_3PO_4 + 10 mM OGP	1hr 40 min
100 mM DCl	> 7 hrs
100 mM DCl + 10 mM OGP	> 7 hrs

Table 4.3: Aggregation onset for different sample conditions.

4.4 Experimental Methods for IR Spectroscopy of Insulin

This section briefly summarizes experimental techniques used for studying human insulin including temperature-dependent FTIR spectroscopy, fast temperature-ramp FTIR spectroscopy, and compact 2D IR spectroscopy. Since all of the instruments are already available without significant rebuilding, I only include relevant details for data collection and spectroscopic analysis.

4.4.1 Temperature-Ramp FTIR Spectroscopy

Temperature-dependent FTIR spectra were collected using a Bruker Tensor 27 FTIR spectrometer with 32 averages at 2 cm^{-1} resolutions unless mentioned otherwise throughout the thesis. Samples were held between two 1 mm thick CaF_2 windows with a 50 μM Teflon spacer to control the IR path length and mounted in a home-built brass sample cell. Solution background spectra were collected to remove solvent contribution to the sample spectra. Dilute HOD spectra at the same solution condition were acquired and used to account for the mismatch of HOD content between sample spectra and background spectra. Temperature was controlled by a recirculating chiller. At each temperature, the sample cell was waited for 90 sec for equilibration prior to each

spectral data collection. For isotope-edited samples, we subtracted the labeled spectrum at each temperature with the UL spectrum at the same condition to remove arginine sidechain vibrations spectrally overlapping with $^{13}C^{18}O$-labeled amide I feature. To account for slight variations of the concentration and path length from sample to sample, we performed normalization of each background-removed IR spectrum before the subtraction based on peak intensity of Tyr sidechain vibration around 1513–1517 cm^{-1}.[21] This mode is insensitive to temperature changes, and therefore is suitable as an internal standard to account for variations in concentration and path length. Normalization based on the integrated area over amide I vibrations from 1560–1700 cm^{-1} is also used to check consistency between the two normalization methods.

4.4.2 Fast T-Ramp FTIR Spectroscopy

The motivation of setting up fast temperature-ramp experiments is to mitigate the irreversible insulin aggregation happening at high temperature, which is kinetically faster to happen at higher ionic strength conditions such as the high salt condition. The setup of the fast temperature-ramp FTIR spectroscopy is the same as the normal temperature-ramp FTIR spectroscopy. However, instead of equilibrating temperature at specified temperature points and then measuring the corresponding IR spectrum, each spectrum was measured on the fly of heating every fixed time interval, with the temperature estimated by pre-determined calibration curve of temperature as a function of heating time in Fig. 4.13. The calibration curve was constructed using a thermocouple in contact with the center of the CaF_2 window with a dry run of fast temperature ramp experiment. In this setup, 4 averages are performed instead of 32 averages, and consecutive spectra are collected spaced by 10 seconds time interval. No apparent decrease of signal-to-noise ratio was observed on the FTIR spectra, but the quality of second derivative got degraded. Initial

temperature is set to zero on the chiller prior to the fast T-ramp. The detailed parameters for running the Macro in OPUS software are summarized in Table 4.4. Note that 9 seconds as the input will approximately give 10 seconds between the two consecutive spectra. Thermodynamics obtained from fast T-ramp is observed to be identical as in the typical temperature ramp experiments as shown in Fig. 4.9. Additionally, the insulin sample can experience the entire dissociation transition without apparent aggregation showing as the peak at 1620 cm^{-1} over the course of fast T-ramp experiment (Fig. 4.13).

Initial Temperature (°C)	Final Temperature (°C)	Time between Spec (sec)	Number of Spec
0	105	9	180

Table 4.4: Parameters for setting up the fast T-ramp FTIR.

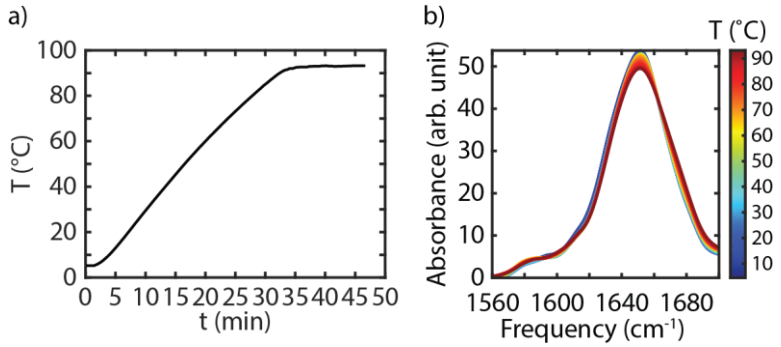

Figure 4.13: (a) Calibration curve of the sample temperature as a function of heating time. (b) Temperature-dependent FTIR spectra of B24B25-labeled insulin in 100 mM NaCl/270 mM DCl.

4.4.3 Compact 2D IR Spectrometer for Equilibrium Experiments

The sample cell setup for 2D IR spectrometer was identical to the FTIR sample cell described above. 2D IR spectra were acquired in a compact pump-probe geometry 2D IR

spectrometer that uses a collinear pump pulse pair using Mach-Zehnder interferometer and a non-collinear probe pulse.[29-30] Since the setup of the spectrometer please find the detailed description of the compact 2D IR spectrometer in Paul Sanstead's thesis.[30] The pulse center frequency of all pulses is 6 μm, resonant with the amide I vibrations. The pulse width is 90 fs in the time domain and has a ~130 cm^{-1} bandwidth in the frequency domain. The waiting time τ_2 between the pump and probe pulses was set to 0.15 ps to minimize pulse overlapping effects, and the coherence time τ_1 was scanned from -0.06 ps to 2.5 ps in 4 fs time step unless otherwise mentioned. The corresponding Fourier-transformed excitation frequency axis has a resolution of 13 cm^{-1}. However, all of the coherence signals decay away around 2 ps. The 2D IR spectra were collected with both parallel polarization and perpendicular polarization. The 2D IR signal is heterodyned by the probe pulse, and spatially dispersed using a grating onto the 64-channel MCT array detector, which has 4 cm^{-1} frequency resolution on the detection frequency axis. Acquired 2D IR spectra are normalized against the pump spectrum to approximately account for the convolution between pump pulses and the intrinsic system 2D IR signal. The same data processing methods were applied as described in Section 3.2 in Paul Sanstead's thesis, including data collection in referenced ΔOD mode, Mertz correction for phasing and timing, and Fast Fourier Transform with zero-padding and windowing, and additional spectral subtraction to reduce long-term laser power drift described in Section 4.2 in Paul Sanstead's thesis.[30]

4.4.4 Undersampling for Efficient Data Acquisition on 2D IR Spectroscopy

For speeding up the 2D IR data acquisition, in particular for transient 2D IR measurements, undersampling along the coherence time τ_1 can be employed to get 2D IR spectra without significant aliasing due to overlapping with the low frequency response or noises. Based on

Nyquist-Shannon Sampling Theorem, the sampling frequency should be at least twice the signal bandwidth in order to reconstruct the original signal that can fit into the undersampled frequency range.[31-32] In other words, we can determine the relation between the sampling frequency F_S, lower bound of the frequency F_L, and higher bound of the frequency F_H. Given the theorem, we can find the following relations.

$$F_H \leq \frac{n}{2} F_S \tag{4.4}$$

$$\frac{n-1}{2} F_S \leq F_L \tag{4.5}$$

$$F_H - F_L \leq \frac{F_S}{2} \tag{4.6}$$

Eqn. (4.6) can be inferred from Eqns. (4.4)–(4.5), where n is a positive integer. Using these equations, one can find out the relation to determine the sampling frequency as follows

$$\frac{2}{n} F_H \leq F_S \leq \frac{2F_L}{n-1} \tag{4.7}$$

$$1 \leq n \leq \frac{F_H}{F_H - F_L} \tag{4.8}$$

The upper bound of the integer n can be obtained from the equality of Eqn. (4.4) and Eqn. (4.6). Note that when n = 1, it is not an undersampling scheme, and one can reduce back to the Nyquist Sampling Theorem, which states that the sampling frequency is required to be twice the signal frequency to avoid aliasing.

This relation is useful to determine the sampling frequency F_S when the frequency range of interest is pre-determined. In the Compact 2D IR spectrometer, the typical center frequency is

around 6 μm, or around 1650 cm^{-1}, and the detection frequency axis obtained from the grating can range from 1480 cm^{-1} to 1780 cm^{-1}. Applying such relation will result in

$$\frac{3560}{n}\ \text{cm}^{-1} \le F_S \le \frac{2960}{n-1}\ \text{cm}^{-1} \tag{4.9}$$

$$1 \le n \le 5.93 \tag{4.10}$$

Or converting into the sampling time step t_S in fs

$$\frac{n-1}{0.0887408} \le t_S \le \frac{n}{0.1067288} \tag{4.11}$$

The range of the sampling time step during the coherence time is summarized in Table 4.5. Also, the relation between the signal frequency F and the aliased frequency F_a can be determined as

$$F = n\left(\frac{F_S}{2}\right) + k \tag{4.12}$$

$$F_a = \begin{cases} -\dfrac{F_S}{2} + k, n \in \text{odd} \\ k, n \in \text{even} \end{cases} \tag{4.13}$$

which is useful to determine the aliased frequency axis.

n	1	2	3	4	5
Upper bound (fs)	9.37	18.74	28.11	37.48	46.85
Lower bound (fs)	0	11.27	22.54	33.81	45.08

Table 4.5: Time step ranges allowed for undersampling during the coherence time τ_1.

The time step for undersampling in practice is set to 24 fs, which can be investigated experimentally or estimated from the model data, with the test script provided in Appendix 4B.

The resulting undersampled spectra with τ_1 = 24 fs show little variation compare to the oversampled spectrum with τ_1 = 4 fs (Fig. 4.14).

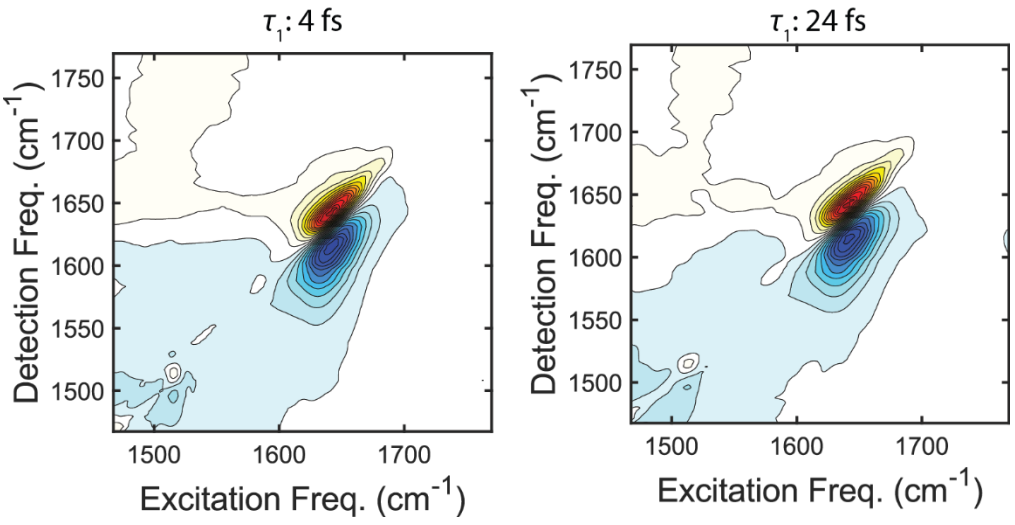

Figure 4.14: Equilibrium ZZZZ-polarized 2D IR spectrum of insulin in low salt condition (10 mM DCl/D_2O). (Left) Oversampled 2D IR spectrum acquired with 4 fs time step along coherence time τ_1. (Right) Undersampled 2D IR spectrum acquired with 24 fs time step along τ_1.

4.5 Acknowledgments

I thank Paul Sanstead for helping me pick up the Compact 2D IR spectrometer, working out and providing the procedure of coating the FEP film onto the CaF_2 windows as the Appendix 4A, and the collaborative efforts to find conditions for insulin. I thank Balamurugan Dhayalan for kindly synthesizing batches of $^{13}C^{18}O$ isotope-edited human insulin and providing synthesis protocol.

4.6 References

1. Sluzky, V.; Klibanov, A. M.; Langer, R., Mechanism of insulin aggregation and stabilization in agitated aqueous solutions. *Biotechnol Bioeng* **1992,** *40* (8), 895-903.

2. Nielsen, L.; Khurana, R.; Coats, A.; Frokjaer, S.; Brange, J.; Vyas, S.; Uversky, V. N.; Fink, A. L., Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. *Biochemistry* **2001,** *40* (20), 6036-46.
3. Whittingham, J. L.; Scott, D. J.; Chance, K.; Wilson, A.; Finch, J.; Brange, J.; Guy Dodson, G., Insulin at pH 2: Structural Analysis of the Conditions Promoting Insulin Fibre Formation. *Journal of Molecular Biology* **2002,** *318* (2), 479-490.
4. Haas, J.; Vohringer-Martinez, E.; Bogehold, A.; Matthes, D.; Hensen, U.; Pelah, A.; Abel, B.; Grubmuller, H., Primary steps of pH-dependent insulin aggregation kinetics are governed by conformational flexibility. *Chembiochem* **2009,** *10* (11), 1816-22.
5. Derewenda, U.; Derewenda, Z.; Dodson, E. J.; Dodson, G. G.; Bing, X.; Markussen, J., X-ray analysis of the single chain B29-A1 peptide-linked insulin molecule. *Journal of Molecular Biology* **1991,** *220* (2), 425-433.
6. Brange, J.; Andersen, L.; Laursen, E. D.; Meyn, G.; Rasmussen, E., Toward understanding insulin fibrillation. *J Pharm Sci* **1997,** *86* (5), 517-25.
7. Ivanova, M. I.; Sievers, S. A.; Sawaya, M. R.; Wall, J. S.; Eisenberg, D., Molecular basis for insulin fibril assembly. *Proc Natl Acad Sci U S A* **2009,** *106* (45), 18990-5.
8. Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy. *Phys Chem Chem Phys* **2010,** *12* (14), 3579-88.
9. Dhayalan, B.; Fitzpatrick, A.; Mandal, K.; Whittaker, J.; Weiss, M. A.; Tokmakoff, A.; Kent, S. B., Efficient Total Chemical Synthesis of (13) C=(18) O Isotopomers of Human Insulin for Isotope-Edited FTIR. *Chembiochem* **2016,** *17* (5), 415-20.
10. Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B., Synthesis of proteins by native chemical ligation. *Science* **1994,** *266* (5186), 776-9.
11. Sohma, Y.; Hua, Q. X.; Whittaker, J.; Weiss, M. A.; Kent, S. B., Design and folding of [GluA4(ObetaThrB30)]insulin ("ester insulin"): a minimal proinsulin surrogate that can be chemically converted into human insulin. *Angew Chem Int Ed Engl* **2010,** *49* (32), 5489-93.
12. Schnölzer, M.; Alewood, P.; Jones, A.; Alewood, D.; Kent, S. B. H., In Situ Neutralization in Boc-chemistry Solid Phase Peptide Synthesis. *International Journal of Peptide Research and Therapeutics* **2007,** *13* (1), 31-44.
13. Avital-Shmilovici, M.; Mandal, K.; Gates, Z. P.; Phillips, N. B.; Weiss, M. A.; Kent, S. B., Fully convergent chemical synthesis of ester insulin: determination of the high resolution X-ray structure by racemic protein crystallography. *J Am Chem Soc* **2013,** *135* (8), 3173-85.
14. Covington, A. K.; Paabo, M.; Robinson, R. A.; Bates, R. G., Use of the glass electrode in deuterium oxide and the relation between the standardized pD (paD) scale and the operational pH in heavy water. *Analytical Chemistry* **1968,** *40* (4), 700-706.
15. Krezel, A.; Bal, W., A formula for correlating pKa values determined in D2O and H2O. *J Inorg Biochem* **2004,** *98* (1), 161-6.
16. Harrison, D. M.; Garratt, C. J., The accurate measurement of insulin molarity. *Biochem J* **1969,** *113* (4), 733-4.
17. Olsen, H. B.; Ludvigsen, S.; Kaarsholm, N. C., Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry* **1996,** *35* (27), 8836-45.
18. DeFlores, L. P.; Tokmakoff, A., Water penetration into protein secondary structure revealed by hydrogen-deuterium exchange two-dimensional infrared spectroscopy. *J Am Chem Soc* **2006,** *128* (51), 16520-1.

19. Andrushchenko, V. V.; Vogel, H. J.; Prenner, E. J., Optimization of the hydrochloric acid concentration used for trifluoroacetate removal from synthetic peptides. *J Pept Sci* **2007,** *13* (1), 37-43.
20. Nielsen, L.; Frokjaer, S.; Brange, J.; Uversky, V. N.; Fink, A. L., Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **2001,** *40* (28), 8397-409.
21. Barth, A., The infrared absorption of amino acid side chains. *Prog Biophys Mol Biol* **2000,** *74* (3-5), 141-73.
22. Zhang, X. X.; Jones, K. C.; Fitzpatrick, A.; Peng, C. S.; Feng, C. J.; Baiz, C. R.; Tokmakoff, A., Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J Phys Chem B* **2016,** *120* (23), 5134-45.
23. Rimmerman, D.; Leshchev, D.; Hsu, D. J.; Hong, J.; Kosheleva, I.; Chen, L. X., Direct Observation of Insulin Association Dynamics with Time-Resolved X-ray Scattering. *J Phys Chem Lett* **2017,** *8* (18), 4413-4418.
24. Tanford, C.; Epstein, J., The Physical Chemistry of Insulin. I. Hydrogen Ion Titration Curve of Zinc-free Insulin1-3. *Journal of the American Chemical Society* **1954,** *76* (8), 2163-2169.
25. Sondergaard, C. R.; Olsson, M. H.; Rostkowski, M.; Jensen, J. H., Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput* **2011,** *7* (7), 2284-95.
26. Olsson, M. H.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **2011,** *7* (2), 525-37.
27. Ghosh, S.; T, D.; Baul, U.; Vemparala, S., Aggregation dynamics of charged peptides in water: Effect of salt concentration. *J Chem Phys* **2019,** *151* (7), 074901.
28. Chattopadhyay, A.; London, E., Fluorimetric determination of critical micelle concentration avoiding interference from detergent charge. *Analytical Biochemistry* **1984,** *139* (2), 408-412.
29. Deflores, L. P.; Nicodemus, R. A.; Tokmakoff, A., Two-dimensional Fourier transform spectroscopy in the pump-probe geometry. *Opt Lett* **2007,** *32* (20), 2966-8.
30. Sanstead, P. J. C. Investigation of DNA Dehybridization through Steady-State and Transient Temperature-Jump Nonlinear Infrared Spectroscopy. The University of Chicago 2018.
31. Whittaker, E. T., XVIII.—On the Functions which are represented by the Expansions of the Interpolation-Theory. *Proceedings of the Royal Society of Edinburgh* **1915,** *35*, 181-194.
32. Shannon, C. E., A mathematical theory of communication. *The Bell System Technical Journal* **1948,** *27* (3), 379-423.

Appendix 4A: Window Coating Procedure for Transient IR Spectroscopy

This appendix listed the procedure for coating the bare CaF_2 windows with fluorinated ethylene propylene (FEP) film, which is useful for extending the pulsed T-jump duration to almost

5 times longer at the final temperature. The material used is the 12.7 μm thick FEP film purchased from CSHyde (Item #23-1/2FEP-24, Type A).

1. Fix FEP sheet to a 0.25” thick acrylic panel by taping the corners in preparation for laser cutting 1” diameter disks. For the cutting template, leave a 1° arc blank in the circumference of the disk in order to maintain a small connection to the sheet. Otherwise the disks will blow away in the laser cutter exhaust once cut. Use the lowest power setting to minimize melting and distortion of the film.
2. After cutting, remove the disks from the sheet by breaking the small amount of material tethering them to the sheet. Wash the disks in DI water to remove residual dust from the cuts. Allow the disks to dry or blow dry with lab air.
3. Clean a 1” diameter CaF_2 window by alternating between H_2O and MeOH rinses. Use lab air to blow the window completely dry and free of dust. Align a 1” diameter FEP disk onto the window such that it lies flat and completely covers the window surface.
4. On a hot plate in the fume hood, heat a vacuum flask to a set point of 300 °C. The melting point of FEP is 260 °C, so this should be sufficient to melt the thin films onto the window. If heated too quickly, the film tends to bubble and break. This step could probably be optimized, but 300 °C seems to work well enough.
5. Using a long pair of tweezers, carefully lower the window with the FEP disk on it into the vacuum flask and place it in the center of the flask, FEP-side up. Immediately stopper the vacuum flask and pull full house vacuum before the film begins to melt.
6. Monitor the film as it melts. Parallel raised ridges will form initially, within about the first minute of heating. Continue to heat the window + film under vacuum until these ridges lie

flat and the film appears smooth. This will typically take 5-10 minutes. If the film is heated for too long, FEP can seep off the edge of the window and leave uncoated areas of CaF_2.

7. Once the film appears flat and clear without visible defects, turn off the house vacuum. Break the vacuum seal at the stopper. Do this step gradually to avoid an inrush of air that can send the window flying.
8. Carefully remove the FEP coated window from the flask using a long tweezers to grab opposite edges of the window and place the window to cool on the lab bench.
9. Once cool, the coated window is ready to use. If done well, it can be difficult to see which side is coated. To load a sample, place a coated window FEP-side up into the brass sample cell. The hydrophobicity of FEP makes loading the sample difficult. If all of the solution is pipetted into the center of the coated window (typical loading procedure), the sample will not spread evenly once sandwiched with a second window. The solution will escape the hydrophobic surfaces and leak out into the sample cell, leaving behind large air gaps.
10. Instead, for a 50 µm spacer, prepare 35 µL of sample. Using a pipet set to 5 µL, place one 5 µL drop in the center of the window. Pipet the 6 remaining 5 µL drops in an even pattern about the central drop spaced 3-4 mm from the center. Sandwich the sample with a second coated window, applying firm even pressure until the cell can be screwed together. This approach tends to result in a much more even distribution of sample and minimizes losses out into the brass.

Typically, FEP coated windows are good for only a single use, but they can be cleaned with a H_2O/MeOH rinse. Do not wipe the film with a Kim wipe since FEP is soft and easily scratched. Instead, use lab air to blow the surface dry between rinses. More often than not, sample

solution starts to penetrate between the CaF_2 and the FEP film during an experiment and the edges of the film start to peel off once the sample cell is disassembled. Old or damaged films are easily pulled off of the window and used windows can be recoated many times once properly cleaned. Soaking coated windows in pure water or acid + NOCHROMIX™ helps remove an old or damaged film so that the window can be reused.

Appendix 4B: Example Script to Test Undersampling

```
% This script is used to test Kubo lineshape FFT on linear response
% as well as how to determine frequency axis using undersampling scheme
%
% Jan. 24, 2019
% Chi-Jui Feng

%% Input options

% Physical constants and parameters
wavenumber2pHz = 2.998e-2;          % Unit conversion from cm-1 to 1/ps
Delta_omega = 10;                   % Magnitude of frequency fluctuations in
cm-1
tau = 1;                            % Pure dephasing time/Spectral diffusion
time/Vibrational frequency correlation time in ps
taup = 1;                           % Vibrational relaxation time constant in
ps
w = [1620 1650];                    % Center frequency of the peaks in cm-1
a = [0.5  1   ];                    % Amplitude of the peaks, arb. unit
dt = 0.005;                         % Oversampled time step in ps
dt_u = 0.024;                       % Undersampled time step in ps
tscan = 2.5;                        % Scan time in ps
n_zp = 2^14;                        % Number of zero-padded points
w1range = [1480 1780];              % Frequency range of interestt

%% Setting up time axis
n_t = tscan/dt+1;                   % Number of time points in oversampling
scheme
t = 0:dt:tscan;                     % Time axis in oversampling scheme
t_u = 0:dt_u:(n_t-1)*dt;            % Time axis in undersampling scheme, the
end time point should be the same as the one in the oversampling scheme
n_tu = size(t_u,2);                 % Number of time points in undersampling
scheme
Fs = (1/dt)/2 / wavenumber2pHz;     % Nyquist frequency of oversampled FID in
cm-1
Fs_u = (1/dt_u)/2 / wavenumber2pHz; % Nyquist frequency of undersampled FID
in cm-1
```

```
%% Setting up frequency axis
% Defining frequency axis for oversampled spectrum
dw = Fs/n_zp;                          % Frequency spacing of oversampled
spectrum, determined by Nyquist frequency and zero-padded length
dw_u = Fs_u/n_zp;                      % Frequency spacing of undersampled
spectrum
waxis = Fs*(-((n_zp-1)/2):(n_zp-1)/2)/(n_zp/2);
% Defining frequency axis for undersampled spectrum
waxis_u = Fs_u*(-((n_zp-1)/2):(n_zp-1)/2)/(n_zp/2); % Original frequency axis
defined by Nyquist frequency

npts = size(waxis_u, 2);

% Determine where the positive frequency should be
m = floor( w(1) / Fs_u );
n = w(1) - m*Fs_u;                      % Equivalent to mod(wi, Fs_u);
if mod(m,2) % The positive (aliased) frequency component lies on the negative
frequency range of the waxis_u
    w_aliased = waxis_u(1) + n;
else        % The positive (aliased) frequency component lies on the positive
frequency range of the waxis_u
    w_aliased = n;
end

% Find the frequency index
[~, wind] = min( abs(waxis_u - w_aliased) );
% Construct the aliased frequency axis of interest
npts_new = diff(w1range) / dw;
w1l =  wind - round( (w(1) - w1range(1)) / dw_u );
if (w1l<0) % The lower bound got wrapped back to the other side of the
frequency axis
    w1l = w1l + npts;
end
w1u =  wind + round( (w1range(2) - w(1)) / dw_u );
if (w1u>npts) % The upper bound got wrapped back to the other size of the
frequency axis
    w1u = w1u - npts;
end
if w1l < w1u
    ind = w1l:w1u;
else
    ind = [w1l:npts 1:w1u];
end

% Construct the new frequency axis
waxis_new = w1range(1) + (0:length(ind)-1) * dw_u;

%% Setting up over-sampled response function
% Kubo lineshape accounting for spectral diffusion/pure dephasing, etc.
g = @(t) (Delta_omega*wavenumber2pHz*2*pi)^2*tau^2 .* ( exp(-t/tau) + t/tau -
1 );
% ad hoc vibrational relaxation modeled by single exponential decay
Gamma = @(t) exp(-t/(2*taup));

R1 = zeros(size(t));
for p = 1:length(w)
```

```
    R1 = R1 + a(p) * cos(w(p)*wavenumber2pHz*2*pi*t) .* exp(-g(t)) .*
Gamma(t); % Positive and negative frequency
    % R1 = R1 + a(p) * exp(1i*w(p)*wavenumber2pHz*2*pi*t) .* exp(-g(t)) .*
Gamma(t); % Single positive frequency
end

% Zero-padding
R1_for_FT = zeros(1, n_zp);
R1_for_FT(1:n_t) = R1;

%% Set up under-sampled response function
R1u = zeros(size(t_u));
for p = 1:length(w)
    R1u = R1u + a(p) * cos(w(p)*wavenumber2pHz*2*pi*t_u) .* exp(-g(t_u)) .*
Gamma(t_u); % Positive and negative frequency
end

% Zero-padding
R1u_for_FT = zeros(1, n_zp);
R1u_for_FT(1:n_tu) = R1u;

%% FFT
S1 = fftshift(fft(R1_for_FT)); % Oversampled
S1u = fftshift(fft(R1u_for_FT)); % Undersampled
S1u_new = S1u(ind); % Picking up the range of interest

%% Plot the oversampled spectrum
figure(1);
clf
subplot(1,2,1)
plot(t, real(R1), 'Color', [0 0 0]);
xlabel('t (ps)');
ylabel('RE[R^{(1)}]');
title('Oversampled FID');
xlim([0 2]);
subplot(1,2,2)
plot(waxis, real(S1), 'Color', [0 0 0])
xlim([-Fs Fs])
xlabel('Freq. (cm^{-1})')
title('Oversampled spectrum');

%%
figure(2);
clf
subplot(1,2,1)
plot(t_u, real(R1u), 'Color', [0 0 0]);
xlabel('t (ps)');
ylabel('RE[R^{(1)}]');
title('Undersampled FID');
xlim([0 2]);
subplot(1,2,2)
plot(waxis_u, real(S1u), 'Color', [0 0 0])
xlim([-Fs_u Fs_u])
xlabel('Original Freq. (cm^{-1})')
title('Undersampled spectrum');
```

```
%%
figure(3)
clf
subplot(1,2,1)
plot(t_u, real(R1u), 'Color', [0 0 0]);
xlim([0 2]);
xlabel('t (ps)');
ylabel('RE[R^{(1)}]');
title('Undersampled FID');
subplot(1,2,2)
plot(waxis_new, real(S1u_new), 'Color', [0 0 0])
xlim(w1range)
xlabel('Aliased Freq. (cm^{-1})')
title('Undersampled spectrum');
```

Chapter 5

Miscellaneous Analysis Methods for IR Spectroscopy of Proteins and Computational Amide I Spectroscopy

5.1 Introduction

In the previous chapters, we discussed the theory of nonlinear spectroscopy that relates optical signals with molecular response functions, practically how to measure and process 2D IR signals and how to simulate a 2D IR spectrum using amide I spectroscopic models. There are still numerous analysis tools and models that help bridge the gap between 2D IR spectra and chemical interpretations of the protein system, including thermodynamics, underlying conformational ensemble and dynamical behaviors. This chapter will discuss models and tools used throughout this thesis that do not belong to typical processing methods or theory regarding IR and 2D IR spectroscopy. Please note that similar models and analysis tools have been applied to different types of problems in IR spectroscopy or some other fields and experiments. What this chapter does is providing some background of these tools with the idea that can be viable to protein IR spectroscopy and computational spectroscopy.

The outline of this chapter is as follows. In section 5.2, basic thermodynamic models of protein unfolding and homodimer dissociation will be covered, which will be useful for characterizing equilibrium thermodynamic behaviors of insulin dimer dissociation and similar

dimer dissociation problems. Section 5.3 discusses a minimally biased approach to reconstruct pure component spectra and underlying populations from a set of equilibrium IR spectra as a function of some experimental variables, or potentially kinetics from a set of transient IR data. Section 5.4 describes methods of ensemble refinement that infers conformational ensemble of a protein system that is consistent with experimental spectra. In later chapter (Chapter 8), the idea of ensemble refinement will also be tested with proof-of-principle peptide system.

5.2 Thermodynamic Models of Protein Unfolding and Dimer Dissociation

Temperature is one of the key variables controlling population of species in experiments, and the corresponding temperature dependence informs us of the chemical processes including protein unfolding and dimer dissociation/association. This section discusses some basic thermodynamic models regarding protein unfolding and dimer dissociation.

5.2.1 Two-State Thermodynamic Model of Protein Unfolding

To start with, thermodynamic two-state reaction describing protein unfolding between folded state F and unfolded state U is given as

$$\mathrm{F} \overset{K_u}{\rightleftharpoons} \mathrm{U} \tag{5.1}$$

In a solution with the total protein concentration $C_{tot} = [\mathrm{F}] + [\mathrm{U}]$, we can define fractions of folded species and unfolded species as θ_F and θ_U such that $\theta_F + \theta_U = 1$, which will be useful to know the population of each species at a given temperature T. Each species is typically assumed to be spectroscopically distinguishable so that measuring spectral change as a function of temperature

can be used as a proxy to the underlying population of some species. The expression of these fractions are given as

$$\begin{aligned}\theta_{\mathrm{F}}(T) &= \frac{[\mathrm{F}]}{C_{\mathrm{tot}}} \\ \theta_{\mathrm{U}}(T) &= 1-\theta_{\mathrm{F}}(T) = \frac{[\mathrm{U}]}{C_{\mathrm{tot}}}\end{aligned} \tag{5.2}$$

We can also express the unfolding constant in terms of the folded fraction as follows

$$K_{\mathrm{u}}(T) = \frac{[\mathrm{U}]}{[\mathrm{F}]} = \frac{1-\theta_{\mathrm{F}}(T)}{\theta_{\mathrm{F}}(T)} \tag{5.3}$$

Or equivalently express the folded fraction and the unfolded fraction in terms of K_{u} as

$$\theta_{\mathrm{F}}(T) = \frac{1}{1+K_{\mathrm{u}}(T)} \tag{5.4}$$

$$\theta_{\mathrm{U}}(T) = \frac{K_{\mathrm{u}}(T)}{1+K_{\mathrm{u}}(T)} \tag{5.5}$$

From Eqns. (5.4) and (5.5), it is straightforward to see when $K_{\mathrm{u}} \gg 1$, the unfolded fraction is close to 1 while it is almost zero when $K_{\mathrm{u}} \ll 1$. In particular, when $K_{\mathrm{u}}=1$, both fractions become 0.5 as expected literally from the definition of equilibrium constant shown in Eqn. (5.3). The corresponding free energy change of unfolding at the standard state is given as

$$\Delta G_{\mathrm{u}}^{\circ}(T) = -RT\ln K_{\mathrm{u}}(T) = \Delta H_{\mathrm{u}}^{\circ}(T) - T\Delta S_{\mathrm{u}}^{\circ}(T) \tag{5.6}$$

Or alternatively, the unfolding equilibrium constant can be written as

$$K_{\mathrm{u}}(T) = \exp\left(\frac{-\Delta G_{\mathrm{u}}^{\circ}(T)}{RT}\right) \tag{5.7}$$

Based on Eqn. (5.7), how free energy changes as a function of temperature determines the equilibrium populations of both folded and unfolded species, which will require knowledge of (temperature-dependent) enthalpy change and entropy change. Typically, the changes of enthalpy and entropy are assumed to be temperature-independent over the temperature range of interest such as DNA dehybridization. Nonetheless, it is well-known that there is non-negligible change of heat capacity in protein folding/unfolding processes, which is highly correlated with change of solvent accessible surface area.[1-2] General expressions of enthalpy and entropy at the standard state are

$$\Delta H^\circ(T_2) = \Delta H^\circ(T_1) + \int_{T_1}^{T_2} \Delta C_P^\circ(T)\,dT \tag{5.8}$$

$$\Delta S^\circ(T_2) = \Delta S^\circ(T_1) + \int_{T_1}^{T_2} \frac{\Delta C_P^\circ(T)}{T}\,dT \tag{5.9}$$

In principle, a precise calorimetry experiment such as differential scanning calorimetry can be used to determine the temperature-dependent change of heat capacity and potentially infer the free energy profile.[3] In practice, however, including temperature dependence to the change of heat capacity is likely to cause over-fitting to a single melting curve extracted from IR spectroscopy. To first order, including a constant change of heat capacity should be the simplest thermodynamic model for calculating equilibrium unfolding constant as a function of temperature. Hence, using a reference temperature T_u, Eqns. (5.8) and (5.9) can be simplified as

$$\Delta H_\mathrm{u}^\circ(T) = \Delta H_\mathrm{u}^\circ(T_\mathrm{u}) + \Delta C_{P,\mathrm{u}}^\circ\left(T - T_\mathrm{u}\right) \tag{5.10}$$

$$\Delta S_\mathrm{u}^\circ(T) = \Delta S_\mathrm{u}^\circ(T_\mathrm{u}) + \Delta C_{P,\mathrm{u}}^\circ \ln\left(\frac{T}{T_\mathrm{u}}\right) = \frac{\Delta H_\mathrm{u}^\circ(T_\mathrm{u})}{T_\mathrm{u}} + \Delta C_{P,\mathrm{u}}^\circ \ln\left(\frac{T}{T_\mathrm{u}}\right) \tag{5.11}$$

In this thermodynamic two-state model, T_u is the unfolding temperature defined at $\theta_F = \theta_U = 0.5$, or equivalently, $K_d = 1$. Then, the free energy change as a function of temperature can be easily written as

$$\Delta G_u^\circ(T) = \Delta H_u^\circ(T_u)\left[1 - \frac{T}{T_u}\right] - \Delta C_{P,u}^\circ\left[T - T_u - T\ln\left(\frac{T}{T_u}\right)\right] \tag{5.12}$$

When the folded fraction is 0.5, the corresponding free energy of unfolding is $\Delta G_u^\circ(T_u) = 0$. In other words, if we try to find a temperature where the free energy is zero in Eqn. (5.12), the temperature corresponds to the unfolding temperature T_u. In addition, if $\Delta C_{P,u}^\circ = 0$, the corresponding enthalpy is known as van't Hoff enthalpy, which is equal to $\Delta H_u^\circ(T_u)$ based on Eqn. (5.10). One can thus interpret the enthalpy change at T_u as the van't Hoff enthalpy, and the slope of the folded fraction at T_u is proportional to van't Hoff enthalpy shown in Eqn. (5.13).

$$\begin{aligned}
&\frac{\partial\left[\Delta G_u^\circ(T)/T\right]}{\partial T} = -\frac{\Delta H_u^\circ(T_u)}{T^2} \\
&\left.\frac{\partial K_u(T)}{\partial T}\right|_{T=T_u} = -\frac{K_u(T_u)}{R}\left.\frac{\partial\left[\Delta G_u^\circ(T)/T\right]}{\partial T}\right|_{T=T_u} = \frac{\Delta H_u^\circ(T_u)}{RT_u^2} \\
&\left.\frac{\partial\theta_F(T)}{\partial T}\right|_{T=T_u} = \frac{-1}{4}\left.\frac{\partial K_u(T)}{\partial T}\right|_{T=T_u} = \frac{-\Delta H_u^\circ(T_u)}{4RT_u^2}
\end{aligned} \tag{5.13}$$

One can in principle use Eqns. (5.4) and (5.12) to determine unknown thermodynamic parameters including T_u, $\Delta H_u^\circ(T_u)$, $\Delta C_{P,u}^\circ$ from an experimental melting curve. In practice, the optical melting curve is usually obtained by picking a frequency slice at a function of temperature or by performing singular value decomposition (SVD) over a frequency range and taking the second component, which reports the dominant changes across the temperature. It is most likely

convolved with additional spectral response not related to the real folding-unfolding transition like slanting baseline(s), which can involve a combination of change of solvation environment due to temperature change, path length variation, and sample evaporation.[4] So in a fitting to a single unfolding curve $I(T)$, one may take the following form for a two-state model fit:

$$I(T) = \left[\left(m_{\mathrm{F}} T - b_{\mathrm{F}} \right) - \left(m_{\mathrm{U}} T - b_{\mathrm{U}} \right) \right] \theta_{\mathrm{F}}(T) + m_{\mathrm{U}} T - b_{U} \tag{5.14}$$

For the fitting, there are 4 additional slant baseline parameters including slope of the species *i*, m_i, and the intercept b_i, in total of 7 fit parameters to a single unfolding curve.

5.2.2 Ionic Strength Dependence of Two-State Unfolding Processes

It is not uncommon to have ionic strength dependence on protein unfolding due to electrostatic screening, some ion-specific effects, *etc*. It is informative to characterize the ionic strength dependence on the thermodynamics of two-state unfolding such that one can understand how ions mediate protein unfolding and use ion parameters to perturb populations of folded and unfolded species. Given the two-state unfolding process in Eqn. (5.1), the simplest possible effect accounting for the ionic strength dependence is the electrostatic screening from Debye-Hückel theory.[5] In the limit of Debye-Hückel regime, electrostatic interactions are effectively screened by surrounding distribution of ions, which consequently depend on the squared root of ionic strength $\sqrt{I}$. Assuming that the electrostatic screening is the dominant effect of perturbing the folding-unfolding equilibrium, which has been observed in a few protein systems,[6-7] the corresponding equilibrium unfolding constant can be given in terms of activity coefficients as

$$K_{\mathrm{u}} = \frac{\gamma_{\mathrm{U}} \left[\mathrm{U} \right]}{\gamma_{\mathrm{F}} \left[\mathrm{F} \right]} = K_{\mathrm{u}} \left(I = 0 \right) \frac{\gamma_{\mathrm{U}}}{\gamma_{\mathrm{F}}} \tag{5.15}$$

$$\Delta G_{\mathrm{u}}^{\circ}(I)=\Delta G_{\mathrm{u}}^{\circ}(I=0)-RT\ln\left(\gamma_{\mathrm{U}}/\gamma_{\mathrm{F}}\right)\equiv\Delta G_{\mathrm{u},0}^{\circ}-RT\ln\left(\gamma_{\mathrm{U}}/\gamma_{\mathrm{F}}\right) \tag{5.16}$$

Debye-Hückel limiting law shows that the natural log of activity coefficient γ is proportional to the square root of the ionic strength. Therefore, the free energy change of unfolding can be written as[6-7]

$$\Delta G_{\mathrm{u}}^{\circ}(I)=\Delta G_{\mathrm{u}}^{\circ}(I=0)+m\sqrt{I}\equiv\Delta G_{\mathrm{u},0}^{\circ}+m\sqrt{I} \tag{5.17}$$

In Eqn. (5.17), $\Delta G_{\mathrm{u},0}^{\circ}$ corresponds to the free energy change at zero ionic strength in kJ/mol, and m is the slope to the square root of ionic strength in unit of kJ/mol.M$^{0.5}$. One can find the relationship between the square root of ionic strength and the fraction of folded state as

$$\theta_{\mathrm{F}}(I)=\frac{1}{1+K_{\mathrm{u}}}=\frac{1}{1+\exp\left(-\Delta G_{\mathrm{u}}^{\circ}(I)/RT\right)} \tag{5.18}$$

$$\theta_{\mathrm{U}}(I)=\frac{K_{\mathrm{u}}}{1+K_{\mathrm{u}}}=\frac{\exp\left(-\Delta G_{\mathrm{u}}^{\circ}(I)/RT\right)}{1+\exp\left(-\Delta G_{\mathrm{u}}^{\circ}(I)/RT\right)} \tag{5.19}$$

These equations indicate that the larger the m value is, the steeper the two-state transition is across the ionic strength. Rearranging Eqns. (5.18) and (5.19) gives

$$\begin{aligned}\ln K_{\mathrm{u}}&=\ln\left(\frac{1-\theta_{\mathrm{F}}(I)}{\theta_{\mathrm{F}}(I)}\right)=\frac{-\Delta G_{\mathrm{u},0}^{\circ}}{RT}-\frac{m\sqrt{I}}{RT}\\ \ln\left(\frac{1}{K_{\mathrm{u}}}\right)&=\ln\left(\frac{1-\theta_{\mathrm{U}}(I)}{\theta_{\mathrm{U}}(I)}\right)=\frac{\Delta G_{\mathrm{u},0}^{\circ}}{RT}+\frac{m\sqrt{I}}{RT}\end{aligned} \tag{5.20}$$

These provide a set of linearized equations, relating the intercept of the fractions with the free energy change at zero ionic strength, and the slope associated with the squared root of the ionic

strength. These two parameters $\Delta G_{\mathrm{u},0}^{\circ}$ and *m* can be used for a fitting to a two-state equilibrium as a function of ionic strength.

Another model to account for the ionic strength dependence is to treat the ion as a chemical denaturant like urea or GdmCl. Then the free energy change can be written based on *m*-value analysis as[2]

$$\Delta G_{\mathrm{u}}^{\circ}\left(I\right)=\Delta G_{\mathrm{u}}^{\circ}\left(I=0\right)+mI\equiv\Delta G_{\mathrm{u},0}^{\circ}+mI \tag{5.21}$$

Analogous to the treatment above,

$$\begin{aligned}\ln K_{\mathrm{u}}&=\ln\left(\frac{1-\theta_{\mathrm{F}}\left(I\right)}{\theta_{\mathrm{D}_1}\left(I\right)}\right)=\frac{-\Delta G_{\mathrm{u},0}^{\circ}}{RT}-\frac{mI}{RT}\\ \ln\left(\frac{1}{K_{\mathrm{u}}}\right)&=\ln\left(\frac{1-\theta_{\mathrm{U}}\left(I\right)}{\theta_{\mathrm{U}}\left(I\right)}\right)=\frac{\Delta G_{\mathrm{u},0}^{\circ}}{RT}+\frac{mI}{RT}\end{aligned} \tag{5.22}$$

Fitting this model will have exactly the same parameters as in the fitting of the model with electrostatic screening, but the dependence of ionic strength is linear instead of the squared root. Examining the dependence of ionic strength using Eqns. (5.20) and (5.22) can help identify which model describes the experimental data better and hence the underlying ion dependence.

5.2.3 Two-State Thermodynamic Model of Homodimer Dissociation

Throughout the thesis, we will focus on homodimer dissociation of insulin, and the simplest possible scheme is two-state dissociation as follows.

$$\mathrm{D}\overset{K_{\mathrm{d}}}{\rightleftharpoons}2\mathrm{M} \tag{5.23}$$

Similar to the treatment as in section 5.2.1, we can derive the relation of fraction of each species to underlying thermodynamic parameters, and use them for fitting to optical dissociation curves. Given the total concentration $C_{\mathrm{tot}} = 2[\mathrm{D}]+[\mathrm{M}]$, we can define fractions of dimer and monomer as θ_{D} and θ_{M} such that $\theta_{\mathrm{D}}+\theta_{\mathrm{M}}=1$. The expression of these fractions are given as

$$\begin{aligned}\theta_{\mathrm{D}} &= \frac{2[\mathrm{D}]}{C_{\mathrm{tot}}} \\ \theta_{\mathrm{M}} &= 1-\theta_{\mathrm{D}} = \frac{[\mathrm{M}]}{C_{\mathrm{tot}}}\end{aligned} \tag{5.24}$$

We can also express the dissociation constant in terms of the dimer fraction as follows.

$$K_{\mathrm{d}} = \frac{[\mathrm{M}]^2}{[\mathrm{D}]} = \frac{2(1-\theta_{\mathrm{D}})^2 C_{\mathrm{tot}}}{\theta_{\mathrm{D}}} \tag{5.25}$$

Or equivalently express the dimer fraction and the monomer fraction in terms of K_{d} as

$$\theta_{\mathrm{D}} = \frac{1}{4C_{\mathrm{tot}}}\left[4C_{\mathrm{tot}} + K_{\mathrm{d}} - \sqrt{K_{\mathrm{d}}^2 + 8C_{\mathrm{tot}}K_{\mathrm{d}}}\right] \tag{5.26}$$

$$\theta_{\mathrm{M}} = 1-\theta_{\mathrm{D}} = \frac{1}{4C_{\mathrm{tot}}}\left[-K_{\mathrm{d}} + \sqrt{K_{\mathrm{d}}^2 + 8C_{\mathrm{tot}}K_{\mathrm{d}}}\right] \tag{5.27}$$

The dissociation constant can be written similarly in terms of the dissociation temperature T_{d}, dissociation enthalpy $\Delta H_{\mathrm{d}}^{\circ}(T_{\mathrm{d}})$, and change of heat capacity $\Delta C_{p,\mathrm{d}}^{\circ}$ as follows.

$$\begin{aligned}K_{\mathrm{d}}(T) &= \exp\left(-\frac{\Delta G_{\mathrm{d}}^{\circ}(T)}{RT}\right) \\ \Delta G_{\mathrm{d}}^{\circ}(T) &= \Delta H_{\mathrm{d}}^{\circ}(T_{\mathrm{d}})\left[1-\frac{T}{T_{\mathrm{d}}}\right] - \Delta C_P^{\circ}\left[T - T_{\mathrm{d}} - T\ln\left(\frac{T}{T_{\mathrm{d}}}\right)\right] + RT\ln(C_{\mathrm{tot}})\end{aligned} \tag{5.28}$$

The dependence of the total concentration C_{tot} originates from mixing entropy assuming ideal solution. In this thermodynamic two-state model, T_{d} is the dissociation temperature defined at $\theta_{\text{D}} = 0.5$, or equivalently, $K_{\text{d}} = C_{\text{tot}}$. The corresponding free energy at dissociation temperature is $\Delta G_{\text{d}}^{\circ}(T_{\text{d}}) = -RT_{\text{d}} \ln(C_{\text{tot}})$ that can be seen from Eqn. (5.28). The slope of the dimer fraction at T_{d} is proportional to the van't Hoff enthalpy shown in Eqn. (5.29).

$$\begin{aligned}
&\frac{\partial\left[\Delta G_{\text{d}}^{\circ}(T)/T\right]}{\partial T} = -\frac{\Delta H_{\text{d}}^{\circ}(T_{\text{d}})}{T^2} \\
&\left.\frac{\partial K_{\text{d}}(T)}{\partial T}\right|_{T=T_{\text{d}}} = \frac{K_{\text{d}}(T_{\text{d}})}{R}\left.\frac{\partial\left[\Delta G_{\text{d}}^{\circ}(T)/T\right]}{\partial T}\right|_{T=T_{\text{d}}} = \frac{C_{\text{tot}}}{RT_{\text{d}}^2}\Delta H_{\text{d}}^{\circ}(T_{\text{d}}) \\
&\left.\frac{\partial \theta_{\text{D}}(T)}{\partial T}\right|_{T=T_{\text{d}}} = \frac{-1}{6C_{\text{tot}}}\left.\frac{\partial K_{\text{d}}(T)}{\partial T}\right|_{T=T_{\text{d}}} = \frac{-\Delta H_{\text{d}}^{\circ}(T_{\text{d}})}{6RT_{\text{d}}^2}
\end{aligned} \tag{5.29}$$

For the fitting to a dissociation curve $I(T)$, one may take the following form for a two-state model fit:

$$I(T) = \left[\left(m_{\text{D}}T - b_{\text{D}}\right) - \left(m_{\text{M}}T - b_{\text{M}}\right)\right]\theta_{\text{D}}(T) + m_{\text{M}}T - b_{\text{M}} \tag{5.30}$$

5.2.4 Estimation of Uncertainties in the Thermodynamic Two-State Models

To estimate uncertainties of the fit parameters and to investigate how reliably a two-state model can be used for interpreting populations, we can apply error propagation and estimate the uncertainty of model results assuming each fit parameter is independent with each other. Note that this analysis provides the lower bound estimate of the uncertainty since we do not account for different types of noise present in the experiments and errors in determining the baselines. Instead, we only use the uncertainties from the fit.

For the two-state unfolding model, given the uncertainties of the unfolding temperature σ_{T_u}, unfolding enthalpy $\sigma_{\Delta H_u^\circ}(T_u)$, and change of heat capacity $\sigma_{\Delta C_{P,u}^\circ}$, one can express the uncertainty of a thermodynamic function *f*(*T*) in the model as

$$\sigma_f(T) = \sqrt{\left[\frac{\partial f(T)}{\partial \Delta H_u^\circ}\right]^2 \sigma^2_{\Delta H_u^\circ}(T_u) + \left[\frac{\partial f(T)}{\partial \Delta C_{P,u}^\circ}\right]^2 \sigma^2_{\Delta C_{P,u}^\circ} + \left[\frac{\partial f(T)}{\partial T_u}\right]^2 \sigma^2_{T_u}} \tag{5.31}$$

Using Eqn. (5.31), the uncertainties of $\Delta H_u^\circ(T)$, $\Delta S_u^\circ(T)$, $\Delta G_u^\circ(T)$, $K_u(T)$, and $\theta_F(T)$ are given as follows.

$$\sigma_{\Delta H_u^\circ}(T) = \sqrt{\sigma^2_{\Delta H_u^\circ}(T_u) + (T - T_u)^2 \sigma^2_{\Delta C_{P,u}^\circ} + \left(\Delta C_{P,u}^\circ\right)^2 \sigma^2_{T_u}} \tag{5.32}$$

$$\sigma_{\Delta S_u^\circ}(T) = \sqrt{\frac{1}{T_u^2}\sigma^2_{\Delta H_u^\circ}(T_u) + \left[\ln\left(\frac{T}{T_u}\right)\right]^2 \sigma^2_{\Delta C_{P,u}^2} + \left\{\frac{\left[\Delta H_u^\circ(T_u)/T_u\right]^2 + \Delta C_{P,u}^{\circ\,2}}{T_u^2}\right\}\sigma^2_{T_u}} \tag{5.33}$$

$$\sigma_{\Delta G_u^\circ}(T) = \sqrt{\sigma^2_{\Delta H_u^\circ}(T) + T^2\sigma^2_{\Delta S_u^\circ}(T)} \tag{5.34}$$

$$\sigma_{K_u}(T) = \frac{K_u(T)}{RT}\sigma_{\Delta G_u^\circ}(T) \tag{5.35}$$

$$\sigma_{\theta_F}(T) = \frac{\sigma_{K_u}(T)}{\left[1 + K_u(T)\right]^2} \tag{5.36}$$

These uncertainty estimates can be obtained once we know σ_{T_u}, $\sigma_{\Delta H_u^\circ}(T_u)$, $\sigma_{\Delta C_{P,u}^\circ}$, which is estimated from the fitting. Note that the uncertainties from the fit should be the lower bound of the uncertainty.

Similarly, we can derive the uncertainty of thermodynamic functions in the two-state model of dimer dissociation.

$$\sigma_{\Delta H_{\mathrm{d}}^{\circ}}(T)=\sqrt{\sigma_{\Delta H_{\mathrm{d}}^{\circ}}^{2}\left(T_{\mathrm{d}}\right)+\left(T-T_{\mathrm{d}}\right)^{2}\sigma_{\Delta C_{P}^{\circ}}^{2}+\left(\Delta C_{P}^{\circ}\right)^{2}\sigma_{T_{\mathrm{d}}}^{2}} \tag{5.37}$$

$$\sigma_{\Delta S_{\mathrm{d}}^{\circ}}(T)=\sqrt{\frac{1}{T_{\mathrm{d}}^{2}}\sigma_{\Delta H_{\mathrm{d}}^{\circ}}^{2}\left(T_{\mathrm{d}}\right)+\left[\ln\left(\frac{T}{T_{\mathrm{d}}}\right)\right]^{2}\sigma_{\Delta C_{P}^{2}}^{2}+\left\{\frac{\left[\Delta H_{\mathrm{d}}^{\circ}\left(T_{\mathrm{d}}\right)/T_{\mathrm{d}}\right]^{2}+\Delta C_{P}^{\circ\,2}}{T_{\mathrm{d}}^{2}}\right\}\sigma_{T_{\mathrm{d}}}^{2}} \tag{5.38}$$

$$\sigma_{\Delta G_{\mathrm{d}}^{\circ}}\left(T\right)=\sqrt{\sigma_{\Delta H_{\mathrm{d}}^{\circ}}^{2}(T)+T^{2}\sigma_{\Delta S_{\mathrm{d}}^{\circ}}^{2}\left(T\right)} \tag{5.39}$$

$$\sigma_{K_{\mathrm{d}}}\left(T\right)=\frac{K_{\mathrm{d}}\left(T\right)}{RT}\sigma_{\Delta G_{\mathrm{d}}^{\circ}}\left(T\right) \tag{5.40}$$

$$\sigma_{\theta_{\mathrm{D}}}\left(T\right)=\frac{1}{4C_{\mathrm{tot}}}\left[1-\frac{K_{\mathrm{d}}(T)+4C_{\mathrm{tot}}}{\sqrt{K_{\mathrm{d}}^{2}\left(T\right)+8C_{\mathrm{tot}}K_{\mathrm{d}}\left(T\right)}}\right]\sigma_{K_{\mathrm{d}}}\left(T\right) \tag{5.41}$$

Note that the only difference between two-state unfolding and two-state dissociation is coming from the expression of the fractions as in Eqns. (5.4) and (5.26).

5.2.5 Sequential Three-State Thermodynamic Models of Dimer Unfolding and Dissociation

The simplest scheme to account for an additional unfolded dimer state into the previous described thermodynamic two-state model is the following sequential three-state model.

$$\mathrm{D}_{1}\overset{K_{\mathrm{u}}}{\rightleftharpoons}\mathrm{D}_{2}\overset{K_{\mathrm{d}}}{\rightleftharpoons}2\mathrm{M} \tag{5.42}$$

In this scheme, the folded dimer state D_1 has to undergo unfolding to the unfolded dimer state D_2, and then further dissociate into two monomers. Given this scheme, equilibrium constant of unfolding process and dissociation process can be written as

$$K_{\mathrm{u}}(T)=\frac{[\mathrm{D}_2]}{[\mathrm{D}_1]}=\exp\left(-\frac{\Delta G_{\mathrm{u}}^{\circ}(T)}{RT}\right) \tag{5.43}$$

$$K_{\mathrm{d}}(T)=\frac{[\mathrm{M}]^2}{[\mathrm{D}_2]}=\exp\left(-\frac{\Delta G_{\mathrm{d}}^{\circ}(T)}{RT}\right) \tag{5.44}$$

Similarly using the argument made in Eqns. (5.10) and (5.11), one can decompose temperature dependence of the free energy change into enthalpic contribution and entropic contribution in Eqns. (5.45) and (5.46).

$$\Delta G_{\mathrm{u}}^{\circ}(T)=\Delta H_{\mathrm{u}}^{\circ}(T_{\mathrm{u}})\left[1-\frac{T}{T_{\mathrm{u}}}\right]+\Delta C_{P,\mathrm{u}}^{\circ}\left[T-T_{\mathrm{u}}-T\ln\left(\frac{T}{T_{\mathrm{u}}}\right)\right] \tag{5.45}$$

$$\Delta G_{\mathrm{d}}^{\circ}(T)=\Delta H_{\mathrm{d}}^{\circ}(T_{\mathrm{d}})\left[1-\frac{T}{T_{\mathrm{d}}}\right]+\Delta C_{P,\mathrm{d}}^{\circ}\left[T-T_{\mathrm{d}}-T\ln\left(\frac{T}{T_{\mathrm{d}}}\right)\right]-RT\ln(C_{\mathrm{tot}}) \tag{5.46}$$

In Eqn. (5.45), the unfolding temperature T_{u} is defined as the temperature at which $[\mathrm{D}_1]=[\mathrm{D}_2]$, and the dissociation temperature T_{d} in Eqn. (5.46) is defined as the temperature where $\Delta G_{\mathrm{d}}^{\circ}(T_{\mathrm{d}})=-RT\ln C_{\mathrm{tot}}$ or $K_{\mathrm{d}}(T_{\mathrm{d}})=C_{\mathrm{tot}}$. However, one difference from the previous two-state model of dimer dissociation is that the monomer fraction is not guaranteed to be 0.5 at T_{d}. Given the total concentration of $C_{\mathrm{tot}}=2[\mathrm{D}_1]+2[\mathrm{D}_2]+[\mathrm{M}]$ and Eqns. (5.43) and (5.44), one can write down the fraction of each species as

$$\theta_{\mathrm{M}}(T)=\frac{[\mathrm{M}]}{C_{\mathrm{tot}}}=1-\theta_{\mathrm{D}_1}(T)-\theta_{\mathrm{D}_2}(T)$$
$$\theta_{\mathrm{D}_2}(T)=\frac{2[\mathrm{D}_2]}{C_{\mathrm{tot}}}=\frac{2C_{\mathrm{tot}}}{K_{\mathrm{d}}(T)}\left(\frac{[\mathrm{M}]^2}{C_{\mathrm{tot}}^2}\right)=\frac{2C_{\mathrm{tot}}}{K_{\mathrm{d}}(T)}\theta_{\mathrm{M}}^2(T) \tag{5.47}$$
$$\theta_{\mathrm{D}_1}(T)=\frac{2[\mathrm{D}_1]}{C_{\mathrm{tot}}}=\frac{1}{K_{\mathrm{u}}(T)}\theta_{\mathrm{D}_2}(T)=\frac{2C_{\mathrm{tot}}}{K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)}\theta_{\mathrm{M}}^2(T)$$

Trying to solve Eqn. (5.47) to obtain the monomer fraction as a function of temperature, one can get the following quadratic equation.

$$2C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]\theta_{\mathrm{M}}^2(T)+K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)\theta_{\mathrm{M}}(T)-K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)=0 \tag{5.48}$$

The corresponding monomer fraction from Eqn. (5.48) can be solved as

$$\theta_{\mathrm{M}}(T)=\frac{-K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)+\sqrt{K_{\mathrm{u}}^2(T)K_{\mathrm{d}}^2(T)+8C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)}}{4C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]} \tag{5.49}$$

One can verify that in the limit of $K_{\mathrm{u}}(T)=0$, there is no monomer fraction at all since there is no way for the folded dimer to dissociate into the monomer. In contrast, in the limit of $K_u(T)\to\infty$, or $[\mathrm{D}_1]\approx 0, \theta_{\mathrm{D}_1}\approx 0$. The expression of Eqn. (5.49) can be further reduced into a two-state picture given in Eqns. (5.26) and (5.27).

$$\theta_{\mathrm{M}}(T)\approx\frac{1}{4C_{\mathrm{tot}}}\left[-K_{\mathrm{d}}(T)+\sqrt{K_{\mathrm{d}}^2(T)+8C_{\mathrm{tot}}K_{\mathrm{d}}(T)}\right] \tag{5.50}$$

$$\theta_{\mathrm{D}_2}(T)=1-\theta_{\mathrm{M}}(T)-\theta_{\mathrm{D}_1}(T)\approx\frac{1}{4C_{\mathrm{tot}}}\left[4C_{\mathrm{tot}}-K_{\mathrm{d}}(T)-\sqrt{K_{\mathrm{d}}^2(T)+8C_{\mathrm{tot}}K_{\mathrm{d}}(T)}\right] \tag{5.51}$$

In this regime, it is simply the two-state dissociation between [D_2] and [M].

On the other hand, if one can find a regime where $[D_2] \approx 0$, then it will look also like an apparent two-state equilibrium between $[D_1]$ and [M]. For example, when $K_u(T) \ll 1$, but it is not equal to zero, then

$$\theta_M(T) \approx \frac{-K_u(T)K_d(T) + \sqrt{K_u^2(T)K_d^2(T) + 8C_{tot}K_u(T)K_d(T)}}{4C_{tot}} \tag{5.52}$$

By defining an apparent equilibrium constant $K_{d1}(T) = K_u(T)K_d(T)$

$$\theta_M(T) \approx \frac{-K_{d1}(T) + \sqrt{K_{d1}^2(T) + 8C_{tot}K_{d1}^2(T)}}{4C_{tot}} \tag{5.53}$$

Since $K_u \ll 1$, the unfolded dimer fraction is negligible, and the fraction of folded dimer will be approximated as

$$\theta_{D_1}(T) = 1 - \theta_M(T) - \theta_{D_2}(T) \approx \frac{1}{4C_{tot}}\left[4C_{tot} - K_{d1}^2(T) - \sqrt{K_{d1}^2(T) + 8C_{tot}K_{d1}(T)}\right] \tag{5.54}$$

We can conclude that as long as one of the dimer species has an extremely low population, the result will always look like a two-state model, but the apparent equilibrium constant depends on which species is the intermediate state.

We can further discuss some other behaviors of the model and obtain some insights, which may be helpful to interpret with experiments. At T_d, the monomer fraction from Eqn. (5.49) can be simplified as

$$\theta_M(T_d) = \frac{-K_u(T_d) + \sqrt{9K_u^2(T_d) + 8K_u(T_d)}}{4\left[1 + K_u(T_d)\right]} \tag{5.55}$$

Consistently with the previous discussion, when $K_{\mathrm{u}}(T) \to \infty$, the monomer fraction becomes 0.5, suggesting that

$$\theta_{\mathrm{M}}(T_{\mathrm{d}}) \le \frac{1}{2} = \theta_{\mathrm{M}}^{\text{2-state}}(T_{\mathrm{d}}) \tag{5.56}$$

In other words, when the dissociation temperatures are the same in both 2-state model and this 3-state unfolding-dissociation model, it is guaranteed the three-state monomer fraction is lower or equal to 0.5. Interestingly, when $K_{\mathrm{u}}(T_{\mathrm{d}}) = 1/5$, the corresponding monomer fraction becomes

$$\theta_{\mathrm{M}}(T_{\mathrm{d}}; K_{\mathrm{d}} = 0.2) = \frac{1}{4} = \frac{1}{2}\theta_{\mathrm{M}}^{\text{2-state}}(T_{\mathrm{d}}) \tag{5.57}$$

Also when $K_{\mathrm{d}}(T) \gg C_{\mathrm{tot}}$ in a situation of T well above T_{d}, expanding $C_{\mathrm{tot}} / K_{\mathrm{d}}(T)$ to second order in Eqn. (5.49) gives

$$\theta_{\mathrm{M}}^{(2)}(T) = 1 - \frac{2C_{\mathrm{tot}}\left[1 + K_{\mathrm{d}}(T)\right]}{K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)} \tag{5.58}$$

In this expansion with the superscript (2) specifying the second-ordered solution, the zeroth-ordered term cancels with $-K_{\mathrm{u}}(T)K_{\mathrm{d}}(T)$ in the Eqn. (5.49), and the first ordered term gives 1 in the Eqn. (5.58). An equivalent and consistent way of deriving Eqn. (5.58) can be done. One can impose the first ordered solution of $\theta_{\mathrm{M}}^{(1)}(T) = 1$. Applying Eqn. (5.47) to obtain $\theta_{\mathrm{D}_1}^{(1)}(T)$ and $\theta_{\mathrm{D}_2}^{(1)}(T)$ gives the form for second-ordered correction on the monomer fraction, which is essentially $-\left[\theta_{\mathrm{D}_1}^{(1)}(T) + \theta_{\mathrm{D}_2}^{(1)}(T)\right]$, and the coefficient of $\theta_{\mathrm{M}}(T)$ matches with the expression in Eqn. (5.58).

From the discussions above, there are some circumstances that the three-state model may look like 2-state model experimentally. First, if one of the dimer species is negligible, it will always reduce to a thermodynamic two-state model. Secondly, if the experimental probe actually cannot distinguish the dimer species such that it is measuring the coarse-grained D state where $[D]=[D_1]+[D_2]$, then it will also reduce to a thermodynamic two-state model.

5.2.6 Parallel Three-State Thermodynamic Model of Dimer Unfolding and Dissociation

For coupled-folding and binding processes, it is not uncommon that there are many association and dissociation pathways. Discussing such a model may be helpful for us to understand the effect of parallel pathways on the thermodynamics. The reaction scheme is written as

$$\begin{aligned} D_1 &\overset{K_u}{\rightleftharpoons} D_2 \\ D_2 &\overset{K_{d2}}{\rightleftharpoons} 2M \\ D_1 &\overset{K_{d1}}{\rightleftharpoons} 2M \end{aligned} \tag{5.59}$$

The difference between this model and the previous three-state model is the additional pathway of dissociating D_1 into two monomers without going through D_2. The equilibrium constants for each step are given as

$$K_{\mathrm{u}}(T)=\frac{[\mathrm{D}_1]}{[\mathrm{D}_2]}=\exp\left(-\frac{\Delta G_{\mathrm{u}}^{\circ}(T)}{RT}\right)$$
$$K_{\mathrm{d1}}(T)=\frac{[\mathrm{M}]^2}{[\mathrm{D}_1]}=\exp\left(-\frac{\Delta G_{\mathrm{d1}}^{\circ}(T)}{RT}\right) \tag{5.60}$$
$$K_{\mathrm{d2}}(T)=\frac{[\mathrm{M}]^2}{[\mathrm{D}_2]}=\exp\left(-\frac{\Delta G_{\mathrm{d2}}^{\circ}(T)}{RT}\right)$$

Similarly, one can write down free energy changes of each process as

$$\Delta G_{\mathrm{u}}^{\circ}(T)=\Delta H_{\mathrm{u}}^{\circ}(T_{\mathrm{u}})\left[1-\frac{T}{T_{\mathrm{u}}}\right]+\Delta C_{P,\mathrm{u}}^{\circ}\left[T-T_{\mathrm{u}}-T\ln\left(\frac{T}{T_{\mathrm{u}}}\right)\right]$$
$$\Delta G_{\mathrm{d1}}^{\circ}(T)=\Delta H_{\mathrm{d1}}^{\circ}(T_{\mathrm{d1}})\left[1-\frac{T}{T_{\mathrm{d1}}}\right]+\Delta C_{P,\mathrm{d1}}^{\circ}\left[T-T_{\mathrm{d1}}-T\ln\left(\frac{T}{T_{\mathrm{d1}}}\right)\right]-RT\ln(C_{\mathrm{tot}}) \tag{5.61}$$
$$\Delta G_{\mathrm{d2}}^{\circ}(T)=\Delta H_{\mathrm{d2}}^{\circ}(T_{\mathrm{d2}})\left[1-\frac{T}{T_{\mathrm{d2}}}\right]+\Delta C_{P,\mathrm{d2}}^{\circ}\left[T-T_{\mathrm{d2}}-T\ln\left(\frac{T}{T_{\mathrm{d2}}}\right)\right]-RT\ln(C_{\mathrm{tot}})$$

In Eqn. (5.61), the unfolding temperature T_{u} is defined as the temperature at which $[\mathrm{D}_1]=[\mathrm{D}_2]$, the same way as before. The dissociation temperature of D_i are defined as the temperature where $\Delta G_{\mathrm{d}i}^{\circ}(T_{\mathrm{d}i})=-RT_{\mathrm{d}i}\ln C_{\mathrm{tot}}$ or $K_{\mathrm{d}i}(T_{\mathrm{d}i})=C_{\mathrm{tot}}$, which is consistent with the way defined in the previous thermodynamic models.

Note that the equilibrium constants in this model have one more stringent constraint than the previous unfolding-dissociation model. At any temperature at equilibrium, the following equation has to be true, which is essentially the detailed balance condition.

$$\frac{[\mathrm{D}_2]}{[\mathrm{D}_1]}\times\frac{[\mathrm{M}]^2}{[\mathrm{D}_2]}\times\frac{[\mathrm{D}_1]}{[\mathrm{M}]^2}=K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T)K_{\mathrm{d1}}^{-1}(T)=1 \tag{5.62}$$

$$K_{\mathrm{d1}}(T)=K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T) \tag{5.63}$$

Or equivalently,

$$\Delta G_{\mathrm{d1}}^{\circ}(T) = \Delta G_{\mathrm{u}}^{\circ}(T) + \Delta G_{\mathrm{d2}}^{\circ}(T) \tag{5.64}$$

A rather strong conclusion from Eqn. (5.63) is that if the two dissociation equilibrium have the same values of dissociation constants at a given temperature T, then the unfolding equilibrium constant must be 1, meaning that $[\mathrm{D}_1] = [\mathrm{D}_2]$.

Expressions for the fractions of each species should be exactly the same as Eqn. (5.47), with the additional constraint described above.

$$\begin{aligned}
\theta_{\mathrm{M}}(T) &= \frac{[\mathrm{M}]}{C_{\mathrm{tot}}} = 1 - \theta_{\mathrm{D}_1}(T) - \theta_{\mathrm{D}_2}(T) \\
\theta_{\mathrm{D}_2}(T) &= \frac{2[\mathrm{D}_2]}{C_{\mathrm{tot}}} = \frac{2C_{\mathrm{tot}}}{K_{\mathrm{d2}}(T)}\left(\frac{[\mathrm{M}]^2}{C_{\mathrm{tot}}^2}\right) = \frac{2C_{\mathrm{tot}}}{K_{\mathrm{d2}}(T)}\theta_{\mathrm{M}}^2(T) \\
\theta_{\mathrm{D}_1}(T) &= \frac{2[\mathrm{D}_1]}{C_{\mathrm{tot}}} = \frac{1}{K_{\mathrm{u}}(T)}\theta_{\mathrm{D}_2}(T) = \frac{2C_{\mathrm{tot}}}{K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T)}\theta_{\mathrm{M}}^2(T) = \frac{2C_{\mathrm{tot}}}{K_{\mathrm{d1}}(T)}\theta_{\mathrm{M}}^2(T)
\end{aligned} \tag{5.65}$$

$$\theta_{\mathrm{M}}(T) = \frac{-K_{\mathrm{d1}}(T) + \sqrt{K_{\mathrm{d1}}^2(T) + 8C_{\mathrm{tot}}\left[1 + K_{\mathrm{d1}}(T)K_{\mathrm{d2}}^{-1}(T)\right]K_{\mathrm{d1}}(T)}}{4C_{\mathrm{tot}}\left[1 + K_{\mathrm{d1}}(T)K_{\mathrm{d2}}^{-1}(T)\right]} \tag{5.66}$$

Hence, the fraction of each species given the same $K_{\mathrm{u}}(T)$ and $K_{\mathrm{d2}}(T)$ will behave exactly the same as the previous three-state model without additional dissociation between D_1 and 2M. Even though the expressions are seemingly the same, there are some difference on the limiting behaviors, $K_{\mathrm{u}}(T)$ in particular. When $K_{\mathrm{u}}(T) = 0$, the mechanism actually reduces back to the thermodynamic two-state model instead of having zero monomer fraction. Note that $K_{\mathrm{d1}}(T) = 0 \times K_{\mathrm{d2}}(T)$, so they are decoupled. Therefore, the limiting behavior between these two thermodynamic three-state models is mostly the same except for the scenario of $K_{\mathrm{u}}(T) = 0$. When the thermodynamic equilibrium

of folding and unfolding is biased toward one dominant species, thermodynamic behavior of dissociation becomes closer to the two-state dissociation between the dominant species and the monomer.

One interesting question is that how does the apparent two-state thermodynamics such as T_d behaves when the two dissociation pathways have different thermodynamics? Let us assume experimentally one cannot resolve D_1 and D_2 properly or one lumps the signal contributions from the two dimer species into construction of an optical dissociation curve, such that the apparent population is reporting [D_1] + [D_2]. Then,

$$\begin{aligned}\theta_{\mathrm{M}}(T) &= \frac{[\mathrm{M}]}{C_{\mathrm{tot}}} = 1-\left[\theta_{\mathrm{D}_1}(T)+\theta_{\mathrm{D}_2}(T)\right] = 1-\theta_{\mathrm{D}'}(T) \\ \theta_{\mathrm{D}'}(T) &\equiv \theta_{\mathrm{D}_1}(T)+\theta_{\mathrm{D}_2}(T) = \frac{2C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]}{K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T)}\theta_{\mathrm{M}}^2(T)\end{aligned} \tag{5.67}$$

From Eqn. (5.67), one can derive the following quadratic relation for the apparent dimer fraction

$$\begin{aligned}&\alpha(T)\theta_{\mathrm{D}'}^2(T)-\left[2\alpha(T)+1\right]\theta_{\mathrm{D}'}(T)+\alpha(T)=0 \\ &\alpha(T)\equiv\frac{2C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]}{K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T)}=\frac{2C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]}{K_{\mathrm{d1}}(T)}\end{aligned} \tag{5.68}$$

Then the solution of the apparent dimer fraction can be calculated as Eqn. (5.69) using $K_{\mathrm{d1}}(T)=K_{\mathrm{u}}(T)K_{\mathrm{d2}}(T)$

$$\theta_{\mathrm{D}'}(T)=\frac{K_{\mathrm{d1}}(T)}{4C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]}\left\{\frac{4C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]}{K_{\mathrm{d1}}(T)}+1-\sqrt{\frac{8C_{\mathrm{tot}}\left[1+K_{\mathrm{u}}(T)\right]+K_{\mathrm{d1}}(T)}{K_{\mathrm{d1}}(T)}}\right\} \tag{5.69}$$

When $K_{\mathrm{u}}(T)=0$, the apparent dimer fraction expectedly becomes

$$\theta_{\mathrm{D}'}(T)=\frac{1}{4C_{\mathrm{tot}}}\left\{4C_{\mathrm{tot}}+K_{\mathrm{d1}}(T)-\sqrt{K_{\mathrm{d1}}^{2}(T)+8C_{\mathrm{tot}}K_{\mathrm{d1}}(T)}\right\}=\theta_{\mathrm{D}_1}(T) \tag{5.70}$$

The apparent dimer fraction describes the two-state thermodynamic dissociation between D_1 and 2*M*. In the other limit, when $K_{\mathrm{u}}(T)\to\infty$, Eqn. (5.69) approaches

$$\theta_{\mathrm{D}'}(T)=\frac{1}{4C_{\mathrm{tot}}}\left[4C_{\mathrm{tot}}+K_{\mathrm{d2}}(T)-\sqrt{K_{\mathrm{d2}}^{2}(T)+8C_{\mathrm{tot}}K_{\mathrm{d2}}(T)}\right]=\theta_{\mathrm{D}_2}(T) \tag{5.71}$$

Other than those two limits, the apparent dimer fraction depends on the unfolding equilibrium constant between D_1 and D_2, and therefore one cannot expect that the apparent dimer fraction does not change with perturbations to the unfolding equilibrium.

When the apparent dimer fraction is 0.5, which will match the apparent dissociation temperature T_{d} in the framework of the two-state model, using Eqn. (5.69) gives

$$K_{\mathrm{d}'}\left(T_{\mathrm{d}'}\right)\equiv K_{\mathrm{d2}}\left(T_{\mathrm{d}'}\right)\frac{4K_{\mathrm{u}}\left(T_{\mathrm{d}'}\right)}{\left[1+K_{\mathrm{u}}\left(T_{\mathrm{d}'}\right)\right]}=K_{\mathrm{d2}}\left(T_{\mathrm{d}'}\right)\frac{4\theta_{\mathrm{D}_2}\left(T_{\mathrm{d}'}\right)}{\left[\theta_{\mathrm{D}_1}\left(T_{\mathrm{d}'}\right)+\theta_{\mathrm{D}_2}\left(T_{\mathrm{d}'}\right)\right]} \tag{5.72}$$

Interestingly, when $K_u(T)\,\forall T$, $K_{d1}(T)=K_{d2}(T)$, the Eqn. (5.72) gives $K_{d'}(T)=2K_{d2}(T)$. If D_1 and D_2 are the same species, this turns out to be one example of the Gibbs paradox that does not take indistinguishability of D_1 and D_2 into account. If D_1 and D_2 are different species, which is extremely unlikely, it basically means there is one additional pathway with exactly the same thermodynamic behaviors such that the apparent dissociation constant is two times more than the individual dissociation constant. From Eqns. (5.71) and (5.72), the apparent equilibrium constant depends nonlinearly on dimer fractions of D_1 and D_2. As long as the thermodynamic dissociations

between the two dimer states and the monomer behave differently from each other, the apparent equilibrium population will be also perturbed in a non-linear way.

5.3 Maximum Entropy Reconstruction on Thermodynamics

In the previous section, we discussed thermodynamic models for protein folding and homodimer dissociation. Practically the application of these models rely heavily on accurately extracting populations of the species, which is usually hindered by fitting additional baselines. Another practical issue is that typical analysis invokes SVD and uses the second SVD component as the only component that contains all the information, which may not be the most accurate treatment. Additionally, protein can exhibit complex thermodynamic equilibria between multiple chemical species such that imposing a simple thermodynamic model may lead to misinterpretation. Having a systematic and minimally biased way to extract population profiles of species can help alleviate such problem. With this goal in mind, this section covers a different approach that uses maximum entropy (MaxEnt) method for reconstructing pure components with associated populations or weights from a series of spectra as a function of experimental variables. This MaxEnt method does not rely on a specific thermodynamic model or fitting to extract the population. Instead, it provides a framework of minimally biased inference to the underlying probability distribution given the data and additional (physical) constraints.[8]

This MaxEnt method utilizes a combination of SVD to distinguish spectral changes along some experimental variable, and remixing SVD vectors by maximizing an objective function that encodes information entropy and additional physical/chemical constraints. The MaxEnt reconstruction results in a set of spectrally distinct species with associated weight or population across the experimental variable. Chemical applications of the MaxEnt method have been

developed originally for separating pure chemical components or species in a mixture of solution.[9] This method, however, is not limited to separating different pure chemical species in a mixture. As long as there is spectral change associated with the experimental variable, this method can be used to resolve the change and identify different components. For instance, it can be changing the protonation state of Gly–Gly that perturbs the vibrational frequency of the amide I vibration,[10] identifying fraying component within the duplex state of DNA oligonucleotide,[11] titration state of 5-carboxylcytosine,[12] ion-pairing in concentrated nitric acid,[13] or solvation environment around the vibrational probe when elevating temperature with insulin dimer dissociation as an example at the end of the section below.

In this section, we describe relevant techniques that constitute the MaxEnt method, including SVD onto temperature-dependent IR spectroscopy, constructing the objective function for iterative optimizations, and practically one application to identifying temperature behavior of insulin dimer dissociation from temperature-dependent IR spectroscopy that is consistent with temperature-jump IR response. Please note that this MaxEnt reconstruction can be used to other experimental variables such as concentration, pH, ionic strength, *etc*, but for convenience, we use temperature as the experimental variable of interest and discuss the method throughout the section.

5.3.1 Singular Value Decomposition (SVD)

Generally speaking, a set of temperature-dependent spectra of S components can be expressed in matrix form of A that contains both temperature trend along the row and frequency response along the column. This matrix A can then be decomposed into temperature component C and spectral component a with additional experimental noise ε with zero means as follows.

$$A_{k\times\nu} = C_{k\times S} a_{S\times\nu} + \varepsilon_{k\times\nu} \tag{5.73}$$

index k refers to the number of temperature points, and v corresponds to the number of frequency points. Identifying physically meaning spectral components a can help us understand thermodynamic equilibria from the corresponding temperature components C. One way of extracting temperature components and spectral components from experimental spectra is SVD. In SVD, the temperature-dependent spectra $A_{k\times\nu}$ are decomposed as

$$A_{k\times\nu} = U_{k\times S}\Sigma_{S\times S}V^T_{S\times\nu} \tag{5.74}$$

In Eqn. (5.74), A is decomposed into two separate orthogonal matrices U and V of S components. All of these S components form an orthonormal basis set, which is helpful for finding the irreducible representation. Each row in the V matrix contains a (difference) spectrum of that particular component, whereas its temperature trend is informed in the corresponding row of U matrix. Note that the first component is referred to as the mean component, which keeps track of the overall population of the average spectrum. In a temperature-dependent IR spectrum of a protein, this first component will have roughly the same magnitude across all temperatures. In contrast, SVD of transient absorption spectra as a function of waiting time/population time (τ_2) will lead to the first component of overall decay of the vibrational spectra. The additional singular matrix Σ contains only diagonal weights of each component in descending order, which indicates how important this component is present in the original spectra.

As a side note, SVD is intimately related to principal component analysis (PCA), which finds the eigenvectors of the covariance matrix from the data. Now, when looking at the spectral components, the PCA of covariance matrix K in frequency can be written as

$$K_{\nu\times\nu} = \boldsymbol{E}\left[\left(A-\bar{A}\right)^T_{\nu\times k}\left(A-\bar{A}\right)_{k\times\nu}\right] = V\Lambda V^T \tag{5.75}$$

$\boldsymbol{E}[A]$ is denoted as the expectation value of the variable A. Given Eqn. (5.74), we can find the following

$$\frac{1}{k-1}A^T_{\nu\times k}A_{k\times\nu} = \left(V_{\nu\times S}\Sigma^T_{S\times S}U^T_{S\times k}\right)\left(U_{k\times S}\Sigma_{S\times S}V^T_{S\times\nu}\right) = V\left(\frac{\Sigma^T\Sigma}{k-1}\right)V^T = V\Lambda V^T \tag{5.76}$$

Therefore, we observe that the V matrix is the eigenvector (principal component) of the normal matrix A^TA, and that the singular matrix Σ contains the squared root of the eigenvalues in Λ. Similarly, when we investigated the covariance matrix L in temperature,

$$L_{k\times k} = \boldsymbol{E}\left[\left(A-\bar{A}\right)_{k\times\nu}\left(A-\bar{A}\right)^T_{\nu\times k}\right] = U\Lambda U^T \tag{5.77}$$

$$\frac{1}{\nu-1}A_{k\times\nu}A^T_{\nu\times k} = \left(U_{k\times S}\Sigma_{S\times S}V^T_{S\times\nu}\right)\left(V_{\nu\times S}\Sigma^T_{S\times S}U^T_{S\times k}\right) = U\left(\frac{\Sigma^T\Sigma}{\nu-1}\right)U^T = U\Lambda U^T \tag{5.78}$$

This indicates that the U matrix is the eigenvector of the normal matrix AA^T.

There are some advantages and limitations of SVD. Firstly, SVD automatically sorts the importance of the components via Σ, which is useful to determine how many components are relevant in the experimental spectra for dimensionality reduction. Secondly, SVD can be used to separate spectral responses from noise by examining both Σ and the spectral component V. Typically, a small value of Σ that contributes less than a few percent of the total singular value has a noisy spectral component. The number of meaningful components r can be determined by choosing a cutoff percentage to the sum of singular values such as 90 %, 99 %, *etc*, and examining the cumulative sum of singular values in order of components such that the first r components reach the cutoff. One caveat of SVD is that it is purely based on mathematical requirements such as orthonormal condition of the components and linear decomposition, it does not necessarily reflect the chemical nature, meaning that the components do not correspond to a physical/chemical

species in the system. For more detailed discussion on SVD, please refer to Kevin's thesis and a detailed review on SVD.[14]

5.3.2 Construction of the MaxEnt Objective Function for IR and 2D IR Spectroscopy

To ensure that the decomposition of a temperature-dependent IR spectra has physical meaning and correspondence to chemical species, additional physical/chemical constraints can be applied and therefore act as biases to remix SVD components. Assuming there are only *r* physically meaningful components and the rest of *S*-*r* components are noise. we can apply a transformation matrix *T* to remix the first *s* SVD components in Eqn. (5.74) such that

$$\begin{aligned} A_{k\times\nu} &\approx U_{k\times r}\Sigma_{r\times r}T^{-1}_{r\times r}T_{r\times r}V^{T}_{r\times\nu} = C_{k\times r}a_{r\times\nu} \\ C_{k\times r} &= U_{k\times r}\Sigma_{r\times r}T^{-1}_{r\times r} \\ a_{r\times\nu} &= T_{r\times r}V^{T}_{r\times\nu} \end{aligned} \tag{5.79}$$

Finding the optimal *T* matrix will help us reconstruct pure components from the experimental data, which requires finding an objective function F_{obj}.

$$F_{\text{obj}} = H + \gamma_a P_a + \gamma_C P_C + \sum_d \lambda_d D_d \tag{5.80}$$

The MaxEnt reconstruction becomes an iterative optimization of finding the minimum of the objective function and the optimal *T* matrix, which can be done using simulated annealing to avoid trapping in the local extrema. This algorithm is readily implemented in Matlab and described elsewhere.[9]

The first term *H* corresponds to an entropy function that describes spectral similarity, analogous to the information entropy as below

$$H = -\sum_{r\nu} h_{r\nu} \ln\left(h_{r\nu}\right) \tag{5.81}$$

$$h_{rv} = \frac{\hat{a}'_{rv}}{\sum_{rv} \left| \hat{a}'_{rv} \right|} \tag{5.82}$$

In Eqn. (5.81), h_{rv} describes the magnitude of the trial spectral first derivative $\hat{a}'_{rv}$. Due to the nature of entropy function, when the spectral component becomes smoother, h_{rv} becomes smaller and effectively increases the entropy. Optimizing the entropy term will find the simplest possible set of spectra, which in essence will recover the SVD components without any additional constraint.

Additional constraints can be supplied through penalty function P and dissimilarity function D in Eqn. (5.80). The penalty function P imposes a non-negativity constraint along the frequency to have non-negative spectra, P_a, along the temperature to have non-negative population profiles P_C, or both. P is given as

$$P_a = \sum_{rv} F\left(\hat{a}_{rv}\right) \hat{a}_{rv}^2 \tag{5.83}$$

$$P_C = \sum_{kr} F\left(\hat{C}_{kr}\right) \hat{C}_{kr}^2 \tag{5.84}$$

$$F(x) = \begin{cases} 0 & , x \geq 0 \\ 1 & , x < 0 \end{cases} \tag{5.85}$$

Step function $F(x)$ ensures to penalize the negative values of given components. The corresponding weight coefficients for these penalty functions are denoted as γ_a, γ_C, respectively. Note that when applying to FTIR spectroscopy, the absorbance spectra are typically non-negative, so the penalty function along frequency P_a is one natural consequence to be applied. However, in 2D IR or transient absorption spectroscopy, such penalty is unphysical since there is always excited state absorption (ESA) or loss feature. Sometime there is loss feature even in difference FTIR spectroscopy. P_a should not be applied in these situations. Application to 2D IR spectroscopy with

penalty terms may be usable with absolute-valued 2D surface and/or DVE spectra. Dissimilarity function D provides an additional heuristic constraint to match certain features with the corresponding weight coefficient λ. This dissimilarity function can be optional, and may be applied depending on specific problems. There are numerous ways of defining the similarity function.[9] One example is to calculate the inner product between two temperature components.

$$D_1 = \sum_{n,m\neq n} \left\langle \hat{C}_n \hat{C}_m \right\rangle \tag{5.86}$$

This dissimilarity function is used to construct orthogonal temperature components during the reconstruction. Minimizing D will ensure the inner product between temperature components is minimized so that they are orthogonal.

Occasionally, it is possible to measure a spectrum of one specific pure component experimentally as a reference spectrum $a_{r\nu}^{\text{ref}}$. In this situation, one can use an additional dissimilarity function D_2 to ensure that one of the reconstructed spectral components resembles this reference spectrum, for simplicity here we choose the first component. The construction of D_2 can be written as a minimization problem of the squared error between the first spectral component and the reference spectrum.

$$D_2 = \arg\min_f \sum_\nu \left(\hat{a}_{1\nu} - f a_\nu^{\text{ref}} \right)^2 \tag{5.87}$$

The additional fit parameter f is the scaling factor undergoing the minimization to account for intensity mismatch between the two spectra. Similarly, when both FTIR spectroscopy and 2D IR spectroscopy of a protein are measured at the same condition, since they measure the same population profile over the temperature range, one can use the reconstructed temperature components of FTIR data to constrain the temperature components of 2D IR spectra, meaning that

$$D_3 = \arg\min \sum_{kr} \left(\hat{C}_{kr} - C_{kr}^{\text{FTIR}} \right)^2 \tag{5.88}$$

The resulting $C_{k\times r}$ is not guaranteed to directly report fractions of the species. One way to transform the components into fractions of each species is performing normalization. Normalizing these weights into the fractions can be recognized as solving the following linear equation.

$$C_{k\times r} \vec{\alpha} = \vec{1} \tag{5.89}$$

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} = C^{-1}\vec{1} \tag{5.90}$$

In Eqns. (5.89) and (5.90), $\vec{\alpha}$ is a column vector containing the normalization constants of each component and $\vec{1}$ is the column vector of 1. The inner product between each weight component and the normalization constant will equal to 1. Then the normalization vector can be solved by computing the matrix inverse of C in Eqn. (5.90). Note that the matrix C needs to be a square matrix, meaning that one has to select S points along the other dimension, which can be somewhat subjective. In practice, one can check which points will yield the most consistent normalization throughout the entire range.

5.3.3 Example of MaxEnt Reconstruction: Additional Spectroscopic Response in Thermal Dissociation of Insulin Dimer

We use thermal dissociation of human insulin dimer as an example to illustrate how to extract equilibrium populations and obtain physical interpretation. Thermal dissociation of insulin dimer is often treated as a two-state dissociation process,[15] and one common practice of extracting

populations from IR spectroscopy of insulin is to use the second SVD component as a proxy to the dimer fraction.[16-17] However, it has been noted both thermodynamically and kinetically that it exhibits non-2-state behaviors to account for unfolding of the monomer or disordering in the dimeric state.[17-19] Additionally, typical second SVD components have sloping baselines at both low temperature and high temperature,[16-17] meaning that fitting to such curves requires in total of 7 parameters to account for sloping baselines. It is also a question of whether these temperature-dependent sloping baseline actually contributes to dimer dissociation or simply reflects spectroscopic solvation response or some other temperature-dependent changes not related to dimer dissociation.[20-22] It is of primary importance to investigate if it is a reasonable assumption to extract dimer fraction from the second SVD component of temperature-dependent IR spectroscopy.

Fig. 5.1 shows the temperature-dependent FTIR spectra of wild-type unlabeled (UL) human insulin in 10 mM DCl (pH* = 1.7), with the sample preparation described in the previous chapter. Amide I vibrations appear as the brightest feature in all spectra within the frequency range of 1600–1700 cm^{-1}. Increasing temperature results in a blue shift of the amide I peak frequency from 1646 cm^{-1} to 1651 cm^{-1}, disappearance of the shoulder at 1682 cm^{-1}, and slight broadening of the entire amide I feature. These subtle spectral changes were associated with β-sheet vibrational modes, and consistently seen in previous IR studies on bovine insulin with both increasing temperature and decreasing insulin concentration,[16-17] which is effectively increasing the monomer population. Hence these spectral changes report primarily on the dimer-monomer transition and it is typically thought to be two-state dissociation.

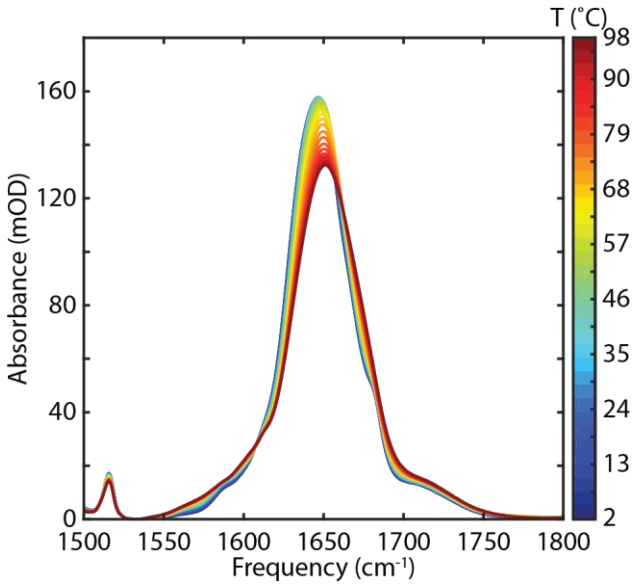

Figure 5.1: Temperature-dependent FTIR spectra of wild-type unlabeled human insulin in 10 mM DCl, ranging from 2 °C to 98 °C.

The corresponding components of SVD over the temperature-dependent IR spectra ranging from 1560 cm^{-1} to 1700 cm^{-1} is shown in Fig. 5.2. Singular values of the first three vectors account for 99 % of the entire spectral variation, indicating that the first three components are sufficient to describe the original spectra. The first spectral component V_1 accounting for 93 % of the singular values corresponds to the mean spectrum over the entire set of the spectra, and the temperature component U_1 changes minimally over the temperature range when comparing to other components. The second spectral component (V_2, 5 % of the entire singular values) shows the most dominant temperature-dependent changes with a gain feature at 1635 cm^{-1} and loss features at 1664 cm^{-1}, 1672 cm^{-1} and 1689 cm^{-1}. This spectral component is consistent with thermal dissociation of the dimer that exhibits loss of β-sheet feature and increasing disordering of the insulin structure in Fig. 5.1 and the temperature-dependent difference spectra of bovine insulin.[17] The corresponding temperature component U_2 shows the dissociation transition with sloping baselines at both low and high temperatures. The third spectral component V_3, which accounts for

only 1 % of the singular values, mostly shows a triplet of gain, loss, and gain feature across the frequency range from 1620 cm^{-1} to 1680 cm^{-1}, which may be indicative of spectral broadening. The temperature component U_3 decreases its magnitude until the minimum at 62 °C, and then starts increasing its magnitude. One question is if the third component is relevant for the dimer dissociation. If so, the thermodynamic two-state model may not be the most appropriate representation.

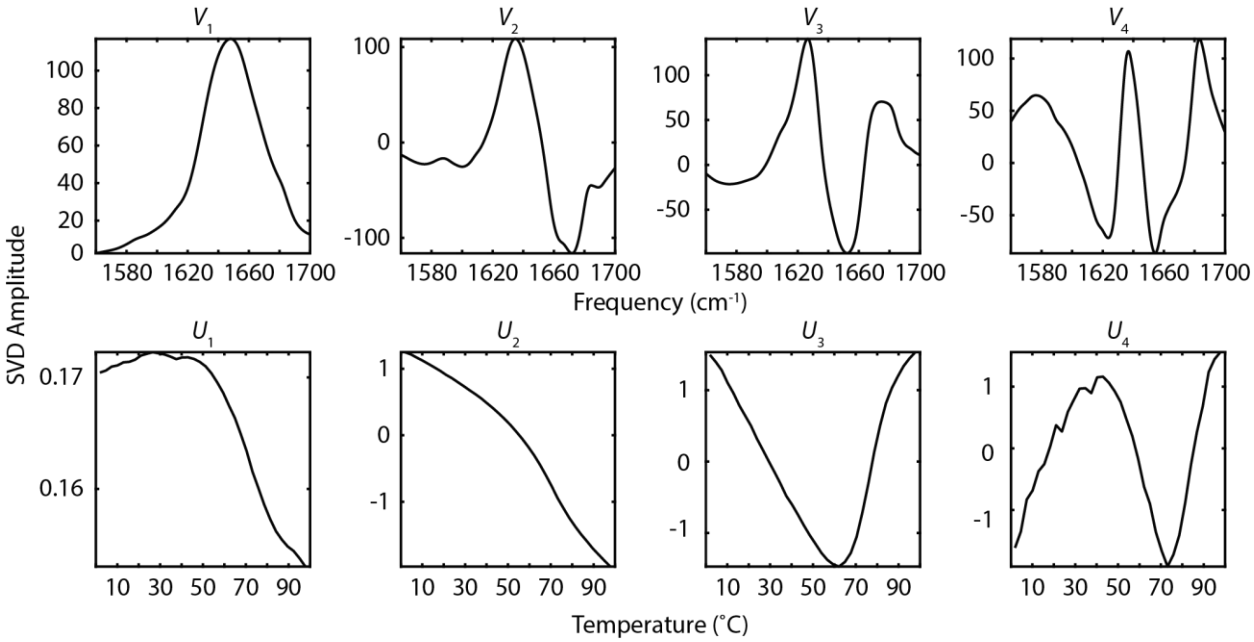

Figure 5.2: SVD of temperature-dependent FTIR spectra. Top: Spectral component (*V*) of the first four vectors. Bottom: Temperature Component (*U*) of the first four vectors.

To evaluate if the third SVD component in FTIR spectroscopy is crucial for the thermodynamic interpretation, temperature-dependent 2D IR spectra of UL insulin were investigated and shown in Fig. 5.3 ranging from 2 °C to 83 °C. The amide I 2D IR spectra at all temperatures show elongated diagonal feature. Low temperature spectra (2.3 °C) of both ZZZZ and ZZYY polarizations show increased cross-peak intensity around 1640 cm^{-1} and 1680 cm^{-1}, which originates from the coupling between $\nu_{\perp}$ and $\nu_{/\!/}$ modes of the β-sheet.[16-17, 23] Increasing temperature results in loss of the cross-peak intensity and elongation of the diagonal feature, but

the peak frequency of the ground state bleach feature does not exhibit significant frequency shift on either frequency axis.

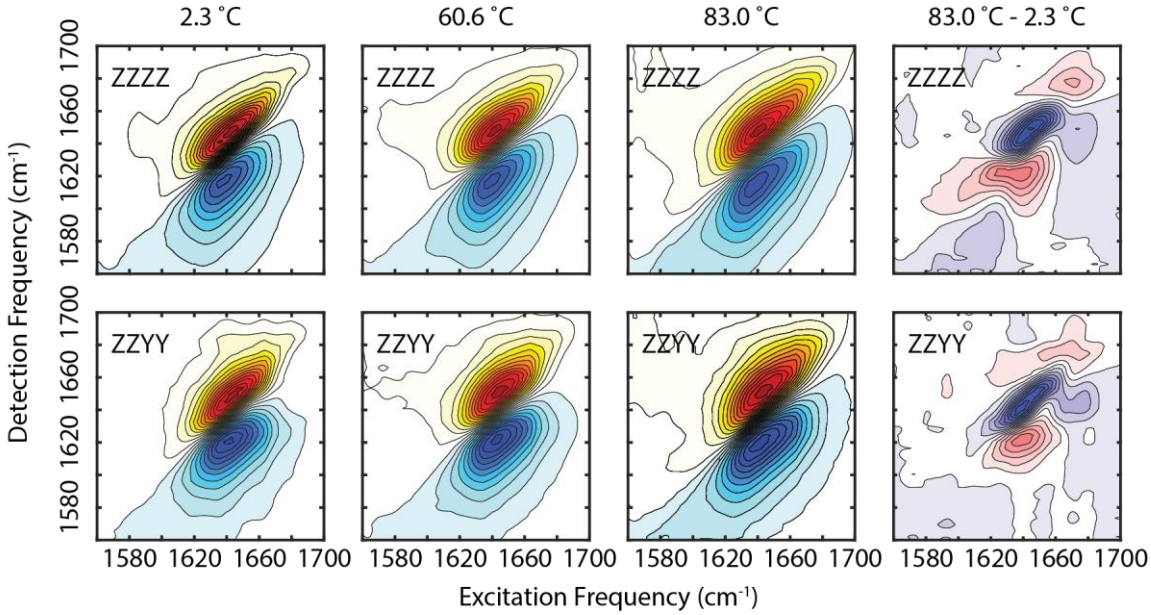

Figure 5.3: Temperature-dependent 2D IR spectra of UL insulin. Top: Parallel-polarized 2D IR spectra. Bottom: Perpendicular-polarized 2D IR spectra. Each spectrum is normalized against the peak intensity of the ground state bleach, and each contour level is spaced by 7.5 % of the maximum or 2% for the difference spectra.

The resulting SVD components are shown in Fig. 5.4. The first three vectors account for 93.5 % of the entire spectral variation across the temperature range, but the fourth spectral component V_4 already appears to be a noisy spectrum, which has 1% of the entire spectral variation. The first spectral component V_1 (85.5% of the total singular values) shows the mean spectra, similar as V_1 in the FTIR spectra. The second spectral component V_2 (5.5% of the total singular values) shows a diagonal gain feature peaked at 1636 cm^{-1}, and a loss feature around 1675 cm^{-1}, consistent with V_2 from FTIR spectra. In addition, there is an off-diagonal gain feature centered around (ω_1, ω_3) = (1637 cm^{-1}, 1665 cm^{-1}), which comes from the coupling of the β-sheet modes. There are some other features which are challenging to be interpreted physically. The temperature component U_2 behaves similarly as U_2 extracted from the FTIR spectra as expected since they in

principle would both report the dimer dissociation. The third temperature component (2.5% of the total singular values) shows similar behavior as in U_3 from FTIR spectra, but the minimum temperature is decreased to below 60 °C. The diagonal feature in V_3 also exhibits the gain, loss and gain triplet with increasing frequency, suggesting that both FTIR and 2D IR spectroscopy reveals consistent spectroscopic response in addition to the second SVD component, which is typically attributed to the thermal dissociation of the dimer.

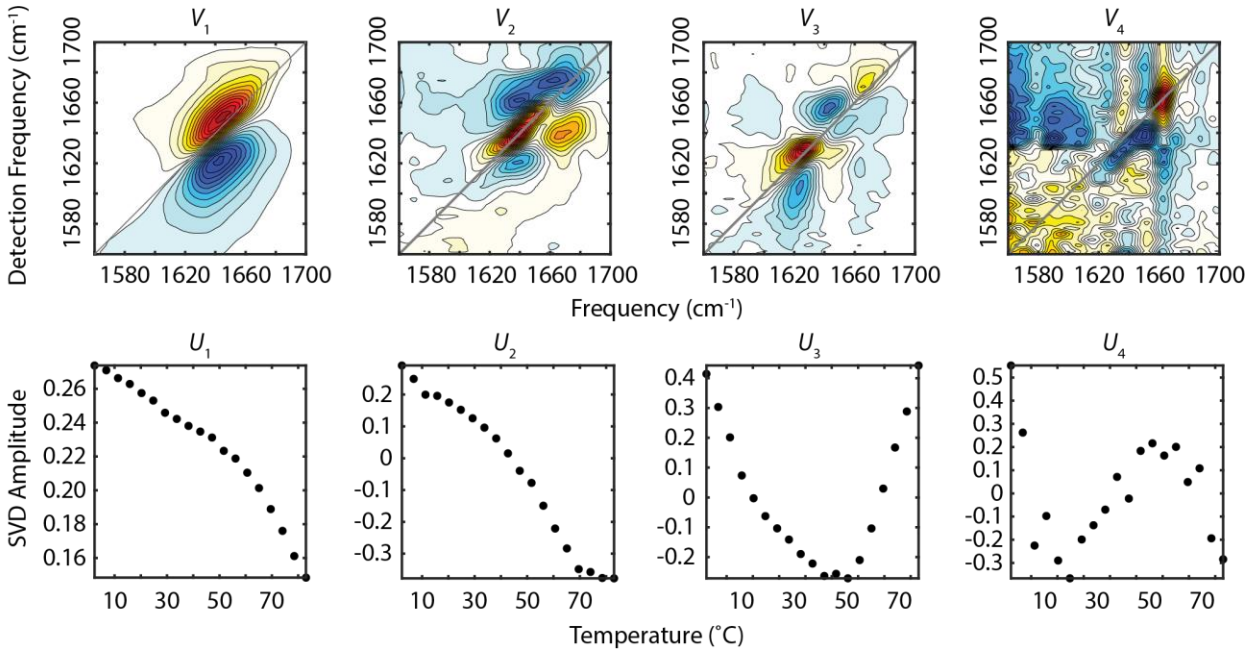

Figure 5.4: SVD of the temperature-dependent ZZYY-polarized 2D IR spectra. Top: Spectral component (V) of the first four vectors. Bottom: Temperature Component (U) of the first four vectors.

MaxEnt reconstruction is used to investigate if the third component contributes to any physically meaningful component with the objective function F_{obs} based on Eqn. (5.80) in Section 5.3.2. Since the 2D IR spectra have ESAs that have negative amplitudes, the coefficient of penalty function on the spectral component, γ_a, was to zero. For the sake of speeding up convergence on MaxEnt reconstruction over 2D IR spectra, additional dissimilarity function D_3 was used in Eqn.

(5.88) to obtain consistent temperature dependence between FTIR components and 2D IR components. The parameters were applied with fixed values outlined in Table 5.1, and the initial guess of the transformation matrix T was a diagonal matrix with 1's on the diagonal, meaning that the initial guess is not mixing the SVD components. Once the optimization was converged, additional optimization with the previous T was performed multiple times to make sure the transformation matrix did not change.

Weight coefficient	γ_a	γ_C	λ_1	λ_2	λ_3
Value	0	10^7	35	0	35

Table 5.1: Weight coefficients used for MaxEnt reconstruction of both FTIR spectra and 2D IR spectra.

The resulting 3-state reconstruction of the FTIR spectra is shown in Fig. 5.5. The first reconstructed component a_1 is dominated at low temperature, resembling the low-temperature dimer spectrum in Fig. 5.1 whereas a_3 is dominated at high temperature, resembling the high-temperature monomer spectrum. Comparing a_3 with a_1 (Figs. 5.5a and 5.5c), a_3 exhibits a blue-shift of the amide I band from 1645 cm^{-1} to 1652 cm^{-1}, loss of the intensity and spectral broadening, which is consistent with major features of dimer dissociation in the experimental FTIR spectra. Comparing to the second SVD component, the difference spectrum shows qualitatively consistent feature except for less loss feature above 1650 cm^{-1}. The spectral component a_2 also shows a minor red-shift of 4 cm^{-1}, but the width of the amide I band is about the same as in a_1. The corresponding temperature component C_2 shows a maximum at 62 °C. Given the linear decrease or increase on the weights of C_1 and C_2 with increasing temperature, the spectroscopic change from a_1 to a_2 exhibits frequency shift of 4 cm^{-1} over the temperature range from 2 °C to 62 °C, which is qualitatively consistent with the linear frequency shift of ~0.07 cm^{-1}/°C. This suggests that the

spectroscopic response from a_2 possibly comes from temperature-dependent solvent effect.[20] As a side note, the two-state fit using $C_1 + C_2$ recovers the global fit results in subsection 9.4.1, with the fitting parameters given in Table 5.2.

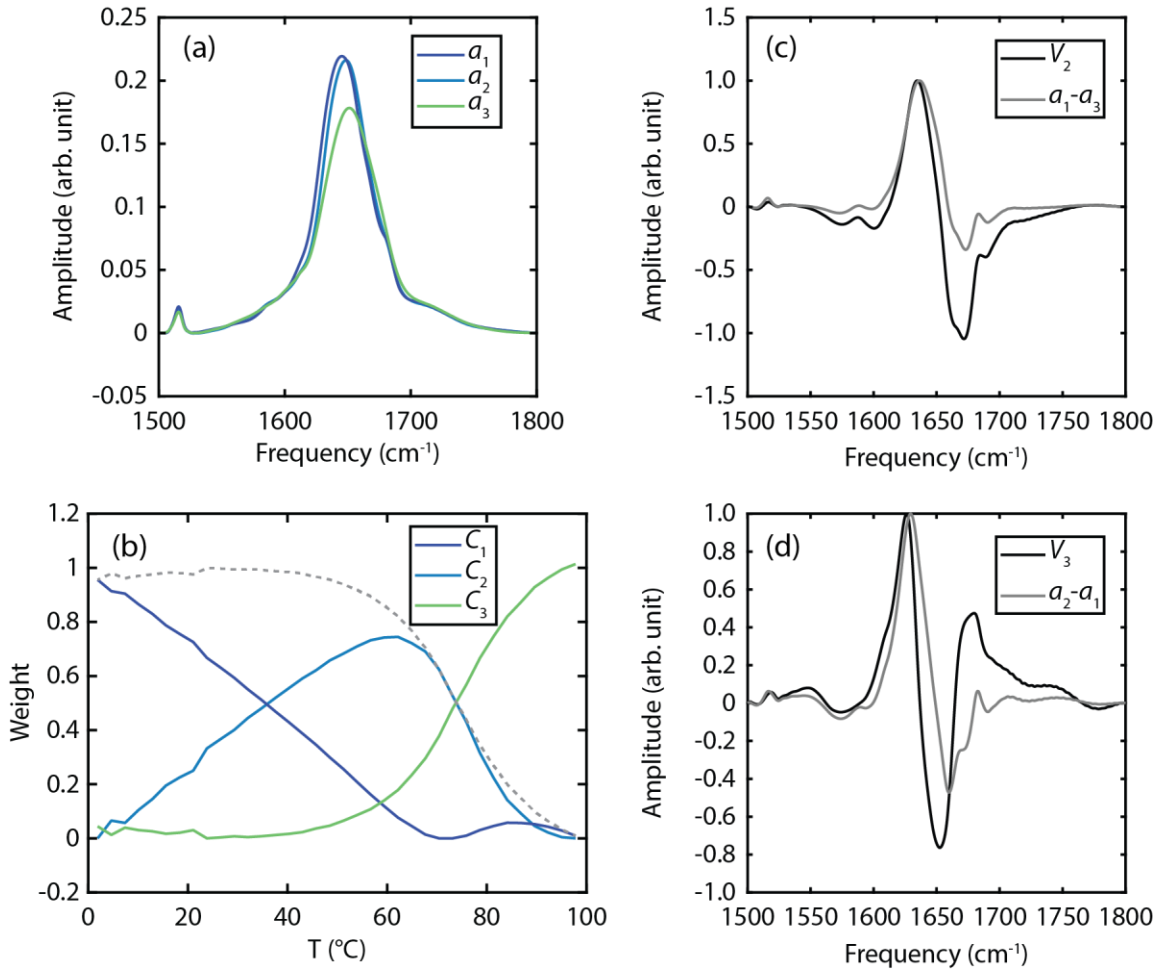

Figure 5.5: MaxEnt 3-state reconstruction of the FTIR spectra in Fig. 5.1 using the first 3 SVD components. (a) Reconstructed spectral components. (b) Reconstructed temperature components normalized to have the sum of the weights to be ~1 using Eqn. (5.90). Gray dashed curve corresponds to the sum of C_1 and C_2. (c) Comparison between the second SVD component V_2 and the difference spectrum between a_1 and a_3. (d) Comparison between the third SVD component V_3 and the difference spectrum between a_2 and a_1.

	T_d (°C)	$\Delta H°(T_d)$ (kJ/mol)	$\Delta C_P°$ (kJ/mol.K)
Two-state Fit from $C_1 + C_2$	68.5 (1.3)	172.7 (23.1)	5.0 (1.66)
Global Fit (Section 9.4)	67.6 (0.03)	184.9 (0.7)	3.5 (0.03)

Table 5.2: Thermodynamic model fit results between the 2-state model to $C_1 + C_2$ and the global fit described in Section 9.4. The standard deviation is indicated in the parenthesis.

The MaxEnt reconstruction of 2D IR spectra is shown in Fig. 5.6. All of the reconstructed spectra from Figs. 5.6a-c appear to be almost identical to the 2D IR spectra at low temperature, medium temperature, and high temperature (Fig. 5.3), and the corresponding temperature components behave similarly as the MaxEnt temperature components in FTIR spectra by using the dissimilarity function D_3. However, once the MaxEnt reconstruction of 2D IR spectra reached convergence, relaxing the additional constraint from D_3 by setting $\lambda_3 = 0$ still gives identical results, meaning that the same reconstruction can be achieved without constraining the temperature profile to be consistent between FTIR and 2D IR spectra.

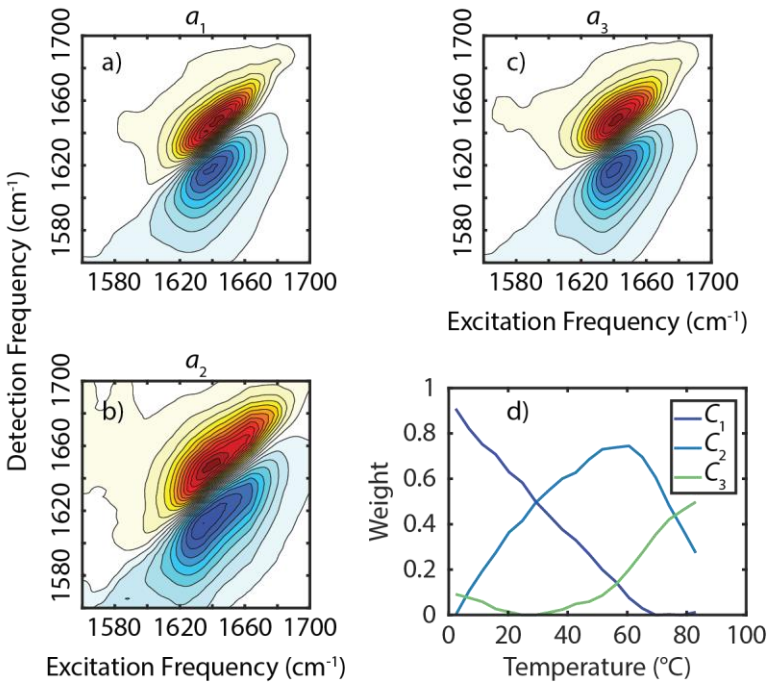

Figure 5.6: MaxEnt reconstruction of the 2D IR spectra. (a) The low temperature component a_1, resembling the dimer spectrum. (b) The intermediate temperature component a_2. (c) The high temperature component a_3, resembling the monomer spectrum. (d) Weights of these reconstructed components along temperature.

Some additional physical insights can be inferred by comparing the equilibrium spectral reconstruction to temperature-jump (T-jump) IR spectroscopy of insulin.[17] Laser-induced T-jump can be used to probe thermally-induced dissociation dynamics and the kinetic information can be obtained by controlling the delay time between 18 °C T-jump after 10 ns heating from 2 μm laser and IR laser at ~6 μm.[24] At sub-microsecond delay time (180 ns), the spectroscopic response comes primarily from fast response related to solvation changes,[17, 25-26] whereas hundreds of microsecond delay time (320 μs), one can see spectroscopic changes of loss of β-sheet, which is intimately related to dimer dissociation.[17] From the observation in the FTIR spectral reconstruction, it may be speculated that the intermediate temperature component is associated with solvation changes, which can be verified by comparing the difference spectrum between a_2 and a_1 with the transient 2D IR (t-2D IR) spectrum measured at sub-microsecond delay time.

Fig. 5.7 shows the head-to-head comparison between difference spectra of MaxEnt spectral components and t-2D IR spectra at sub-microsecond delay time (180 ns) and sub-ms delay time (320 μs). It shows good agreement between reconstructed difference spectra and t-2D IR spectra, meaning that the spectroscopic change from a_1 to a_2 can be assigned to mostly solvation changes due to temperature change such as weakening of the water protein hydrogen bonds or change, change of D_2O dielectric properties, or a combination of both.[20, 25] The dimer dissociation occurs thermodynamics from a_2 to a_3. Investigating equilibrium IR spectroscopy of insulin using MaxEnt shows its ability to resolve additional spectroscopic response associated with solvation changes, and the results of 3-state reconstruction suggest that the 2-state assumption is still viable for extracting the thermodynamic properties of dimer dissociation from equilibrium IR spectroscopy. Also the temperature dependent solvation response may also occur in other protein and peptide, it would be interesting to examine if similar behaviors happen and potentially apply this MaxEnt approach. Please note that other experimental probes may give additional thermodynamic information such as differential scanning calorimetry (DSC) so that non-2-state behaviors can be resolved.

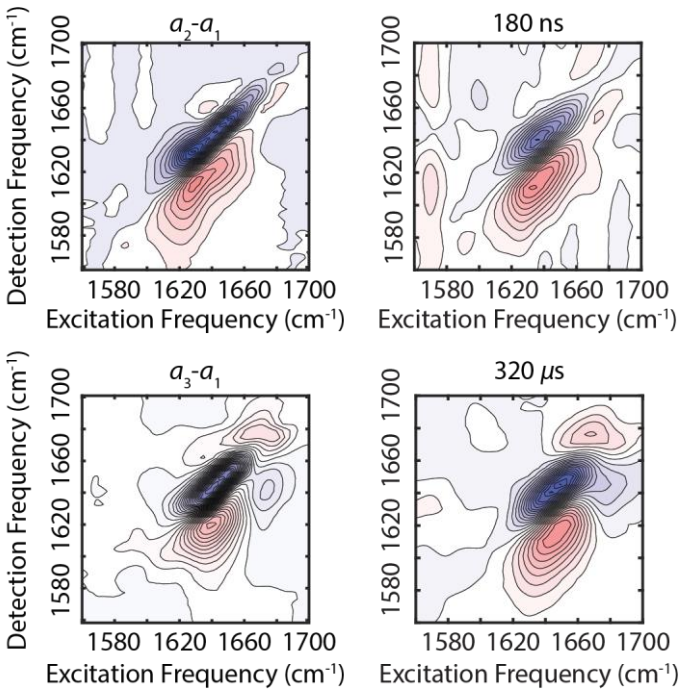

Figure 5.7: Head-to-head comparison between difference spectra of reconstructed components and t-2D IR spectra of human insulin in 10 mM DCl. Top left: difference spectrum between a_2 and a_1. Bottom left: difference spectrum between a_3 and a_1. Top right: t-2D IR spectrum of human insulin in 10 mM DCl at the delay time of 180 ns. Bottom right: t-2D IR spectrum at the delay time of 320 μs.

5.4 Ensemble Refinement

5.4.1 Introduction

One essential question of protein 2D IR spectroscopy is how to identify the structure-spectrum correlation, which is in essence an assignment problem of structurally interpreting spectroscopic signals. Experimentally, one can systematically investigate some controllable variables such as pH, temperature, ionic strength, mutations, or even using different experimental probes to infer underlying structures. This experimental approach provides a good basis or "evidence" for understanding the physical properties of the protein system, but it usually gives

indirect information of the atomistic structural details. Also, structural characterization of proteins exhibiting conformational disorder requires an ensemble description. Such ensemble structure determination is naturally an ill-posed problem where the degrees of freedom in experimentally relevant conformational states far exceeds the limited number of measurements and information content of experiments.[27]

Computational Amide I IR spectroscopy has been a useful tool to investigate the structure-spectrum correlation with the advantage of directly simulating IR spectrum given an atomistic protein structure derived from molecular dynamics simulations.[28] Computational spectroscopy also provides opportunities to determine the experimental conformation distributions of a protein system. However, molecular simulations inevitably rely on the accuracy of the force fields. Even though force fields capture most of the relevant conformations and dynamical processes, developing accurate force fields that can quantitatively determine conformational distributions is still an active research area over decades.[29-30] Other issues such as limited sampling of rare events due to the gap between computationally accessible time scales and the time scale of conformational interconversions may also hinder the accuracy of determining ensemble distributions.

The problem of structure determination has received numerous attention and studies on various proteins using restraint MD to determine a single protein structure or structures that match experimental data.[31-32] However, the concept of ensemble refinement using Bayesian inference against experimental data was firstly presented from the seminal paper in 2005,[33] which is essentially reweighting existing ensemble populations from simulations against experimental data so that the resulting computed observable like J coupling, X-ray scattering profile, or IR spectrum is consistent with the experiments. This ensemble refinement approach has been proven successful for inferring the protein conformational ensembles consistent with experiments.[27] Frameworks of

ensemble refinement have been developed mostly based on Maximum Entropy (MaxEnt) principle and Bayesian statistics. In this section below, we briefly describe these two approaches on applications to IR spectroscopy, discuss key differences between MaxEnt principle and Bayesian statistics, and eventually use a simple toy model to demonstrate the nature of MaxEnt and Bayesian approaches and their key differences.

5.4.2 Maximum Entropy-Based Ensemble Refinement

Maximum Entropy (MaxEnt) method on ensemble refinement provides a minimally biased inference to the ensemble populations that reproduce the experimental observations or information.[8, 34] This MaxEnt framework reweighs the initial populations of the ensemble p_n^{o} into a new set of populations p_n that matches the given set of (ensemble-averaged) experimental observables f_i. In other words,

$$\langle f_i \rangle = \sum_n p_n f_{n,i} = f_i^{\exp} \tag{5.91}$$

To achieve these experimental constraints, the MaxEnt method maximizes the relative entropy functional, or Kullback-Leibler divergence given as[35-36]

$$S_{\mathrm{KL}}\left[\boldsymbol{p} \,||\, \boldsymbol{p}^{\mathrm{o}} \right] = -\sum_n p_n \ln\left(\frac{p_n}{p_n^{\mathrm{o}}} \right) \tag{5.92}$$

subject to the normalization constraint

$$\sum_n p_n = 1 \tag{5.93}$$

as well as the constraint from the experimental observables given in Eqn. (5.91).

Eqn. (5.92) measures the difference of information entropy between the final populations and the initial populations. When the final set of the populations p_n is the same as the initial set of the populations p_n°, it indicates that the given experimental observation provides no new information. For instance, throwing a fair coin in principle gives probability of 0.5 for both head and tail. Someone threw the same fair coin for 100 times, and the result gives exactly head 50 times and tail 50 times. Then this observation does not provide any new information such that the MaxEnt probability is the same as before.

This MaxEnt optimization in Eqn. (5.92) subjective to the constraints in Eqns. (5.91) and (5.93) is equivalently to a problem of Lagrange multipliers, in which the corresponding solution can be found analytically with the form

$$p_n \propto p_n^{\circ} \prod_i \exp\left(\lambda_i f_{n,i}\right) \tag{5.94}$$

In Eqn. (5.94), λ_i is a Lagrange multiplier for the experimental observable $f_i^{\exp}$, which acts as a scaling factor of the probability of state *n*. These Lagrange multipliers can be determined numerically using the Levenberg-Marquardt algorithm.[35]

$$\lambda_i^{m+1} = \lambda_i^m - \left[M^T M + \alpha \cdot \mathrm{diag}\left(M^T M\right)\right]^{-1} \left\langle f_i \right\rangle_{\boldsymbol{p}(\boldsymbol{\lambda};m)} \tag{5.95}$$

α is an input stability parameter to balance efficiency and stability, and *M* is the derivative matrix with the matrix element

$$M_{ij} = \left\langle f_i \right\rangle_{\boldsymbol{p}(\boldsymbol{\lambda};m)} \left\langle f_j \right\rangle_{\boldsymbol{p}(\boldsymbol{\lambda};m)} - \left\langle f_i f_j \right\rangle_{\boldsymbol{p}(\boldsymbol{\lambda};m)} \tag{5.96}$$

computed with the probability distribution vector $\boldsymbol{p}(\boldsymbol{\lambda};m)$ at the m^{th} iteration

$$p_n\left(\boldsymbol{\lambda};m\right) \propto p_n^{\circ} \prod_i \exp\left(\lambda_i^m f_{n,i}\right) \tag{5.97}$$

This optimization iteratively changes the values of λ_i and the probability vector $\boldsymbol{p}$ until convergence. Practically, initial guess of the probability vector is treated as a uniform distribution across all N states such that $p_n^o = 1/N$, which is in essence the maximum entropy distribution without any additional information. In contrast, λ values can be chosen somewhat arbitrarily. For an example of how MaxEnt performs on simple problems, please refer to the nicely presented work by Lindorff-Larsen and co-workers.[36]

For the application to IR spectroscopy, one can define certain experimental observables for the MaxEnt method. For instance, in the application to the IR spectroscopy of disordered elastin-like peptide (ELP)[37], the isotope-labeled amide I peak at the proline turn was used for determining the conformational ensemble of the ELP mutants. This isotope label on the proline turn is sensitive to the turn distance, and mutation on the adjacent residue sidechains effectively change the average turn distance such that the label can spectroscopically distinguish on average different ELP mutants. In the construction of MaxEnt optimization, the isotope-labeled peak was treated as a probability density, and the corresponding experimental observables were defined as the first moment and the second moment of the labeled peak. The input structural basis for the probability vector were constructed using long trajectories from MD simulation of ELP, which were then divided into many short trajectories as the basis. These states can be used for calculating the corresponding isotope-labeled simulated spectra and performing MaxEnt optimization. The results of the MaxEnt optimization shows consistent probability distribution of the same ELP mutant across examined force fields, and demonstrates that isotope-edited IR spectroscopy can be a useful tool to perform quantitative ensemble refinement.[37]

A few notes on the MaxEnt refinement. First, MaxEnt refinement imposes “hard constraints”, meaning that the reweighted ensemble will reproduce the expected values from the

constraints whenever possible, meaning that systematic errors or noise will cause the reweighted ensemble to deviate from the true ensemble.[8] On the flip side, when the accuracy of the spectroscopic model and the experimental outcome is reliable enough, MaxEnt refinement will give the minimally biased answer. Secondly, MaxEnt refinement does not take experimental uncertainty into account, meaning that it cannot explicitly treat the noise contribution from the experiment, although it is possible to investigate the effect of systematic errors in the model.[36-37] Also, the current MaxEnt refinement approach used the first and second moment of the given linear IR spectrum, which cannot be directly applied to 2D IR spectroscopy that has ESA as negative features. Some new metrics and observables have to be proposed to be able to extend the MaxEnt ensemble refinement onto 2D IR spectroscopy, such as the absolute-valued 2D IR surface or DVE spectra.

5.4.3 Bayesian Ensemble Refinement

The fundamental concept of Bayesian ensemble refinement is that one can infer the most likely probability distribution based on given (experimental) data and some knowledge of the prior distribution, which is inherently a kind of probabilistic interpretation. In Bayesian ensemble refinement formulated by Hummer and Köfinger,[38] Bayes theorem states that the posterior distribution $p(\boldsymbol{x}\,|\,\text{data})$ based on experimental data follows

$$p(\boldsymbol{x}\,|\,\text{data}) \propto p(\text{data}\,|\,\boldsymbol{x})\,p(\boldsymbol{x}) \tag{5.98}$$

In Eq. (5.98), the prior distribution $p(\boldsymbol{x})$ corresponds to a sub-ensemble $\boldsymbol{x}$ that contains structures generated from a molecular simulation with uniformly distributed probabilities or probabilities based on the Boltzmann factor. The likelihood function $p(\text{data}\,|\,\boldsymbol{x})$ describes the probability of

reproducing experimental data by $\boldsymbol{x}$. Assuming Gaussian errors $\{\sigma_i\}$ of the experimental observables $\{y_i^{\exp}\}$, the likelihood function can be written as

$$p(\text{data} \mid \boldsymbol{x}) = \exp\left(-\sum_i \frac{\left[y_i(\boldsymbol{x}) - y_i^{\exp} \right]^2}{2\theta\sigma_i^2} \right) \tag{5.99}$$

In Eqn. (5.99), it measures the squared error of the observable $y_i(\boldsymbol{x})$ to the experimental observable scaled by the Gaussian error σ_i, implying that the more uncertain the experimental data is due to noise, the less reweighting on the ensemble probability will be performed due to the deviation of the simulation data from the experimental observable. Note that $\theta \leq 1$ is an adjustable parameter expressing the level of confidence in the (spectroscopic) model, and $1/\theta$ is equivalent to the Lagrange multiplier in the MaxEnt formalism.[38] However, the influence of the noise from the uncertainty σ_i is not taken into account from the previous MaxEnt approach. Large θ reflects high confidence in the model, whereas smaller θ would refine the prior distribution against the experimental data more. The optimal value of θ can be determined by the L-curve method,[38-39] illustrated in the toy model below and in the Appendix 7B. Also, there is no *a priori* expectation in Bayesian ensemble refinement that the squared errors have to minimized like in MaxEnt refinement. A short summary of the difference between MaxEnt ensemble refinement and Bayesian ensemble refinement is given in Table 5.3.

	MaxEnt	Bayesian
Adjustable parameter	λ	$1/\theta$
Experimental observables	Hard constraints	Soft constraints, no *a priori* expectations
Errors and uncertainties	No explicit terms accounting for errors	Random error treated with Gaussian distribution
Application to 2D IR spectroscopy	New metric needed to properly define the observables	Squared error or spectral overlap readily applicable

Table 5.3: Summary of comparison between MaxEnt ensemble refinement and Bayesian ensemble refinement.

For the application to linear IR spectroscopy, Eqn. (5.99) can be rewritten as the following

$$p\left(\text{data}\,|\,\boldsymbol{x}\right)=\exp\left(-\sum_{i}\frac{\int d\omega\left[I_i\left(\omega,\boldsymbol{x}\right)-I_i^{\text{exp}}\left(\omega\right)\right]^2}{2\theta\sigma_i}\right) \tag{5.100}$$

in which $I_i^{\text{exp}}\left(\omega\right)$ corresponds to the i^{th} experimental linear spectrum. In Eqn. (5.100), it actually has an issue of determining the unknown scaling factor between the experimental spectrum and the simulated spectra. In practice, one can assume that the integration of the spectrum over a certain frequency range is proportional to the protein concentration such that both experiments and simulations can be normalized, meaning that.

$$\tilde{I}_i^{\text{exp}}\left(\omega\right)=I_i^{\text{exp}}\left(\omega\right)\Big/\int_{\omega_1}^{\omega_2}d\omega I_i^{\text{exp}}\left(\omega\right) \tag{5.101}$$

$$\tilde{I}_i\left(\omega,\boldsymbol{x}\right)=I_i\left(\omega,\boldsymbol{x}\right)\Big/\int_{\omega_I}^{\omega_F}d\omega I_i\left(\omega,\boldsymbol{x}\right) \tag{5.102}$$

One can use the normalized spectra given in Eqns. (5.101) and (5.102) to calculate the likelihood function in Eqn. (5.100). For a given absorptive 2D IR spectrum $S(\omega_1, \omega_3)$, the presence of the ESA makes the normalization not as straightforward as integration of the spectrum. Instead, one

can possibly reconstruct the absolute-valued spectrum of the 2D surface $S_{abs}(\omega_1, \omega_3)$, and use that for the normalization. In other words,

$$p\left(S_i^{\exp}\left(\omega_1,\omega_3\right)\mid \boldsymbol{x}\right)=\exp\left(-\frac{\iint d\omega_1 d\omega_3\left[\tilde{S}_i\left(\omega_1,\omega_3,\boldsymbol{x}\right)-\tilde{S}_i^{\exp}\left(\omega_1,\omega_3\right)\right]^2}{2\theta\sigma_i}\right) \tag{5.103}$$

$$\tilde{S}_i^{\exp}\left(\omega_1,\omega_3\right)=S_i^{\exp}\left(\omega_1,\omega_3\right)\Big/\int_{\omega_{1,I}}^{\omega_{1,F}} d\omega_1\int_{\omega_{3,I}}^{\omega_{3,F}} d\omega_3\, S_{i,\text{abs}}^{\exp}\left(\omega_1,\omega_3\right) \tag{5.104}$$

$$\tilde{S}_i\left(\omega_1,\omega_3,\boldsymbol{x}\right)=S_i\left(\omega_1,\omega_3,\boldsymbol{x}\right)\Big/\int_{\omega_{1,I}}^{\omega_{1,F}} d\omega_1\int_{\omega_{3,I}}^{\omega_{3,F}} d\omega_3\, S_{i,\text{abs}}\left(\omega_1,\omega_3,\boldsymbol{x}\right) \tag{5.105}$$

Computationally, the absolute-valued 2D spectrum in the impulsive limit can be simply obtained from taking the absolute value of the entire third-order response function. Experimentally in the Boxcar geometry, since one can independently measure rephasing and non-rephasing spectra, and obtain both the real part and imaginary part during post-processing, the absolute-valued spectra of the absorptive surface can be computed. In the pump-probe geometry, one additional assumption of causality needs to be imposed such that one can estimate the absolute-valued surface from the absorptive 2D surface. Please refer to section 2.4.1 in Becky's thesis for how to estimate the absolute-valued surface.[40]

Alternatively, another way of defining the likelihood function in linear spectroscopy is given below.[41]

$$p(\text{data}\mid \boldsymbol{x})=\exp\left(-\sum_i\frac{1-s_i}{2\theta\sigma_i^2}\right)=\exp\left(-\chi^2(\theta)\right) \tag{5.106}$$

$$s_i(\boldsymbol{x})=\frac{\int d\omega I_i^{x}(\omega)I_i^{\exp}(\omega)}{\sqrt{\left(\int d\omega\left(I_i^{x}(\omega)\right)^2\right)\times\left(\int d\omega\left(I_i^{\exp}(\omega)\right)^2\right)}} \tag{5.107}$$

$s_i(\boldsymbol{x})$ is the spectral overlap quantifying the similarity between the simulated spectra from $\boldsymbol{x}$, $I_i^x(\omega)$, and the spectrum from experiment, $I_i^{\exp}(\omega)$.[42-43] Representative values of $s_i(\boldsymbol{x})$ are 1 for identical spectra, 0 for non-overlapping spectra, and -1 for identical but opposite-signed spectra. If $I_i^x(\omega)$ is identical to $I_i^{\exp}(\omega)$, then the measure of error $\chi^2(\theta)=0$ and $p(\text{data}\,|\,\boldsymbol{x})=1$, meaning that there is no need to refine the probability at all. All the other cases would reduce $p(\text{data}\,|\,\boldsymbol{x})$ depending on the value of θ. For 2D IR spectroscopy, the corresponding spectral overlap is given as

$$s_i(\boldsymbol{x})=\frac{\iint d\omega_1 d\omega_3 S_i^x(\omega_1,\omega_3)S_i^{\exp}(\omega_1,\omega_3)}{\sqrt{\left(\iint d\omega_1 d\omega_3\left(S_i^x(\omega_1,\omega_3)\right)^2\right)\times\left(\int\int d\omega_1 d\omega_3\left(S_i^{\exp}(\omega_1,\omega_3)\right)^2\right)}} \tag{5.108}$$

This likelihood function using the metric of spectral overlap does not require additional data processing on the 2D IR spectra such as extracting the absolute-valued 2D surface.

5.4.4 Proof-of-Concept Toy Model

This sub-section illustrates how the MaxEnt and Bayesian frameworks for ensemble refinement can be performed on a toy model. The setup of the toy model is shown in Fig. 5.8. The experimental linear absorption spectrum is a single Gaussian peak shown in Fig. 5.8a with the center frequency of 1650 cm^{-1}, and the FWHM of 23.55 cm^{-1}, meaning that the standard deviation σ_ω is 10 cm^{-1}. Random noise is added when needed with the amplitude defined by the signal-to-noise (S/N) ratio. In this toy model, molecular structures are represented by a single variable, x, which is the one-dimensional "molecular" coordinate ranging from -10 to 10. The initial probability distribution of the structures $p(x)$ in Fig. 5.8b is set to be

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-(x-3)^2/2\right) \tag{5.109}$$

And the relationship between the coordinate x and the spectrum is given as

$$I(\omega, x) = \exp\left(-\left(\omega - \omega_{\text{peak}}(x)\right)^2 / 2\sigma_\omega^2\right) \tag{5.110}$$

$$\omega_{\text{peak}}(x) = 1650 + 5x \tag{5.111}$$

Based on Eqn. (5.111), the peak frequency of the simulated spectra can vary from 1600 cm^{-1} to 1700 cm^{-1}, but the width is independent of the coordinate for simplicity. Based on this relation, the experimental spectrum will imply the true distribution with probability 1 at x = 0 (Fig. 5.8b), and the corresponding prior spectrum shown in Fig. 5.8a.

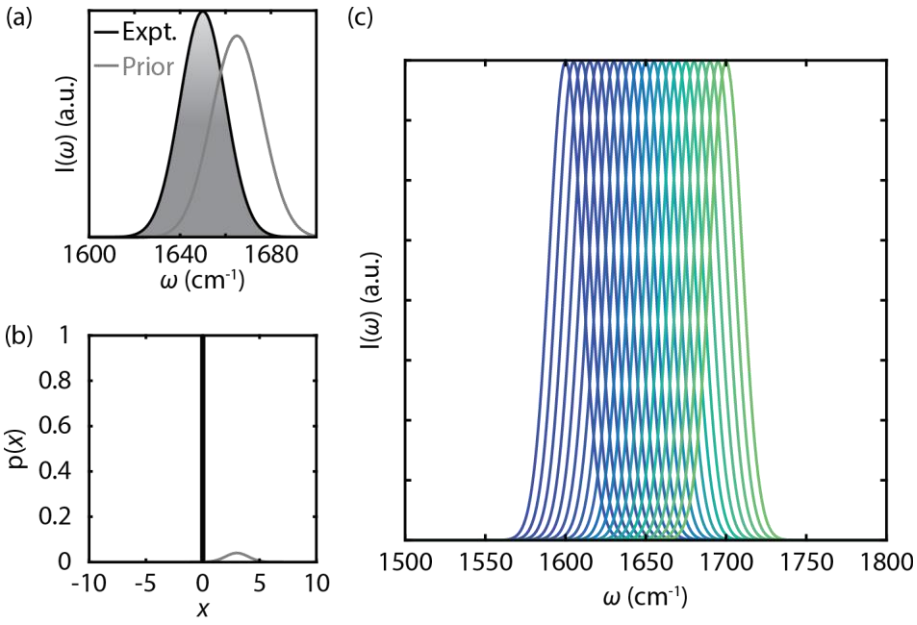

Figure 5.8: Setup of the toy model. (a) Linear absorption spectra of the experiment (black) and the simulation prior (gray). (b) Probability distribution as a function of x. Black: The experiment. Gray: The simulation prior. (c) Simulated spectra as a function of x, which is used for constructing the prior spectrum in (a) and subsequent ensemble refinement.

The result of MaxEnt ensemble refinement is shown in Fig. 5.9. The same experimental constraints were used as in the ELP study, meaning the first moment and the second moment of the spectrum.[37] Without any noise (Figs. 5.9a–b), the refinement gives very similar spectrum as in the experiment, but there is a subtle difference on spectral lineshape. The refined probability distribution shows the peak position the same as in the true experimental distribution, but the width is wider, meaning that the MaxEnt refinement does not give exactly the same answer, but qualitatively it captures the correct average.

When there is significant amount of noise present in the spectrum as in Figs. 5.9c–d, the refinement spectrum actually shows a 3 cm^{-1} red-shift compared to the true peak frequency, and the refined probability distribution also shows the incorrect peak position of x, indicating that hard constraints imposed by MaxEnt ensemble refinement biases the results due to the noise contribution. Please note that this amount of noise can most likely be alleviated in real experiments, but here is used for the sake of demonstrating the nature of MaxEnt ensemble refinement.

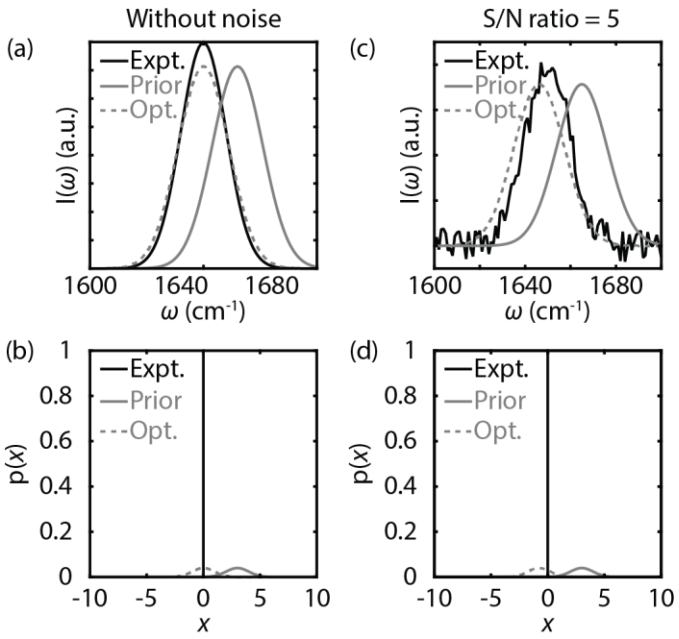

Figure 5.9: The result of ensemble refinement based on MaxEnt principle without noise (a–b) and with noise (c–d). (a) Refined spectrum (dashed gray) without noise compared to the experiment (black) and the prior spectrum (solid gray). (b) Refined probability distribution. (c) Refined spectrum with the presence of significant noise (S/N ratio = 5) (d) Refined probability distribution with the presence of the noise.

Bayesian ensemble refinement takes the noise contribution from experiments into account and it does not enforce the simulated observables to match the experimental ones, so one would expect it will be less prone to biases due to the noise. The dependence of θ in Bayesian ensemble refinement using the squared error term in Eqn. (5.100) is shown in Fig. 5.10. Large θ values reflect more confidence of the model such that the refined probability better resembles the prior distribution, and the refined spectrum also appears more similar to the prior spectrum. Once the θ value gets smaller, the experimental evidence becomes more important (less confidence on the model), and the refined spectrum can eventually match the experimental spectrum with the correct probability distribution whereas MaxEnt does depend on how to define the observable, which may not give quantitatively the correct answer. The other case of adding random noise still gives qualitatively the same dependence on θ (Bottom of Fig. 5.10). More importantly, the refined

spectra at small θ values still show pretty good agreement with the experimental data, which demonstrates the nature of Bayesian ensemble refinement that uses experimental data as soft constraints and takes the experimental noise contribution into account. In contrast, MaxEnt ensemble refinement can lead to biased results as shown in Fig. 5.9.

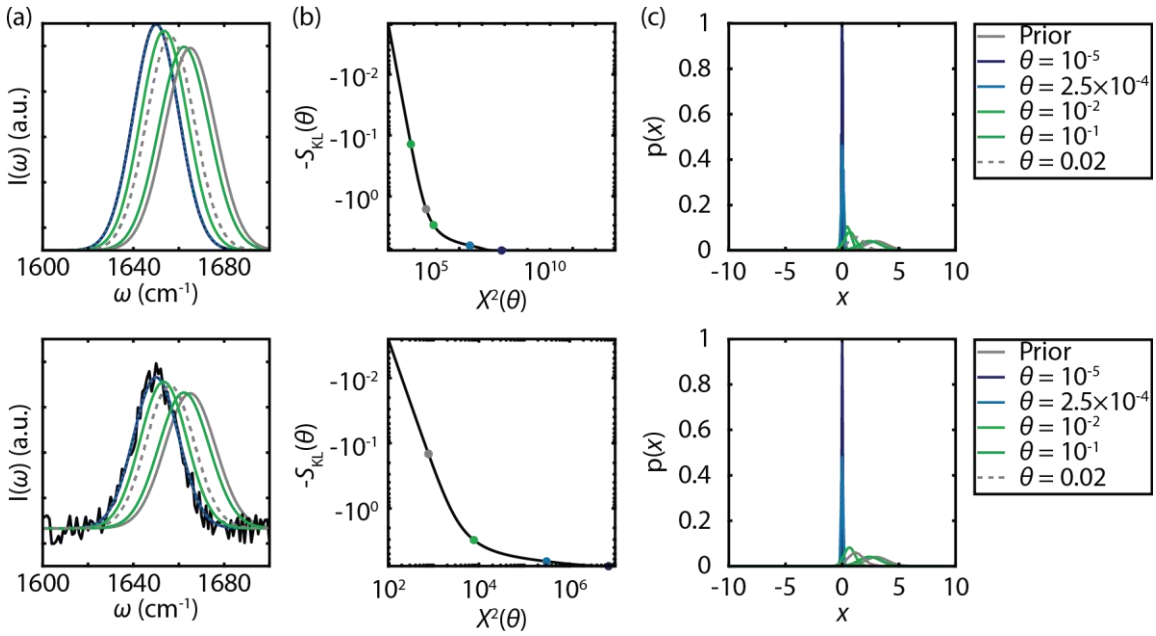

Figure 5.10: Results of Bayesian ensemble refinement based on the squared error in Eqn. (5.100). (a) Refined spectra as a function of θ without noise (top) and with signal-to-noise ratio of 5 (bottom) Dashed line corresponds to the θ value determined by the maximum curvature of the L-curve in (b). (b) L-curve that plots the relative entropy $-S_{KL}(\theta)$ against the squared error $\chi^2(\theta)$. (c) Refined probability distribution $p(x)$ as a function of θ.

In practice, it is possible to have systematic errors in the experimental data and aiming for matching the experiments perfectly can easily lead to over-fitting. To determine the optimal θ value that balances the maximum possible information content in the experimental data and minimizing the possibility of over-fitting, one can use the L-curve method with the characteristic log-log plot shown in Fig. 5.10b. Generally, both $S_{KL}(\theta)$ as defined in Eqn. (5.92) and $\chi^2(\theta)$ increase with decreasing θ, although the slope varies. In the steep regime of Fig. 5.10b (large θ), a small change of θ leads to a large decrease of $-S_{KL}(\theta)$, indicating that increasing θ refines the

posterior distribution to agree better with the experiments. Conversely in the flat regime (small θ), the distribution changes little with increasing θ, but the measure of error $\chi^2(\theta)$ increases dramatically, implying that the distribution is very sensitive to the experimental data including noise. Thus the plot has a characteristic "L" shape, and the optimal θ is found at the point of maximum curvature. One can see that this optimal θ gives the refined spectrum somewhere between actual experimental spectrum and the original spectrum, which reflects the nature of Bayesian statistics that uses experimental observables as soft constraints, in contrast to the MaxEnt refinement. One can either more aggressively choosing smaller θ to see it matches better without sacrificing the quality of the refinement, or adding more experimental inputs that provide similar information at the same condition to help determine the optimal spectrum that resembles the experimental spectrum more (Please see Chapter 7 for more details). In practice of finding the maximum curvature, one can use cubic spline fitting to the L-curve or manually selecting the point around the corner of the L-curve. Note that sometimes determining the maximum curvature may not lead to the point close to the corner of the L-curve possibly because L-curve is not sampled properly or not smooth enough, so it is recommended to double check manually to make sure the θ is selected properly.

To evaluate the dependence of the likelihood functional form, the ensemble refinement using spectral overlap as the metric in Eqn. (5.106) is shown in Fig. 5.11. Overall, the dependence of θ behaves qualitatively the same as the refinement using squared error as the metric. Given the same θ value, spectral overlap performs the refinement more aggressively than the squared error in this model. Also, the optimal θ values determined by the maximum curvatures present in the L-curves actually gives identical refined spectra across the two metrics, suggesting that either metric can be used for ensemble refinement against IR spectroscopy. In practice a single experiment may

not provide enough information content so that Bayesian refinement gives an estimate between the correct distribution and the initial distribution as in Fig. 5.11. Multiple experimental data may be needed to get a converged distribution or a distribution with high confidence.[41] A real application to protein IR spectroscopy using Ala–Ala–Ala is presented in detail in Chapter 7.

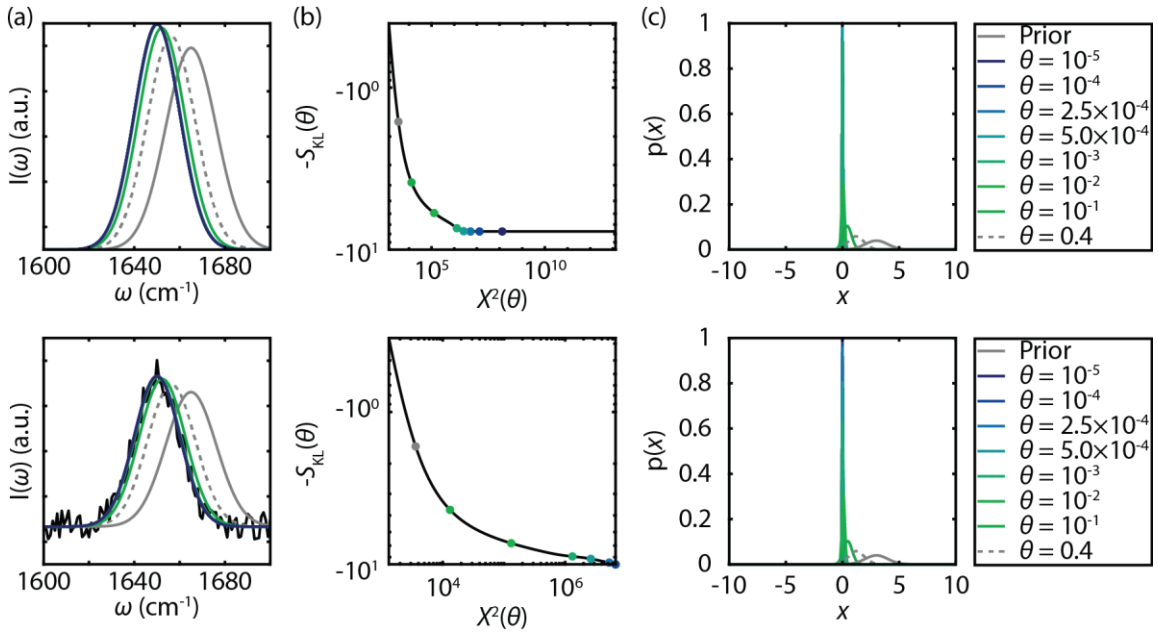

Figure 5.11: Results of Bayesian ensemble refinement based on the spectral overlap in Eqn. (5.106). (a) Refined spectra as a function of θ without noise (top) and with signal-to-noise ratio of 5 (bottom) Dashed line corresponds to the θ value determined by the maximum curvature of the curve. (b) L-curve that plots the relative entropy $-S_{KL}(\theta)$ against the squared error $\chi 2(\theta)$. (c) Refined probability distribution as a function of θ.

5.5 Acknowledgments

I thank Brennan Ashwood, Luis Busto de Moner, and Yumin Lee for carefully reading through this chapter and providing valuable feedbacks.

5.6 References

1. Gomez, J.; Hilser, V. J.; Xie, D.; Freire, E., The heat capacity of proteins. *Proteins* **1995,** *22* (4), 404-12.
2. Myers, J. K.; Pace, C. N.; Scholtz, J. M., Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **1995,** *4* (10), 2138-48.
3. Sanchez-Ruiz, J. M., Probing free-energy surfaces with differential scanning calorimetry. *Annu Rev Phys Chem* **2011,** *62*, 231-55.
4. Amunson, K. E.; Anderson, B. A.; Kubelka, J., Temperature effects on the optical path length of infrared liquid transmission cells. *Appl Spectrosc* **2011,** *65* (11), 1307-13.
5. Debye, P.; Hückel, E., Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. *Physikalische Zeitschrift* **1923,** *24* (185), 305.
6. de Los Rios, M. A.; Plaxco, K. W., Apparent Debye-Huckel electrostatic effects in the folding of a simple, single domain protein. *Biochemistry* **2005,** *44* (4), 1243-50.
7. Religa, T. L.; Markson, J. S.; Mayor, U.; Freund, S. M.; Fersht, A. R., Solution structure of a protein denatured state and folding intermediate. *Nature* **2005,** *437* (7061), 1053-6.
8. Jaynes, E. T., On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **1982,** *70* (9), 939-952.
9. Widjaja, E.; Garland, M., Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-species data set. *J Comput Chem* **2002,** *23* (9), 911-9.
10. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
11. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J Am Chem Soc* **2016,** *138* (36), 11792-801.
12. Dai, Q.; Sanstead, P. J.; Peng, C. S.; Han, D.; He, C.; Tokmakoff, A., Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS Chem Biol* **2016,** *11* (2), 470-7.
13. Lewis, N. H. C.; Fournier, J. A.; Carpenter, W. B.; Tokmakoff, A., Direct Observation of Ion Pairing in Aqueous Nitric Acid Using 2D Infrared Spectroscopy. *J Phys Chem B* **2019,** *123* (1), 225-238.
14. Jones, K. C. Temperature-jump 2D IR spectroscopy to study protein conformational dynamics. Massachusetts Institute of Technology, Massachusetts Institute of Technology, 2012.
15. Antolikova, E.; Zakova, L.; Turkenburg, J. P.; Watson, C. J.; Hanclova, I.; Sanda, M.; Cooper, A.; Kraus, T.; Brzozowski, A. M.; Jiracek, J., Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J Biol Chem* **2011,** *286* (42), 36968-77.
16. Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy. *Phys Chem Chem Phys* **2010,** *12* (14), 3579-88.
17. Zhang, X. X.; Jones, K. C.; Fitzpatrick, A.; Peng, C. S.; Feng, C. J.; Baiz, C. R.; Tokmakoff, A., Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J Phys Chem B* **2016,** *120* (23), 5134-45.

18. Huus, K.; Havelund, S.; Olsen, H. B.; van de Weert, M.; Frokjaer, S., Thermal dissociation and unfolding of insulin. *Biochemistry* **2005,** *44* (33), 11171-7.
19. Rimmerman, D.; Leshchev, D.; Hsu, D. J.; Hong, J.; Kosheleva, I.; Chen, L. X., Direct Observation of Insulin Association Dynamics with Time-Resolved X-ray Scattering. *J Phys Chem Lett* **2017,** *8* (18), 4413-4418.
20. Amunson, K. E.; Kubelka, J., On the temperature dependence of amide I frequencies of peptides in solution. *J Phys Chem B* **2007,** *111* (33), 9993-8.
21. Ashwood, B.; Sanstead, P. J.; Dai, Q.; He, C.; Tokmakoff, A., 5-Carboxylcytosine and Cytosine Protonation Distinctly Alter the Stability and Dehybridization Dynamics of the DNA Duplex. *J Phys Chem B* **2020,** *124* (4), 627-640.
22. Sanstead, P. J.; Ashwood, B.; Dai, Q.; He, C.; Tokmakoff, A., Oxidized Derivatives of 5-Methylcytosine Alter the Stability and Dehybridization Dynamics of Duplex DNA. *J Phys Chem B* **2020,** *124* (7), 1160-1174.
23. Demirdoven, N.; Cheatum, C. M.; Chung, H. S.; Khalil, M.; Knoester, J.; Tokmakoff, A., Two-dimensional infrared spectroscopy of antiparallel beta-sheet secondary structure. *J Am Chem Soc* **2004,** *126* (25), 7981-90.
24. Chung, H. S.; Khalil, M.; Smith, A. W.; Tokmakoff, A., Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *Rev Sci Instrum* **2007,** *78* (6), 063101.
25. Chung, H. S.; Ganim, Z.; Jones, K. C.; Tokmakoff, A., Transient 2D IR spectroscopy of ubiquitin unfolding dynamics. *Proc Natl Acad Sci U S A* **2007,** *104* (36), 14237-42.
26. Jones, K. C.; Peng, C. S.; Tokmakoff, A., Folding of a heterogeneous beta-hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc Natl Acad Sci U S A* **2013,** *110* (8), 2828-33.
27. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M., Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **2017,** *42*, 106-116.
28. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
29. Best, R. B.; Buchete, N. V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophys J* **2008,** *95* (1), L07-9.
30. Huang, J.; MacKerell, A. D., Jr., Force field development and simulations of intrinsically disordered proteins. *Curr Opin Struct Biol* **2018,** *48*, 40-48.
31. Nilges, M.; Gronenborn, A. M.; Brunger, A. T.; Clore, G. M., Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng* **1988,** *2* (1), 27-38.
32. Hua, Q. X.; Shoelson, S. E.; Kochoyan, M.; Weiss, M. A., Receptor binding redefined by a structural switch in a mutant human insulin. *Nature* **1991,** *354* (6350), 238-41.
33. Rieping, W.; Habeck, M.; Nilges, M., Inferential structure determination. *Science* **2005,** *309* (5732), 303-6.
34. Jaynes, E. T., Information Theory and Statistical Mechanics. *Physical Review* **1957,** *106* (4), 620-630.
35. Roux, B.; Weare, J., On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys* **2013,** *138* (8), 084107.
36. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K., Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* **2014,** *10* (2), e1003406.

37. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
38. Hummer, G.; Kofinger, J., Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys* **2015,** *143* (24), 243150.
39. Hansen, P. C.; O'Leary, D. P., The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM Journal on Scientific Computing* **1993,** *14* (6), 1487-1503.
40. Nicodemus, R. A. Hydrogen bond reorganization and vibrational relaxation in water studied with ultrafast infrared spectroscopy. Massachusetts Institute of Technology, 2011.
41. Feng, C. J.; Dhayalan, B.; Tokmakoff, A., Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to AlaAlaAla. *Biophys J* **2018,** *114* (12), 2820-2832.
42. Kruiger, J. F.; van der Vegte, C. P.; Jansen, T. L., Suppressing sampling noise in linear and two-dimensional spectral simulations. *J Chem Phys* **2015,** *142* (5), 054201.
43. Bondarenko, A. S.; Jansen, T. L., Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins. *J Chem Phys* **2015,** *142* (21), 212437.

Chapter 6

Dynamics of Peptide-Water Interactions in Dialanine: An Ultrafast Amide I 2D IR and Computational Spectroscopy Study

The work presented in this chapter has been published and is reprinted with permission from:
Feng, C.-J.; Tokmakoff, A. The Dynamics of Peptide-Water Interactions in Dialanine: An Ultrafast Amide I 2D IR and Computational Spectroscopy Study. *Journal of Chemical Physics.* **2017**, 147 (8), 085101.

6.1 Abstract

We present a joint experimental and computational study of the dynamic interactions of dialanine (Ala–Ala) with water, comparing the results of ultrafast 2D IR and infrared transient absorption spectroscopy of its amide I vibration with spectra modeled from molecular dynamics (MD) simulations. The experimental data are analyzed to describe vibrational frequency fluctuations, vibrational energy relaxation, and chemical exchange processes. The origin of these processes in the same underlying fluctuating forces allows a common description in terms of the fluctuations and conformational dynamics of the peptide and associated solvent. By comparing computational spectroscopy from MD simulations with multiple force fields and water models, we describe how the dynamics of water hydrogen bond fluctuations and switching processes act as a source of friction that governs the dephasing and vibrational relaxation, and provide a description of coupled water and peptide motions that give rise to spectroscopic exchange processes.

6.2 Introduction

Protein-water interactions mediate protein conformational changes in processes such as folding,[1-4] protein-protein recognition,[5-7] ligand and drug binding,[8-9] and enzymatic catalysis.[10-12] Although water's role in such processes commonly manifest themselves as large scale collective solvation effects, protein-water interactions are ultimately rooted in numerous local hydrogen bond interactions and the local structural rearrangements of the solvent around exposed residues. The interplay of local solvent dynamics with protein conformational changes is highly coupled, and even simple conformational transitions can involve non-intuitive water dynamics on a complex energy landscape. For instance, even a small peptide such as alanine dipeptide executes conformational transitions by moving along a solvent coordinate that has little correlation with rotating backbone dihedral angles.[13-14]

Currently, molecular dynamics (MD) simulations offer a detailed atomistic picture of protein-water interactions, and have served as our primary source of understanding molecular scale protein solvation.[15-17] However, experiments that access this level of information have lagged behind, in part because of the fast ps–ns time-scales of these processes, and partially because experiments typically offer indirect information on the water. Dielectric relaxation spectroscopy,[18] ultrafast fluorescence spectroscopy,[19] and terahertz spectroscopy[20-21] have provided information on the dynamics of protein-water interactions and thickness of hydration layer, but often it is difficult to build molecular interpretations from the spectra. NMR relaxation methods provide perhaps the most detailed site-specific information on picosecond-nanosecond water interactions.[22-24] Despite the success of extracting site-specific information from these experiments, the time scale of hydrogen bond fluctuations of water usually lies in fs–ps time scale,[25] which

cannot be captured by these experiments. Ultrafast infrared spectroscopy serves as a unique tool with structural sensitivity and fs–ps temporal resolution.[25-26]

Of growing interest are methods in computational spectroscopy of amide I vibrations which allow for direct comparison of the infrared spectroscopy of proteins and peptides with structure and trajectories from MD simulations.[27] Amide I vibrational modes are primarily C=O stretching motion, and sense hydrogen bonding interactions to the amide oxygen and hydrogen. To identify the molecular origin of amide I spectral features, spectroscopic frequency maps have been developed to translate local electrostatics from MD simulation such as electrostatic potential, electric field, and gradient into an amide I frequency.[28-34] Maps for vibrational coupling between different amide I oscillators or between different amide modes are used to calculate the interaction of multiple backbone amide groups in peptides and proteins.[35-39] These methods are now reaching a point of quantitative accuracy at which MD simulation data can be used to analyze experimental observables in terms of detailed solvation structure and dynamics, and comparison of experimental and simulated spectra can help assess the accuracy of force fields and water models.

At the most basic level, such studies have the ability to reveal the dynamical interaction of water with the individual peptide groups of the protein backbone. The influence of water on amide I vibrations has been extensively investigated with IR spectroscopy of N–methylacetamide (NMA), one of the simplest molecules with a single amide group.[26, 40-42] Computational spectroscopy of NMA has shown that water dynamics plays a key role in vibrational dephasing of the amide I mode.[29, 42-46] In particular time-resolved IR spectroscopy experiments have been successfully described in terms of the time-dependent shifts in the electrostatic potential or electric field that results from water hydrogen bonding dynamics.[32, 47-48] Experimental and computational studies have demonstrated that vibrational relaxation of NMA in D_2O is also influenced by hydrogen bond

breaking and reforming dynamics. Relaxation proceeds in a biphasic manner with a fast sub-picosecond intra-molecular component and a picosecond intermolecular relaxation process mediated by the solvent.[41-42, 49-52]

These extensive studies of NMA have provided considerable physical insight into the role of water interactions on amide I vibrational spectroscopy. The current study is motivated to build on this foundation and the recent advances in amide I computational modeling to develop a detailed molecular picture of the structure and dynamics of water associating with peptide linkages that form a protein backbone. What are the hydrogen bonding patterns present? What are the time-scales for water hydrogen-bond fluctuations and hydrogen bond exchange? How is water hydrogen-bonding dynamics coupled with the local peptide conformational dynamics? With an accurate spectroscopic model, this information can be experimentally addressed with the help of MD simulations. At the same time, the correspondence between experimental observables and computational predictions allows a way of testing the influence of the specific force fields and water models used in simulating the experiments.

In this work, we present a joint experimental-computational study of the relationship between the amide I vibrational dynamics and the solvation of the amide unit in dialanine (NH_3^+–AA–COOH, Ala–Ala). Ala–Ala was chosen as a simple peptide with one amide group (rather than the amide terminated alanine dipeptide[53-55]), with linkages to charged terminal groups that are more representative of a solvent-exposed protein environment than NMA. We perform 2D IR spectroscopy and transient absorption experiments, and compare these with MD simulation and spectral modeling to help us elucidate the molecular origin of the observed dynamics such as spectral diffusion, variation of vibrational lifetime, and water hydrogen-bond exchange times in Ala–Ala.

6.3 Materials and Methods

Ala–Ala was purchased from Sigma-Aldrich and used without further purification. The peptide was dissolved to a concentration of 200 mM (30 mg/mL) in 1 M DCl in D_2O to avoid spectral overlap between the amide I vibration and the water bend vibration, and to protonate the carboxyl-terminus to shift its carbonyl vibration to ~1720 cm^{-1}. Although we report amide I' spectra, for simplicity we use the terms amide I and amide I' interchangeably throughout this study. For all of the IR measurements, the sample was held between two CaF_2 windows spaced by a 50 μm Teflon spacer to have OD ~0.4.

FTIR spectra were collected at room temperature using a Bruker Tensor 27 FTIR spectrometer with 64 averages at 2 cm^{-1} resolution. A background spectrum of 1 M DCl in D_2O was measured for subtracting the solvent vibrational profile from the sample spectrum. A linear baseline correction from 1550 cm^{-1} to 1800 cm^{-1} was applied to flatten the baseline of the subtracted sample spectrum.

IR transient absorption spectra and 2D IR spectra were acquired in a pump-probe geometry 2D IR spectrometer at room temperature described elsewhere.[56] The waiting time was scanned from –0.1 ps to 5.0 ps for transient absorption spectra, and from 0.15 ps to 5.0 ps for 2D IR spectra. The 2D IR spectra were collected with both parallel (ZZZZ) polarization and perpendicular (ZZYY) polarization. The magic angle spectra were reconstructed from the parallel-polarized spectra and the perpendicular-polarized spectra by $I_{\mathrm{MA}} = I_{\mathrm{ZZZZ}} + 2I_{\mathrm{ZZYY}}$.

MD simulations of protonated Ala–Ala in water box were performed using GROMACS 4.6.7 package.[57] Force fields used in this study were CHARMM27[58-59] (C27), CHARMM36[60]

(C36), and CHARMM36m[61] (C36m). Water models used in C27 and C36 were SPC/E and TIP3P while in C36m we used SPC, SPC/E, TIP3P, TIP4P, TIP5P, and the deuterated SPC/E model modified by changing the mass of hydrogen to the mass of deuterium. The simulation boxes were set to a dodecahedron geometry, with the walls set at least 1 nm away from the peptide. One chloride ion was added to balance the charge of the Ala–Ala. The energy of the protein and solvent configuration was minimized to guarantee a reasonable starting structure for further equilibrations. A 100 ps temperature equilibration of solvent and ions around the position-restrained peptide at 300 K with the Berendsen thermostat[62] was performed. After the temperature equilibration, the density of the box was adjusted by a 1 ns NPT equilibration at 1 bar with the Berendsen thermostat and barostat.[62] The subsequent 10 ns preparation runs on the unrestrained peptide were performed using the Nosé−Hoover thermostat[63-64] under NVT conditions, and the 100 ns production runs were simulated with 1 fs integration step, and 20 fs/frame sampling rate for spectral simulations. The MD data analysis such as backbone dihedral angles was performed using PLUMED 2.[65]

Spectroscopic simulations were carried out by a mixed quantum/classical mapping approach described elsewhere.[27, 34] Briefly this method maps a time-dependent collective variable (such as the electric field acting on the amide carbonyl oxygen) from a classical MD simulation to a time-dependent quantum mechanical transition energy trajectory or transition dipole moment trajectory, which is then used for time-domain calculations of IR spectra. Amide I spectral parameters were computed along the MD trajectories using g_amide, an open source program available on GitHub.[66] The spectroscopic maps used in this study were the one-site field map (1F) that has been parametrized against experimental dipeptide FTIR data,[34] and the four-site potential map (4P) optimized against experimental spectra of the isotope-labeled NuG2b protein.[34] This 1F map uses modified glycine charge and TIP3P water model charges instead of SPC/E charges,

different from our previous 1F map.[33] Note that all of these mapping approaches do not include the carbonyl stretch of the carboxyl-terminus into spectral simulations.

The exciton Hamiltonian trajectories and corresponding dipole moment trajectories were used to compute linear IR and 2D IR spectra by the dynamic wavefunction propagation method,[67] using the home-built spectral simulation program, g_spec.[68] The window time for calculating response functions was set to 11 ps, which is equivalent to 3 cm^{-1} frequency resolution. For calculating time-averaged response functions, starting structures were separated by 2 ps, resulting in 50,000 realizations for a 100 ns trajectory. The model includes a vibrational lifetime for the amide I mode, set to the 1.0 ps value measured in our transient absorption experiment.

To better sample the solvation environment around Ala–Ala at particular backbone dihedrals, 2D umbrella sampling is performed using PLUMED 2 and Gromacs. The starting structures are chosen at $(\phi,\psi) = (-80,-180)$, $(-80,-140)$, and $(-50,-130)$. The force constant of the harmonic biased potential is set to 327.5 kJ/mol on both ϕ and ψ, corresponding to 5º standard deviation at 300 K. Both the equilibration run and the production run are 1 ns long, with the same parameters used in the unbiased MD simulations except for the biased potential applied.

6.4 Results

In Fig. 6.1a, the experimental FTIR spectrum of Ala–Ala in D_2O shows an amide I peak frequency centered at 1670 cm^{-1}, and an asymmetric lineshape with a width of 31 cm^{-1} FWHM. The frequency is 50 cm^{-1} higher than the resonance frequency for the commonly studied NMA in D_2O,[42] which suggests a significant difference in the solvation environment for Ala–Ala primarily due to the protonated terminal amino group.[27, 30-32, 43] The 2D IR spectrum at early waiting time

($\tau_2 = 0.15$ ps) is peaked at the same frequency and has an asymmetric lineshape that is diagonally elongated, characteristic of inhomogeneous broadening. The less intense but broader peak at 1720 cm^{-1} originates from the terminal carboxylic acid C=O stretch. The 2D spectrum reveals a weak cross-peak between the two carbonyl resonances at the lowest contours of our spectrum. However, the coupling from Fig. 6.1a is estimated to be only ~1 cm^{-1},[69] which we interpret to mean that the coupling is weak enough that they do not need to be explicitly included in our amide I spectral modeling.

For comparison, simulated FTIR and 2D IR amide I spectra using the same C36 force field but different water models and spectral maps are shown in Figs. 6.1b–d. Almost identical FTIR lineshapes are found between SPC/E and TIP3P simulations, but different peak frequencies and subtle lineshape changes can be observed in the 2D spectra (Figs. 6.1b–c). However, none of these simulations reproduce the asymmetry of the FTIR experiment, and the experimental 2D lineshape. The simulated spectra from the 1F map consistently show less elongation along the diagonal than the experimental spectra (similar to another study[70]), while the simulated spectra from the 4P map have comparable diagonal width to the experimental 2D spectrum (Fig. 6.1d). Recognizing that solvation configurations near the amide group vary for different force fields and water models, it is possible that simulations overestimate the population of more hydrogen-bonded species. The amide I FTIR peak frequencies and FWHM from experiment and C36 simulations are summarized in Table 6.1.

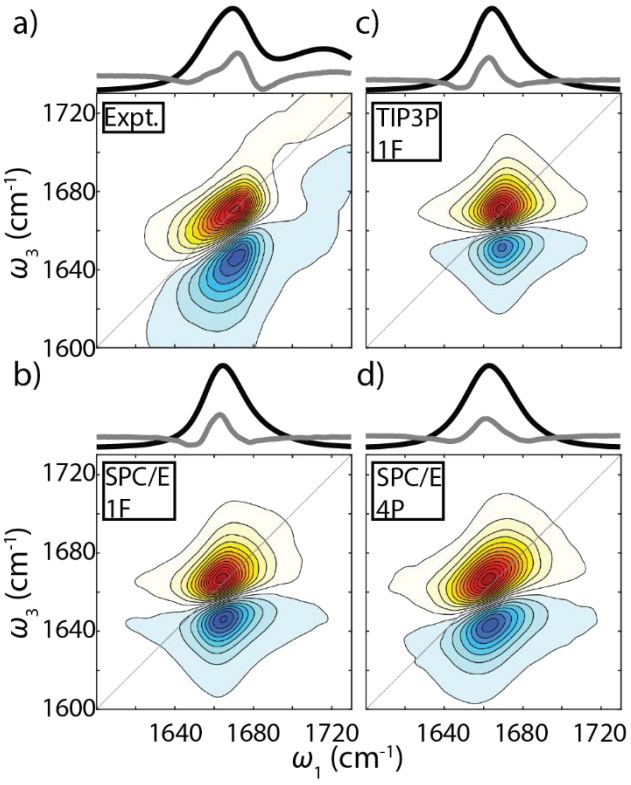

Figure 6.1: Amide I FTIR spectrum (black curve), second derivative of the FTIR spectrum (gray, scaled by -25) and ZZZZ polarized 2D IR spectrum of Ala–Ala. The waiting time of these 2D spectra is 0.15 ps. (a) Experimental spectrum. (b) Simulated spectrum from C36 SPC/E trajectory and the 1F map. (c) Simulated spectrum from C36 TIP3P trajectory and the 1F map. (d) Simulated spectrum from C36 SPC/E trajectory and the 4P map.

Examples of the waiting time dependence of experimental 2D IR spectra are shown in Fig. 6.2. The 2D lineshape changes between $\tau_2 = 0$ and 5 ps show a characteristic evolution of the lineshape from diagonally elongated to symmetric, and a rotation of the node between positive and negative features. These characteristics are commonly associated with vibrational spectral diffusion and can be extracted using the center line slope (CLS) method.[71-72] The results of this analysis are shown in Fig. 6.2, and exhibit single exponential decay with a 1.3 ps timescale. Experimental CLS decays of different frequency regions show similar dynamics, and nonlinearity of the center lines across the waiting time series is not significant (Fig. 6.3). The experimental CLS decay is compared with the CLS decays from simulated spectra in Fig. 6.2b. The simulation decays

are both much faster than the experimental CLS, and we find that TIP3P water decays with a 0.55 ps timescale, even faster than SPC/E water, with a 0.9 ps time scale. The time scale of CLS decays are summarized in Table 6.2.

	ω_{peak} (cm^{-1})	FWHM (cm^{-1})
Exp.	1670	31
SPC/E, 4P	1663	30
TIP3P, 4P	1666	28
SPC/E, 1F	1665	26
TIP3P, 1F	1671	21

Table 6.1: FTIR Peak frequency ω_{peak} and FWHM from the experiment and the C36 simulations.

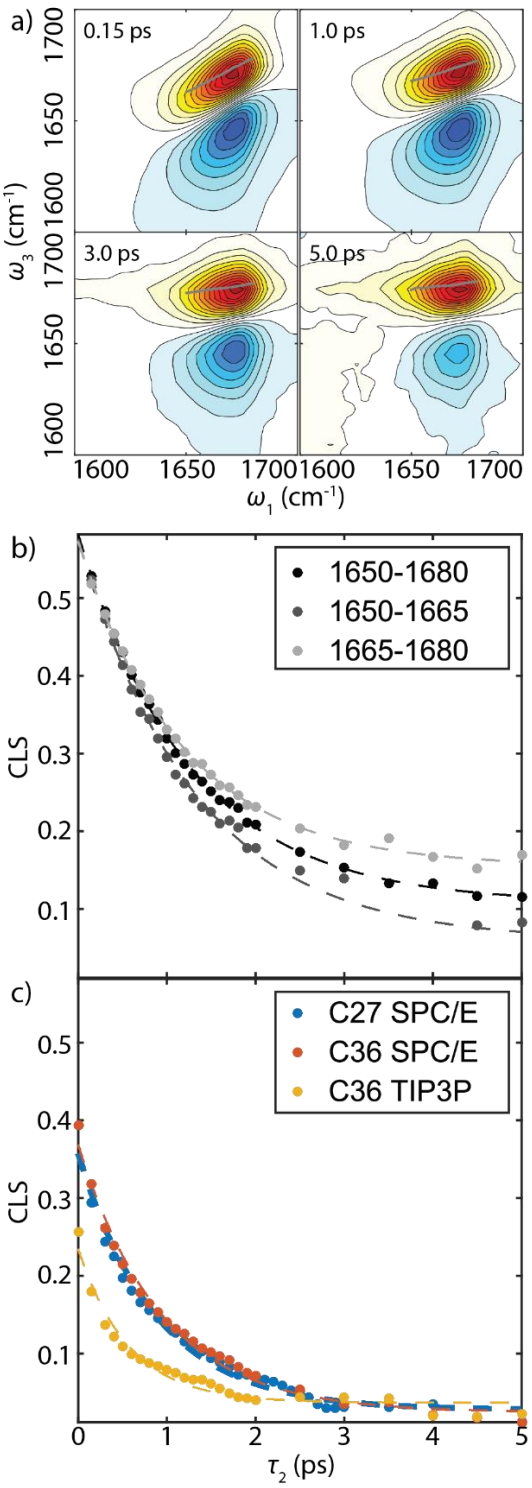

Figure 6.2: (a) Selected experimental waiting time series of magic angle 2D IR spectra from 0.15 ps to 5 ps. Center line derived from ω_1 slices are shown in gray dots. (b–c) CLS decays for varying frequency ranges (dots) and the corresponding fit curves (dashed lines) from (b) the experiment, and (c) from the simulations using the 1F map.

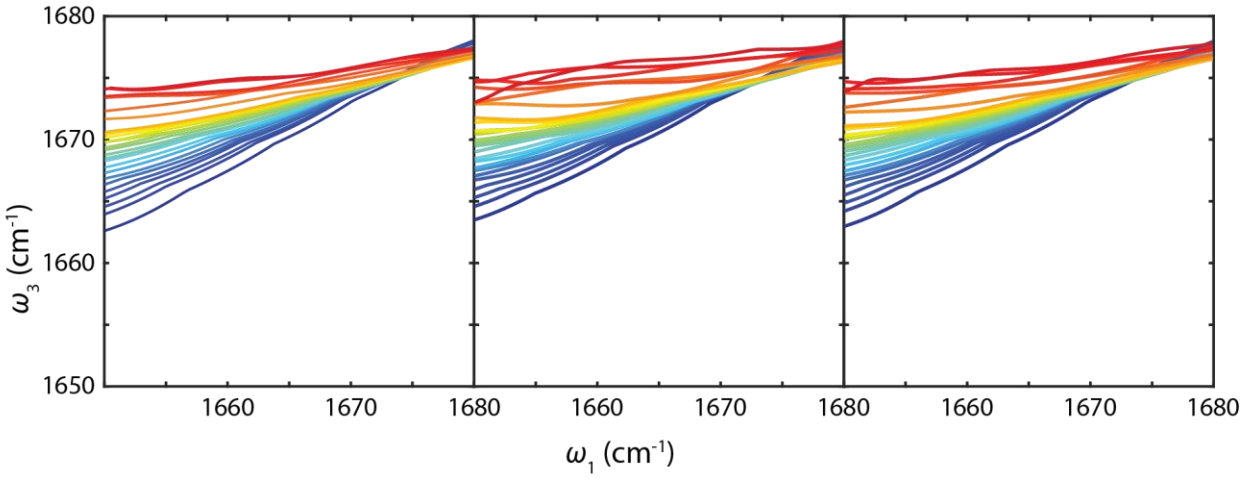

Figure 6.3: (a) Selected experimental waiting time series of magic angle 2D IR spectra from 0.15 ps to 5 ps. Center line derived from ω_1 slices are shown in gray dots. (b–c) CLS decays for varying frequency ranges (dots) and the corresponding fit curves (dashed lines) from (b) the experiment, and (c) from the simulations using the 1F map.

	1650-1680	1650-1665	1665-1680	C27 SPC/E	C36 SPC/E	C36 TIP3P
a	0.47	0.52	0.42	0.33	0.34	0.20
b (ps)	1.27	1.31	1.18	0.88	0.94	0.57
c	0.11	0.059	0.15	0.027	0.024	0.037

Table 6.2: (a) Selected experimental waiting time series of magic angle 2D IR spectra from 0.15 ps to 5 ps. Center line derived from ω_1 slices are shown in gray dots. (b–c) CLS decays for varying frequency ranges (dots) and the corresponding fit curves (dashed lines) from (b) the experiment, and (c) from the simulations using the 1F map.

Vibrational lifetime measurements using transient absorption are shown in Fig. 6.4. From the bleach and induced absorption of the pump-probe measurements, the amide I intensity decay of Ala–Ala is observed to be bi-exponential with time scales of 0.23 ps and 1.1 ps, comparable with the reported 1.2 ps from Hamm *et al.*[26] As a comparison, the relaxation time scales of N-methylacetamide-d_7 in D_2O have been found to be 0.38 ps and 2.1 ps, and the time-scales were attributed to energy exchange between anharmonically coupled amide I and II vibrations followed

by solvent-mediated dissipation of vibrational energy.[49] The decay times of the carboxylic acid C=O stretch are about 0.23 and 0.78 ps (Fig. 6.4b).

As a comparison, we integrated spectral regions of the waiting time series of 2D IR spectra (Fig. 6.4c), choosing two frequency windows that correspond to the peaks in the experimental second derivative spectrum (Fig. 6.1a). Integrating the upper diagonal at 1670–1675 cm^{-1} gives a 1.0 ps exponential decay, while the decay time of the peak integration at 1660–1665 cm^{-1} is 0.8 ps. This difference suggests a vibrational lifetime that varies with frequency, but the comparison must account for the influence of lineshape changes due to spectral diffusion, or chemical exchange. In addition, the integrated peak intensity of the off-diagonal region in Fig. 6.4c which might reveal exchange processes has the decay time scale of 1.14 ps—longer than the two diagonal blocks. Since there is no intermolecular energy transfer between the dilute amide groups in our experiment, this indicates that a chemical exchange between different hydrogen bonding solvation environments may also be present. Thus, vibrational lifetime, vibrational dephasing, and chemical exchange processes should all be considered in explaining the 2D waiting time dependence.

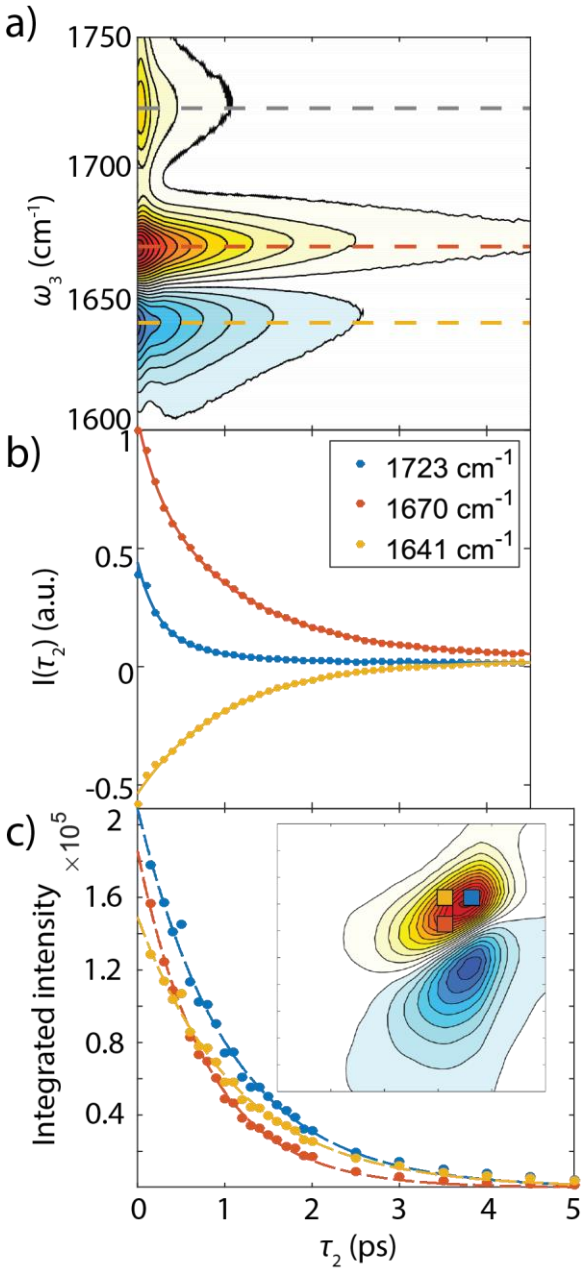

Figure 6.4: (a) Transient absorption spectrum of Ala–Ala. Dashed lines indicate the frequency slices for fitting in (b). (b) Intensity decays (Colored dots) and fit curves (solid colored lines) with respect to waiting time τ_2. (c) Integrated peak intensities as a function of τ_2. Integrated areas are represented by the colored rectangles.

To better isolate the frequency dependence of amide I vibrational lifetime, vibrational lifetime decays were obtained from single exponential fits to each point of the waiting time series of absolute value magic angle 2D IR spectra. The resulting fits are presented as a 2D lifetime

heatmap in Fig. 6.5a. Looking along the diagonal, the heatmap indicates a clear trend of increasing vibrational lifetime with frequency, with a variation from ~1.0–1.3 picoseconds (~25%) over the amide I resonance.

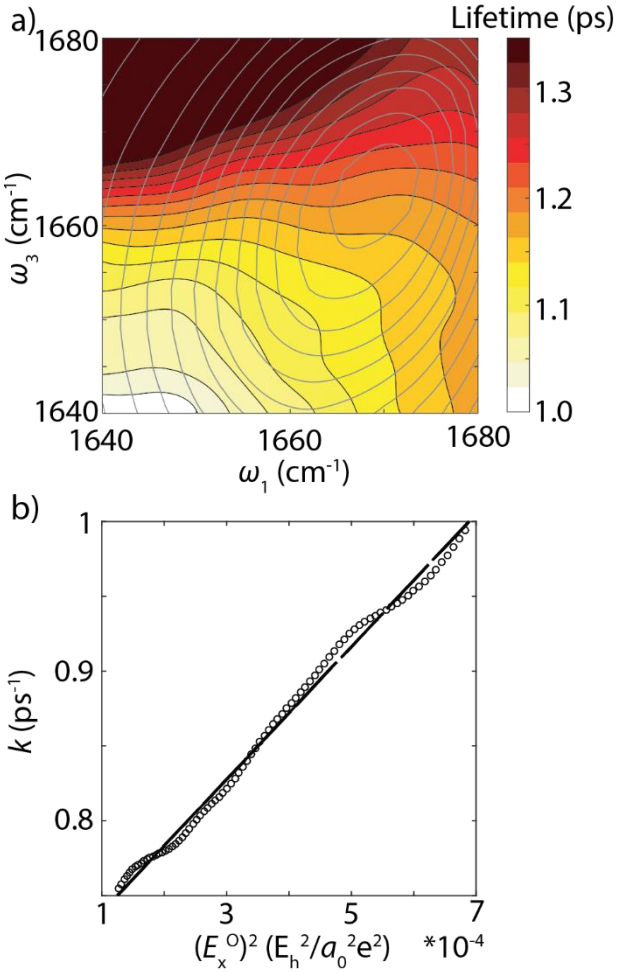

Figure 6.5: (a) Amide I lifetime heat map (solid contours) of Ala–Ala from magic angle absolute value surface (gray contours). Each colored contour line is space by 25 fs. (b) Scatter plot of relaxation rate and squared electric field exerted on amide oxygen atom along the C=O axis. Black line is the least square fit, $\left(E_x^O\right)^2 = 0.0023k - 0.001$, $R^2 = 0.99$.

The time scale of chemical exchange can be estimated from the slow chemical exchange model,[73-74] using data from Fig. 6.5. We selected frequency ranges for the two states involved in the chemical exchange process using the peaks at 1672 cm^{-1} (*U*) and 1659 cm^{-1} (*L*) in the second derivative FTIR spectrum (Fig. 6.1a). From Fig. 6.5, we found the vibrational relaxation rates at

these frequencies are $k_{\mathrm{U}} = (1.28\,\mathrm{ps})^{-1}$ and $k_{\mathrm{L}} = (1.16\,\mathrm{ps})^{-1}$. The decay rate of the cross-peak intensity $k_{CP} = (1.35\,\mathrm{ps})^{-1}$ will depend both on vibrational relaxation and chemical exchange rates as

$$k_{CP} \cong \frac{k_{\mathrm{L}} + k_{\mathrm{U}}}{2} - \sqrt{\left(\frac{k_{\mathrm{L}} - k_{\mathrm{U}}}{2}\right)^2 + k_{\mathrm{LU}} k_{\mathrm{UL}}} \tag{6.1}$$

Here k_{LU} is the chemical exchange rate from state *L* to state *U*, and *vice versa*. Using the experimental values of k_{L} and k_{U}, and enforcing detailed balance results in an upper bound of the chemical exchange rate $k_{\mathrm{LU}} < (15\,\mathrm{ps})^{-1}$.

6.5 Discussions

The time-resolved IR spectroscopy of the amide I vibration of dialanine indicates that a variety of spectrally varying dynamical processes are present, including spectral diffusion, chemical exchange, and vibrational relaxation. Each of these have similar time-scales and have closely related molecular origins, which we can investigate with the help of MD simulations and amide I spectral modeling. Spectral diffusion originates from vibrational frequency fluctuations, which are likely due to the fluctuating hydrogen bond environment in the vicinity of the amide carbonyl, while chemical exchange would reflect the coupled changes in peptide and water hydrogen bonding configurations. From a classical perspective, both vibrational frequency fluctuations and the vibrational lifetime can be described in the context of time correlation function of fluctuating forces experienced by the amide oscillator $C_F(t) \equiv \langle F(t)F(0)\rangle$. In this case, the

fluctuating forces are predominantly of electrostatic origin and depend intimately on hydrogen bonding to the amide group.

6.5.1 Vibrational Frequency Fluctuations

The frequency fluctuations characterized by the CLS are proportional to the amide I vibration frequency correlation function, $C_{\delta\omega}(t) \equiv \langle \delta\omega(t)\delta\omega(0)\rangle$, which depends on the time-dependent interaction potential of the amide I vibration with its environment.[27, 30-32, 43] The amide I spectral models used here assume a linear relationship between vibrational frequency and local electrostatic parameters such as field or potential. For Ala–Ala in water, we expect that the amide group is highly solvated, and therefore the frequency fluctuations would be most influenced by interactions with surrounding water molecules.

We investigated the role of water dynamics on the amide I vibrational dynamics using amide I spectral simulations with different water models. As shown in Fig. 6.6a, the decays of the simulated $C_{\delta\omega}(t)$ directly computed from the amide I frequency trajectories indicate that they are influenced by the water model and do not appear dependent on the force field used. The time-scales are faster than the experiment, but this is likely also a reflection of the water model. For instance, similar effects are observed in simulations of vibrational dephasing of the O–H stretching vibration of HOD in D_2O, with the finding that polarizable water models tend to predict longer, more-accurate frequency correlation functions than fixed charge models, such as the water models we have employed here.[75-76]

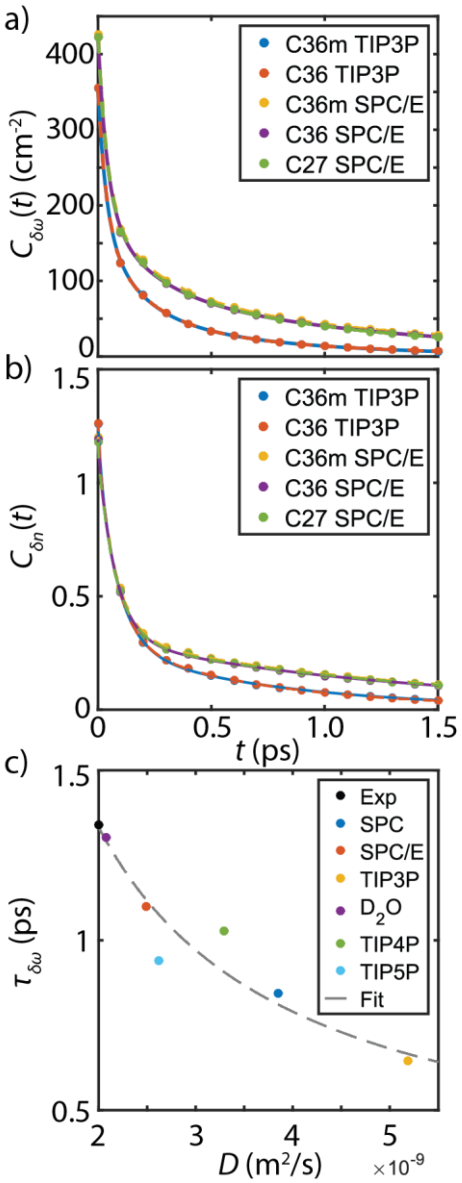

Figure 6.6: (a) $C_{\delta\omega}(t)$ computed from the frequency trajectories computed by the 1F map. The raw data points are presented as solid circles whereas the fits are represented as solid curves. The fit function is tri-exponentials with constant offset. (c) $C_{\delta n}(t)$ computed from MD trajectories. The raw data points are presented as solid circles whereas the fits are represented as solid curves. The fit function is tri-exponentials with constant offset. (d) Scatter plot of the frequency correlation time, $\tau_{\delta\omega}$, against the water/D_2O self-diffusion coefficient from the experiment and C36m simulations. Gray curve: the fit curve of the three-site water models by $aD^{-1}+b$.

The key role of the dynamics of water hydrogen bonds to the amide carbonyl is also shown in Fig. 6.6b. We analyzed hydrogen bonding dynamics using geometric criteria to define the

number of hydrogen bonds between water and the C, O, N, and H atoms of the amide group, n. When the distance between hydrogen bond donor and acceptor is ≤3.5 Å, and the donor-hydrogen-acceptor angle is ≥150°, the donor-acceptor pair is considered hydrogen-bonded. We found that the time correlation function for the hydrogen bond number, $C_{\delta n}(t) \equiv \langle \delta n(t) \delta n(0) \rangle$, drops in amplitude by ~2/3 in the first 200 fs as a result of fast fluctuations in water hydrogen bonding configurations. $C_{\delta n}(t)$ has a long time decay of 1.3 ps that matches the slow time scale of the $C_{\delta\omega}(t)$ decay (Fig. 6.6b). This indicates that the frequency fluctuations depend on the hydrogen bond fluctuation dynamics around the amide group, similar to prior observations.[45-46] The fit results are summarized in Table 6.3 and Table 6.4.

	a_1 (cm^{-2})	τ_1 (ps)	a_2 (cm^{-2})	τ_2 (ps)	a_3 (cm^{-2})	τ_3 (ps)	b (cm^{-2})
C36m TIP3P	176.3	0.036	115.8	0.16	59.9	0.65	0.84
C36 TIP3P	185.8	0.038	115.3	0.19	50.8	0.69	1.4
C36m SPC/E	232.6	0.042	109.4	0.26	79.6	1.28	4.1
C36 SPC/E	223.7	0.041	103.8	0.21	91.1	1.10	2.8
C27 SPC/E	226.8	0.041	100.6	0.22	92.1	1.08	2.4

Table 6.3: Parameters for the fit of $C_{\delta\omega}(t)$ in Fig. 6.5. The model used is $C_{\delta\omega}(t) = \sum_{i=1}^{3} a_i \exp(-t/\tau_i) + b_i$.

	a_1	τ_1 (ps)	a_2	τ_2 (ps)	a_3	τ_3 (ps)	b
C36m TIP3P	0.15	0.014	0.83	0.087	0.28	0.74	0.005
C36 TIP3P	0.12	0.011	0.84	0.081	0.30	0.67	0.006
C36m SPC/E	0.06	0.007	0.82	0.077	0.31	1.27	0.017
C36 SPC/E	0.05	0.006	0.83	0.077	0.30	1.26	0.016
C27 SPC/E	0.06	0.007	0.81	0.078	0.30	1.30	0.015

Table 6.4: Parameters for the fit of $C_{\delta n}(t)$ in Fig. 6.5. The model used is $C_{\delta n}(t) = \sum_{i=1}^{3} a_i \exp(-t/\tau_i) + b_i$.

Since TIP3P water is known to have a higher diffusion coefficient than the SPC/E water,[77] it allows us to test for a correlation between the frequency correlation time $\tau_{\delta\omega}$ and the diffusion coefficients D of various water models. The diffusion coefficients of several different water models and the corresponding correlation time for the decay of $C_{\delta\omega}(t)$ are plotted in Fig. 6.6c. We find that $\tau_{\delta\omega}$ obeys a D^{-1} dependence on the water model diffusion coefficient, with a trend that also describes the experimental data.[78] In combination with our previous observations, this indicates that experimental observations of spectral diffusion are tracking the same hydrogen bond reorganization about the amide group that governs self-diffusion in water. TIP4P and TIP5P water models deviate somewhat from the curve, probably due to different treatment of electrostatic interactions between the amide group and water.

6.5.2 Vibrational Lifetime

Since the IR spectroscopy of the amide I vibration can be effectively described through a fluctuating electric field generated by the environment, and recognizing that the electric field is an

electrostatic force acting on the vibration, the vibrational dephasing described by $C_{\delta\omega}(\tau)$ should be proportional to a time correlation function for the fluctuating force acting on the amide I coordinate: $C_F(\tau) = \langle F(\tau)F(0)\rangle$. In a classical representation, the same correlation function can be related to the amide I vibrational relaxation rate as follows.[79-80]

$$k_{\text{obs}} = \sum_{i,j} k(\omega_{ij}) \propto \sum_{i,j} \text{Re} \int_{-\infty}^{\infty} dt e^{i\omega_{ij}t} C_F(t) \tag{6.2}$$

Here ω_{ij} is the vibrational energy gap defining the relaxation initial and final states. The vibrational relaxation rate is proportional to the Fourier component of the fluctuating force correlation function evaluated at the vibrational energy gap, or equivalently proportional to the spectral density at ω_{ij} in the high temperature limit. Note that the vibrational population relaxation of the amide I band will have contributions from many relaxation channels beyond simple v=1→0 relaxation, such as amide I to II intramolecular vibrational energy transfer, solvent-mediated dissipation, and other anharmonic relaxation channels.[42, 49, 81]

Assuming the electric field experienced by the carbonyl oxygen of the amide group is the dominant source of the fluctuating force, then the relaxation rate would also be proportional to the corresponding electric field time correlation function: $C_F(t) \propto \langle E_x^{\text{O}}(t)E_x^{\text{O}}(0)\rangle$. If the $C_F(t)$ correlation time also does not vary much with amide I frequency, a reasonable assumption from our experiments, then the rate is approximately proportional to the magnitude of the squared electric field strength. To investigate the relationship between the experimental vibrational relaxation rate and the electric field strength exerted on the amide oxygen along the C=O axis E_x^{O} , instead of directly evaluating $\langle E_x^{\text{O}}(t)E_x^{\text{O}}(0)\rangle$ from the simulation and estimating the relaxation rate, we estimate the corresponding $E_x^{\text{O}}(\omega)$ at each experimental frequency in the heatmap (Fig. 6.5a)

using the 1F map, and correlate $\left(E_x^{\mathrm{O}}(\omega)\right)^2$ with the relaxation rate $k(\omega)$ shown in Fig. 6.5b. These values are strongly linearly correlated, suggesting that the electrostatic interactions can account for the vibrational relaxation of the amide I mode. Increasing force exerted along the C=O axis effectively reduces the vibrational lifetime. Note that the 4P map has an equivalent translation to the field-based 3F map.[34] Although translating experimental frequency back to electric field by the 3F map would be ambiguous, the 3F map samples similar fluctuation dynamics around the amide group as the 1F map (Fig. 6.7). The assumption that electric field experienced by the carbonyl oxygen of the amide group is the dominant source of the fluctuating force should be still reasonable.

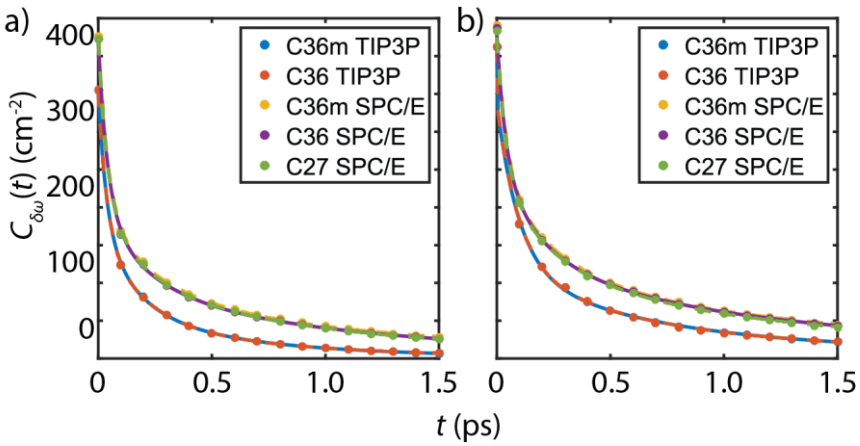

Figure 6.7: (a) $C_{\delta\omega}(t)$ computed from the frequency trajectories computed by the 1F map. (b) $C_{\delta\omega}(t)$ from the frequency trajectories computed by the 4P map, or the equivalent 3F map.

For the classical description of vibrational relaxation and vibrational dephasing, this fluctuating force can be thought of as effective friction acting on the amide vibration. From the fluctuation-dissipation theorem, the friction coefficient for such processes is related to the force-force correlation function as $\gamma = (2mk_{\mathrm{B}}T)^{-1}\int_{-\infty}^{\infty} dt C_F(t)$, which is proportional to the zero-frequency component in the Eq. (6.2). The friction coefficient can be expressed in terms of the diffusion coefficient in water through $D = k_{\mathrm{B}}T/\gamma$. In the simplest picture the direct

proportionality of these quantities predicts that the amide I frequency correlation function should be inversely related to water's diffusion coefficient, as observed in Fig. 6.6c. Therefore, we conclude that fluctuating forces originating in water's hydrogen bond fluctuations and switching appear to be the primary origin of both experimental relaxation processes we observe.

6.5.3 Peptide Solvation Environment and Chemical Exchange

To characterize the relationship between peptide conformation and amide group solvation patterns, we calculated the probability distribution for the average hydrogen bond number $\langle n \rangle$ as a function of the backbone dihedral angles (ϕ, ψ). This is presented as a color-coded potential of mean force (PMF) in Fig. 6.8a. It shows clear variation of $\langle n \rangle$ with peptide backbone conformation between values of 1 and 2, indicating that the solvation structure around the amide group is coupled to the peptide configuration. A similar, though less prominent, hydrogen bond number distribution is observed when only counting those water hydrogen bonds made to the amide oxygen atom (Fig. 6.8b). Although Figs. 6.8a–b appear to have only one free energy basin, it is well established that the dihedral angles are not effective coordinates for describing peptide conformational dynamics such as alanine dipeptide.[13-14] Therefore one cannot use it to discern dynamical behavior, for instance whether changes of $\langle n \rangle$ are due to fast fluctuations in solvent hydrogen bonding configurations or slower global conformational changes. Similar considerations apply to distinguishing solvation structures, as seen when we plot the probability distribution as a function of $\langle n \rangle$ (Fig. 6.8c). The Ala–Ala conformational distribution has two peaks corresponding to solvent configurations with $\langle n \rangle$ = 1.4 and 1.8.

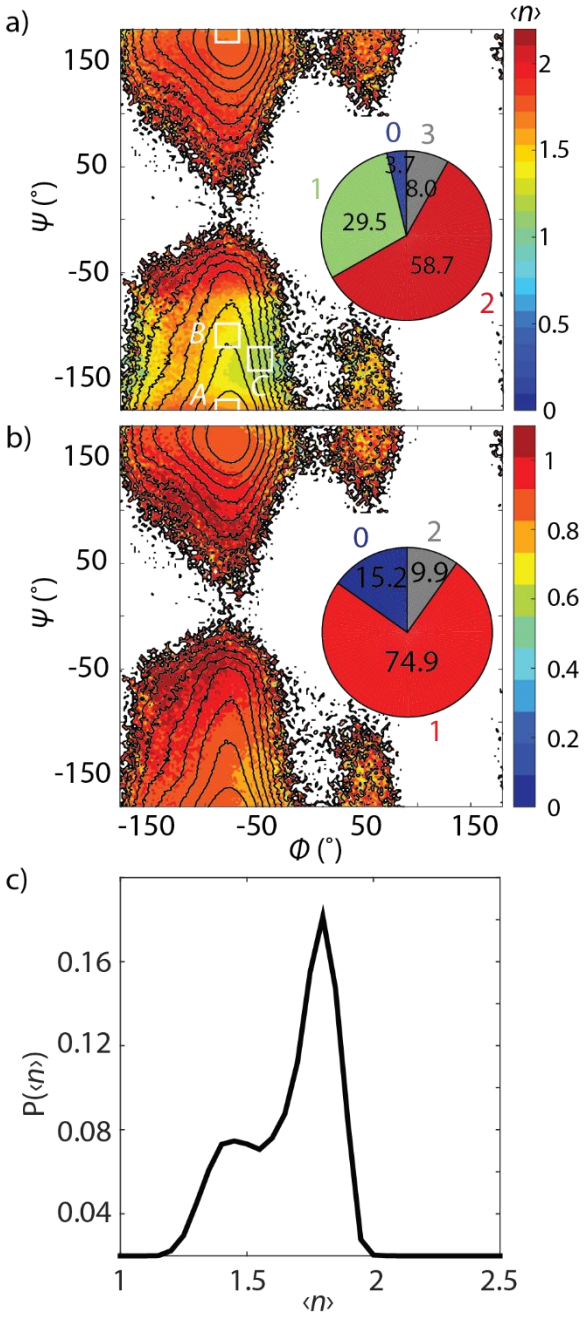

Figure 6.8: (a) Colored contours: $\langle n \rangle$ as a function of backbone dihedrals ϕ and ψ from C36 SPC/E trajectory. Black contour lines: PMF spaced by $k_B T$ up to 10 $k_B T$ at 300 K. PMF is computed by $\mathrm{PMF}(\phi,\psi) = -k_B T \ln P(\phi,\psi)$, where $P(\phi,\psi)$ is the probability of observing Ala–Ala at (ϕ,ψ). Inset: Population of different hydrogen bonding configurations from SPC/E water to the amide group. The black rectangular boxes represent the states for estimating the first passage time in Fig. 6.12. (b) Colored contours: average hydrogen bond number to the amide carbonyl $\langle n_{C=O} \rangle$ as a function of ϕ and ψ. (c) Probability distribution as a function of $\langle n \rangle$.

To investigate the relationship between peptide configuration and solvation structure, we investigated the structure and dynamics about three Ala-Ala configurations chosen on the basis of their dihedral angle and water hydrogen bond count. State *A* is the low energy configuration on the PMF in Fig. 6.8a at $(\phi_A, \psi_A) = (-80°, -180°)$, and is characterized by $\langle n_A \rangle = 1.73$. At an energy of ~$2k_BT$ above state *A*, state *B* was chosen as $(\phi_B, \psi_B) = (-80°, -140°)$ to correspond to the lower value of $\langle n_B \rangle = 1.4$ observed for the minor peak in Fig. 6.8c. State *C*, at $(\phi_C, \psi_C) = (-50°, -130°)$, corresponds to the lowest average value of the hydrogen bond counts, $\langle n_C \rangle = 1.25$, and lies ~$4k_BT$ in energy above state *A*.

In Fig. 6.9 we present visualizations of solvation structures of water around the Ala–Ala amide unit for states *A*, *B*, and *C*. Although the dihedral angles vary little between states, there are noticeable effects on hydrogen bonding to water. While *A* and *B* have on average ~1 water hydrogen bond to the oxygen of the amide carbonyl, we find that *A* differs from *B* on average by the presence of more hydrogen bonds from the amide hydrogen to a water oxygen (Figs. 6.8a–b). Although the solvent accessible surface area of the amide group across the entire PMF is uniform (Fig. 6.10), we find a clear correlation between the hydrogen bond number from the amide hydrogen to water, $\langle n_{\mathrm{N-H}} \rangle$, and the distance between the carbon of the N-terminal methyl side-chain and the amide hydrogen, $\langle d_{\mathrm{C\cdots H}} \rangle$ (Figs. 6.11a and 6.11c). This indicates that the configuration of the methyl side chain influences if water can hydrogen bond to the amide hydrogen. State *C* differs from *A* and *B* by rotating ψ, resulting in a shorter distance between the COOH group and amide carbonyl, $\langle d_{\mathrm{C\cdots O}} \rangle$ (Fig. 6.11b). A decrease in $\langle d_{\mathrm{C\cdots O}} \rangle$ leads to lower hydrogen bond number to the amide carbonyl $\langle n_{\mathrm{C=O}} \rangle$ (Fig. 6.11d), indicating that the configuration

of the COOH group also can act to influence the water hydrogen bonded to the amide carbonyl. The differences in solvent configurations between states are also observed as the water density becoming more diffuse with decreasing $\langle n \rangle$. We conclude that water hydrogen bonding patterns to the amide group are intimately coupled to the peptide conformation generally, and the configuration of the methyl sidechain and the COOH group in particular.

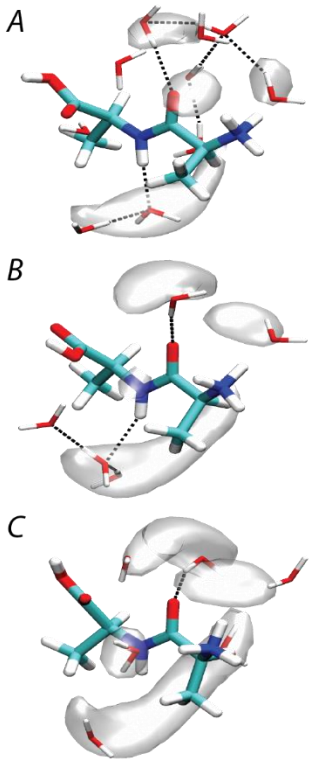

Figure 6.9: Solvation structures and water probability density (transparent isosurface) of Ala–Ala at states *A*, *B* and *C* from 2D umbrella sampling. (A) $(\phi_A, \psi_A) = (-80^\circ, -180^\circ)$, $\langle n_A \rangle = 1.8$. (B) $(\phi_B, \psi_B) = (-80^\circ, -140^\circ)$, $\langle n_B \rangle = 1.4$, and (C) $(\phi_C, \psi_C) = (-50^\circ, -130^\circ)$, $\langle n_C \rangle = 1.25$. Representative solvent configurations are plotted on top of mass-weighted isosurfaces for water that are within 3.5 Å of the amide group. The isosurfaces are plotted with isovalue at 40% of the maximum. Black dashed lines correspond to hydrogen bonds.

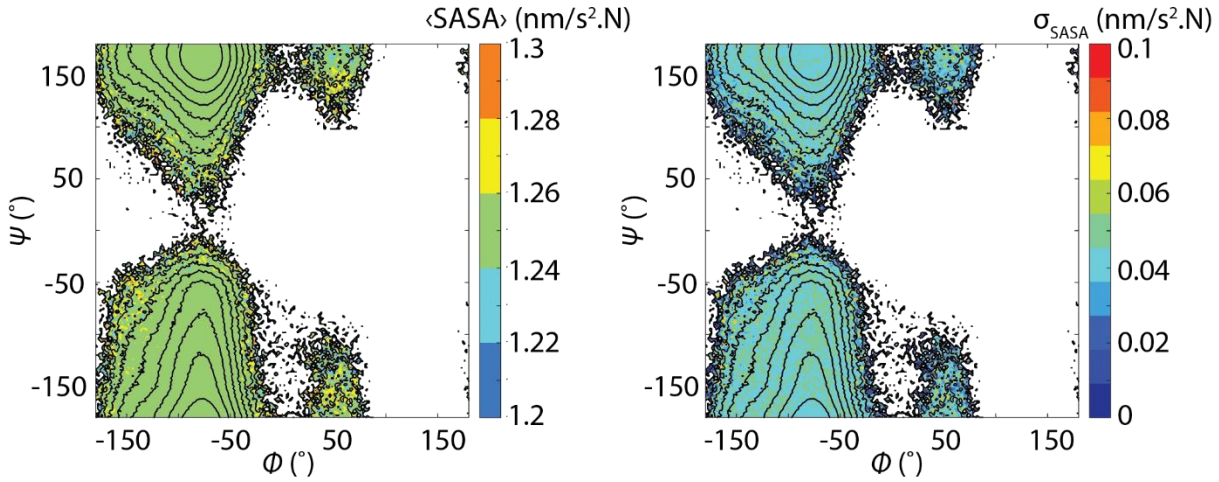

Figure 6.10: Mean (left) and standard deviation (right) of the solvent accessible surface area (SASA) of the amide group as a function of ϕ and ψ from C36 SPC/E trajectory. Black contour lines: Potential of mean force (PMF) spaced by k_BT up to 10 k_BT at 300 K. C36 TIP3P trajectory has an identical distribution.

Turning to dynamics, to estimate the chemical exchange time scale from the simulations, we computed the first passage time (FPT) distributions between states *A*, *B* and *C* (Fig. 6.12). We find that the FPT distributions are well described by the asymmetric form with a $t^{-3/2}$ tail expected for random walk diffusion. The mean FPTs between pairs of states are summarized as follows.

$$A \underset{2.6\text{ ps}}{\overset{7.8\text{ ps}}{\rightleftarrows}} B \underset{1.3\text{ ps}}{\overset{15\text{ ps}}{\rightleftarrows}} C \underset{50\text{ ps}}{\overset{9.3\text{ ps}}{\rightleftarrows}} A$$

All mean FPTs are found to lie between 1–50 ps, which are long compared to the time-scales of local hydrogen bonding fluctuations. The longest time-scales are associated with the $A \rightleftarrows C$ equilibrium, and are of similar time-scale to the estimated chemical exchange rate from 2D IR experiments. Given the correlation we find between solvation structures and peptide dihedral angles, these exchange processes are expected to involve peptide conformational changes coupled with solvent reorganization. We also observe that the most probable FPT for all transitions involving state *B* are <1 ps, indicating that configurational changes associating with this state may

in some cases involve fast fluctuations in solvent structure and peptide conformation, without an irreversible conformational exchange.

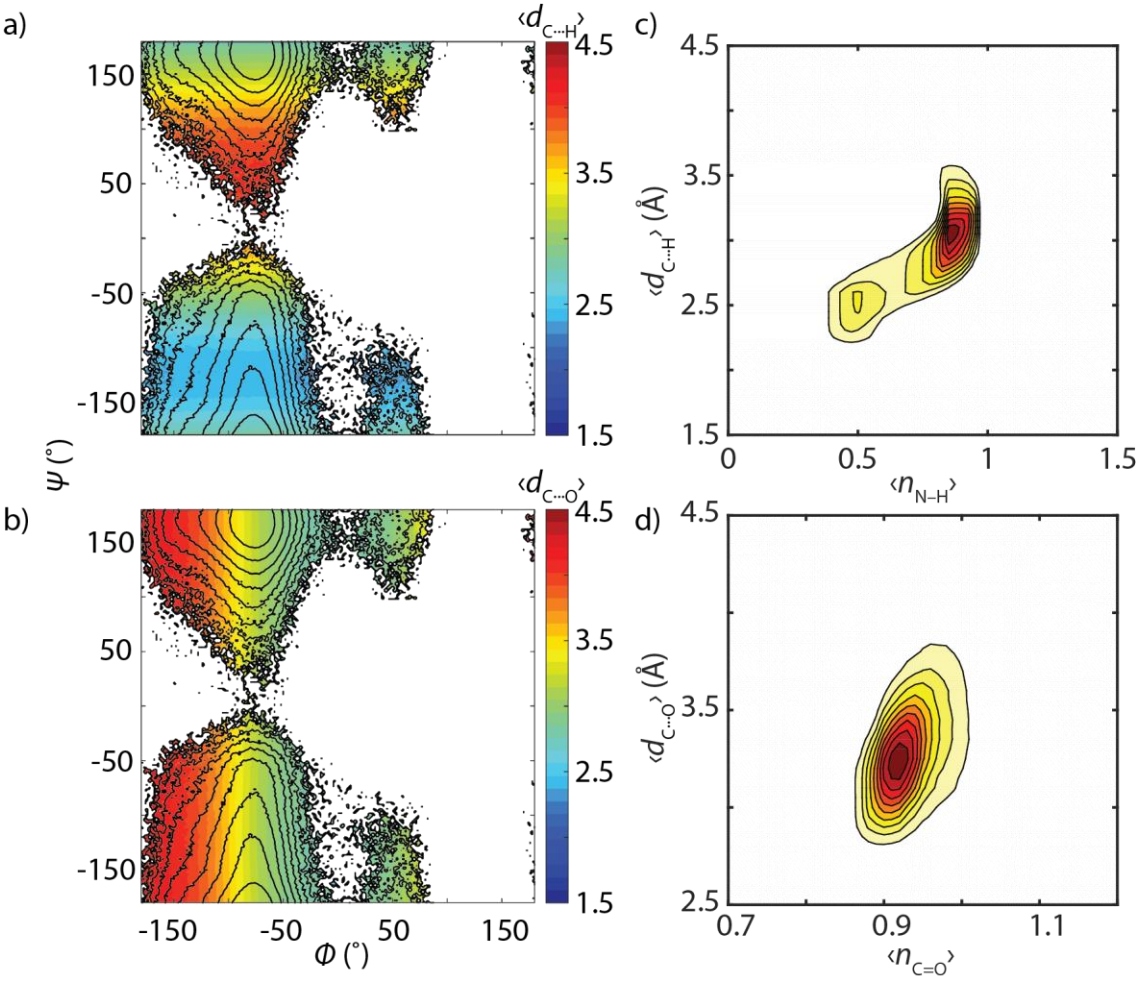

Figure 6.11: (a) Colored contours: Mean distance between N–terminal side chain methyl carbon and amide hydrogen $\langle d_{\mathrm{C\cdots H}} \rangle$ as a function of backbone dihedrals. Black contour lines: PMF spaced by $k_{\mathrm{B}}T$ up to 10 $k_{\mathrm{B}}T$ at 300 K. (b) Colored contours: Mean distance between carboxylic acid carbon and amide oxygen $\langle d_{\mathrm{C\cdots O}} \rangle$ as a function of backbone dihedrals. Black contour lines: PMF spaced by $k_{\mathrm{B}}T$ up to 10 $k_{\mathrm{B}}T$ at 300 K. (c) Joint probability distribution of average hydrogen bond number from amide hydrogen to water $\langle n_{\mathrm{N-H}} \rangle$ and $\langle d_{\mathrm{C\cdots H}} \rangle$ (d) Joint probability distribution of $\langle n_{\mathrm{C=O}} \rangle$ and $\langle d_{\mathrm{C\cdots O}} \rangle$.

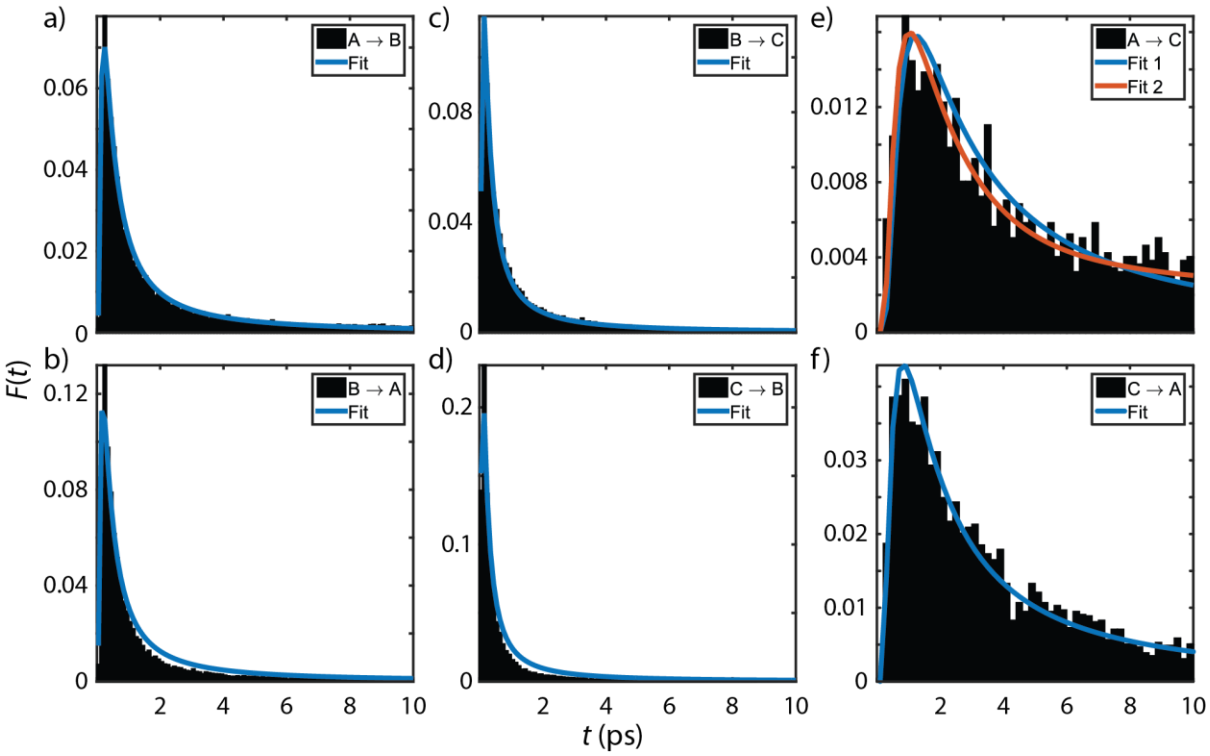

Figure 6.12: First passage time distribution (a) from state *A* to state *B*, (b) from *B* to *A*, (c) from *B* to *C*, (d) from *C* to *B*, (e) from *A* to *C*, (f) from *C* to *A*. The model used for the fits is $\sum_i a_i t^{-3/2} \exp[-b_i / t]$. Fit and Fit 1 refers to single component (i=1) whereas Fit 2 corresponds to two components (i=1, 2).

6.5.4 Spectroscopic Model and Water Model

As found in Fig. 6.1 and Table 6.1, all of the simulations calculate a redshift in peak frequency relative to the experiment. This suggests that either the frequency map predicts incorrect frequencies, the water models give a distribution of solvation structures that do not correspond to experiment, or both models have problems. Additionally, the spectral line shapes differ between experiment and simulations, which similarly could have roots in the spectral model, or in the dynamics used to calculate the spectra. Since both models have issues of concern, identifying the origin of the mismatch between experiment and simulations is difficult; however, with the help of testing by spectroscopic modeling, we can estimate the effects and rule out some potential origins.

To investigate the correlation between amide I frequency and the structure of peptide and associated solvent, we computed the joint probability distribution between $\langle n \rangle$ and $\langle \omega \rangle$ for the 1F and 4P spectral models (Figs. 6.13a–b). We find a correlation between $\langle n \rangle$ and $\langle \omega \rangle$, indicating that frequency changes do provide a means to probe the hydrogen-bonding environment around the amide group. However, the slope of varying $\langle \omega \rangle$ with $\langle n \rangle$ indicates that it is uncertain to assign a set of structure based only on a given amide I frequency. Similarly, when viewing the distribution of $\langle \omega \rangle$ as a function of backbone dihedrals (Figs. 6.13c–d), inferring peptide structures from given frequencies can lead to different conclusions.

However, this analysis does not account for possible differences between the experiment and the simulations due to the water model. While state *B* is predicted to be 1670 cm^{-1} by both maps, the more dominant state *A* has lower frequency regardless of the maps, resulting in a different asymmetry of the frequency distribution from the experimental spectra. Given the error bar estimate of these maps as ±2 cm^{-1},[33-34] it remains possible that the redshift in peak frequency stems from the water model over-populating more hydrogen-bonded environments rather than originating from the uncertainty of the frequency maps. We also find that different water models can predict some variation of hydrogen bond populations (Fig. 6.14), also seen in other studies regarding peptide solvation structures.[82-83] The effect of state *C* would be little due to Boltzmann weighting. Meaningful conclusions on this question would be better addressed with a higher level of theory and a water model more appropriate for IR spectroscopy simulations than the fixed charged models examined in this study such as *ab initio* MD,[45, 84] MB-pol,[85-87] *etc*.

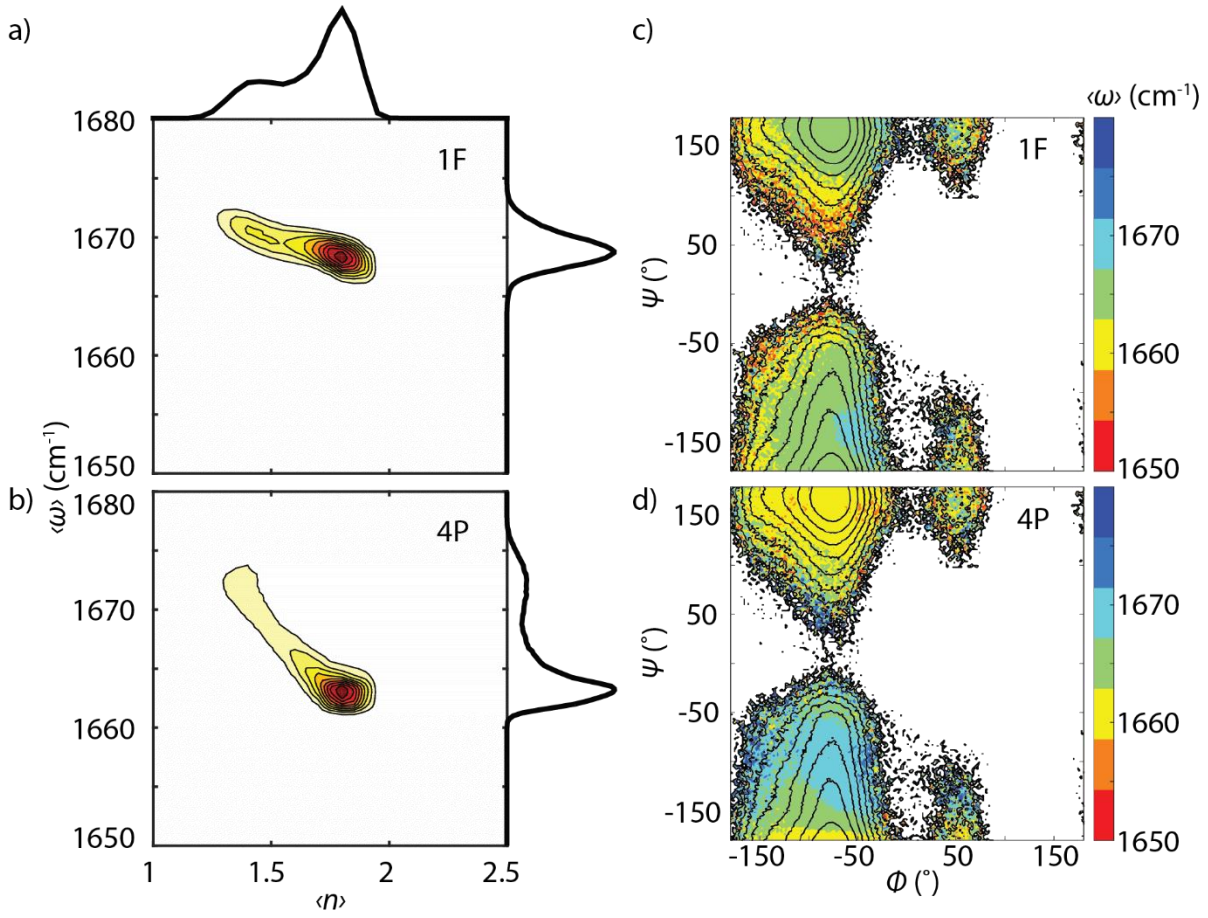

Figure 6.13: Probability distribution as a function of $\langle n \rangle$ taken from the distribution in Fig. 6.8, and along $\langle \omega \rangle$ (a) from the 1F map shown in Fig. 6.13c, and (b) from the 4P map shown in Fig. 6.13d. Tiny populated data points were removed, but the remaining data points still reaches 90 % of the entire data points. The 1D projections along $\langle n \rangle$ and $\langle \omega \rangle$ are next to the 2D contour map. (c) Colored contours: $\langle \omega \rangle$ as a function of ϕ and ψ from C36 SPC/E trajectory and the 1F map. Black contour lines: PMF spaced by $k_B T$ up to 10 $k_B T$ at 300 K. (d) Colored contours: $\langle \omega \rangle$ as a function of ϕ and ψ from C36 SPC/E trajectory and the 4P map. Black contour lines: PMF spaced by $k_B T$ up to 10 $k_B T$ at 300 K.

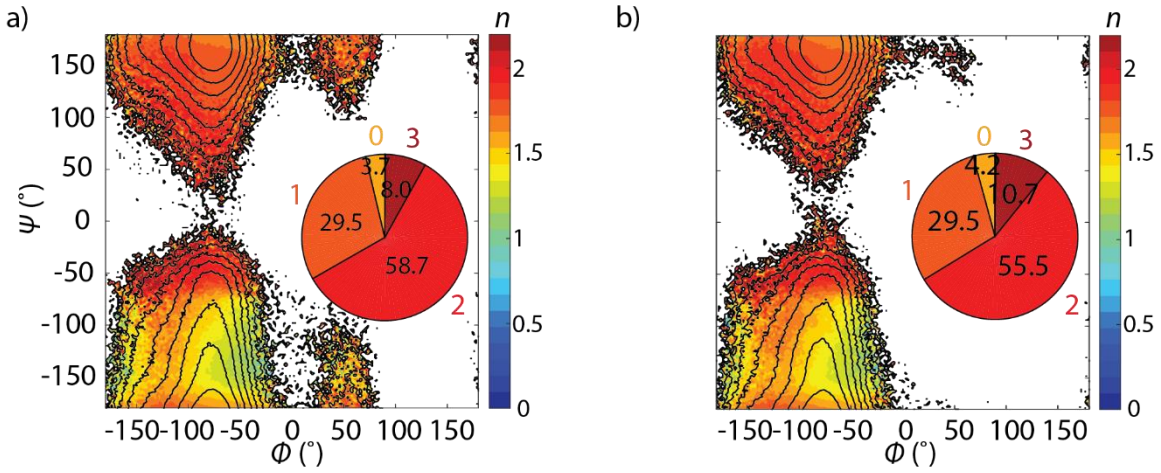

Figure 6.14: Average hydrogen bond number distribution $\langle n \rangle$ for (a) SPC/E and (b) TIP3P as a function of backbone dihedrals. Black contour lines: PMF spaced by k_BT up to 10 k_BT at 300 K. Insets show population percentage of different water hydrogen bonds to the amide group.

To estimate the effect of the variation of vibrational lifetime on the spectral lineshape, we incorporated a fluctuating vibrational relaxation rate given in Fig. 6.5 into our spectral simulation by adding an additional population decay factor to the transition dipole correlation function of the form $P(\tau_2) = \exp\left(-\int_0^{\tau_2} k(\tau) d\tau\right)$. We found that the only noticeable effect is a change of intensity (Fig. 6.15), and concluded that the variation of vibrational lifetime contributes little to the difference of vibrational lineshape. We also simulated the 2D spectrum with slower spectral diffusion rate by deuterating the SPC/E water molecules shown in Fig. 6.16. The effect is also subtle on the intensity while the spectral lineshape remains similar. Therefore, we can rule out variation of vibrational lifetime and faster spectral diffusion as the origin.

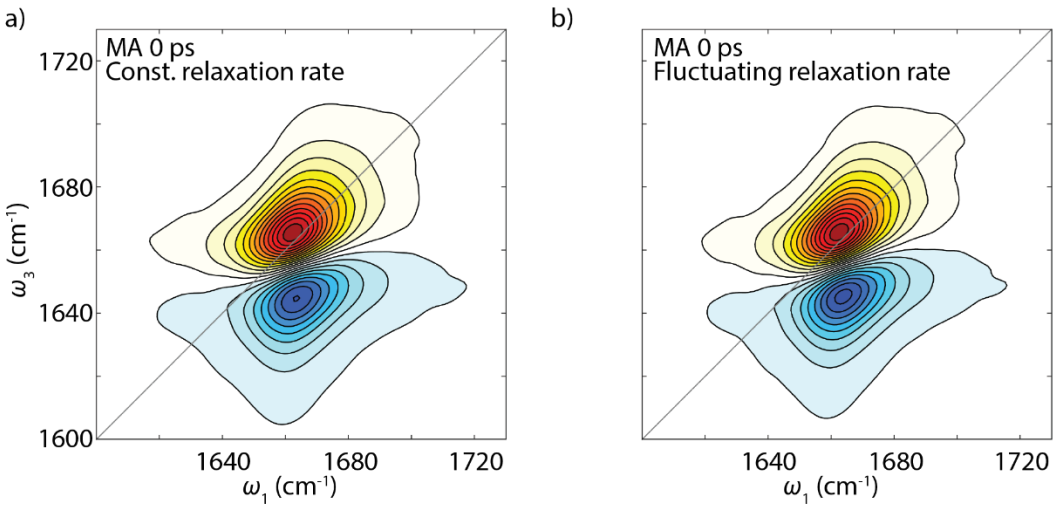

Figure 6.15: The effect of vibrational lifetime variations on the spectral simulation. (a) MA 2D IR surface at $\tau_2 = 0$ with constant relaxation rate set to 1 ps. (b) MA 2D IR surface at $\tau_2 = 0$ with fluctuating relaxation rate.

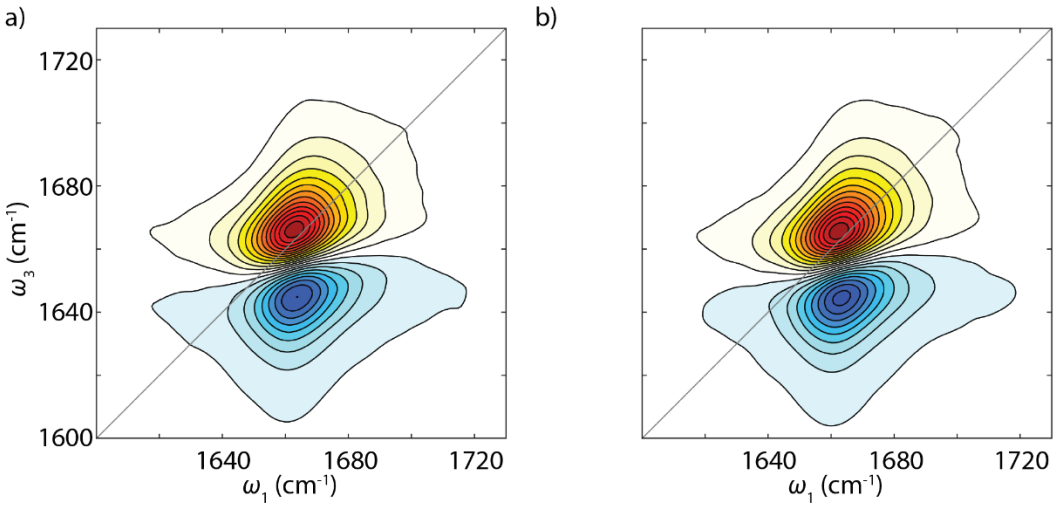

Figure 6.16: The effect of different spectral diffusion rates on the spectral simulation. (a) MA 2D IR surface at $\tau_2 = 0.15$ ps from C36m SPC/E trajectory. (b) MA 2D IR surface at $\tau_2 = 0.15$ ps from C36m Deuterated SPC/E trajectory.

6.6 Conclusions

In this study, we performed a detailed analysis of peptide-water interactions by ultrafast amide I vibrational spectroscopy and computational spectroscopy based on MD simulations. We observed spectral diffusion, variation of vibrational relaxation time scale, and chemical exchange processes in Ala–Ala. We found that the spectral diffusion of the amide I vibration is dictated by water solvation dynamics. This effect can be explained classically as friction experienced by the amide I vibration, or hydrogen bond fluctuation dynamics around the amide group microscopically. For vibrational relaxation, we observed strong linear correlation between relaxation time and squared electric field strength exerted along the amide carbonyl, showing that the electric field from the local solvation environment serves as an effective friction force in response to amide I vibration. On the basis of simulations, we conclude that the origin of chemical exchange observed in the experiment requires solvent reorganization coupled with peptide conformational motions. Specifically, we found that the configuration of peptide side chain and protonated COOH group influences the access of water to the N–H and C=O of the amide group. In a more general sense, we hope that this study illustrates the high level of atomistic information that can be gleaned on solvation structure and dynamics from head-to-head comparisons between advanced vibrational spectroscopy and structure-based spectral modeling rooted in MD simulations.

6.7 Acknowledgments

I thank Paul Sanstead as a helpful go-to person to learn about how to take quality 2D IR spectra, and I thank Paul Stevenson for helpful discussions on how to analyze waiting time series data.

6.8 References

1. Cheung, M. S.; Garcia, A. E.; Onuchic, J. N., Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci U S A* **2002,** *99* (2), 685-90.
2. Head-Gordon, T., Minimalist models for protein folding and design. *Current Opinion in Structural Biology* **2003,** *13* (2), 160-167.
3. Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G., Water in protein structure prediction. *Proc Natl Acad Sci U S A* **2004,** *101* (10), 3352-7.
4. Bagchi, B., *Water in Biological and Chemical Processes*

From Structure and Dynamics to Function. Cambridge University Press 2013.
5. Papoian, G. A.; Ulander, J.; Wolynes, P. G., Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc* **2003,** *125* (30), 9170-8.
6. Levy, Y.; Onuchic, J. N., Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct* **2006,** *35*, 389-415.
7. Lim, V. I.; Curran, J. F.; Garber, M. B., Hydration shells of molecules in molecular association: A mechanism for biomolecular recognition. *J Theor Biol* **2012,** *301*, 42-8.
8. Levinson, N. M.; Boxer, S. G., A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity. *Nat Chem Biol* **2014,** *10* (2), 127-32.
9. Poornima, C. S.; Dean, P. M., Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *Journal of Computer-Aided Molecular Design* **1995,** *9* (6), 521-531.
10. Pocker, Y., Water in enzyme reactions: biophysical aspects of hydration-dehydration processes. *Cell Mol Life Sci* **2000,** *57* (7), 1008-17.
11. Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G., How enzymes work: analysis by modern rate theory and computer simulations. *Science* **2004,** *303* (5655), 186-95.
12. Adkar, B. V.; Jana, B.; Bagchi, B., Role of water in the enzymatic catalysis: study of ATP + AMP --> 2ADP conversion by adenylate kinase. *J Phys Chem A* **2011,** *115* (16), 3691-7.
13. Bolhuis, P. G.; Dellago, C.; Chandler, D., Reaction coordinates of biomolecular isomerization. *Proc Natl Acad Sci U S A* **2000,** *97* (11), 5877-82.
14. Ma, A.; Dinner, A. R., Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* **2005,** *109* (14), 6769-79.
15. Fogarty, A. C.; Laage, D., Water dynamics in protein hydration shells: the molecular origins of the dynamical perturbation. *J Phys Chem B* **2014,** *118* (28), 7715-29.
16. Sterpone, F.; Stirnemann, G.; Laage, D., Magnitude and molecular origin of water slowdown next to a protein. *J Am Chem Soc* **2012,** *134* (9), 4116-9.
17. Nerukh, D.; Karabasov, S., Water-Peptide Dynamics during Conformational Transitions. *J Phys Chem Lett* **2013,** *4* (5), 815-9.
18. Oleinikova, A.; Sasisanker, P.; Weingärtner, H., What Can Really Be Learned from Dielectric Spectroscopy of Protein Solutions? A Case Study of Ribonuclease A. *The Journal of Physical Chemistry B* **2004,** *108* (24), 8467-8474.
19. Yang, J.; Wang, Y.; Wang, L.; Zhong, D., Mapping Hydration Dynamics around a beta-Barrel Protein. *J Am Chem Soc* **2017,** *139* (12), 4399-4408.
20. Born, B.; Kim, S. J.; Ebbinghaus, S.; Gruebele, M.; Havenith, M., The terahertz dance of water with the proteins: the effect of protein flexibility on the dynamical hydration shell of ubiquitin. *Faraday Discuss* **2009,** *141*, 161-73; discussion 175-207.

21. Born, B.; Weingartner, H.; Brundermann, E.; Havenith, M., Solvation dynamics of model peptides probed by terahertz spectroscopy. Observation of the onset of collective network motions. *J Am Chem Soc* **2009,** *131* (10), 3752-5.
22. Armstrong, B. D.; Choi, J.; Lopez, C.; Wesener, D. A.; Hubbell, W.; Cavagnero, S.; Han, S., Site-specific hydration dynamics in the nonpolar core of a molten globule by dynamic nuclear polarization of water. *J Am Chem Soc* **2011,** *133* (15), 5987-95.
23. Nucci, N. V.; Pometun, M. S.; Wand, A. J., Site-resolved measurement of water-protein interactions by solution NMR. *Nat Struct Mol Biol* **2011,** *18* (2), 245-9.
24. Lewandowski, J. R.; Halse, M. E.; Blackledge, M.; Emsley, L., Protein dynamics. Direct observation of hierarchical protein dynamics. *Science* **2015,** *348* (6234), 578-81.
25. Fecko, C. J.; Eaves, J. D.; Loparo, J. J.; Tokmakoff, A.; Geissler, P. L., Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water. *Science* **2003,** *301* (5640), 1698-702.
26. Hamm, P.; Lim, M.; Hochstrasser, R. M., Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *The Journal of Physical Chemistry B* **1998,** *102* (31), 6123-6138.
27. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
28. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.
29. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.
30. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
31. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
32. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.
33. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
34. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
35. Torii, H.; Tasumi, M., Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *Journal of Raman Spectroscopy* **1998,** *29* (1), 81-86.
36. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
37. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
38. Hayashi, T.; Mukamel, S., Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides. *J Phys Chem B* **2007,** *111* (37), 11032-46.
39. Maekawa, H.; De Poli, M.; Moretto, A.; Toniolo, C.; Ge, N. H., Toward detecting the formation of a single helical turn by 2D IR cross peaks between the amide-I and -II modes. *J Phys Chem B* **2009,** *113* (34), 11775-86.

40. Woutersen, S.; Mu, Y.; Stock, G.; Hamm, P., Hydrogen-bond lifetime measured by time-resolved 2D-IR spectroscopy: N-methylacetamide in methanol. *Chemical Physics* **2001,** *266* (2-3), 137-147.
41. Zanni, M. T.; Asplund, M. C.; Hochstrasser, R. M., Two-dimensional heterodyned and stimulated infrared photon echoes of N-methylacetamide-D. *The Journal of Chemical Physics* **2001,** *114* (10), 4579.
42. DeCamp, M. F.; DeFlores, L.; McCracken, J. M.; Tokmakoff, A.; Kwac, K.; Cho, M., Amide I vibrational dynamics of N-methylacetamide in polar solvents: the role of electrostatic interactions. *J Phys Chem B* **2005,** *109* (21), 11016-26.
43. Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. I. Adiabatic approximations. *The Journal of Chemical Physics* **2003,** *118* (8), 3480-3490.
44. Cazade, P. A.; Bereau, T.; Meuwly, M., Computational two-dimensional infrared spectroscopy without maps: N-methylacetamide in water. *J Phys Chem B* **2014,** *118* (28), 8135-47.
45. Yadav, V. K.; Chandra, A., First-Principles Simulation Study of Vibrational Spectral Diffusion and Hydrogen Bond Fluctuations in Aqueous Solution of N-Methylacetamide. *J Phys Chem B* **2015,** *119* (30), 9858-67.
46. Woutersen, S.; Pfister, R.; Hamm, P.; Mu, Y.; Kosov, D. S.; Stock, G., Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations. *The Journal of Chemical Physics* **2002,** *117* (14), 6833-6840.
47. Kwac, K.; Cho, M., Molecular dynamics simulation study of N-methylacetamide in water. I. Amide I mode frequency fluctuation. *The Journal of Chemical Physics* **2003,** *119* (4), 2247-2255.
48. Kwac, K.; Cho, M., Molecular dynamics simulation study of N-methylacetamide in water. II. Two-dimensional infrared pump–probe spectra. *The Journal of Chemical Physics* **2003,** *119* (4), 2256-2263.
49. DeFlores, L. P.; Ganim, Z.; Ackley, S. F.; Chung, H. S.; Tokmakoff, A., The anharmonic vibrational potential and relaxation pathways of the amide I and II modes of N-methylacetamide. *J Phys Chem B* **2006,** *110* (38), 18973-80.
50. Bastida, A.; Soler, M. A.; Zuniga, J.; Requena, A.; Kalstein, A.; Fernandez-Alberti, S., Instantaneous normal modes, resonances, and decay channels in the vibrational relaxation of the amide I mode of N-methylacetamide-D in liquid deuterated water. *J Chem Phys* **2010,** *132* (22), 224501.
51. Jeon, J.; Cho, M., Redistribution of carbonyl stretch mode energy in isolated and solvated N-methylacetamide: kinetic energy spectral density analyses. *J Chem Phys* **2011,** *135* (21), 214504.
52. Farag, M. H.; Bastida, A.; Ruiz-Lopez, M. F.; Monard, G.; Ingrosso, F., Vibrational energy relaxation of the amide I mode of N-methylacetamide in D(2)O studied through Born-Oppenheimer molecular dynamics. *J Phys Chem B* **2014,** *118* (23), 6186-97.
53. Kim, Y. S.; Hochstrasser, R. M., Dynamics of amide-I modes of the alanine dipeptide in D2O. *J Phys Chem B* **2005,** *109* (14), 6884-91.
54. Kim, Y. S.; Wang, J.; Hochstrasser, R. M., Two-dimensional infrared spectroscopy of the alanine dipeptide in aqueous solution. *J Phys Chem B* **2005,** *109* (15), 7511-21.
55. Jansen, T.; Knoester, J., Nonadiabatic effects in the two-dimensional infrared spectra of peptides: application to alanine dipeptide. *J Phys Chem B* **2006,** *110* (45), 22910-6.

56. Deflores, L. P.; Nicodemus, R. A.; Tokmakoff, A., Two-dimensional Fourier transform spectroscopy in the pump-probe geometry. *Opt Lett* **2007,** *32* (20), 2966-8.
57. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-54.
58. Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* **2004,** *25* (11), 1400-15.
59. Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E., Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *J Chem Theory Comput* **2010,** *6* (2), 459-66.
60. Huang, J.; MacKerell, A. D., Jr., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* **2013,** *34* (25), 2135-45.
61. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2017,** *14* (1), 71-73.
62. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
63. Nosé, S., A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **1984,** *81* (1), 511-519.
64. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A Gen Phys* **1985,** *31* (3), 1695-1697.
65. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014,** *185* (2), 604-613.
66. Reppert, M. *g_amide*, 2017.
67. Torii, H., Effects of intermolecular vibrational coupling and liquid dynamics on the polarized Raman and two-dimensional infrared spectral profiles of liquid N,N-dimethylformamide analyzed with a time-domain computational method. *J Phys Chem A* **2006,** *110* (14), 4822-32.
68. Reppert, M.; Feng, C. J. *g_spec*, 2017.
69. Hamm, P.; Lim, M.; DeGrado, W. F.; Hochstrasser, R. M., The two-dimensional IR nonlinear spectroscopy of a cyclic penta-peptide in relation to its three-dimensional structure. *Proc Natl Acad Sci U S A* **1999,** *96* (5), 2036-41.
70. Bondarenko, A. S.; Jansen, T. L., Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins. *J Chem Phys* **2015,** *142* (21), 212437.
71. Kwak, K.; Park, S.; Finkelstein, I. J.; Fayer, M. D., Frequency-frequency correlation functions and apodization in two-dimensional infrared vibrational echo spectroscopy: a new approach. *J Chem Phys* **2007,** *127* (12), 124503.
72. Fenn, E. E.; Fayer, M. D., Extracting 2D IR frequency-frequency correlation functions from two component systems. *J Chem Phys* **2011,** *135* (7), 074502.
73. Kim, Y. S.; Hochstrasser, R. M., Chemical exchange 2D IR of hydrogen-bond making and breaking. *Proc Natl Acad Sci U S A* **2005,** *102* (32), 11185-90.

74. Kwak, K.; Zheng, J.; Cang, H.; Fayer, M. D., Ultrafast two-dimensional infrared vibrational echo chemical exchange experiments and theory. *J Phys Chem B* **2006,** *110* (40), 19998-20013.
75. Harder, E.; Eaves, J. D.; Tokmakoff, A.; Berne, B. J., Polarizable molecules in the vibrational spectroscopy of water. *Proc Natl Acad Sci U S A* **2005,** *102* (33), 11611-6.
76. Schmidt, J. R.; Roberts, S. T.; Loparo, J. J.; Tokmakoff, A.; Fayer, M. D.; Skinner, J. L., Are water simulation models consistent with steady-state and ultrafast vibrational spectroscopy experiments? *Chemical Physics* **2007,** *341* (1-3), 143-157.
77. Mahoney, M. W.; Jorgensen, W. L., Diffusion constant of the TIP5P model of liquid water. *The Journal of Chemical Physics* **2001,** *114* (1), 363.
78. Drost-Hansen, W., The Structure and Properties of Water. D. Eisenberg and W. Kauzmann. Oxford University Press, New York, 1969. xiv + 300 pp., illus. Cloth, $10; paper, $4.50. *Science* **1969,** *166* (3907), 861-861.
79. Oxtoby, D. W., Vibrational Population Relaxation in Liquids. **2007**, 487-519.
80. Egorov, S. A.; Skinner, J. L., A theory of vibrational energy relaxation in liquids. *The Journal of Chemical Physics* **1996,** *105* (16), 7047-7058.
81. Bloem, R.; Dijkstra, A. G.; Jansen, T.; Knoester, J., Simulation of vibrational energy transfer in two-dimensional infrared spectroscopy of amide I and amide II modes in solution. *J Chem Phys* **2008,** *129* (5), 055101.
82. Hajari, T.; Bandyopadhyay, S., Water structure around hydrophobic amino acid side chain analogs using different water models. *J Chem Phys* **2017,** *146* (22), 225104.
83. Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M., Explicit Water Models Affect the Specific Solvation and Dynamics of Unfolded Peptides While the Conformational Behavior and Flexibility of Folded Peptides Remain Intact. *J Chem Theory Comput* **2010,** *6* (11), 3569-79.
84. Gaigeot, M. P.; Vuilleumier, R.; Sprik, M.; Borgis, D., Infrared Spectroscopy of N-Methylacetamide Revisited by ab Initio Molecular Dynamics Simulations. *J Chem Theory Comput* **2005,** *1* (5), 772-89.
85. Babin, V.; Leforestier, C.; Paesani, F., Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J Chem Theory Comput* **2013,** *9* (12), 5395-403.
86. Babin, V.; Medders, G. R.; Paesani, F., Development of a "First Principles" Water Potential with Flexible Monomers. II: Trimer Potential Energy Surface, Third Virial Coefficient, and Small Clusters. *J Chem Theory Comput* **2014,** *10* (4), 1599-607.
87. Medders, G. R.; Babin, V.; Paesani, F., Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties. *J Chem Theory Comput* **2014,** *10* (8), 2906-10.

Chapter 7

Refinement of Peptide Conformational Ensemble by 2D IR Spectroscopy: Application to Ala–Ala–Ala

The work presented in this chapter has been published and is reprinted with permission from:

Feng, C.-J.; Dhayalan, B.; Tokmakoff, A. Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophysical Journal* **2018**, 114 (12), 2820–2832.

7.1 Abstract

Characterizing ensembles of intrinsically disordered proteins is experimentally challenging due to the ill-conditioned nature of ensemble determination with limited data, and the intrinsic fast dynamics of the conformational ensemble. Amide I 2D IR spectroscopy has picosecond time resolution to freeze structural ensembles as needed for probing disordered protein ensembles and conformational dynamics. Also, developments in amide I computational spectroscopy now allow a quantitative and direct prediction of amide I spectra based on conformational distributions drawn from molecular dynamics (MD) simulations, providing a route to ensemble refinement against experimental spectra. We performed Bayesian ensemble refinement method on Ala–Ala–Ala (AAA) against isotope-edited FTIR and 2D IR spectroscopy and tested potential factors affecting the quality of ensemble refinements. We found that isotope-edited 2D IR spectroscopy provides a stringent constraint on AAA conformations, and returns consistent conformational ensembles with

the dominant ppII conformer across varying prior distributions from many MD force fields and water models. The dominant factor influencing ensemble refinements is the systematic frequency uncertainty from spectroscopic maps. However, the uncertainty of conformer populations can be significantly reduced by incorporating 2D IR spectra in addition to traditional FTIR spectra. Bayesian ensemble refinement against isotope-edited 2D IR spectroscopy thus provides a route to probe equilibrium complex protein ensembles and potentially non-equilibrium conformational dynamics.

7.2 Introduction

Intrinsically disordered proteins (IDP) exhibit a variety of thermally accessible conformers reflecting basins on a complex energy landscape, that are also characterized by dynamics such as conformational fluctuations and activated kinetics of interconversion between free energy basins.[1-3] As a result, structural characterization of IDPs or proteins with intrinsic disordered regions (IDRs) requires an ensemble description, which creates numerous experimental challenges.[4-5] In particular, ensemble structure determination is naturally an ill-posed problem in which the degrees of freedom in relevant conformational states far exceeds the limited number of measurements and information content of experiments.[6]

The dynamics nature of the ensemble also means that structural variation and conformational dynamics cannot be decoupled. Traditional structural tools are often limited by their intrinsic time resolution, which prohibits one from accessing conformational fluctuations and interconversion of conformers with time scales spanning from picoseconds to microseconds.[4, 7-8] For example, measuring chemical shifts or J couplings in NMR spectroscopy is limited by the coalescence time scale of ms such that faster conformational dynamics are average.[9] Optical

spectroscopies do carry the advantage of femtosecond time-scales for their light-matter interaction, but in most cases have little or no structural information content. On the other hand, IR and 2D IR spectroscopies probe structure sensitive to molecular vibrations with fs–ps time resolution, which can be used to carry out structural characterization on a peptide or protein structure that is essentially frozen.[10-13] 2D IR spectrum represents a correlation map between different vibrational modes by spreading the spectrum onto independent excitation and detection frequency axes. This enhances the structural information content of vibrational spectra, providing higher contrast and characterizing inhomogeneous distributions that result from a static structural distribution.

With this goal in mind, we have been developing tools for protein and peptide structural characterization using 2D IR spectroscopy of amide I vibrations, which result primarily from the C=O stretching vibration of the protein backbone amide group. Amide I spectroscopy can be used to sense local electrostatics, hydrogen bonding to the carbonyl, and secondary structures of proteins.[12] However, the vibrational lifetime of 1–1.3 ps significantly broadens the amide I peaks,[14-15] lowering the structural resolution with highly congested amide I spectra.[16] Enhanced structural information can be achieved by using various polarizations of ultrafast infrared pulses[17] and introducing site-specific isotope labeling such as ^{13}C or $^{13}C^{18}O$ on the amide carbonyl, which provides an additional frequency shift to isolate a specific C=O bond from a congested spectrum, allowing the extraction of local structural details.[18-19] The structural sensitivity of amide I vibrational and site-specific isotope labeling has assisted in addressing conformational distributions of peptides and proteins,[5, 20-22] helix-coil transition dynamics,[23-25] and ion-permeation mechanism of ion channels.[26-27]

Despite intensive experimental advances to investigating protein structures, all experimental methods are challenged to interpret their measurement. Therefore, it remains of great

interest to make use of atomistic models such as molecular dynamics (MD) simulation to offer atomistic or coarse-grained descriptions of protein structures and motions, which can rationalize experimental evidence, predict experimental outcomes, and even help design suitable experiments. However, these computational tools often suffer from a separate set of challenges, including limited sampling of rare events due to the gap between computationally accessible time scales and the time scale of conformational dynamics, which hinder the accuracy of predicted ensemble distributions. Also, recent efforts of force field (FF) developments have improved the ability to study disordered proteins,[28] but the question of how much the uncertainty from FFs and water models affects the ensemble predictions of IDPs and proteins with IDRs quantitatively still remains.[29-33] A practical approach using ensemble refinement, which reweights existing ensemble populations from simulations against experimental data, has proven successful for facilitating the inference of protein conformational ensembles consistent with experiments,[6] including developing different frameworks such as maximum entropy (ME) principle,[34-38] Bayesian statistics,[39-41] and biological applications against experimental data such as NMR,[6, 42] SAXS,[43-44] and IR spectroscopy.[5, 45] A recent detailed review of ensemble structure determination can be found.[4]

Making direct comparisons of a protein or peptide structure with IR experiments is now possible using computational amide I spectroscopy. This method can be used to predict IR and 2D IR spectra for a single structure or simulated conformational distributions drawn from MD trajectories, providing a route to ensemble refinement against IR experiments.[46] Specifically, amide I spectroscopic maps predict amide I vibrational frequencies to high accuracy using local electrostatics calculated from MD simulations such as electrostatic potential or electric field at the site of interest.[47-54] Maps for vibrational coupling between different amide I vibrations are used to calculate the interaction of multiple backbone amide groups.[55-59] These maps have reached the

point of predicting amide I spectroscopic observables to a high level of accuracy with 2 cm^{-1} frequency uncertainty[53] and provided a direct way to refine conformational ensembles of proteins.

As part of our effort to develop tools for refining protein and peptide conformational ensembles, we recently applied the ME method to elastin-like peptides (ELPs) to refine the ELP ensembles against isotope-edited FTIR spectroscopy[5] drawing structures from multiple FFs. Although this proved effective for describing the extension in a type-I turn in these peptides, the ME method is not readily extended to more complex features, such as multiple overlapping resonances and asymmetric spectral lineshapes. Additionally, the ME framework requires that the constraints are strictly satisfied after the refinement, which may lead to biased refinement or poor convergence if the experimental constraints are chosen improperly or if uncertainties exist in the experiment (such as noise or signal bias). To make use of the added information content of 2D IR spectroscopy and generalize this method to account for arbitrary spectral lineshape, we have implemented a Bayesian ensemble refinement framework against multiple experimental IR spectra using FTIR and 2D IR spectroscopy with the capability of integrating other experimental techniques.

As a proof-of-principle study, ensemble refinement against isotope-edited FTIR and 2D IR spectroscopy is performed on Ala–Ala–Ala (AAA), because the backbone conformational variation of AAA has been well-characterized to contain mostly ppII conformer (Fig. 7.1) using various experimental approaches including NMR, vibrational circular dichroism, and 2D IR spectroscopy.[60-65] Additionally, ensemble determination based on Bayesian framework has been applied on AAA using NMR data, and it has been demonstrated that sufficient experimental data can lead to converged populations of conformers from various FFs.[39-40] AAA conformational variation is simple enough to investigate the capability of ensemble refinement against 2D IR

spectroscopy, and the factors influencing the quality of refining conformational ensemble of peptides and proteins. This study will help us consolidate the foundation of methods of refining protein conformational ensembles against isotope-edited 2D IR spectroscopy, and enable the potential of globally constraining a description of an IDP ensemble against multiple isotope-edited samples.

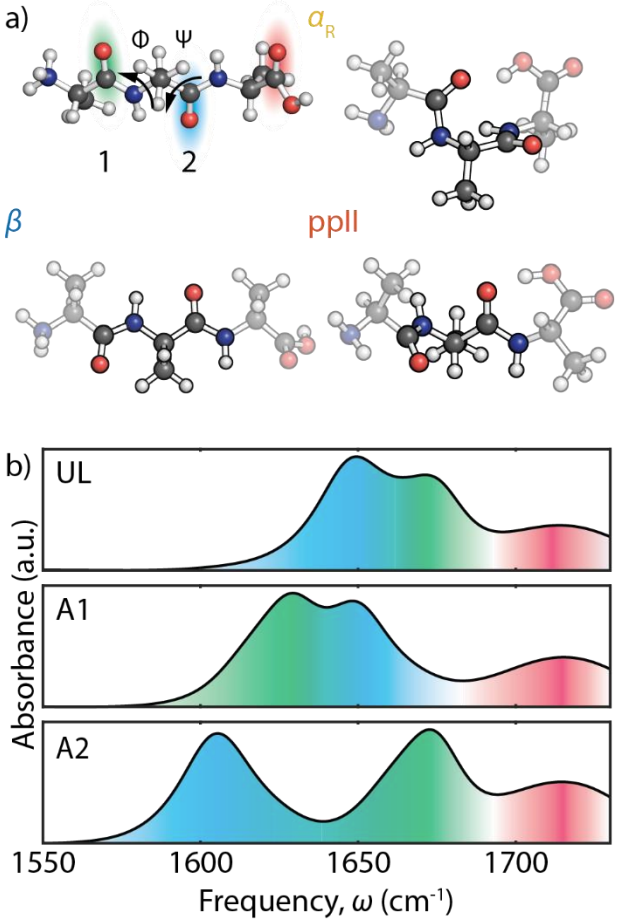

Figure 7.1: (a) Structure of cationic AAA and the dominant conformers α_R ((ϕ,ψ)=(–60°,–40°)), β ((ϕ,ψ)=(–135°,135°)), and ppII ((ϕ,ψ)=(–70°,150°)). The amide I vibrations of the A1 and A2 sites are color-coded green and blue, respectively, while the C=O stretch of the COOD group is color-coded red. (b) FTIR spectra of UL AAA, A1-labeled AAA, and A2-labeled AAA, using the same color coding.

7.3 Materials and Methods

7.3.1 Solid Phase Peptide Synthesis of Ala–Ala–Ala

1–^{13}C labeled Alanine was purchased from Cambridge Isotope Laboratories and Fmoc-protected using Fmoc-OSu in the presence of $NaHCO_3$ in Dioxane-water mixture. These building blocks were used in synthesizing (1–^{13}C)Ala–Ala–Ala (A1-labeled AAA) and Ala– (1–^{13}C)Ala–Ala (A2-labeled AAA) by manual Fmoc solid phase peptide synthesis. All peptides were synthesized on a 0.1 mmol scale with 5.5-fold excess of Fmoc-protected amino acids (0.55 mmol) and HBTU (0.5 mmol) in presence of DIEA in DMF. N^{α}-Fmoc protecting groups were removed by treating the resin-attached peptide with piperidine (20% v/v) in DMF. Fmoc-(1-^{13}C)Ala was coupled by using a minimum amount of the isotope-labeled amino acid: Fmoc-AA (0.35 mmol), HBTU (0.3 mmol), and DIEA (0.6 mmol) in DMF for 1 h. After peptide chain assembly was completed, the peptide-resin was subjected to N^{α}-Fmoc deprotection by treating with 20% v/v piperidine/DMF. The crude peptide was then cleaved from the 2-Chlorotrityl-(S-DVB)resin by treatment with TFA/TIPS/water (95:2.5:2.5 v/v) conditions at ambient temperature, and worked up by precipitation from ice-cold ether. Unlabeled (UL) AAA was synthesized by the same procedure except that UL alanine was used as building blocks. Detailed synthesis, purification and characterizations were described in the Appendix 7A.

7.3.2 Sample Preparation

Trialanines were iteratively dissolved in 1M DCl/D_2O and lyophilized to remove residual TFA, whose absorption overlaps with amide I absorption. For IR measurements, AAAs were dissolved to a concentration of 90 mM (30 mg/mL) in 1 M DCl/D_2O to avoid spectral overlap between amide I vibrations and H_2O bend vibration, and to protonate the carboxyl-terminus to

shift its carbonyl vibration to ~1720 cm^{-1}. Under this condition, AAA is cationic, or ND_3^+–Ala–Ala–Ala–COOD. Although we report amide I' spectra, for simplicity we use the terms amide I and amide I' interchangeably throughout this study. For all of the IR measurements, samples were held between two 1mm thick CaF_2 windows spaced by a 50 μm Teflon spacer.

7.3.3 FTIR Spectroscopy

FTIR spectra were collected at room temperature using a Bruker Tensor 27 FTIR spectrometer with 64 averages at 2 cm^{-1} resolution. A background spectrum of 1 M DCl in D_2O was measured for subtracting the solvent vibrational profile from the sample spectrum. A linear baseline correction from 1550 cm^{-1} to 1800 cm^{-1} was applied to flatten the baseline of the subtracted sample spectrum. For comparing spectra with simulations, we also subtracted the C=O resonance from the terminal COOD group by fitting all peaks with Gaussians and subtracting the COOD peak. To avoid bias due to baseline drift, a ~70 cm^{-1} logistic window smoothing is applied on spectra from both experiments and simulations. The subtracted experimental spectra and the window function are shown in Fig. 7.2.

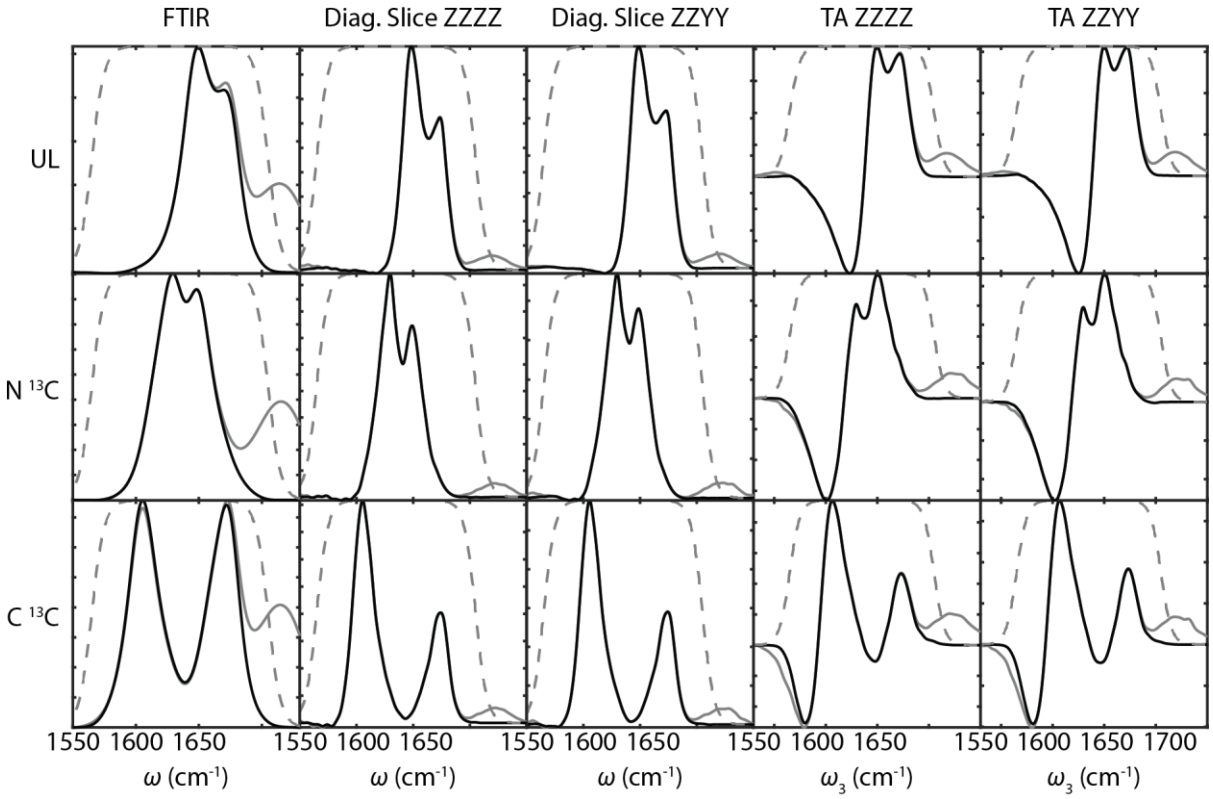

Figure 7.2: Subtraction of C=O stretch of the COOH group and windowing on experimental spectra. The gray solid curves and the black curves are the experimental spectra before subtraction, and after subtraction, respectively. The logistic window function is represented by the dashed gray curves.

7.3.4 2D IR Spectroscopy

Absorptive 2D IR spectra record the change of optical density at the detection frequency corresponding to a particular excitation frequency, serving as a correlation map between vibrational modes. 2D IR spectra were acquired in a pump-probe geometry 2D IR spectrometer,[66] with a collinear pump pulse pair, and a non-collinear probe pulse to acquire the 2D IR signal. The pulse center frequency of all pulses is 6 μm, resonant with the amide I vibrations. The pulse width is 90 fs in the time domain and has a ~130 cm^{-1} bandwidth in the frequency domain. The coherence time between the two pump pulses is scanned from -0.16 ps to 2.5 ps with 4 fs time step. The

corresponding Fourier-transformed excitation frequency axis has a resolution of 13 cm^{-1}. However, all of the coherence signals decay away around 1 ps. The 2D IR signal is heterodyned by the probe pulse, and dispersed onto the 64-channel MCT array detector, which has the resolution of 4 cm^{-1}. The waiting time between the second pump pulse and the probe pulse is set to 0.15 ps to avoid the effect of pulse overlap on the spectral lineshape. All of the 2D IR spectra are collected with both parallel polarization and perpendicular polarization between excitation and detection pulses. We also present transient absorption (TA) spectra, which are obtained by projecting the absorptive 2D IR spectrum onto the detection frequency axis.

7.3.5 MD Simulation

Details of the MD simulations are described in the Supporting Material. Briefly, cationic Ala–Ala–Ala simulations were performed using GROMACS 4.6.7 package.[67] The FFs used were CHARMM27 (C27),[68-69] CHARMM36 (C36),[70] CHARMM36m (C36m),[71] OPLS-AA,[72-73] OPLS-AA/M,[74] AMBER99sb–ildn,[75] AMBER14sb,[76] and AMBERfb15.[77] Water models used were SPC/E[78] and TIP3P[79] for all FFs. Additional TIP3Pfb[80] water model was used specifically for AMBERfb15, resulting in 17 combinations of FFs and water models. Since AMBER FFs do not have parameters for a protonated COOH group, we instead use a $CONH_2$ group at the C-terminus, which affects the electrostatics and the frequencies of the two amide groups (See the Appendix 7B), but may not influence notably on the conformational ensemble[81]. 100 ns production runs using the Nosé−Hoover thermostat[82-83] were simulated with 1 fs integration step, and 20 fs/frame sampling rate for amide I spectral simulations. The sampling quality is investigated by the Block Averaging method described below. The MD structural data analysis of backbone dihedral angles ϕ and ψ was performed using PLUMED 2.[84] The potential of mean force (PMF) of each MD

trajectory is computed by using $\mathrm{PMF}(\phi,\psi) = -k_{\mathrm{B}}T \ln P(\phi,\psi)$ in which $P(\phi,\psi)$ is the probability distribution of AAA as a function of (ϕ,ψ) at T=300 K.

The population distribution of conformational state is analyzed by partitioning the potential of mean force (PMF) as a function of ϕ and ψ angles into boxes defined below.

$$\begin{aligned}
\alpha_{\mathrm{R}} &: -160^{\circ} < \phi < -20^{\circ} \text{ and } -120^{\circ} < \psi < 50^{\circ} \\
\beta &: \begin{cases} -180^{\circ} < \phi < -110^{\circ} \text{ and } 50^{\circ} < \psi < 180^{\circ} \\ -180^{\circ} < \phi < -110^{\circ} \text{ and } -180^{\circ} < \psi < -120^{\circ} \\ 160^{\circ} < \phi < 180^{\circ} \text{ and } 110^{\circ} < \psi < 180^{\circ} \end{cases} \\
\mathrm{ppII} &: \begin{cases} -110^{\circ} < \phi < -20^{\circ} \text{ and } 50^{\circ} < \psi < 180^{\circ} \\ -110^{\circ} < \phi < -20^{\circ} \text{ and } -180^{\circ} < \psi < -120^{\circ} \end{cases}
\end{aligned} \tag{7.1}$$

This type of definition is also used in some studies.[17, 29, 39, 85] However, we adjusted the boundary of ϕ angle between β and ppII states to ensure the boundary would not cut into the ppII basin in the AMBER FFs and OPLS FFs, and lie between the two basins (Fig. 7.3).

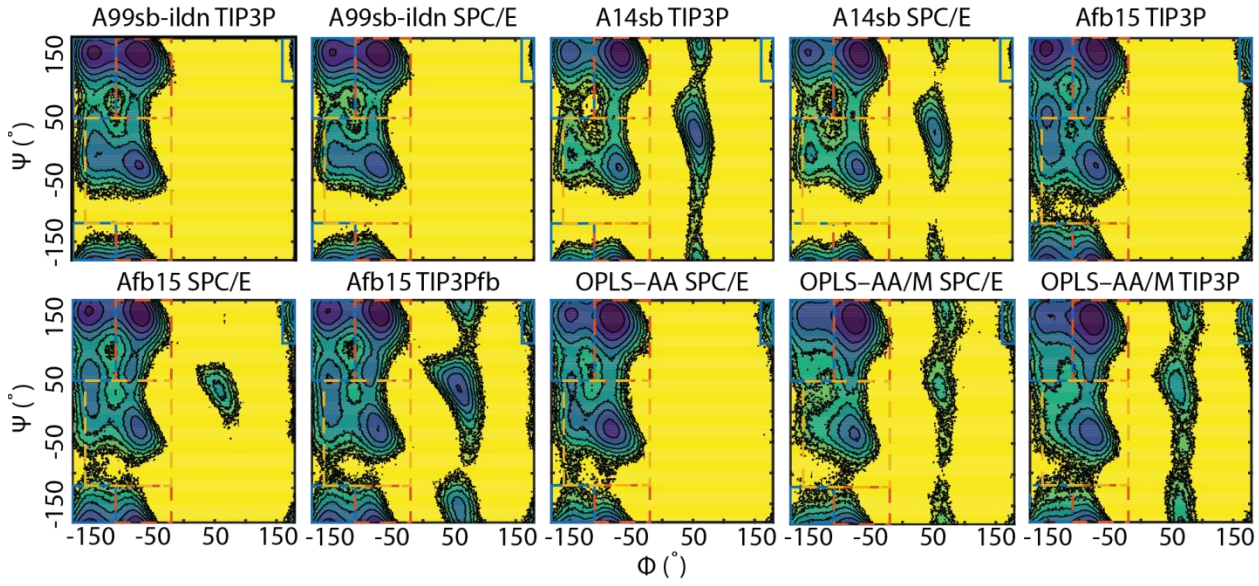

Figure 7.3: PMF(ϕ,ψ) computed for AMBER and OPLS–AA FFs. Colored contour lines are spaced by $k_{\mathrm{B}}T$ up to 6 $k_{\mathrm{B}}T$ at T=300 K. Blue boxes: β conformer. Red boxes: ppII conformer. Yellow box: α_{R} conformer.

7.3.6 Block Averaging

Block averaging is used to quantify the quality of sampling conformational ensembles.[86] We estimate the blocked standard error (BSE) of the population of β, ppII, and α_R as a function of block size n, which is increased by 10 ps until the block size reaches 10 ns in length. At each block size, we perform the estimate of BSE of the conformational state j using the Eq. (7.2).

$$\mathrm{BSE}_j(n) = \sigma_n(\langle p_j \rangle) / \sqrt{M} \tag{7.2}$$

in which j runs through β, ppII, and α_R while M is the number of blocks given size n and the trajectory length of 100 ns. All of combinations of FFs and water models reach the plateau on the order of 1 ns, indicating that 100 ns sampling should be sufficient for visiting dominant basins.

7.3.7 Amide I Spectral Simulation

IR vibrational spectra of the two coupled amide I vibrations of the AAA peptide backbone were calculated from the Fourier transform of a dipole time-correlation function using a mixed quantum-classical model. Details of this spectral simulation strategy using coupled oscillator (exciton) model have been described previously,[15, 46, 53] and the home-built programs used in these calculations, g_amide and g_spec, are freely available.[87-88] Collective electrostatic variables are used to translate, or "map", a series of instantaneous structures along MD trajectories into a time-dependent Hamiltonian and transition dipole moment that describes the two amide I vibrations. The two amide oscillators, or "sites", are identified by the atomic positions of the CONH peptide backbone linkages. The vibrational frequency of each site is generated using one of two empirical frequency maps that correlate the collective variable with a specific vibrational frequency: the 1–site electric field map (1F) that uses the electric field created by the local environment at the amide oxygen projected along the C=O bond, and the 4–site potential map (4P) which uses the

electrostatic potential at the C, O, N and H positions.[53] Unless mentioned, the spectroscopic map used throughout this study is the 1F map, which has a frequency prediction accuracy of σ=2 cm^{-1}.[53] To calculate these electrostatic parameters across multiple FFs and water models, we use CHARMM FF charges with modified glycine charges and TIP3P charges in this study.[53]

In addition to the frequency of each site, the vibrational coupling between the two amide oscillators is obtained with a second map. Through-bond coupling between adjacent sites is generated by the DFT-based nearest-neighbor coupling map, and through-space coupling is computed by a transition charge coupling map.[57] Note that our calculations do not account for C=O stretch of the terminal COOH group.

We performed amide I spectral simulations by calculating a response function from dipole correlation functions using a dynamic wavefunction propagation method.[89] In this study we implemented a new Trotter expansion to reduce computation time,[88, 90] while maintaining errors at less than 1% (Fig. 7.4). The window time for calculating response functions was set to 11 ps, equivalent to 3 cm^{-1} frequency resolution, which is comparable to our frequency map errors of 2 cm^{-1}.[53] The anharmonicity of the amide I oscillator is set to 16 cm^{-1}, determined experimentally.[91] The model includes a vibrational lifetime for amide I modes, which is set as a 1.0 ps exponential decay to match the lifetime measured in our transient absorption experiment of AA.[15] The 1–^{13}C isotope frequency shift is set to 40 cm^{-1} obtained from FTIR experiments of AAA.

For ensemble refinement, we assume a separation of time-scales between the large amplitude conformational dynamics of the peptide and the fast fluctuations that give rise to the spectral lineshape. Spectra were calculated for 1000 conformational sub-states obtained by splitting the full 100 ns trajectories into 100 ps short trajectories. For the time-averaged response

function of each sub-ensemble, a moving average was applied by separating starting frame every 0.5 ps, resulting in 200 realizations for each sub-ensemble.

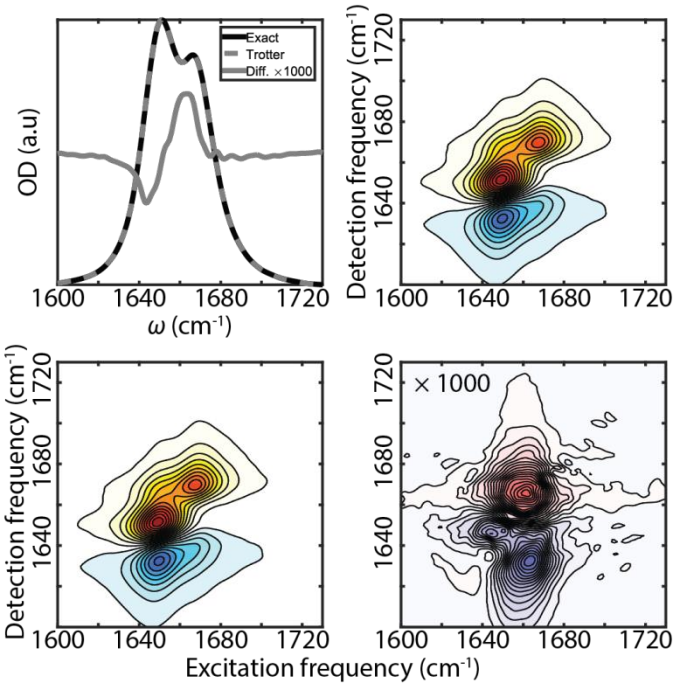

Figure 7.4: The error due to Trotter expansion. (a) Simulated FTIR spectra from C36 TIP3P trajectory (b) Simulated ZZZZ 2D IR spectrum using common matrix diagonalization scheme (exact) at waiting time set to 0.15 ps (c) Simulated ZZZZ 2D IR spectrum using Trotter expansion scheme at waiting time set to 0.15 ps (d) Difference spectrum of the two ZZZZ 2D IR spectra. All of the 2D IR spectra are normalized by the maximum peak intensity of the numerically exact spectrum. The color scale ranges from -0.002 to 0.002.

To investigate the correlation between conformations of AAA and amide I spectral features, we decomposed the ensemble-averaged spectra into spectra of conformers based on the structural definition described in MD simulation section. The spectral decomposition is done by splitting the 100 ns trajectory into 100 sub-ensembles. For each sub-ensemble, it would be assigned to a conformer when the corresponding population percentage is more than 75% in the entire sub-ensemble. Otherwise, the sub-ensemble would not be counted into any conformer.

7.3.8 Bayesian Ensemble Refinement

Under the assumption that MD simulations should provide a reasonable sampling of possible configurations, but that the corresponding distribution may deviate from the true ensemble due to inaccuracies in FFs and water models., we used a Bayesian ensemble refinement scheme to reweight existing ensemble populations to be consistent with experimental data.[92] Bayes theorem states that the posterior distribution based on experimental data $p(\boldsymbol{x}\,|\,\text{data})$ follows

$$p(\boldsymbol{x}\,|\,\text{data}) \propto p(\text{data}\,|\,\boldsymbol{x})\,p(\boldsymbol{x}) \tag{7.3}$$

In Eq. (7.3), the prior distribution $p(\boldsymbol{x})$ corresponds to a sub-ensemble $\boldsymbol{x}$ generated from a spectral simulation with uniformly distributed probabilities. The likelihood function $p(\text{data}\,|\,\boldsymbol{x})$ describing the probability of reproducing experimental data by $\boldsymbol{x}$ is formulated as follows.

$$p(\text{data}\,|\,\boldsymbol{x}) = \exp\left(-\sum_i \frac{1-s_i}{2\theta\sigma_i^2}\right) = \exp\left(-\chi^2(\theta)\right) \tag{7.4}$$

$$s_i(\boldsymbol{x}) = \frac{\int d\omega I_i^{x}(\omega) I_i^{\exp}(\omega)}{\sqrt{\left(\int d\omega \left(I_i^{x}(\omega)\right)^2\right) \times \left(\int d\omega \left(I_i^{\exp}(\omega)\right)^2\right)}} \tag{7.5}$$

$s_i(\boldsymbol{x})$ is the spectral overlap quantifying the similarity between the simulated spectra from $\boldsymbol{x}$, $I_i^{x}(\omega)$, and the spectrum from experiment, $I_i^{\exp}(\omega)$.[93] Here i refers to a specific type of experimental spectrum. We draw from a total of 15 different types of linear and nonlinear IR experimental spectra, including traditional FTIR spectra, diagonal slices through 2D IR spectra, and TA spectra (a projection of the 2D spectrum). Representative values of $s_i(\boldsymbol{x})$ are 1 for identical spectra, 0 for non-overlapping spectra, and -1 for identical but opposite-signed spectra. If $I_i^{x}(\omega)$ is identical to $I_i^{\exp}(\omega)$, then the measure of error $\chi^2(\theta) = 0$ and $p(\text{data}\,|\,\boldsymbol{x}) = 1$, meaning that there is no need to refine the probability at all. All the other cases would reduce $p(\text{data}\,|\,\boldsymbol{x})$ depending on θ, an

adjustable parameter expressing the level of confidence in the model, and $1/\theta$ is equivalent to the Lagrange multiplier in the ME formalism.[92] Large θ reflects high confidence in the model, whereas smaller θ would refine the prior distribution against the experimental data more. The optimal value of θ can be determined by the L-curve method,[92, 94] with detailed descriptions in Appendix 7B. The correlation between the optimal value of θ and the relative accuracy of the FFs and water models is discussed in Appendix 7B. The uncertainty σ_i is set to 2 cm^{-1} to account for errors from the spectroscopic maps.[53]

7.4 Results

7.4.1 Experimental Amide I Spectra

Infrared spectra were acquired on three isotopologues of AAA: the natural abundance unlabeled form, Ala–Ala–Ala (UL), and two singly ^{13}C labelled peptides: (1–^{13}C)Ala–Ala–Ala (A1) and Ala–(1–^{13}C)Ala–Ala (A2). At low pD in deuterated solvent, the peptides exist in the fully protonated form: ND_3^+–Ala–Ala–Ala–COOD. In Fig. 7.1b, the experimental FTIR spectrum of UL shows two distinct amide I peaks centered at 1650 and 1671 cm^{-1}, and a weak peak centered at 1714 cm^{-1} from the C=O stretch of the carboxyl group. Based on the spectra of A1 and A2 shown in Fig. 7.1b, the ^{13}C label shifts one of the amide I peaks while the other amide I peak is virtually unchanged, but the relative intensities of the two peaks change upon isotopic substitutions. Based on this observation, we conclude that the peaks at 1671 cm^{-1} and 1650 cm^{-1} in the UL spectrum originate from the amide I mode at the A1 and A2 positions, respectively. The frequency difference between these amide I modes in the UL spectrum results from a blue-shift of A1 stemming from its proximity to the positively charged ND_3^+ group, as suggested by previous studies.[61, 95] The

slight intensity variations of these peaks can be rationalized on the basis of a coupled oscillator model, described in the Appendix 7C.

The isotopic frequency shift to the amide I vibration resulting from the ^{13}C label can also be unambiguously determined from these spectra. From the coupled oscillator model, the ^{12}C-to-^{13}C isotopic frequency shift Δ can be expressed as $\Delta = 2\left(\bar{\omega}^* - \bar{\omega}\right)$, where $\bar{\omega}^*$ and $\bar{\omega}$ refer to the average of the two peak frequencies from the singly-^{13}C-labeled spectrum and the UL spectrum, respectively. Both A1 and A2 spectra give a consistent isotope frequency shift of 40 cm^{-1} consistent with previous measurements,[18-19] which is used in our spectral simulations.

Fig. 7.5a presents 2D IR spectra of the three AAA isotopologues under parallel and perpendicular polarization conditions. A 2D IR spectrum is a correlation map between the vibrational excitation frequency and detection frequency, and thus provides more information than FTIR spectra, such as a 2D lineshape that reflects the underlying broadening mechanisms and cross-peaks indicating coupling between vibrational modes. Each resonance in the 2D spectrum is a positive/negative (red/blue) doublet, which represent the ground state bleach (GSB) of the 0–1 quantum transition and excited state absorption (ESA) from the 1–2 quantum transition. The detection frequencies of these transitions differ as a result of the anharmonicity of vibrations. The 2D spectra in Fig. 7.5a are diagonally elongated, characteristic of inhomogeneous broadening arising from variations in the conformation of AAA and the variable solvation environments around the amide groups.

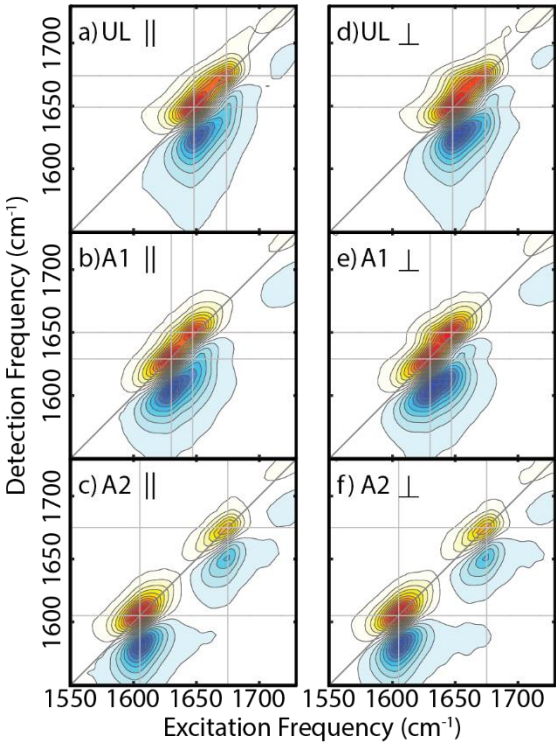

Figure 7.5: (a–c) Experimental parallel-polarized (‖) 2D IR spectra of (a) UL AAA, (b) A1-labeled AAA, and (c) A2-labeled AAA. **(**d–f) Experimental perpendicular-polarized (⊥) 2D IR spectra of (a) UL AAA, (b) A1-labeled AAA, and (c) A2-labeled AAA. The intensity of each spectrum is normalized to the maximum peak intensity.

Additional features of the 2D spectra are found in the cross peaks between amide I modes and the C=O stretch, which are sensitive to the relative orientation of the vibrational transition dipoles, and hence the conformation of the peptide backbone.[60-61] While the cross peaks between amide I modes at (1649 cm^{-1}, 1674 cm^{-1}) in the UL spectra and at (1630 cm^{-1}, 1650 cm^{-1}) in the A1-labeled spectra are not apparent in the parallel polarization between the excitation pulses and the detection pulse relative to the diagonal, they appear slightly more intense in the perpendicular polarization. From the cross-peaks, the coupling between amide I modes is estimated to be <8 cm^{-1} assuming the weak-coupling limit,[14] consistent with previous 2D IR studies of UL and 1–^{13}C labeled AAA.[60-61] Since there is no significant cross peak between the amide I vibration and the

C=O stretch of the terminal COOD in any 2D IR spectra, we conclude that this coupling is weak enough to be neglected in our amide I spectral model, consistent with our previous finding for dialanine.[15]

7.4.2 Effect of Conformational Variations on Amide I Spectra

To investigate how the variation of conformational distributions affects the amide I spectra, we simulated amide I spectra of AAA using C27, C36, and C36m FFs with SPC/E water. One difference between these FFs is the energy correction map, CMAP, which adjusts dihedral angle preferences. C27 is known for a significant bias toward α-helical conformations[29] while C36 corrects this bias. Additional refinement in C36m adjusts the propensity of the left-handed α-helical basin (α_L) to better describe intrinsically disordered proteins.[70-71] Thus, this series of CHARMM FFs provides a useful exploration of how differences in α-helical content affect the amide I spectra. Potentials of mean force (PMFs) of AAA for the three FFs are shown as a function of backbone dihedrals in Fig. 7.6, identifying the dominant conformational basins, β around $(\phi,\psi)=(-150°,150°)$, ppII around $(\phi,\psi)=(-60°,150°)$, α_R around $(\phi,\psi)=(-70°,-50°)$, and α_L around $(\phi,\psi)=(50°, 50°)$. A listing of population fractions in these states for all FF/solvent model combinations is given in Table 7.2. From Fig. 7.6 we see that α conformations are noticeably shallower in C36 relative to C27, corresponding to a decrease of α_R population from 29% in C27 to 7% in C36 and 2% in C36m. Total populations of α_L are always less than 4%, indicating that they contribute little to spectral simulations. Simulated IR spectra for the three isotopologues using the full trajectory from these FFs are shown in Fig. 7.6. While these peak frequencies show only a subtle 1–2 cm^{-1} blue shift, we observe a decrease in intensity of the higher frequency peak from

C27 to C36m, indicating that α_R conformers contribute to the intensity of this peak, and that the amide I spectra are sensitive to the underlying conformational distribution.

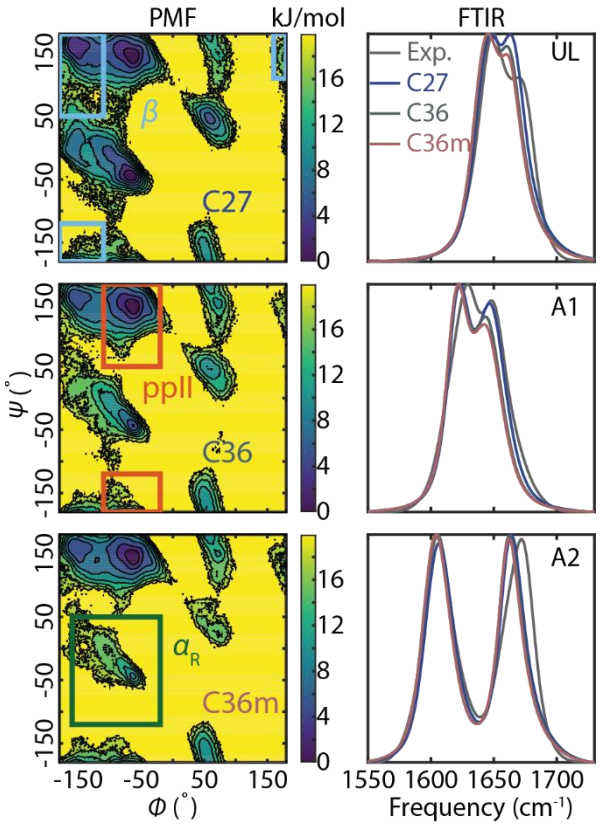

Figure 7.6: (Left) PMF(ϕ,ψ) computed from C27 SPC/E, C36 SPC/E, and C36m SPC/E trajectories. Contours are spaced by k_BT up to 6 k_BT. Colored boxes represent the definitions of conformer basins β (light blue), ppII (red), and α_R (Green). (Right) FTIR spectra of UL, A1, and A2-labeled AAA from the experiment (gray) and from the C27, C36, and C36m trajectories. The intensities of the simulated spectra are normalized to the maximum peak intensity of the experimental spectrum.

To describe the correlation between AAA conformations and amide I spectra, we decomposed the ensemble-averaged spectra into spectra for conformers β, ppII, and α_R. The conformational states, defined by the colored box boundaries in Fig. 7.6, correspond to common definitions,[17, 29, 85] with small FF-dependent shifts to the β-ppII boundary (Fig. 7.3). The resulting averaged conformer spectra (Fig. 7.7) show relatively small differences in their FTIR spectra,

mainly a difference in the relative intensities of the two amide I peaks. By defining the intensity ratio of the higher to the lower frequency peak, $R = I(\omega_+) / I(\omega_-)$, we find a clear trend in the variation of peak intensity ratio with structure as $R_\beta < R_{\text{ppII}} < R_{\alpha_R}$. The effects are much clearer in 2D IR spectra, in which the conformers are clearly distinguished either by the frequency of one dominant peak at 1648 cm^{-1} for β or 1669 cm^{-1} for α_R, or as the presence of two peaks for ppII. Head-to-head comparisons of peak intensities between the conformer spectra and the experimental spectra in Fig. 7.7 lead to the qualitative conclusion that the AAA conformational ensemble consists mostly of conformers in the ppII basin, with some population in the β basin, but no substantial population in the α_R state.

Conformers can also be distinguished through differences in their peak frequencies, as summarized in Table 7B.1. Although the relationship of peak frequency with structure is not trivial, we observe that the peak frequency from the A2 amide follows the trend $\omega_{\text{ppII}} \leq \omega_\beta \leq \omega_{\alpha_R}$. Also, for the peak from the A1 unit, the frequency is always highest for the α_R conformers.

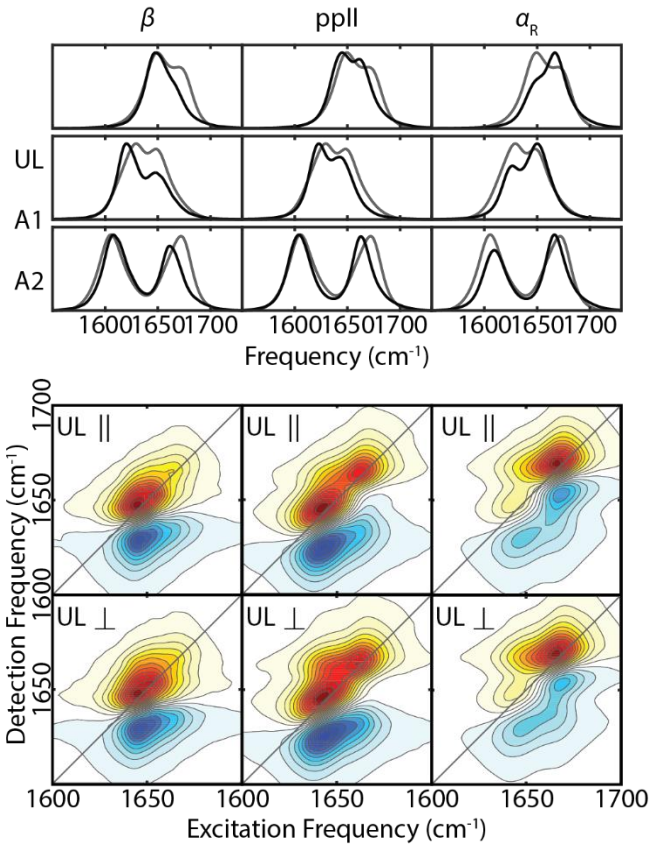

Figure 7.7: (Top) FTIR spectra from the experiments (gray), and FTIR spectra of conformers from the C27 SPC/E simulation (black). The intensity of each conformer spectrum is normalized to the corresponding experimental spectrum. (Bottom) parallel-polarized and perpendicular-polarized 2D IR spectra of the conformers. The intensity is normalized to the maximum peak intensity.

7.4.3 Ensemble Refinement against Amide I Spectroscopy

Our objective in this study is to test an ensemble refinement scheme that can be used to account for complex spectral features such as those found in FTIR and 2D IR spectra. In our recent study of ELPs, the ME refinement is proved effective for describing the extension in a type-I turn across multiple FFs.[5] However, the constraints of mean frequency and variance of a spectrally isolated peak is not easily extended to more complex features, such as multiple overlapping resonances and asymmetric spectral lineshapes. Additionally, the ME framework requires that the

constraints are strictly satisfied after the refinement, which may lead to biased refinement or poor convergence if the experimental constraints are chosen improperly or if uncertainties exist in the experiment (such as noise or signal bias). To incorporate information from multiple experiments and generalize the refinement method to arbitrary spectral lineshapes, we apply a Bayesian framework using the spectral overlap function defined in Eq. (7.5) as the new refinement metric.[93] The Bayesian framework is naturally suitable for updating the posterior probability distribution when given new experimental information, and constraints need not be matched exactly.

Although 2D IR spectra are sensitive to the underlying conformations, refining AAA conformational ensembles against spectra in two full frequency dimensions is computationally intensive. Therefore, for refinement, we reduced the dimensionality of the spectra in two ways: (1) taking diagonal slices through the 2D IR spectrum in which the excitation and detection frequencies are equal, and (2) using TA spectra, a projection of the 2D IR spectrum onto the detection frequency axis. Since the cross-peaks in this AAA case are insensitive to the underlying conformations (Fig. 7.7), using TA spectra and diagonal slices are reasonable simplifications. These spectra still contain constraints that are unique to the 2D IR spectrum and distinct from the FTIR spectrum, but in the case where cross-peaks were more pronounced, additional slices including the cross-peaks could be used.

An example of ensemble refinement of the C36m TIP3P trajectory simultaneously against all forms of IR spectra for all three isotopologues is shown in Fig. 7.8. Qualitatively, the frequency and intensity changes in refined FTIR spectra (Fig. 7.8c), 2D IR diagonal slices (Fig. 7.8d), and TA spectra (Fig. 7.8e) generally agree better with experiments, and this is borne out in the calculated spectral overlap changes in refinement (Fig. 7.9). Comparing the PMFs before and after the refinement (Figs. 7.8a–b) indicates that the refined ensemble is mostly conformers in the ppII

basin (86%) with a smaller fraction in the β basin (14%), and a negligibly small amount in other states. However, even with the constraints of 15 independent spectra, simulated spectra do not match the experiments exactly. There are many contributing factors to this mismatch, including errors or uncertainty in the spectroscopic map and FF, and inadequate structural sampling. While the spectral overlap values of FTIR spectra do not increase much after the refinement (Fig. 7.9), the overlap of simulated 2D IR diagonal slices and TA spectra with the experiments improve significantly, indicating that 2D spectra provide more stringent ensemble refinement constraints than FTIR.

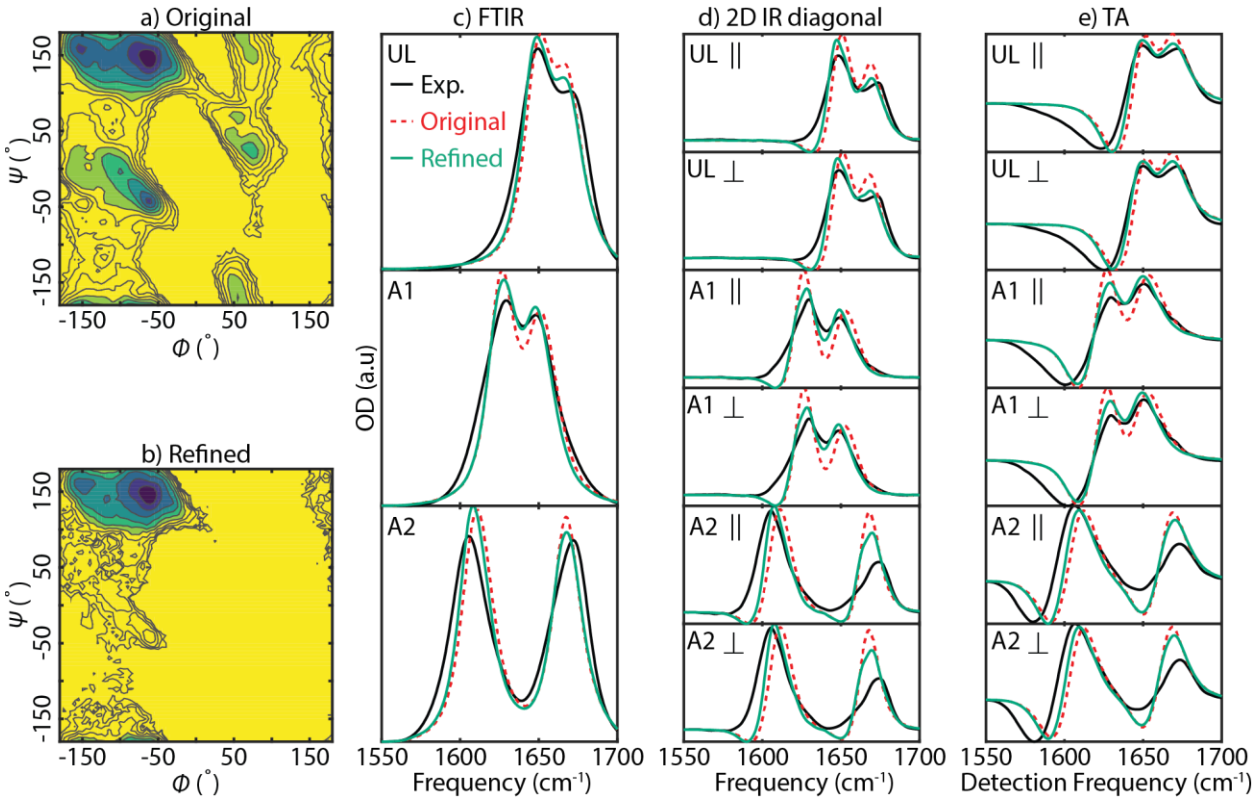

Figure 7.8: AAA ensemble refinement of C36m/TIP3P trajectory against infrared spectra. PMF(ϕ,ψ) of C36m TIP3P trajectory before (a) and after (b) ensemble refinement. The colored contours are spaced by k_BT up to 6 k_BT, while black contour lines extend to 10 k_BT. (c-e) Spectra used for the ensemble refinement, including FTIR spectra (c), diagonal slices (d), and TA spectra (e) from the experiments (black), simulation before refinement (dashed red), and simulation after refinement (solid green).

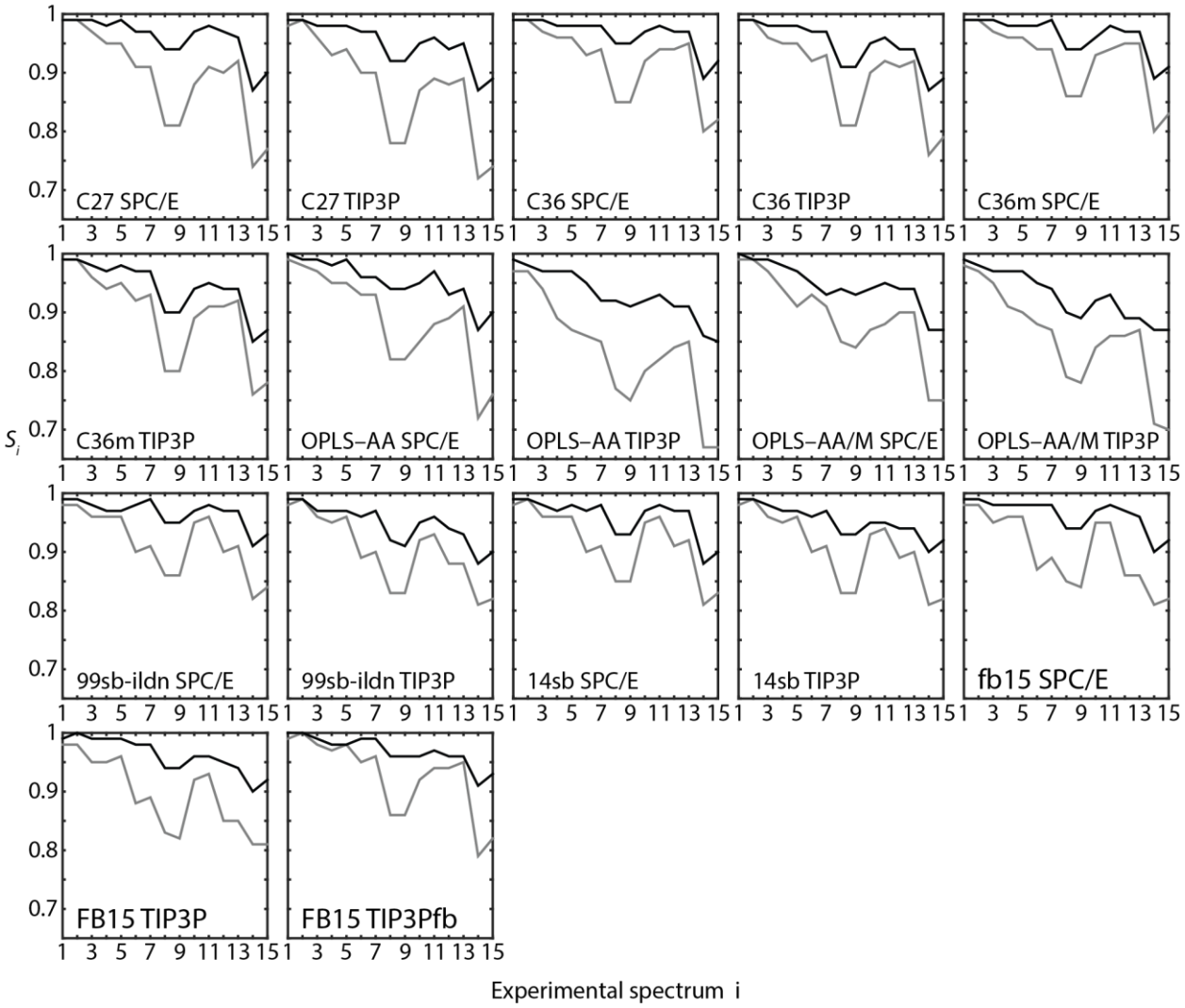

Figure 7.9: Spectral overlap value before (gray line) and after (black line) Bayesian ensemble refinement of different experimental spectra, including FTIR spectra (*i*=1–3), diagonal slices (*i*=4–9), and transient absorption spectra (*i*=9–15).

The consistency of Bayesian ensemble refinement against IR spectra across different prior distributions was examined by comparing the results from 17 combinations of FFs and water models. Fig. 7.10 illustrates the fraction of the population in ppII, β, and α_R states for all FF/water models before refinement, and the corresponding population fractions obtained after refinement. Prior to refinement, all FFs consistently predict the highest populations in ppII, but otherwise the original ensemble populations vary by nearly 30% among these FFs and water models, and predict

an average of 12% of the population in the α_R basin. However, the refined ensembles are more consistent, predicting ppII as the largest population (85% on average) with the rest mostly β conformer, and negligible α_R population in any ensemble. Thus, all FF/water combinations overestimate the population of the α_R basin and underestimate the population of the ppII basin based on comparison with our IR spectra. Also, the refined distributions of populations among these FFs and water models become narrower than the original distributions, indicating that Bayesian ensemble refinement against amide I spectra gives a consistent trend across many combinations of FFs and water models. The mean populations and the standard deviations of these conformers are summarized in Table 7.1 with other structural studies of AAA, and a complete list of populations before and after refinement is given in Tables 7.2–7.3.

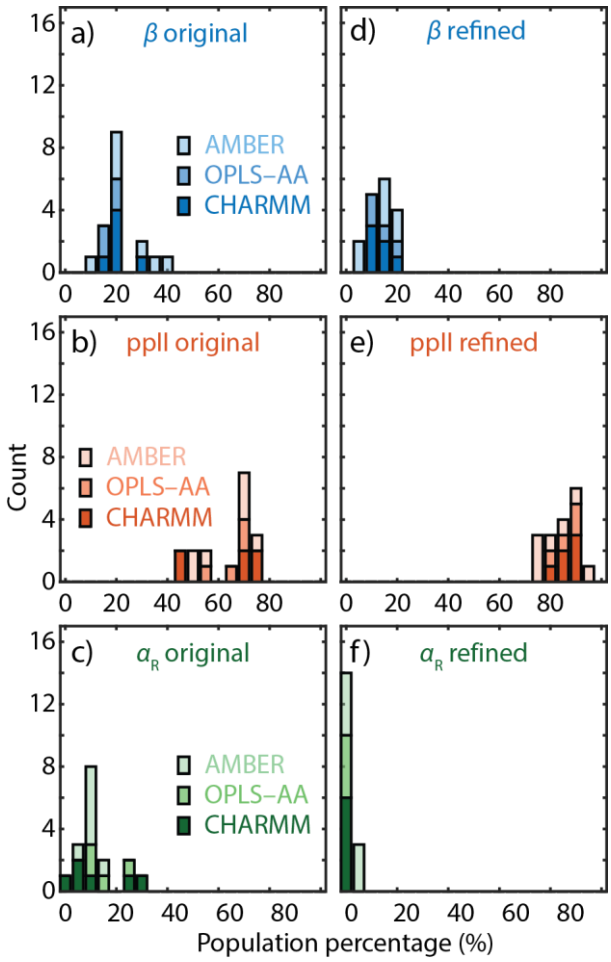

Figure 7.10: (a–c) Histogram of the original population percentage of (a) β conformer, (b) ppII conformer, and (c) α_R conformer. (d–f) Histogram of the refined population percentage of (d) β conformer, (e) ppII conformer, and (f) α_R conformer. The histograms are constructed from 17 combinations of FFs and water models listed in Table 7.2.

	Conformational State Populations				Method
	β	ppII	α_R	α_L	
Present study					
Original	22% (7%)	63% (11%)	12% (8%)	3% (2%)	MD simulations
Refined	14% (5%)	85% (6%)	1% (2%)	<0.1%	Bayesian ensemble refinement against FTIR and 2D IR
Hamm[62]	0%	80%	20%	0%	Fitting 2D IR spectra
Schweitzer-Stenner[63]	16%	84%	0%	0%	Fitting VCD, Raman, FTIR, and J coupling
Stock and Schwalbe[64]	8%	92%	0%	0%	Fitting to NMR
Cho[65]	12%	88%	0%	0%	NMR with Gromos 43A1
Zhang[40]	2.0% (1.8%)	85.8% (4.9%)	5.5% (4.1%)	3.5% (2.7%)	NMR data with Integrated Bayesian approach
Pande and Das[39]	23% (6%)	67% (9%)	10% (8%)	–	Bayesian Energy Landscape Tilting

Table 7.1: Average and standard deviation record in parentheses of the conformer population distributions shown in Fig. 7.10 and Table 7.2 before refinement, after refinement, and other studies of AAA.

	β		ppII		α_R	
	Original	**Refined**	**Original**	**Refined**	**Original**	**Refined**
C27 SPC/E	0.22 (0.03)	0.19 (0.15/0.25)	0.44 (0.04)	0.81 (0.75/0.85)	0.29 (0.07)	0.00
C27 TIP3P	0.28 (0.02)	0.11 (0.11/0.22)	0.44 (0.03)	0.89 (0.77/0.89)	0.24 (0.03)	0.00 (0.00/0.02)
C36 SPC/E	0.16 (0.01)	0.12 (0.10/0.14)	0.73 (0.03)	0.88 (0.86/0.90)	0.07 (0.03)	0.00
C36 TIP3P	0.20 (0.01)	0.15 (0.13/0.18)	0.68 (0.03)	0.85 (0.82/0.87)	0.06 (0.01)	0.00
C36m SPC/E	0.19 (0.01)	0.11 (0.05/0.20)	0.77 (0.02)	0.89 (0.80/0.95)	0.02 (0.01)	0.00
C36m TIP3P	0.20 (0.01)	0.14 (0.13/0.15)	0.69 (0.02)	0.86 (0.84/0.86)	0.09 (0.02)	0.00 (0.00/0.01)
OPLS–AA SPC/E	0.17 (0.01)	0.08 (0.07/0.15)	0.66 (0.03)	0.92 (0.84/0.93)	0.17 (0.03)	0.00 (0.00/0.01)
OPLS–AA TIP3P	0.16 (0.01)	0.16 (0.15/0.21)	0.56 (0.02)	0.84 (0.77/0.85)	0.27 (0.02)	0.00 (0.00/0.02)
OPLS–AA/M SPC/E	0.20 (0.01)	0.11 (0.09/0.15)	0.69 (0.02)	0.89 (0.85/0.91)	0.09 (0.02)	0.00 (0.00/0.01)
OPLS–AA/M TIP3P	0.20 (0.01)	0.21 (0.21/0.23)	0.69 (0.02)	0.79 (0.76/0.79)	0.09 (0.02)	0.00 (0.00/0.01)
Amber99sb-ildn SPC/E	0.20 (0.01)	0.16 (0.15/0.20)	0.69 (0.02)	0.81 (0.76/0.82)	0.09 (0.02)	0.03 (0.01/0.04)
Amber99sb-ildn TIP3P	0.31 (0.01)	0.22 (0.18/0.34)	0.56 (0.02)	0.75 (0.66/0.75)	0.13 (0.02)	0.03 (0.00/0.15)
Amber14sb SPC/E	0.11 (0.01)	0.07 (0.05/0.09)	0.75 (0.06)	0.91 (0.78/0.94)	0.06 (0.02)	0.01 (0.00/0.10)
Amber14sb TIP3P	0.20 (0.01)	0.04 (0.04/0.10)	0.69 (0.02)	0.95 (0.87/0.96)	0.09 (0.02)	0.00 (0.00/0.06)
Amberfb15 SPC/E	0.39 (0.01)	0.15 (0.13/0.19)	0.51 (0.02)	0.84 (0.67/0.85)	0.09 (0.01)	0.01 (0.01/0.14)
Amberfb15 TIP3P	0.35 (0.01)	0.17 (0.09/0.23)	0.52 (0.01)	0.76 (0.72/0.84)	0.12 (0.01)	0.07 (0.04/0.07)
Amberfb15 TIP3Pfb	0.20 (0.02)	0.20 (0.09/0.24)	0.69 (0.04)	0.77 (0.72/0.90)	0.09 (0.02)	0.02 (0.01/0.03)

Table 7.2: Population fractions of conformers before and after ensemble refinement with frequency corrections if any. Parentheses in the original populations represent the range of population due to standard error computed from block averaging while parenthesis in the refined ensemble represent the lowest and highest populations found by varying the systematic frequency shift from -4 to 4 cm^{-1}.

	β	ppII	α_R
Original	22% (7%)	63% (11%)	12% (8%)
Refined CHARMM	14% (3%)	86% (6%)	0% (0%)
Refined OPLS–AA	14% (6%)	86% (6%)	0% (0%)
Refined AMBER	14% (7%)	83% (8%)	2% (2%)
Refined Total	14% (5%)	85% (6%)	1% (2%)

Table 7.3: Average and standard deviation record in parentheses of the conformer population distributions shown in Fig. 7.10 before refinement, after refinement from CHARMM FFs, OPLS–AA FFs, AMBER FFs, and all of the FFs.

7.5 Discussions

We have tested a Bayesian ensemble refinement protocol that draws from MD trajectories to simulate amide I spectra of site-specifically isotope-labelled peptides. Structure-based spectral modeling of amide I vibrations draws from MD simulations to sample structures of peptide and solvent that predict IR frequencies, intensities and lineshapes. Combining a series of linear and nonlinear IR spectra on several peptide isotopologues provides multiple constraints that can be self-consistently analyzed. Our test of this procedure on AAA shows that the changes of amide I peak frequency and intensity can distinguish conformational basins within the AAA energy landscape, and be used to describe the underlying conformational distribution in heterogeneous ensembles. IR and 2D IR spectra reveal that AAA contain mostly conformations in the ppII basin and some portion of conformations in the β basin, which is consistent with many previous studies.[39-40, 62-65] This is also quantitatively supported by Bayesian ensemble refinement across 17 combinations of FFs and water models. The results highlight the potential of amide I IR and 2D IR spectroscopy as a tool for ensemble refinement of peptides and proteins, and provide a rigorous statistical framework for interpreting the underlying conformational distribution.

This study also identified several avenues for improvement, and various challenges for describing conformational distributions quantitatively and accurately with IR spectroscopy. In implementing this refinement strategy, results may be influenced by multiple possible sources of error or bias beyond uncertainty in the experimental spectra, including inaccuracy in FFs and water models, inadequate sampling, and errors in the underlying spectroscopic models. Additionally, the outcome may also be affected by the choice of which type of experimental spectra used. In the following, we describe our analysis of how these factors affect the Bayesian ensemble refinement and interpretation of the underlying conformational ensemble.

7.5.1 Effect of Experimental Input on Ensemble Refinement

Using different experimental inputs will influence the regularization used to refine the simulated prior ensemble. To investigate the correlation between the refinement quality and experimental input, we compared ensemble refinements of C36m TIP3P and C27 SPC/E trajectories against (1) a single UL FTIR spectrum, (2) three FTIR spectra of all isotopologues, (3) six 2D IR diagonal slices for all isotopologue/polarization combinations, (4) six TA spectra of all isotopologue/polarization combinations, and (5) the full set of fifteen spectra in Fig. 7.8. Fig. 7.11 summarizes the refined population ratios obtained by these five sets of restraints, showing a clear decrease in the error bars for the refined populations as more restraints are added. Refinement purely on FTIR data gave the poorest agreement, with the worst case being 55% error bars in the ppII and α_R populations using the C27 SPC/E initial ensemble. However, significant improvements are obtained by using nonlinear spectra (2D diagonal slices or TA spectra), indicating that 2D IR spectroscopy provides a more stringent constraint than FTIR. TA spectra generally give a narrower

uncertainty than the diagonal slice because of additional information of the ESA present in the TA spectra.

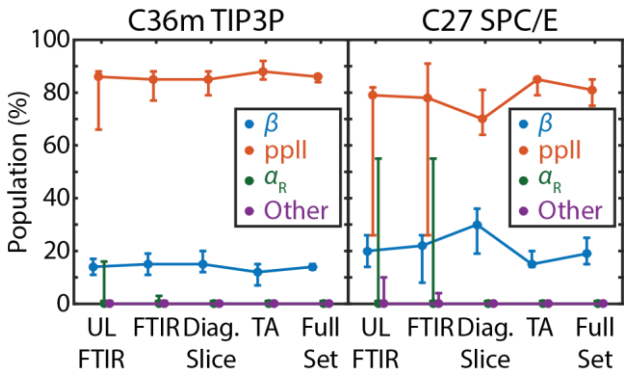

Figure 7.11: Effect of different experimental information on the ensemble refinement. Refined population percentage of conformers against various inputs of experiment spectra from (Left) C36m TIP3P trajectory, and (Right) C27 SPC/E trajectory with post-frequency correction. Inputs for refining ensemble includes UL FTIR, full FTIR spectra, all of the diagonal slices, all of the TA spectra, and the entire set of spectra. Colored error bars reflect the $\pm 2\sigma$ uncertainty due to the spectroscopic map, determined by the maximum and the minimum populations with systematic frequency error ranging from -4 to 4 cm^{-1}.

7.5.2 Water Model Dependence

SPC/E, TIP3P, and TIP3Pfb are different in structure, electrostatic charges, and van der Waals parameters, which result in different intermolecular interactions and diffusion coefficients.[80, 96] These differences noticeably affect the amide I frequency and spectral lineshape, as illustrated in our recent study of Ala–Ala.[15] SPC/E water was shown to better reproduce dynamical behavior and spectral lineshape in 2D IR spectroscopy[15, 97] whereas TIP3P water has a benefit of speeding up conformational sampling of proteins,[98] and it is commonly used with many FFs.

To illustrate the effect of different water models on the amide I frequency, Figs. 7.12a and 7B.2 presents the average vibrational frequencies of the two amide vibrations as a function of

peptide backbone dihedral angles using the TIP3P and SPC/E water models with the C36m FF. One clear observation is that the frequency depends on AAA conformations, and the α_R basin has the highest frequency, consistent with the comparisons of conformer spectra in Fig. 7.7. The other observation is that SPC/E water results in a uniform shift of -2 to -6 cm^{-1} relative to TIP3P. This shift is on the order of the uncertainty of our spectroscopic model, but can significantly influence the ensemble refinements.

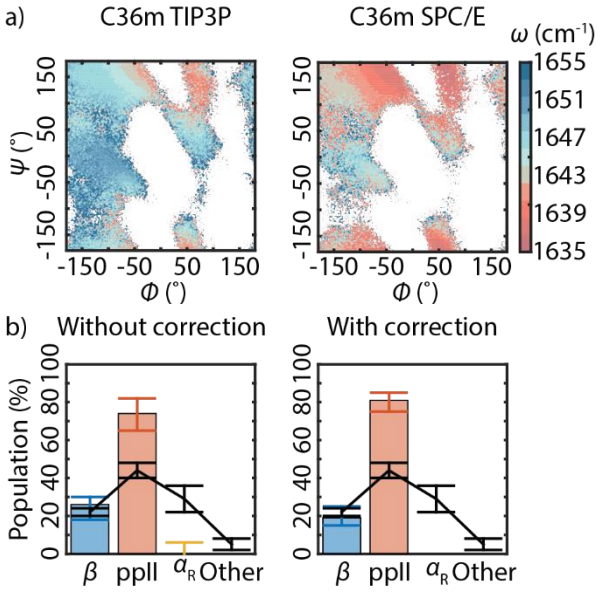

Figure 7.12: (a) Mean lower peak (exciton) frequency distribution of UL AAA from C36m TIP3P ensemble (left) and C36m SPC/E ensemble (right). Note that the TIP3P atomic charges are applied to C36m SPC/E trajectory. (b) Population percentage of conformers of C27 SPC/E ensemble before refinement (black line), and after refinement against the entire set of the spectra (colored bars). Frequency correction is applied on the right panel. Black error bars represent the uncertainty of the original distribution estimated from block averaging. Colored error bars reflect the $\pm 2\sigma$ uncertainty due to the spectroscopic map, determined by the maximum and the minimum populations with systematic frequency error ranging from -4 to 4 cm^{-1}.

The effect of the red shift of SPC/E water on the refined population changes of the C27 SPC/E ensemble can be seen in Fig. 7.12b. Without frequency correction to have the same average site frequencies as the CHARMM TIP3P trajectories (See the Appendix 7B), the population of

ppII is 74%, 11% lower than the average value with the frequency correction (Table 7.1), and the corresponding uncertainty is as high as 17%. In contrast, the refinements with the frequency correction have better consistency against the whole set of IR spectra. Although the correction applied here is only an approximation for correcting the frequency shift due to structural difference between water models, it helps avoid false positive interpretations by using FTIR spectra as the only constraint without frequency correction (Fig. 7B.3) and achieves better consistency among the various FFs shown in Fig. 7.8.

7.5.3 Errors in Spectroscopic Maps

Spectroscopic maps certainly influence the quality of ensemble refinement with sources of error such as choice of electrostatic variables used in the frequency map, and errors in the frequency or coupling maps. Among these factors, we found that the most significant factor is the error due to the frequency map, as partly seen in the systematic bias present in different water models above. Taking 2σ (± 4 cm^{-1}) of the spectroscopic map error as the estimated confidence interval,[5, 53] the highest uncertainty of determining the ensemble populations against the full set of spectra is found to be 18% in AMBERfb15 ensembles (Table 7.2). These high uncertainties indicate that the frequency error is the most influential source of degrading quality of the ensemble refinement. Increased uncertainties originate from introducing different water models, and potentially the $CONH_2$ group, but the range of uncertainty is still narrower than ±15% reported in the ELP ensemble refinement study against FTIR spectroscopy,[5] supporting the conclusion that 2D IR spectroscopy serves as an additional and stringent constraint on refining conformational ensemble.

Errors due to the nearest-neighbor coupling map are difficult to address because there are no experimental standards for evaluating the quality of coupling maps. However, at the DFT level calculations, the unsigned coupling map uncertainty is 1.7 cm^{-1}.[99] Based on this value, we estimated the error bars in frequency prediction due to coupling uncertainty to be 1.3 cm^{-1} (See Appendix 7C), smaller than the confidence interval of site frequency uncertainty (2 cm^{-1}), suggesting that the coupling uncertainty may be less of a factor than the errors in the frequency map. However, we cannot conclude that the uncertainty in the coupling map is insignificant for the ensemble refinements. Also, different choices of electrostatic variables are found to give consistent ensemble refinement trends (See Appendix 7B).

7.5.4 Comparison to Other Studies of AAA

Our refinements of 17 combinations of FFs and water models indicate that the conformational ensemble of AAA consists of 85% ppII, 14% β, and negligible population of α_R, mostly consistent with other studies shown in Table 7.1 except for the result from BELT, and some studies having larger α_R population than the β population. There are a few differences among these studies such as different definitions of the conformational states, various approaches of determining structural ensembles, and different experimental techniques employed. First, how to cluster structures into conformational states varies from study to study, which would lead to different values of conformational populations. Note that our boundary of distinguishing ppII and β conformers has been shifted to a smaller value of ϕ to reside in the middle of the two basins, resulting in a larger box of ppII and a larger population compared to the BELT study. Second, the previous studies of refining or restraining ensemble are against NMR data such as chemical shift and J coupling while in our study we use FTIR and 2D IR spectra, which provide different

information, and have different uncertainty. From the simulated conformer spectra in Fig. 7.7, the IR spectroscopy of AAA strongly disfavors α_R, but small amount may still be possible within the sensitivity of the experiment.[29] The early 2D IR study of AAA suggested 20% α_R. They excluded β conformer based on the angle derived from ^3J coupling between C_α and N proton and performed fitting based on the two-state model of ppII and α_R.[62] However, all of the studies suggest the dominance of the ppII conformation, and the Bayesian ensemble refinement against IR spectroscopy is consistent with most of the studies.

7.6 Conclusions

We studied a Bayesian ensemble refinement scheme using structure-based amide I spectral modeling of site-specifically labelled peptide isotopologues against FTIR and 2D IR spectroscopy. This proof-of-principle study on Ala–Ala–Ala drawing on multiple force fields and water models demonstrates the practical capability of 2D IR experiments to constrain quantitative ensemble refinement protocols, and in this case consistently resulted in ensembles consisting of 85% population in the ppII basin, 14% in the β basin, and negligible α_R population, consistent with most previous studies on AAA. The nature of Bayesian statistics allows us to improve refinement by integrating other complementary experimental constraints. Investigating potential sources of uncertainty, we find the dominant factor influencing results is systematic frequency uncertainty from the amide I frequency spectroscopic map. However, with the structural information given by 2D IR spectroscopy, the upper bound of the uncertainty range can be narrowed down to ~15% in the worst case. We believe this study helps lay the groundwork for general methods of refining protein conformational ensembles against isotope-edited 2D IR spectroscopy. With additional steps to advance this refinement approach, including further improvements to amide I

spectroscopic models and incorporating conformational binning strategies better suited to proteins such as Markov State Models,[13, 21-22] we hope to effectively describe conformational ensembles in complex, disordered systems at equilibrium and in non-equilibrium dynamic processes.

7.7 Acknowledgments

I thank Paul Sanstead as a helpful go-to person to learn about how to take quality 2D IR spectra as this dataset was taken around the same time as the waiting time series data of Ala–Ala. I thank Paul Stevenson for helpful discussions on different candidate inputs for ensemble refinements.

7.8 References

1. Papoian, G. A., Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proc Natl Acad Sci U S A* **2008,** *105* (38), 14237-8.
2. Burger, V.; Gurry, T.; Stultz, C., Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **2014,** *6* (10), 2684-2719.
3. Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M., Controlling entropy to tune the functions of intrinsically disordered regions. *Curr Opin Struct Biol* **2014,** *26*, 62-72.
4. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M., Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **2017,** *42*, 106-116.
5. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
6. Rieping, W.; Habeck, M.; Nilges, M., Inferential structure determination. *Science* **2005,** *309* (5732), 303-6.
7. Fleming, G. R.; Wolynes, P. G., Chemical Dynamics in Solution. *Physics Today* **1990,** *43* (5), 36-43.
8. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450* (7172), 964-72.
9. Bryant, R. G., The NMR time scale. *Journal of Chemical Education* **1983,** *60* (11), 933.
10. Zanni, M. T.; Hochstrasser, R. M., Two-dimensional infrared spectroscopy: a promising new method for the time resolution of structures. *Current Opinion in Structural Biology* **2001,** *11* (5), 516-522.
11. Hamm, P.; Zanni, M., *Concepts and methods of 2D infrared spectroscopy*. Cambridge University Press: 2011.
12. Baiz, C.; Reppert, M.; Tokmakoff, r., An Introduction to Protein 2D IR Spectroscopy. **2013**, 361-404.

13. Baiz, C. R.; Lin, Y. S.; Peng, C. S.; Beauchamp, K. A.; Voelz, V. A.; Pande, V. S.; Tokmakoff, A., A molecular interpretation of 2D IR protein folding experiments with Markov state models. *Biophys J* **2014,** *106* (6), 1359-70.
14. Hamm, P.; Lim, M.; DeGrado, W. F.; Hochstrasser, R. M., The two-dimensional IR nonlinear spectroscopy of a cyclic penta-peptide in relation to its three-dimensional structure. *Proc Natl Acad Sci U S A* **1999,** *96* (5), 2036-41.
15. Feng, C. J.; Tokmakoff, A., The dynamics of peptide-water interactions in dialanine: An ultrafast amide I 2D IR and computational spectroscopy study. *J Chem Phys* **2017,** *147* (8), 085101.
16. Bredenbeck, J.; Hamm, P., Peptide structure determination by two-dimensional infrared spectroscopy in the presence of homogeneous and inhomogeneous broadening. *The Journal of Chemical Physics* **2003,** *119* (3), 1569-1578.
17. Feng, Y.; Huang, J.; Kim, S.; Shim, J. H.; MacKerell, A. D., Jr.; Ge, N. H., Structure of Penta-Alanine Investigated by Two-Dimensional Infrared Spectroscopy and Molecular Dynamics Simulation. *J Phys Chem B* **2016,** *120* (24), 5325-39.
18. Torres, J.; Kukol, A.; Goodman, J. M.; Arkin, I. T., Site-specific examination of secondary structure and orientation determination in membrane proteins: The peptidic13C?18O group as a novel infrared probe. *Biopolymers* **2001,** *59* (6), 396-401.
19. Decatur, S. M., Elucidation of residue-level structure and dynamics of polypeptides via isotope-edited infrared spectroscopy. *Acc Chem Res* **2006,** *39* (3), 169-75.
20. Baiz, C. R.; Peng, C. S.; Reppert, M. E.; Jones, K. C.; Tokmakoff, A., Coherent two-dimensional infrared spectroscopy: quantitative analysis of protein secondary structure in solution. *Analyst* **2012,** *137* (8), 1793-9.
21. Smith, A. W.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A.; Roy, S.; Jansen, T. L.; Knoester, J., Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J Phys Chem B* **2010,** *114* (34), 10913-24.
22. Baiz, C. R.; Tokmakoff, A., Structural disorder of folded proteins: isotope-edited 2D IR spectroscopy and Markov state modeling. *Biophys J* **2015,** *108* (7), 1747-1757.
23. Huang, C. Y.; Getahun, Z.; Zhu, Y.; Klemke, J. W.; DeGrado, W. F.; Gai, F., Helix formation via conformation diffusion search. *Proc Natl Acad Sci U S A* **2002,** *99* (5), 2788-93.
24. Tucker, M. J.; Abdo, M.; Courter, J. R.; Chen, J.; Brown, S. P.; Smith, A. B., 3rd; Hochstrasser, R. M., Nonequilibrium dynamics of helix reorganization observed by transient 2D IR spectroscopy. *Proc Natl Acad Sci U S A* **2013,** *110* (43), 17314-9.
25. Meuzelaar, H.; Marino, K. A.; Huerta-Viga, A.; Panman, M. R.; Smeenk, L. E.; Kettelarij, A. J.; van Maarseveen, J. H.; Timmerman, P.; Bolhuis, P. G.; Woutersen, S., Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations. *J Phys Chem B* **2013,** *117* (39), 11490-501.
26. Stevenson, P.; Gotz, C.; Baiz, C. R.; Akerboom, J.; Tokmakoff, A.; Vaziri, A., Visualizing KcsA conformational changes upon ion binding by infrared spectroscopy and atomistic modeling. *J Phys Chem B* **2015,** *119* (18), 5824-31.
27. Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D.; Skinner, J. L.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016,** *353* (6303), 1040-1044.

28. Huang, J.; MacKerell, A. D., Jr., Force field development and simulations of intrinsically disordered proteins. *Curr Opin Struct Biol* **2018,** *48*, 40-48.
29. Best, R. B.; Buchete, N. V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophys J* **2008,** *95* (1), L07-9.
30. Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A., Accuracy of current all-atom force-fields in modeling protein disordered states. *J Chem Theory Comput* **2015,** *11* (1), 2-7.
31. Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmuller, H., Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J Chem Theory Comput* **2015,** *11* (11), 5513-24.
32. Henriques, J.; Cragnell, C.; Skepo, M., Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput* **2015,** *11* (7), 3420-31.
33. Levine, Z. A.; Shea, J. E., Simulations of disordered proteins and systems with conformational heterogeneity. *Curr Opin Struct Biol* **2017,** *43*, 95-103.
34. Pitera, J. W.; Chodera, J. D., On the Use of Experimental Observations to Bias Simulated Ensembles. *J Chem Theory Comput* **2012,** *8* (10), 3445-51.
35. Roux, B.; Weare, J., On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys* **2013,** *138* (8), 084107.
36. Cavalli, A.; Camilloni, C.; Vendruscolo, M., Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J Chem Phys* **2013,** *138* (9), 094112.
37. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K., Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* **2014,** *10* (2), e1003406.
38. Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noe, F., Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc Natl Acad Sci U S A* **2017,** *114* (31), 8265-8270.
39. Beauchamp, K. A.; Pande, V. S.; Das, R., Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys J* **2014,** *106* (6), 1381-90.
40. Xiao, X.; Kallenbach, N.; Zhang, Y., Peptide Conformation Analysis Using an Integrated Bayesian Approach. *J Chem Theory Comput* **2014,** *10* (9), 4152-4159.
41. Brookes, D. H.; Head-Gordon, T., Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J Am Chem Soc* **2016,** *138* (13), 4530-8.
42. Fisher, C. K.; Huang, A.; Stultz, C. M., Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc* **2010,** *132* (42), 14919-27.
43. Rozycki, B.; Kim, Y. C.; Hummer, G., SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011,** *19* (1), 109-16.
44. Shevchuk, R.; Hub, J. S., Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput Biol* **2017,** *13* (10), e1005800.
45. Sethi, A.; Anunciado, D.; Tian, J.; Vu, D. M.; Gnanakaran, S., Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics. *Chem Phys* **2013,** *422*.
46. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
47. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.

48. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.
49. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
50. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
51. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.
52. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
53. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
54. Torii, H., Amide I Vibrational Properties Affected by Hydrogen Bonding Out-of-Plane of the Peptide Group. *J Phys Chem Lett* **2015,** *6* (4), 727-33.
55. Torii, H.; Tasumi, M., Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *Journal of Raman Spectroscopy* **1998,** *29* (1), 81-86.
56. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
57. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
58. Hayashi, T.; Mukamel, S., Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides. *J Phys Chem B* **2007,** *111* (37), 11032-46.
59. Maekawa, H.; De Poli, M.; Moretto, A.; Toniolo, C.; Ge, N. H., Toward detecting the formation of a single helical turn by 2D IR cross peaks between the amide-I and -II modes. *J Phys Chem B* **2009,** *113* (34), 11775-86.
60. Woutersen, S.; Hamm, P., Structure Determination of Trialanine in Water Using Polarization Sensitive Two-Dimensional Vibrational Spectroscopy. *The Journal of Physical Chemistry B* **2000,** *104* (47), 11316-11320.
61. Woutersen, S.; Hamm, P., Isotope-edited two-dimensional vibrational spectroscopy of trialanine in aqueous solution. *The Journal of Chemical Physics* **2001,** *114* (6), 2727-2737.
62. Woutersen, S.; Pfister, R.; Hamm, P.; Mu, Y.; Kosov, D. S.; Stock, G., Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations. *The Journal of Chemical Physics* **2002,** *117* (14), 6833-6840.
63. Schweitzer-Stenner, R., Distribution of conformations sampled by the central amino acid residue in tripeptides inferred from amide I band profiles and NMR scalar coupling constants. *J Phys Chem B* **2009,** *113* (9), 2922-32.
64. Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H., Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *J Am Chem Soc* **2007,** *129* (5), 1179-89.
65. Oh, K. I.; Lee, K. K.; Park, E. K.; Yoo, D. G.; Hwang, G. S.; Cho, M., Circular dichroism eigenspectra of polyproline II and beta-strand conformers of trialanine in water: Singular value decomposition analysis. *Chirality* **2010,** *22 Suppl 1*, E186-201.

66. Deflores, L. P.; Nicodemus, R. A.; Tokmakoff, A., Two-dimensional Fourier transform spectroscopy in the pump-probe geometry. *Opt Lett* **2007,** *32* (20), 2966-8.
67. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-54.
68. Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* **2004,** *25* (11), 1400-15.
69. Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E., Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *J Chem Theory Comput* **2010,** *6* (2), 459-66.
70. Huang, J.; MacKerell, A. D., Jr., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* **2013,** *34* (25), 2135-45.
71. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2017,** *14* (1), 71-73.
72. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.
73. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides†. *The Journal of Physical Chemistry B* **2001,** *105* (28), 6474-6487.
74. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *J Chem Theory Comput* **2015,** *11* (7), 3499-509.
75. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-8.
76. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015,** *11* (8), 3696-713.
77. Wang, L. P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martinez, T. J.; Pande, V. S., Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J Phys Chem B* **2017,** *121* (16), 4023-4039.
78. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P., The missing term in effective pair potentials. *The Journal of Physical Chemistry* **1987,** *91* (24), 6269-6271.
79. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983,** *79* (2), 926-935.
80. Wang, L. P.; Martinez, T. J.; Pande, V. S., Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J Phys Chem Lett* **2014,** *5* (11), 1885-91.

81. Toal, S.; Meral, D.; Verbaro, D.; Urbanc, B.; Schweitzer-Stenner, R., pH-Independence of trialanine and the effects of termini blocking in short peptides: a combined vibrational, NMR, UVCD, and molecular dynamics study. *J Phys Chem B* **2013,** *117* (14), 3689-706.
82. Nosé, S., A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **1984,** *81* (1), 511-519.
83. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A Gen Phys* **1985,** *31* (3), 1695-1697.
84. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014,** *185* (2), 604-613.
85. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D., Jr., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput* **2012,** *8* (9), 3257-3273.
86. Grossfield, A.; Zuckerman, D. M., Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. **2009,** *5*, 23-48.
87. Reppert, M. *g_amide*, 2017.
88. Reppert, M.; Feng, C. J. *g_spec*, 2017.
89. Torii, H., Effects of intermolecular vibrational coupling and liquid dynamics on the polarized Raman and two-dimensional infrared spectral profiles of liquid N,N-dimethylformamide analyzed with a time-domain computational method. *J Phys Chem A* **2006,** *110* (14), 4822-32.
90. Liang, C.; Jansen, T. L., An Efficient N(3)-Scaling Propagation Scheme for Simulating Two-Dimensional Infrared and Visible Spectra. *J Chem Theory Comput* **2012,** *8* (5), 1706-13.
91. Hamm, P.; Lim, M.; Hochstrasser, R. M., Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *The Journal of Physical Chemistry B* **1998,** *102* (31), 6123-6138.
92. Hummer, G.; Kofinger, J., Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys* **2015,** *143* (24), 243150.
93. Kruiger, J. F.; van der Vegte, C. P.; Jansen, T. L., Suppressing sampling noise in linear and two-dimensional spectral simulations. *J Chem Phys* **2015,** *142* (5), 054201.
94. Hansen, P. C.; O'Leary, D. P., The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM Journal on Scientific Computing* **1993,** *14* (6), 1487-1503.
95. Sieler, G.; Schweitzer-Stenner, R.; Holtz, J. S. W.; Pajcini, V.; Asher, S. A., Different Conformers and Protonation States of Dipeptides Probed by Polarized Raman, UV−Resonance Raman, and FTIR Spectroscopy. *The Journal of Physical Chemistry B* **1999,** *103* (2), 372-384.
96. Mahoney, M. W.; Jorgensen, W. L., Diffusion constant of the TIP5P model of liquid water. *The Journal of Chemical Physics* **2001,** *114* (1), 363.
97. Schmidt, J. R.; Roberts, S. T.; Loparo, J. J.; Tokmakoff, A.; Fayer, M. D.; Skinner, J. L., Are water simulation models consistent with steady-state and ultrafast vibrational spectroscopy experiments? *Chemical Physics* **2007,** *341* (1-3), 143-157.
98. Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M., Explicit Water Models Affect the Specific Solvation and Dynamics of Unfolded Peptides While the Conformational Behavior and Flexibility of Folded Peptides Remain Intact. *J Chem Theory Comput* **2010,** *6* (11), 3569-79.
99. Carr, J. K.; Zabuga, A. V.; Roy, S.; Rizzo, T. R.; Skinner, J. L., Assessment of amide I spectroscopic maps for a gas-phase peptide using IR-UV double-resonance spectroscopy and density functional theory calculations. *J Chem Phys* **2014,** *140* (22), 224111.

Appendix 7A: Synthesis, Purification, and Characterization of Ala–Ala–Ala

The work presented in this chapter has been published and is reprinted with permission from:

Feng, C.-J.; Dhayalan, B.; Tokmakoff, A. Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophysical Journal* **2018**, 114 (12), 2820–2832.

7A.1 Analytical LC-MS

Analytical reverse-phase HPLC and LC-MS were performed using an Agilent 1100 series HPLC system equipped with an online MSD ion trap. All chromatographic separations were performed on a C8 (4.6×150 mm) Phenomenex column at 40 ^{o}C, using a linear gradient (0.5–8% of solvent B in solvent A over 15 min for trialanines; 5–65% of solvent B in solvent A over 20 min for Fmoc-Ala; solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile) at a flow rate of 1.0 mL/min, with detection by UV absorption at 214 nm. Masses were obtained by online electrospray mass spectrometry. All MS data shown were collected across the entire principal UV absorbing peak in each chromatogram.

7A.2 Preparative Reverse-Phase HPLC Purifications

Crude peptides were dissolved in water, acidified to pH 2–3, and filtered (0.22μ). The clear solution was then loaded onto a C18 (9.4×250 mm) Zorbax column, and the peptide components eluted with flow rate of 5 mL/min using a shallow gradient of solvent B in solvent A (solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile). Fractions containing the desired purified peptides were identified by analytical LC and mass spectrometry, then combined and lyophilized.

7A.3 Synthesis of Unlabeled Ala–Ala–Ala Tripeptide

The Fmoc chemistry Solid-phase peptide synthesis (SPPS) protocol used was: scale 0.1 millimol H-Ala-O-2-Chlorotrityl-(S-DVB)resin; DMF washes: 10 sec flow, 1×1min batch; Fmoc-AA (0.55 mmol), 0.5M HBTU in DMF (1 mL, 0.5 mmol); 0.75 mmol DIEA (131 µL); activation 30 sec; coupling 30 min. N^{α}Fmoc removal: 20% v/v piperidine/DMF 2×5min batch treatments.

After final N^{α}Fmoc deprotection at the N-terminus, the peptide-resin was washed with DMF followed by DCM. Then the product peptide was cleaved from the H-Ala-O-2-Chlorotrityl-(S-DVB)resin by subjecting it to TFA/TIPS/water (95:2.5:2.5 v/v) conditions at ambient temperature. After 2.5 h, the cleavage mixture was collected in a 50 mL Falcon tube and the resin was washed once again with 1mL TFA. The filtrate was evaporated under the stream of Nitrogen, and the residue was triturated with 1:1 ice-cold ether and hexane (2×). The resultant white precipitate was then dissolved directly in water and purified using RP prep-HPLC on C18 column (9.4×250 mm) to give the target peptide in 38% yield (8.8 mg). LC-MS (ESI): [M+H]+ calculated: 231.1; found: 230.7±0.1 Da shown in Fig. 7A.1.

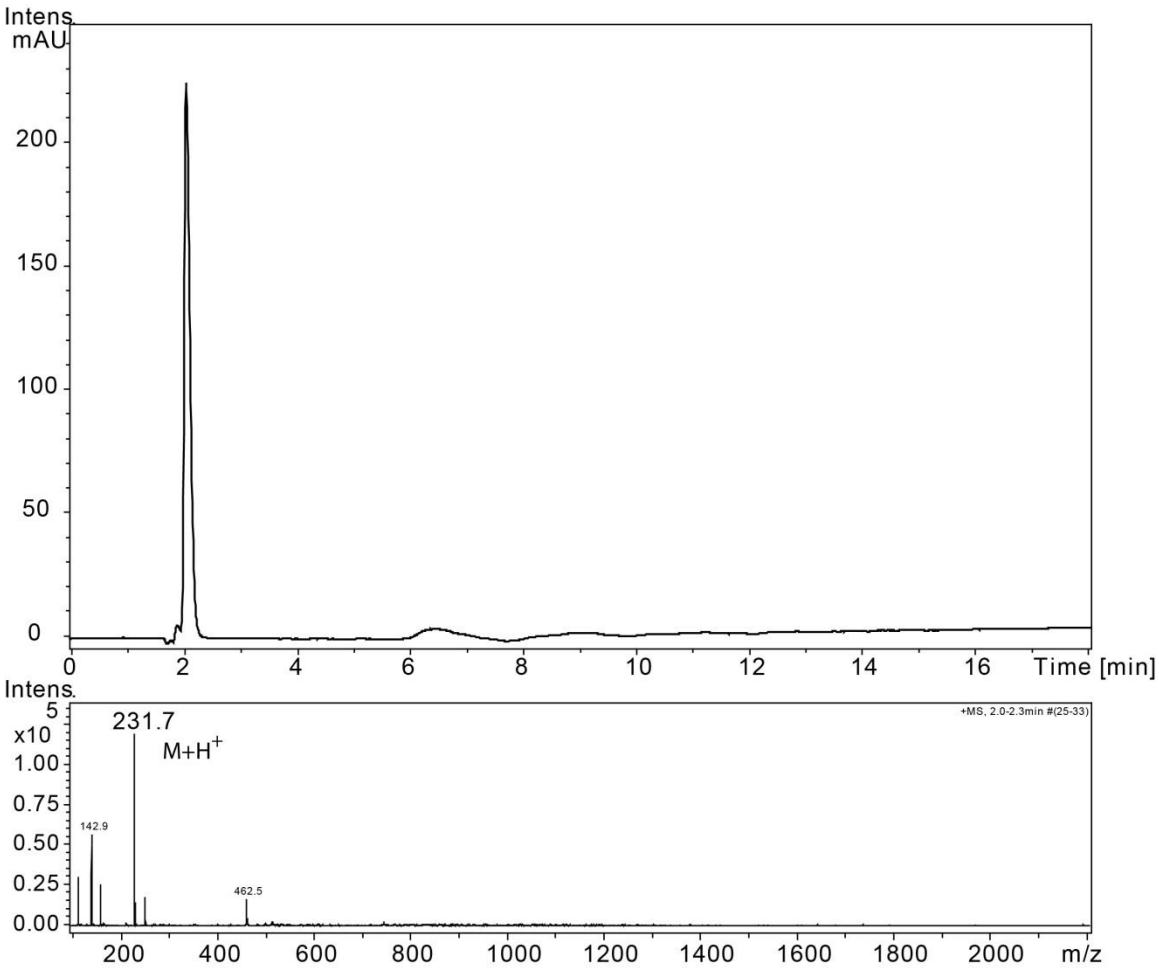

Figure 7A.1: Analytical LC-MS data for unlabeled Ala–Ala–Ala. (Top panel) Reverse phase HPLC – The chromatographic separations were performed on a C_8 (4.6x150 mm) column at 40 ^oC, using a linear gradient (0.5−8%) of solvent B in solvent A over 15 mins (solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile) at a flow rate of 1.0 mL/min with detection by UV absorption at 214 nm. (Bottom panel) Online ESI-MS spectra, taken across the whole of the UV peak.

7A.4 Fmoc-Protection of (1–^{13}C)Alanine

(1–^{13}C)Alanine (100 mg, 1.11 mmol) was dissolved in 5 mL water, followed by addition of 186.7 mg of $NaHCO_3$. This was mixed with 3mL of dioxane and cooled to 0–5^o C. Then, Fmoc-OSu was dissolved in 3mL dioxane and added dropwise to the above mixture. This was stirred at the same temperature for 2 hours and then at room temperature overnight. After evaporating all

the solvents in reduced pressure, the residue was taken in water (25 mL) and washed with EtOAc (2×5mL). Combined organic layers were once back extracted with satd. $NaHCO_3$ solution (1×5 mL) and discarded the organic layer. Combined aqueous layers were cooled to ~5 ^oC and acidified using 1N HCl to pH ~2. The acidified solution was extracted with ethyl acetate three times (20 mL each). The combined EtOAc layers were dried on Na_2SO_4 and evaporated under reduced pressure to give Fmoc-(1-^{13}C)alanine as a white solid (325 mg, 94%). ESI-MS (ESI): [M+Na] calculated: 335.1 Da; found: 335.1±0.1 Da shown in Fig. 7A.2.

7A.5 Synthesis of (1–^{13}C)Ala–Ala–Ala and Ala–(1-^{13}C)Ala–Ala

The Fmoc chemistry SPPS of (1–^{13}C)Ala–Ala–Ala and Ala–(1–^{13}C)Ala–Ala followed the same protocol as the unlabeled version. Note: Fmoc-(1–^{13}C)Ala was coupled by using a minimum amount of the isotope-labeled amino acid: Fmoc-(1–^{13}C)Ala (0.35 mmol), HBTU (0.3 mmol), and DIEA (0.6 mmol) for 1 h. The isolated yields for the labeled trialanines respectively are 9.0 mg (38.8%) and 9.1 mg (39.2%) shown in Figs. 7A.3–7A.4.

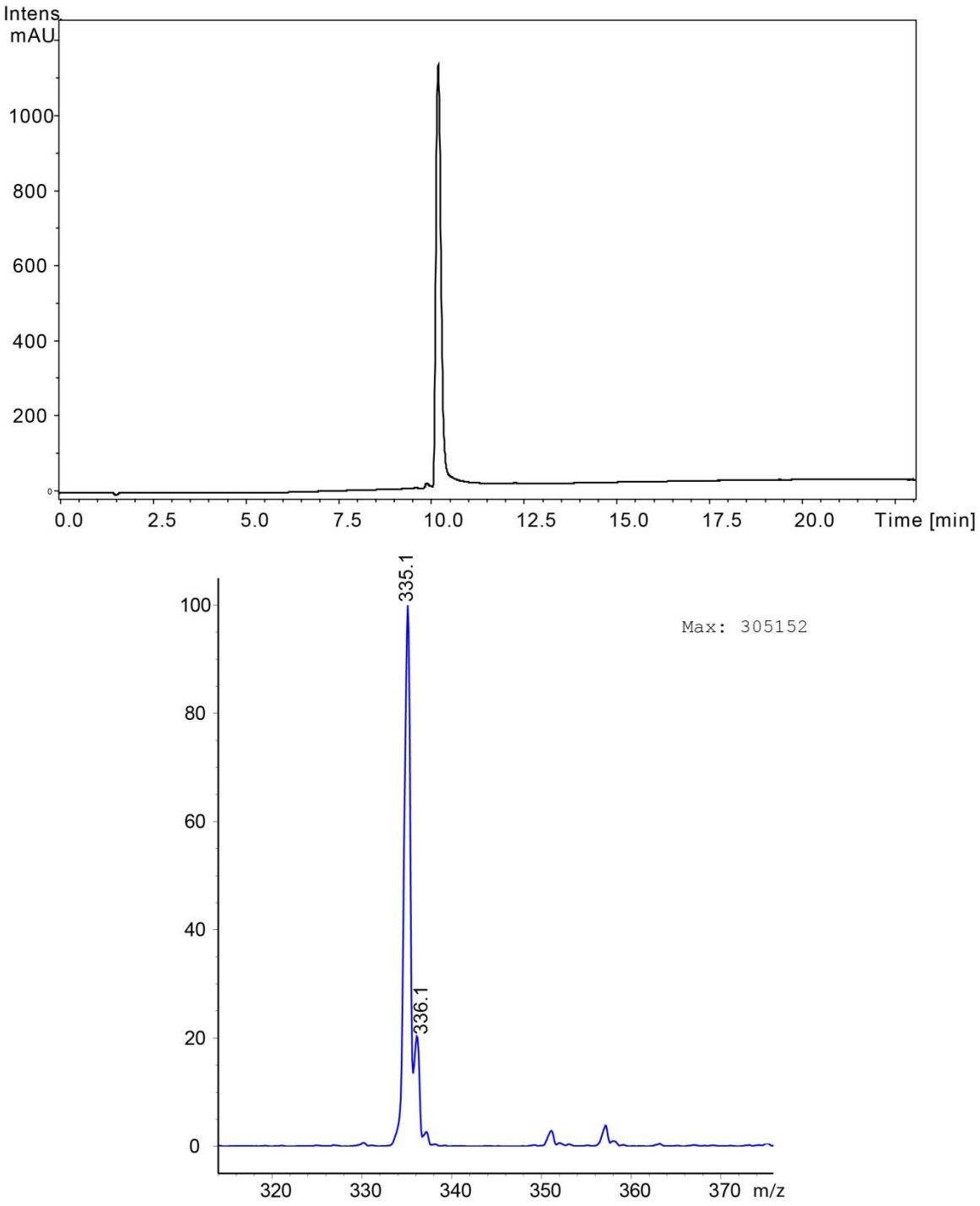

Figure 7A.2: (Top panel) Analytical HPLC data for Fmoc-(1–^{13}C)Ala. Reverse phase HPLC – The chromatographic separations were performed on a C_8 (4.6x150 mm) column at 40 ºC, using a linear gradient (5−65%) of solvent B in solvent A over 20 mins (solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile) at a flow rate of 1.0 mL/min with detection by UV absorption at 214 nm. (Bottom panel) ESI-MS spectra obtained by direct infusion into mass spectrometer.

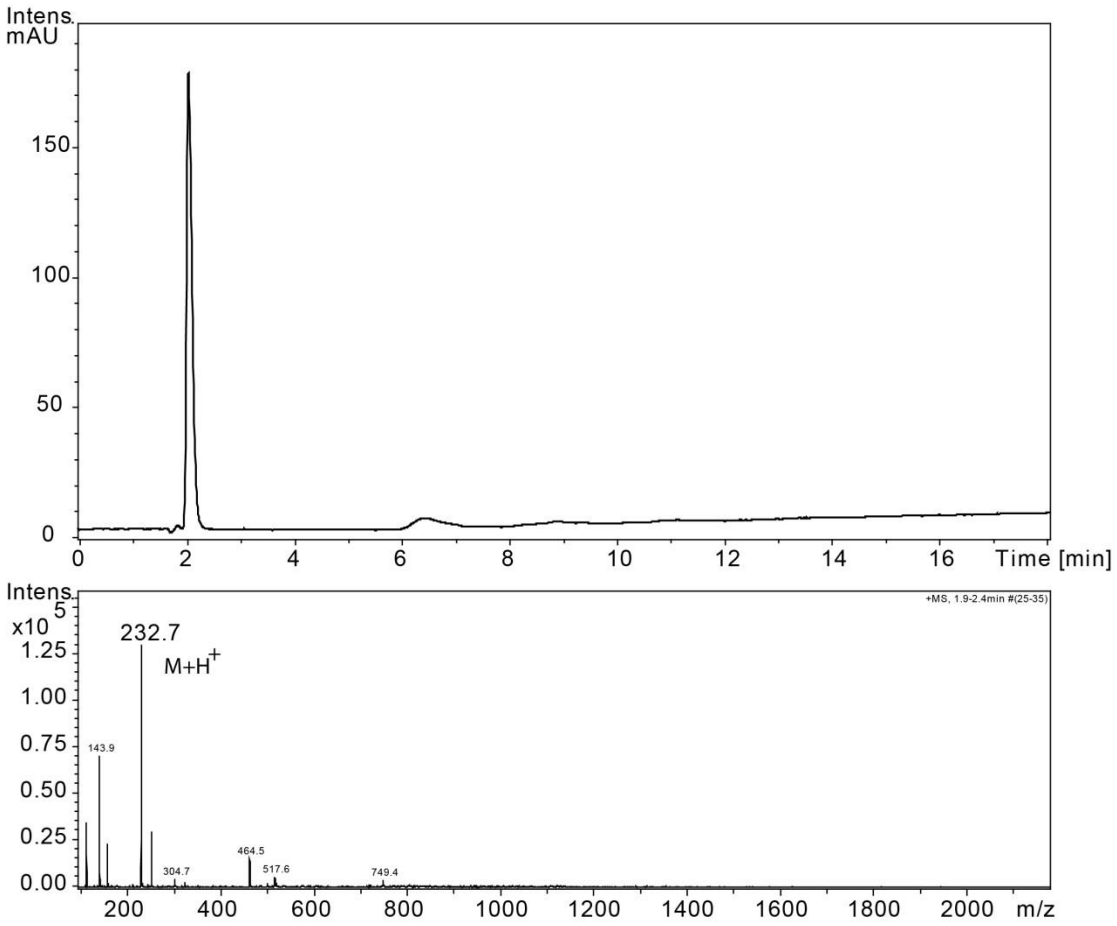

Figure 7A.3: Analytical LC-MS data for (1–^{13}C)Ala–Ala–Ala. (Top panel) Reverse phase HPLC – The chromatographic separations were performed on a C_8 (4.6×150 mm) column at 40 ^{o}C, using a linear gradient (0.5−8%) of solvent B in solvent A over 15 mins (solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile) at a flow rate of 1.0 mL/min with detection by UV absorption at 214 nm. (Bottom panel) Online ESI-MS spectra, taken across the whole of the UV peak. LC-MS (ESI): [M+H]+ calculated: 232.1; found: 231.7±0.1 Da.

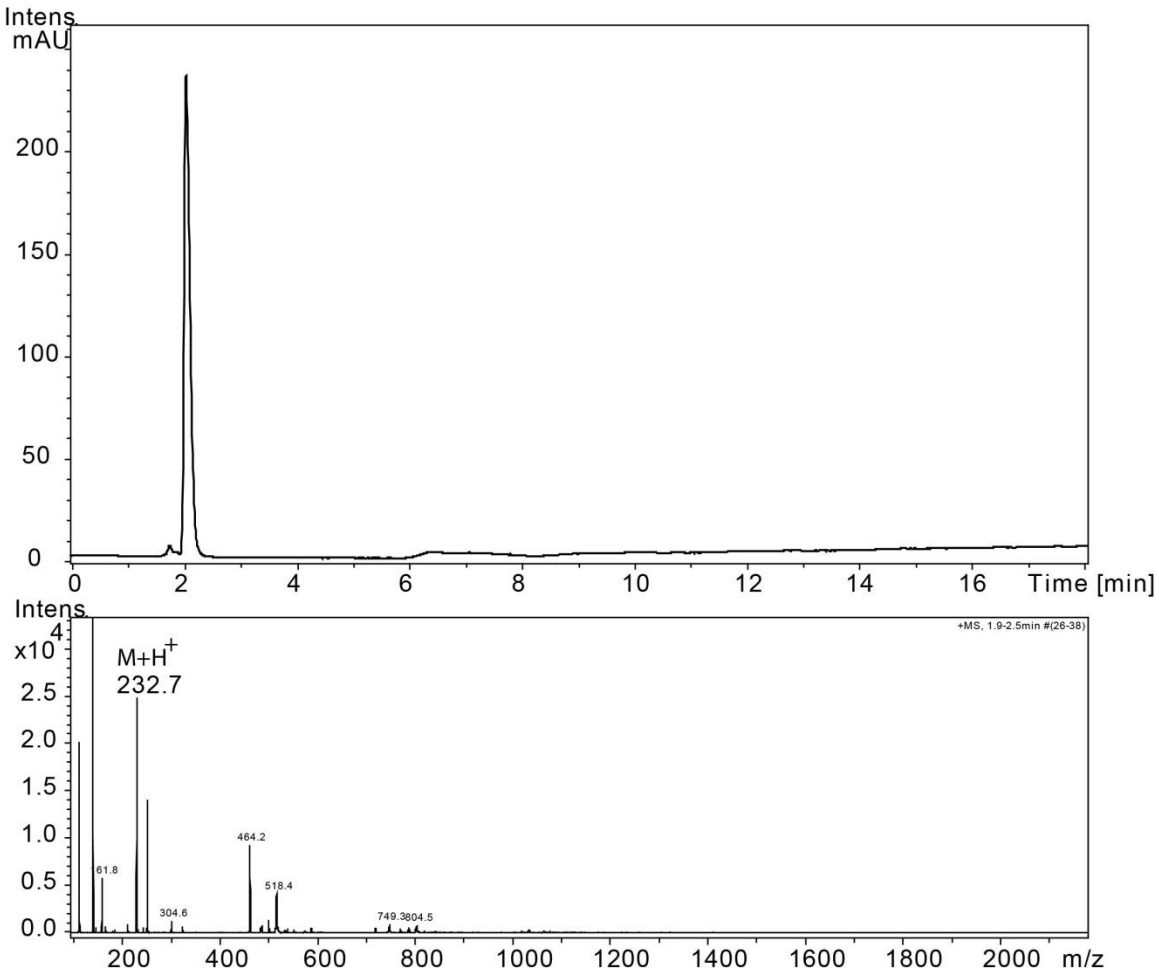

Figure 7A.4: Analytical LC-MS data for Ala–(1–^{13}C)Ala–Ala. (Top panel) Reverse phase HPLC – The chromatographic separations were performed on a C_8 (4.6×150 mm) column at 40 ^{o}C, using a linear gradient (0.5−8%) of solvent B in solvent A over 15 mins (solvent A = 0.1% TFA in water, solvent B = 0.08% TFA in acetonitrile) at a flow rate of 1.0 mL/min with detection by UV absorption at 214 nm. (Bottom panel) Online ESI-MS spectra, taken across the whole of the UV peak. LC-MS (ESI): [M+H]+ calculated: 232.1; found: 231.7±0.1 Da.

Appendix 7B: Additional Descriptions of Bayesian Ensemble Refinement

The work presented in this chapter has been published and is reprinted with permission from:

Feng, C.-J.; Dhayalan, B.; Tokmakoff, A. Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophysical Journal* **2018**, 114 (12), 2820–2832.

7B.1 Effect of θ value on the ensemble refinement

Based on the Bayesian ensemble refinement approach,[92] the likelihood function $p(\text{data} \mid \boldsymbol{x})$, describing the quality of experimental IR spectra reproduction by $\boldsymbol{x}$ is formulated as follows (Eq. (7.4)).

$$p(\text{data} \mid \boldsymbol{x}) = \exp\left(-\sum_i \frac{1-s_i}{2\theta\sigma_i^2}\right) = \exp(-\chi^2(\theta)) \tag{7.6}$$

$s_i(\boldsymbol{x})$ is the spectral overlap between the simulated spectra from $\boldsymbol{x}$ and the spectrum from experiment i, defined in Eq. (7.5). θ is an adjustable parameter expressing the level of confidence in the model, and $1/\theta$ is equivalent to the Lagrange multiplier in the maximum entropy formalism.

Optimal value of θ can be determined by plotting the relative entropy $-S_{\text{KL}}(\theta)$ against $\chi^2(\theta)$ (Fig. 7B.1), with the definition of the relative entropy as

$$-S_{\text{KL}}(\theta) = \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \text{data}) \ln\left(\frac{p(\boldsymbol{x} \mid \text{data})}{p(\boldsymbol{x})}\right) \tag{7.7}$$

The relative entropy increases if experimental data provide information to refine the posterior distribution. Generally, both $S_{\text{KL}}(\theta)$ and $\chi^2(\theta)$ increase with decreasing θ, although the slope varies. In the steep regime of Fig. 7B.1 (large θ), a small change of θ leads to a large decrease of $-S_{\text{KL}}(\theta)$, indicating that increasing θ refines the posterior distribution to agree better with experiments. Conversely in the flat regime (small θ), the distribution changes little with increasing θ, but the measure of error $\chi^2(\theta)$ increases dramatically, implying that the distribution is very sensitive to the

experimental data including noise. Thus the plot has a characteristic "L" shape, and the optimal θ is found at the point of maximum curvature.[92, 94] The maximum curvature and the corresponding optimal θ throughout this study are derived from the cubic spline fitting to the L-curve.

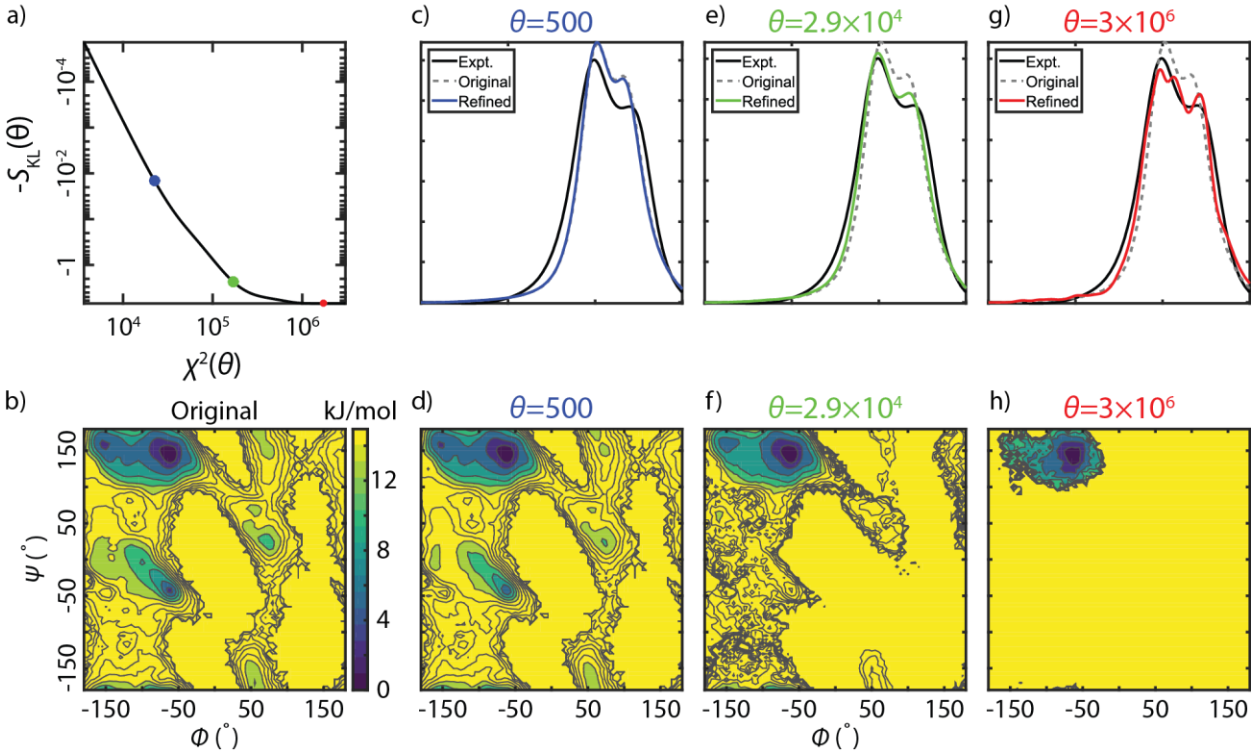

Figure 7B.1: The effect of different θ values on the ensemble refinement. (a) L-curve of the C36m TIP3P trajectory against the UL FTIR spectrum (b) PMF before Bayesian ensemble refinement. (c–d) FTIR spectrum and the refined PMF at θ = 500 (e–f) FTIR spectrum and the refined PMF at $\theta = 2.9\times10^4$, the point of maximum curvature in the L-curve. (g–h) FTIR spectrum and the refined PMF at $\theta = 3\times10^6$. Each contour line in the PMFs is spaced by k_BT at 300 K up to 15 k_BT.

7B.2 Frequency Corrections for Ensemble Refinement

The quality of ensemble refinement against IR spectroscopy may be significantly influenced by using different water models which vary in their potential and electrostatics, and using a $CONH_2$ terminal group in the AMBER force fields instead of the COOH group. Our 1-site

amide I electric field map (1F) was originally parametrized using TIP3P atomic charges on the SPC/E water model and CHARMM FF charges.[53] However, we found that simulated spectra using SPC/E water have a systematic red shift relative to the experiments and the TIP3P water model in AAA (Fig. 7B.2), and it enhances the α_R population in the refinement against FTIR spectra (Figs. 7B.3), which may cause false positive interpretations. Since the average frequency of α_R conformers is the highest (Table 7B.1, Fig. 7.7 and 7B.2), an increased population of α_R compensates for the SPC/E water red shift. In contrast, TIP3P water model from all CHARMM force fields give more consistent refinement populations and narrower range of uncertainty (Table 7.). Therefore, we choose frequencies obtained from CHARMM TIP3P trajectories as our reference point for ensemble refinements.

To account for the frequency bias from water models and the $CONH_2$ group, we calculated the frequency shift of each amide site due to different water models and $CONH_2$ group, with the results summarized in Table 7B.2. We found both SPC/E and TIP3Pfb water models exhibit red shift compared to TIP3P water. The corresponding frequency corrections are applied prior to the spectral simulations to compensate the frequency bias. One example of ensemble refinement with frequency correction is shown in Fig. 7B.3, indicating that the frequency correction can help reduce the uncertainty of populations from the systematic bias, and avoid false positive interpretations. However, even with the frequency correction, the SPC/E water and TIP3Pfb water generally gives larger uncertainty shown in Table 7.2, suggesting that this simple frequency correction cannot fully rescue the structural difference between water models.

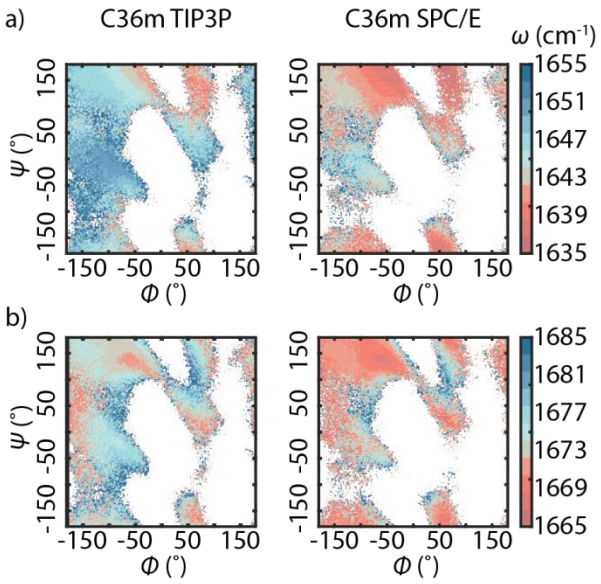

Figure 7B.2: (a–b) Mean peak (exciton) frequency distributions of UL AAA from C36m TIP3P (left) and C36m SPC/E (right). Note that the TIP3P atomic charges are already applied to C36m SPC/E trajectory.

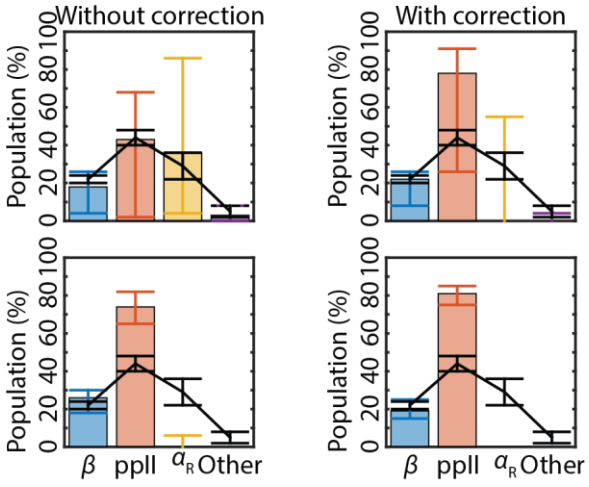

Figure 7B.3: Population of representative conformers before refinement (black line), and after refinement (colored bars) of C27 SPC/E trajectory without frequency correction (left), and with frequency correction (right). Top row and bottom row show the refinements against the experimental FTIR spectra, and the full set of spectra, respectively. Black error bars represent the uncertainty of original distribution estimated from block averaging. Colored error bars reflect the uncertainty of spectroscopic map, determined by the maximum and the minimum populations with systematic frequency error ranging from -4 to 4 cm^{-1}.

	β		ppII		α_R	
	A1	A2	A1	A2	A1	A2
UL	1667	1648	1661	1645	1667	1647
^{13}C–A1	1620	1648	1623	1643	1627	1650
^{13}C–A2	1661	1608	1663	1604	1667	1610

Table 7B.1: Peak frequencies (in cm^{-1}) of the conformer FTIR spectra shown in Fig. 7.7. Frequencies were determined from FTIR spectra and second derivative of the FTIR spectra.

	TIP3P/COOH	SPC/E	TIP3Pfb	$CONH_2$
A1	0.00	-2.06	-9.16	-0.11
A2	0.00	-4.62	-8.07	1.53

Table 7B.2: Average site frequency shift of various water models relative to the average of C27 TIP3P, C36 TIP3P, and C36m TIP3P trajectories for post-frequency correction. Average frequency shifts of different water models are computed by switching off all of the FF charges except for water molecules while the average frequency shift of $CONH_2$ relative to COOH group is computed by switching off all of the FF charges except for CONH2 and COOH.

7B.3 Effect of the Choice of Electrostatic Variables in frequency maps on Ensemble Refinement

The method used for mapping the local electrostatic variables around the amide group calculated from the FF into an amide I frequency may also influence the quality of ensemble refinement, including number of sites used to sample electrostatics, and the electrostatic variables used for translating into frequency. Our empirical 4P map can examine how different mapping impacts the ensemble refinements. We found the 4P map consistently gives the same refined ensemble populations as the 1F map in C36m TIP3P ensemble but with larger uncertainty (Fig. 7B.4), indicating that different choice of electrostatic variables may have impact on the refinement, but it would not be as dominant as the frequency systematic bias when applying different water models on the refinement (Fig. 7.12).

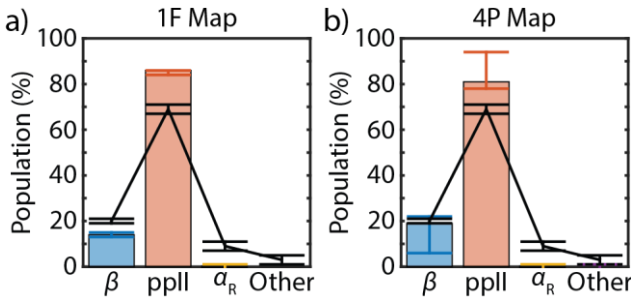

Figure 7B.4: Population of representative conformers before refinement (black line), and after refinement (colored bars) from C36m TIP3P trajectory using (a) 1F map (b) 4P map. Black error bars represent the uncertainty of the original distribution estimated from block averaging method. Colored error bars reflect the uncertainty of spectroscopic map, determined by the maximum and the minimum populations with systematic map error ranging from -4 to 4 cm^{-1}.

7B.3 Correlation between θ and the Accuracy of Force Fields and Water Models

The θ parameter, which is described in the previous section and in the manuscript, is an adjustable parameter expressing the level of confidence in the model, and $1/\theta$ is equivalent to the Lagrange multiplier in the maximum entropy formalism. Since θ is optimally determined by the L-curve method for each prior distribution drawn from FFs and water models, one of the interesting questions arising from θ is if the optimal values of θ reflect the relative accuracy of these FFs and water models.

To address this question, Fig. 7B.5 shows the optimal θ values and the difference of ppII population percentage between the original population and the average refined population among all of the FFs and water models, with the corresponding correlation coefficients listed in Table 7B.3. We found that there is a weak correlation among all of the FFs and water models (Fig. 7B.5a). However, the correlations within the CHARMM FFs and OPLS–AA FFs are much stronger while there is very weak anti-correlation in the AMBER FFs. Two sources of inducing different

correlation behaviors are different terminal $CONH_2$ group used in the AMBER FFs, and the frequency corrections for ensemble refinements, which both try to alleviate the systematic frequency error in the spectroscopic map when comparing different FF/water models. Fig. 7B.5 suggests that the optimal value of θ may indicate the relative accuracy of the employed FFs/water models, but it is also sensitive to how we treat the systematic frequency error.

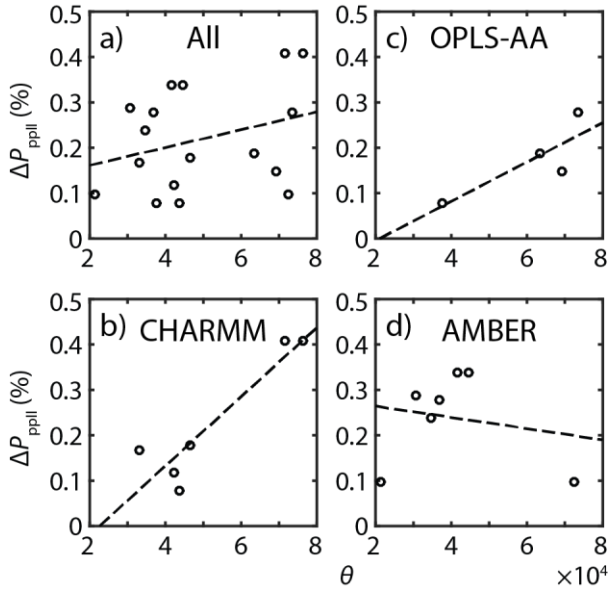

Figure 7B.5: Correlation between θ and the population percentage change of the ppII conformer between the original population and the average ppII population among all FFs/water models, with the data drawn from (a) all FFs and water models, (b) all CHARMM FFs and water models, (c) all OPLS-AA FFs and water models, and (d) all AMBER FFs and water models. Black open circles represent the raw data points while the dashed black lines represent the corresponding linear regression lines.

	All	CHARMM	OPLS-AA	AMBER
R	0.31	0.90	0.84	-0.19
R^2	0.10	0.81	0.70	0.04

Table 7B.3: Correlation coefficients between θ and the population percentage change of the ppII conformer drawn from Fig. 7B.5.

Appendix 7C: Vibrational Exciton Hamiltonian of Two Level Systems and Estimation of Coupling Uncertainty on Amide I FTIR Spectra

The work presented in this chapter has been published and is reprinted with permission from:

Feng, C.-J.; Dhayalan, B.; Tokmakoff, A. Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophysical Journal* **2018**, 114 (12), 2820–2832.

7C.1 Vibrational Exciton Hamiltonian of Two Level Systems

To facilitate understandings of the FTIR amide I spectra and correlate structures to spectra using spectroscopic maps described in the main manuscript, vibrational exciton Hamiltonian accounting for 0–1 transitions can be written as

$$\hat{H}_{\mathrm{AmI}}^{1\mathrm{Q}} = \sum_{n=1}^{2} \omega_n \,|\, n\rangle\langle n \,| + \sum_{n=1,m\neq n}^{2} J_{mn} \,|\, m\rangle\langle n \,| \tag{7.8}$$

The indices *m* and *n* indicate the site of vibration excitation. $|\, n\rangle$ refers to a single excitation of site *n* with excitation energy ω_n, and J_{mn} is the coupling between the sites *n* and *m*. For AAA, J_{mn} is the coupling between two amide I modes. The coupled two-level system Hamiltonian of AAA has the analytical solution with eigenvalues $\omega_{\pm}$ as

$$\begin{aligned} \omega_{\pm} &= \frac{\omega_1 + \omega_2}{2} \pm \frac{1}{2}\sqrt{\left(\omega_1 - \omega_2\right)^2 + 4J_{12}^2} \\ &= \bar{\omega} \pm \frac{1}{2}\sqrt{\omega_{12}^2 + 4J_{12}^2} \end{aligned} \tag{7.9}$$

Assuming the energy gap between the two sites $\omega_{12} \geq 0$, the corresponding eigenvectors can be transformed from the site basis by Eqs. (7.10) and (7.11).

$$\begin{pmatrix} |+\rangle \\ |-\rangle \end{pmatrix} = \begin{pmatrix} C_1^{+} & C_2^{+} \\ C_1^{-} & C_2^{-} \end{pmatrix} \begin{pmatrix} |1\rangle \\ |2\rangle \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} |1\rangle \\ |2\rangle \end{pmatrix} \tag{7.10}$$

$$\tan 2\theta = \frac{J_{12}}{\omega_{12}}, 0 \le \theta \le \frac{\pi}{4} \tag{7.11}$$

Eq. (7.10) shows that the exciton modes are delocalized between the two sites. The corresponding exciton transition dipole moment describing intensity of observed peak in FTIR spectrum can be written as $\vec{\mu}_{\pm 0} = \sum_{n=1}^{2} C_n^{\pm} \vec{\mu}_{n0}$. Assuming the strength of the transition dipole moment μ is the same for both sites, which is also applied in the spectral simulations in this study, the corresponding peak intensity of the FTIR spectra would be

$$\left|\vec{\mu}_{\pm 0}\right|^2 = \mu^2 \left(1 + 2C_1^{\pm} C_2^{\pm} \cos\Theta_{12}\right) = \mu^2 \left[1 \pm \frac{J_{12}^2}{\sqrt{\omega_{12}^2 + J_{12}^2}} \cos\Theta_{12}\right] \tag{7.12}$$

Therefore, the intensities of the FTIR spectra depend on ω_{12}, J_{12}, and the strength of the transition dipole moment and angle between the two dipoles Θ_{12}. When one of the amide group is 1–^{13}C labeled to have isotope frequency shift Δ, the corresponding average frequency would be $\bar{\omega}_{^{13}\mathrm{C}} = \bar{\omega} + \Delta/2$, and we can determine the isotopic frequency shift by computing the average peak frequency from UL FTIR spectrum and ^{13}C–label FTIR spectrum. In principle, we can also determine J_{12} from the IR experiments,[60-61] but the coupling value depends on peptide conformations, which is non-trivial to be extracted by this approach. The two-quantum Hamiltonian describing the 1–2 transition is constructed by weakly anharmonic exciton model.[60]

7C.2 Estimation of Coupling Uncertainty on Amide I FTIR Spectra

Errors due to nearest-neighbor coupling map are difficult to address because there are no experimental standards for evaluating the quality of coupling maps. However, at the DFT level calculations, the unsigned coupling map uncertainty is 1.7 cm^{-1}.[99] Based on this value, we can estimate the uncertainty of amide I frequency splitting and the peak intensity in the FTIR spectra. The uncertainty of the exciton frequency based on Eq. (7.9) can be written as

$$\delta\omega = \frac{\sqrt{\omega_{12}^2 + 4(2\sigma_J + J)^2} - \sqrt{\omega_{12}^2 + 4J^2}}{2} \tag{7.13}$$

with $\omega_{12} = 16.5\,\mathrm{cm}^{-1}$ computed from the C36m TIP3P trajectory, and the coupling value estimated as 8 cm^{-1}, the upper bound taken from the 2D IR cross peak in the manuscript. By taking $2\sigma_J = 3.4$ cm^{-1}, the uncertainty of $\delta\omega$ gives 2.6 cm^{-1}, that is smaller than the confidence interval (4 cm^{-1}) of site frequency uncertainty, suggesting that the coupling uncertainty may affect less than the uncertainty of the site frequency. However, we cannot conclude that it is not significant for the ensemble refinements.

A similar strategy can be applied to estimate the peak intensity uncertainty in the FTIR spectra. Given Eqs. (7.10)–(7.12), we can derive the relative uncertainty of FTIR peak intensity $\delta\left|\vec{\mu}_{\pm 0}\right|^2 \big/ \left|\vec{\mu}_{\pm 0}\right|^2$ as a function of the uncertainty of the coefficients $C_n^{\pm}$ in Eq. (7.12), which intimately depend on σ_J.

$$\frac{\delta\left|\vec{\mu}_{\pm 0}\right|^2}{\left|\vec{\mu}_{\pm 0}\right|^2} = \frac{2\left(\delta C_1^{\pm} C_2^{\pm} + C_1^{\pm}\delta C_2^{\pm} + \delta C_1^{\pm}\delta C_2^{\pm}\right)\cos\Theta_{12}}{1 + 2C_1^{\pm}C_2^{\pm}\cos\Theta_{12}} \tag{7.14}$$

$\delta\left|\vec{\mu}_{\pm 0}\right|^2$ corresponds to the difference of peak intensity with and without $2\sigma_J$. Taking ppII conformer as an example with representative Θ_{12} angle of ~110°, the relative uncertainties of

$\delta|\vec{\mu}_{+0}|^2/|\vec{\mu}_{+0}|^2$ and $\delta|\vec{\mu}_{-0}|^2/|\vec{\mu}_{-0}|^2$ are +6% and -7%, respectively. Comparing to our conformer spectra shown in Fig. 7.7, the uncertainty of the peak intensity may not dramatically distort the entire spectral lineshape to be similar as β conformer spectrum or even α_R conformer spectrum. To conclude, the effect of coupling on vibrational frequency dominates the quality of ensemble refinement over its influence on peak intensity.

Chapter 8

Computational IR Spectroscopy of Insulin Dimer and Conformational Heterogeneity

8.1 Abstract

We have investigated the structure and conformational dynamics of insulin dimer using a Markov state model (MSM) built from extensive unbiased atomistic MD simulations, and performed infrared spectral simulations of the insulin MSM to describe how structural variation within the dimer can be experimentally resolved. Our model reveals two significant conformations to the dimer: a dominant native state consistent with other experimental structures of the dimer, and a twisted state with a structure that appears to reflect a ~55° clockwise rotation of the native dimer interface. The twisted state primarily influences the contacts involving the C-terminus of insulin's B chain, shifting the registry of its intermolecular hydrogen bonds and reorganizing its sidechain packing. The MSM kinetics predict that these configurations exchange on a 14 μs timescale, largely passing through two Markov states with a solvated dimer interface. Computational amide I spectroscopy of site-specifically $^{13}C^{18}O$ labeled amides indicates that the native and twisted conformation can be distinguished through a series of single and dual labels involving the B24F, B25F, and B26Y residues. Additional structural heterogeneity and disorder is observed within the native and twisted states, and amide I spectroscopy can also be used to gain

insight into this variation. This study will provide important interpretive tools for IR spectroscopic investigations of insulin structure, and transient IR kinetics experiments studying the conformational dynamics of insulin dimer.

8.2 Introduction

Protein structure-function relationships have been rethought over the past two decades to account for the functional role of conformational disorder, which appears in intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs).[1] Proteins with structural heterogeneity and conformational disorder contain a variety of thermally accessible conformers undergoing rapid structural fluctuations or activated interconversion kinetics between free-energy basins on a complex free energy landscape.[2] Such IDPs and proteins with IDRs have been observed to be involved in many biological processes including regulation, signaling, and coupled-folding and binding to functional partners.[1, 3]

Structural characterization of IDPs and proteins with IDRs requires an ensemble description, which creates numerous experimental challenges. Ensemble-structure determination is naturally an ill-posed problem in which degrees of freedom in relevant conformational states far exceed limited number of measurements and information content of experiments.[4] Also, fast conformational fluctuations and activated conformational dynamics cannot be decoupled,[5] meaning that conformational characterization inherently requires experimental probes with both high structural and temporal resolutions. Traditional structural tools are often limited by their intrinsic time resolution, which prohibits one from accessing conformational fluctuations and interconversion of conformers with time scales spanning many decades from picoseconds to microseconds.[6-9] NMR spectroscopy that measures chemical shifts and J couplings is limited by

the coalescence time scale of ms such that faster conformational dynamics are averaged whereas relaxation experiments provide indirect structural information.[10] Optical spectroscopies do carry the advantage of femtosecond time-scales for their light-matter interaction, but in most cases have little structural information content. Advances are being made with IR and 2D IR spectroscopies which probe structure sensitive molecular vibrations with fs–ps time resolution, and thereby have the capability of structural characterization on a peptide or protein structure that is essentially static.[11-21]

Despite intensive experimental advances to investigating protein structures, all experimental methods face challenges of structural interpretation, which inevitably require structure-based models. Molecular dynamics (MD) simulation offers atomistic descriptions of protein structures and motions. Also, recent efforts of force field (FF) developments have improved the ability to study disordered proteins.[22] Clustering methods and kinetic frameworks have been developed to analyze conformational dynamics such as time-structure independent component analysis (tICA)[23-26] and Markov State Models (MSMs),[27-28] which provides a natural structural basis for computing experimental properties and studying interconversion kinetics. Computational advances have laid foundations of rationalizing experimental evidence, predicting experimental outcomes, and even helping design suitable experiments.

Direct quantitative comparison of atomistic protein structures and IR experiments is now possible using computational amide I spectroscopy.[29] Amide I IR spectroscopy, which probes the C=O stretching vibration of the polypeptide backbone in the 1600–1700 cm^{-1} frequency range, can be used to interrogate local hydrogen bonding contacts to the carbonyl oxygen and structure-sensitive couplings between different amide carbonyls. This method can be used to predict traditional IR absorption and 2D IR spectra for simulated conformational distributions drawn from

MD trajectories or MSMs, providing a unique route to investigate structure-spectrum correlations. Specifically, amide I vibrational frequencies can be predicted to high accuracy using spectroscopic "maps" that relate frequency to the local electrostatic potential or electric field from MD force fields at specific amide carbonyl sites.[30-37] Similarly, maps for vibrational coupling between different amide I vibrations are used to calculate the interaction of multiple backbone amide groups.[38-41] These maps have reached the point of predicting amide I spectroscopic observables to a high level of accuracy with 2 cm^{-1} frequency uncertainty and provided a direct way to structurally interpret experimental IR spectra.[36] This approach has additional power when interpreting site-specific isotope-edited IR spectroscopy using computational spectroscopy with MD simulations and MSMs. This approach has been used for a number of peptides and proteins, such as spectroscopic investigation of various secondary structures,[42-44] conformational characterization of TrpZip2,[14] disordered peptides and amyloid fibrils,[20, 45] structural disorder of NTL9[15] and CD3ζ transmembrane domain,[46] and investigating permeation mechanism of potassium ion in KcsA.[18, 47-48]

In this study, we investigate the structural heterogeneity of insulin homodimer using a MSM built on MD simulations, and show how this structural variation can be interrogated using amide I IR spectroscopy of site-specifically isotope-labelled amide carbonyls. Insulin dimer dissociation is a necessary step prior to binding to insulin receptor, which is regarded as a coupled unfolding and unbinding process studied both computationally and experimentally.[49-53] Insulin monomer is a 51-residue peptide composed of two disulfide bonded chains with 21 residues on chain A and 30 residues on chain B. It has three α-helical segments, a β-turn located from B20Gly to B23Gly, and the B-chain C-terminus ranging from B24Phe–B30Thr is known to be disordered with the extent depending on mutations and solution environment.[54-61] This intrinsically disordered

region folds and binds into a well-defined inter-monomer β-sheet in the native dimer structure stabilized by hydrogen bonding and sidechain packing of the aromatic triplet of B24Phe, B25Phe, and B26Tyr.[49, 61-63] This B-chain C-terminus is also involved in the insulin receptor binding, exhibiting detachment of the B-chain C-terminus, a significant dihedral rotation on B24Phe, and hinging motion upon recognition with the insulin receptor,[64-67] which may share similar conformational transitions as in the dimerization.[52]

Insulin dimer has also become a useful model system for investigating coupled folding and binding dynamics. Computational insulin dimerization studies have focused almost entirely on the conformational characterization in the monomeric state, with the current viewpoint that the dimer structure resembles published crystal structures. However, recent simulation study on mutant dimer, showed that mutation of B24Phe to Gly resulted in additional dimer conformations including strongly interacting dimer and weakly interacting dimer, which involves conformational change between B10His and B13Glu, and increased solvation of the dimer interface.[68] A detailed description of the dimer conformational distribution still requires investigation, in particular accounting for changes of solvation environment such as pH, ionic strength, and temperature that can mediate dimer conformational changes and in turn dynamics of dimer dissociation and association.[69-70]

Here, we present a computational study to characterize equilibrium conformational ensemble of insulin dimer is aqueous solution using MSMs built off extensive MD simulations, and computational amide I spectroscopy to predict how conformational substates can be observed in experiments. The MSM of insulin dimer provides an all-atom, high-resolution structural basis with associated interconversion rates to compare to IR structural and kinetics measurements. In addition to structures that are similar to previous experimental results, the MSM reveals a second

major conformational state, the twisted dimer state with register-shifted β-strands, that has not been observed experimentally and computationally. Conformational exchange between twisted state and native state is predicted to occur on a 10-20 μs time scale. With the help of computational amide I spectroscopy, we proposed site-specific isotope labels that can effectively distinguish these two major conformational states, as well as intermediate configurations visited as they interconvert. This study forms a computational basis for experimental investigation of insulin dimer structures and kinetic experiments that can resolve conformational transition between these states.

8.3 Methods

8.3.1 Molecular Dynamics Simulations

MD sampling was initiated from the native crystal structure of wild-type human insulin dimer (PDB: 3W7Y) using the AMBER99sb-ildn force field (FF) and TIP3P water model.[71-72] To match the preferred low pH conditions for infrared spectroscopy that increase solubility, titratable side chains including HIS, GLU, and N-terminal NH_3^+ were protonated. The COOH group was replaced by $CONH_2$ group because the AMBER FF does not have parameters for COOH. The protonated insulin dimer was solvated in a cubic water box, having 9000 water molecules and additional Na^+ and Cl^- ions to represent the ionic strength of 0.15 M.

Potential energy minimization was performed to ensure a reasonable starting structure for further temperature equilibration. To equilibrate temperature, the system was then gradually heated to 310 K in the NVT ensemble for 20 ps. Subsequent density equilibration was performed in the NPT ensemble at 310 K and 1 atm for 100 ps. Production runs were performed in the NPT ensemble at 310 K and 1 atm using OpenMM on Folding@home.[73-74] Several rounds of MD

simulations were performed, with starting structures reseeded from previous rounds to accumulate more statistics on rare conformational transitions and to explore new configurations faster than a direct single MD simulation. The aggregate sampling of insulin dimer consisted of 409 MD trajectories with the total sampling of 1.71 ms.

8.3.2 Construction of Dimer Markov State Model

The Markov State Model (MSM) of insulin dimer was constructed using MSMBuilder.[75] For clustering structures from MD sampling into conformational states, the collective variables (CVs) were inter-atomic distances between 102 C_α atoms in the dimer, resulting in 102×101/2 = 5151 C_α–C_α pairs. To ensure robustness of the results to outliers in the data, the quantile range was used instead of standard deviations. The distances were scaled with RobustScaler module to remove the medians and to the quantile range between the 25th quartile and the 75th quartile. Centering and scaling were performed independently on each pair. After scaling, time-structure independent components analysis (tICA) was applied on the time series data of all C_α–C_α pairs with the lag time of 125 ns.[25-26] The first 20 independent components (tICs) from tICA were selected as a subspace for subsequent k-medoids clustering, resulting in k = 100 states.[76-77] These states were used as the basis for building the MSM with the lag time of 150 ns, with structures in each state drawn from the original MD sampling.

8.3.3 Visualization of the Markov State Model network

To visualize the MSM, network graphs were generated using Gephi 0.9.2,[78] and the corresponding network layout was produced using the ForceAtlas algorithm.[78] Markov states are treated as nodes with the radius proportional to its equilibrium population. Each node is connected

with edges whose thicknesses reflect the sum of forward and backward transition probabilities between nodes. The ForceAtlas algorithm treats this network as a coupled spring-mass system, in which the spring constants correspond to the sum of the transition probabilities. Repulsive forces are added between each node to avoid spatial overlap of the nodes. The algorithm minimized the overall energy of the system by rearranging the layout of the network such that states stay in proximity to each other when they interconvert rapidly.

8.3.4 Simulations of Amide I Spectra

Amide I spectral simulations were performed using a mixed quantum-classical model that builds on atomistic structures drawn from classical MD simulations.[29] Simulations of explicitly solvated structures for all 100 configurations within each of the 100 Markov states were performed using GROMACS 4.6.7,[79] using the AMBER99sb-ildn force field and TIP3P water.[71, 80] For each configuration, the solvent and ions were equilibrated around the position-restrained peptide for 100 ps at 300 K using the Berendsen thermostat.[81] Subsequent 1 ns production runs on the unrestrained protein were performed using the Nosé−Hoover thermostat under NVT conditions with a 1 fs integration step, and 20 fs/frame sampling rate for spectral simulations.[82-83] The final spectra for each Markov state was obtained by averaging the calculated spectra over all 100 initial configurations.

IR vibrational spectra were calculated from the Fourier transform of a transition dipole time-correlation function obtained from a mixed quantum-classical model implemented in the freely available g_amide and g_spec programs.[84-85] The model treats the amide I vibrations as a set of coupled oscillators assigned to each amide group of the backbone.[29, 86] Collective electrostatic variables are used to translate, or “map”, a series of instantaneous structures along a

trajectory onto a time-dependent Hamiltonian and transition dipole moment for the amide I vibrations. The amide oscillators, or "sites", are identified by atomic positions of the CONH peptide backbone linkages. The vibrational frequency of each site is generated from a 4–site potential map (4P) which evaluates the electrostatic potential at the C, O, N and H positions and maps it to a vibrational frequency. The 4P map used in this study, 4PN-150, has a frequency prediction accuracy of σ=2.25 cm^{-1}.[36] In addition to the frequency of each site, vibrational coupling between amide I oscillators is obtained using coupling maps for mechanical through-bond coupling and electrostatic through-space coupling.[40] Additional details are provided in the Chapter 3.

Amide I transition dipole correlation functions were calculated using a dynamic wavefunction propagation method,[87] using a Trotter expansion to reduce computation time described in detail in Chapter 3.[88-89] The window time for calculations was set to 11 ps, equivalent to 3 cm^{-1} frequency resolution. The model includes a 1.0 ps vibrational lifetime for amide I modes, to match experiments.[86, 89] The isotope frequency shift of $^{13}C^{18}O$ isotope labels relative to $^{12}C^{16}O$ is set to –65 cm^{-1}.[90-91]

8.4 Results

8.4.1 Insulin Dimer MSM

We performed 1.7 ms of aggregated unbiased equilibrium MD sampling of insulin dimer and constructed an all-atom, 100-state MSM using time-structure independent components analysis (tICA)[23, 25-26] and k-medoids clustering. tICA variationally combines coordinates to maximize their auto-correlation time such that the resulting independent components (tICs) encode

structural and kinetic information ordered by decreasing implied time scale.[25-26] K-medoids clustering is chosen to construct the MSM to identify statistically well-sampled states and to obtain a real-structure medoid instead of an average structure for the cluster.

The MSM contains dimer structures that are mostly compact, with a negligibly small number of configurations that are fully dissociated or loosely bound structures with non-specific contacts. Monomer conformations in the medoid structures are mostly within the range of folds observed in experimental structures, and individual Markov states vary in the degree of the structural disorder. Structural variation between dimers primarily reflects non-native contacts at the dimer interface between monomers, conformational variation, and unfolded segments. The time scales for exchange between states vary from a few μs to tens of μs. A summary of structural variables describing all 100 states is provided in the Appendix 8A.

An overview of the MSM is presented as a network plot in Fig. 8.1a. Each Markov state is represented as a node (circle) whose radius is proportional to its equilibrium population in the MSM. Edges connecting these nodes correspond to pathways of state interconversions, with line thickness proportional to the sum of interconversion rates between pairs of states. The network layout is optimized such that states in proximity show fast state interconversion, giving a coordinate-free visualization of the equilibrium population and kinetics of the MSM.

The nodes of the network diagram in Fig. 8.1a are color-mapped to tIC1 values, illustrating that tIC1, the slowest kinetic process of the system, describes shifts in population between two large groupings of states. Color coding the network plot by the state's average RMSD of heavy atoms relative to the dimer crystal structure (Fig. 8.1b), we observe that it is correlated with tIC1 (the correlation coefficient $\rho = -0.64$). The RMSD values vary from 3.7 Å to 8.1 Å across all 100 states, but the distribution is bimodally separated by states in the lower left with low RMSD

configurations closest to x-ray dimer structures with values of 4–5 Å, and high RMSD values of >5.5 Å in the top right. This indicates that tIC1 is related to a significant conformational change in the dimer. We also calculated correlations of tIC1 to several collective variables (CVs), and found strong correlations (>0.8) to torsion angles, distances, and hydrogen bonds involving the B chains at the dimer interface (see Table 8.1 and the Appendix 8A).

We used tIC1 values to group these two dominant clusters in order to understand the conformational changes that it describes. These coarse-grained states are shown as dashed circles in Fig. 8.1, and with the structural differences illustrated in Fig. 8.2. These exhibit twisted motion of the two monomers at the dimer interface, which leads to a disruption of native β-sheet contacts and a reconfiguration of sidechains at the dimer interface. Structural changes along tIC1 are described by several CVs used in previous simulation studies,[51-52] which are summarized in Table 8.1.

	Native (*N*)	**Twisted (*T*)**	ρ
p (%)	66	27	
RMSD (Å)	4.4 (0.3)	5.6 (0.3)	–0.64
$\overline{\langle \Phi_\alpha \rangle}$ (°)	115.2 (15.3)	58.8 (14.7)	0.94
$\overline{\langle n_{\mathrm{HB}}^{\mathrm{Amide}} \rangle}$	3.4 (0.2)	1.4 (0.1)	0.96
$\overline{\langle n_{\mathrm{HB}}^{\mathrm{water}} \rangle}$	3.0 (0.3)	4.2 (0.5)	–0.71
$\overline{\langle n_{\mathrm{MM}} \rangle}$	50.45 (0.65)	48.31 (1.84)	0.79
$\overline{\langle n_\alpha \rangle}$	3.28 (0.23)	2.47 (0.24)	0.69
$\overline{\langle n_\beta \rangle}$	5.79 (0.06)	2.93 (0.25)	0.97
B22R ϕ/ψ (°)	–99.2/4.6	–83.2/130.2	–0.82/0.95
B23G ϕ/ψ (°)	59.1/120.5	–65.0/–66.8	0.95/0.94
$\overline{\langle d_\alpha \rangle}$ (Å)	7.1 (0.33)	8.0 (0.42)	–0.80
$\overline{\langle d_\beta \rangle}$ (Å)	5.1 (0.09)	8.5 (0.69)	–0.83
$\overline{\langle n_{\mathrm{HB}}^{\mathrm{A19\cdots B24}} \rangle}$	0.60 (0.44)	0.80 (0.38)	–0.51

Table 8.1: Physical properties of the native (*N*) and twisted (*T*) dimer structures from the MSM, and their correlation coefficient to tIC1 (ρ): population percentage *p*, RMSD of heavy atoms with respect to the crystal structure, average pseudo-dihedral angle between the B-chain helices $\langle \Phi_\alpha \rangle$, average number of amide HBs between inter-monomer β residues including B23G, B24F, B25F, B26Y $\langle n_{\mathrm{HB}}^{\mathrm{Amide}} \rangle$, average number of water-amide HBs of these β residues $\langle n_{\mathrm{HB}}^{\mathrm{water}} \rangle$, average number of inter-monomer contacts $\langle n_{\mathrm{MM}} \rangle$, α contacts $\langle n_\alpha \rangle$, and β contacts $\langle n_\beta \rangle$, backbone torsion angles ϕ/ψ for B22R and B23G, average distance of α contacts $\langle d_\alpha \rangle$ and β contacts $\langle d_\beta \rangle$, and average number of HBs between A19 amide unit and B24 amide unit $\langle n_{\mathrm{HB}}^{\mathrm{A19\cdots B24}} \rangle$. Bracket average indicates the average over all structure within the same Markov state whereas the bar average refers to the weighted average over states based on their equilibrium population.

The dominant coarse-grained state, the native dimer (*N*), has the largest population of 66%, an average RMSD value of 4.4 Å, and tIC1 values from 1.0 to 0.77. Configurations within this state are similar to the crystal structure, exhibiting an intact inter-monomer β-sheet with an average of 3.4 inter-monomer hydrogen bonds (HBs) $\overline{\langle n_{\mathrm{HB}}^{\mathrm{Amide}} \rangle}$ of β residues including B23G, B24F, B25F

and B26Y. The native state also has two pairs of intermolecular sidechain contacts between B24F and B26F (Fig. 8.2), as well as contacts between the two B25F sidechains and between the B16Y and B26Y sidechains, all of which contribute significantly to stabilizing the dimer.[92] Most of the conformational variation within the native state arises from disorder away from the dimer interface in the N-termini of the B chain and in the fold of the N-terminal helix of the A chain.

The other dominant coarse-grained state, the twisted dimer (*T*), has a population of 27%, a larger RMSD of 5.8 Å, and tIC1 values from –0.83 to –1.26. Structures within this state appear as if one twisted the monomers of the native state clockwise (Fig. 8.2) around the inter-monomer axis, resulting in several conformational changes at the dimer interface. The β-sheet residues of the B-chain remain aligned but have a registry shift relative to the native structure by one amide unit and a decrease of 2 HBs between strands ($\overline{\langle n_{\mathrm{HB}}^{\mathrm{Amide}} \rangle}$ = 1.4). These strands appear wrapped around the dimer rather than the flat sheet observed in the native state. The unusual B20G–B23G turn of the native structure adopts a new configuration with canonical torsion angles. The native β-strand contacts between F and Y residues are replaced with non-native contacts between the two B24F sidechains. Most prominently the twisting motion is seen through the change in the relative orientation of the two B-chain helices, which can be quantified through an α-helix pseudo-dihedral angle $\overline{\langle \Phi_\alpha \rangle}$.[52] The helices rotate relative to each other by –55° on average from 115° to 59° between *N* and *T* states. Fig. 8.1d shows a network plot color-coded by the pseudo-dihedral angle, illustrating how this conformation change identifies the coarse-grained states.

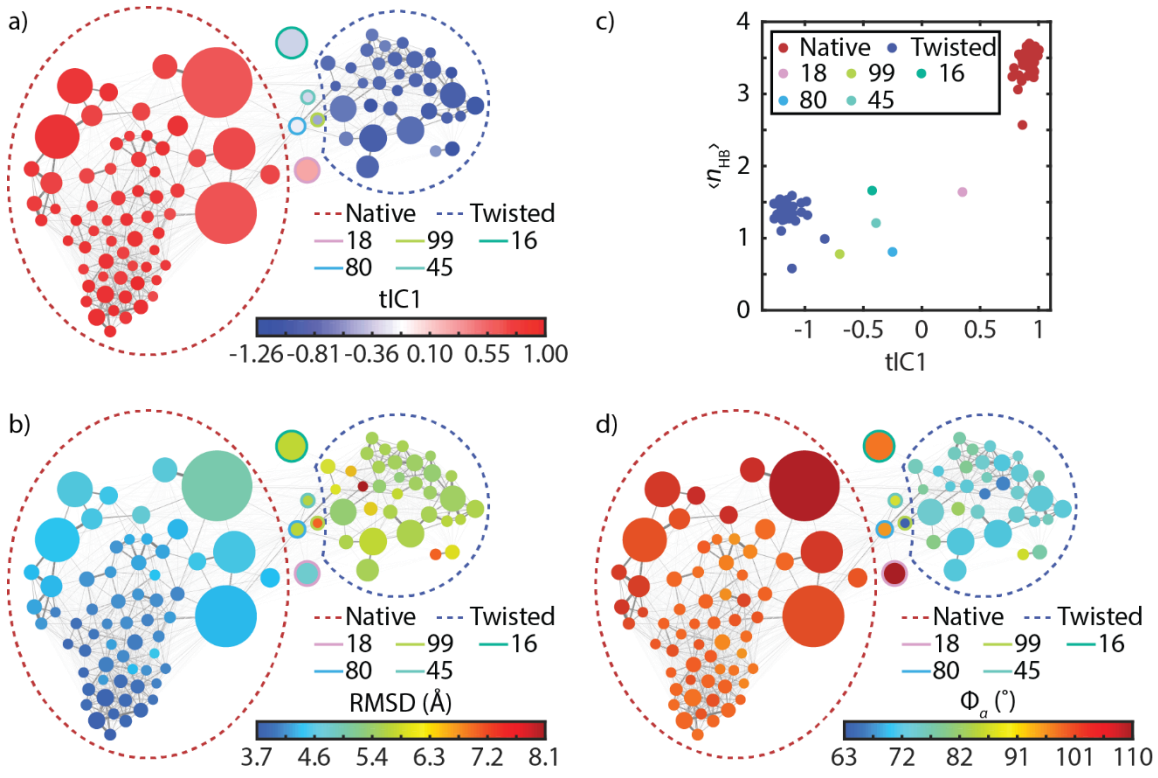

Figure 8.1: (a) Network representation of the dimer MSM with nodes color-coded by tIC1 values accounting for the slowest global process in the transition matrix. Thickness and color of the edges connecting nodes are proportional to the interconversion probabilities. Colored dashed circles identify the native and twisted states identified from a coarse-grained 3-state k-medoids clustering along tIC1. A full assignment of nodes to specific states is found in the Appendix 8B. (b) Network plot color-coded by heavy atom RMSD with respect to crystal structure (PDB: 3W7Y). (c) Correlation of tIC1 with the average number of β-sheet HBs for the Markov state (ρ = 0.96). (d) Network plot color-coded by α pseudo-dihedral angles (ρ = 0.94).

We quantified changes to the average number of inter-monomer contacts involving the B-chain α-helix (n_α), the β-sheet residues (n_β), and all residues including non-native contacts (n_{MM}), as described in the Appendix 8A. This showed that there was only a slight decrease of the number of α contacts from the native state to the twisted state, whereas the number of β contacts decreases by ~3 contacts, accounting for majority of loss of inter-monomer contacts (Table 8.1). This observation indicates that the conformational change along tIC1 perturbs mostly on the local structure along β-sheet residues and sidechain packing while minimally disrupting other contacts.

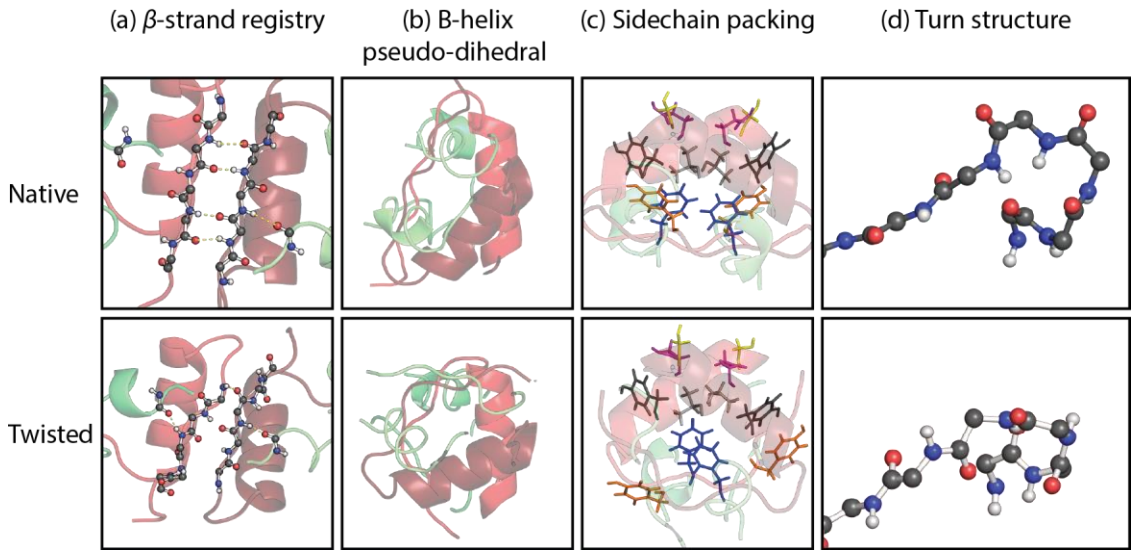

Figure 8.2: Structural differences between the native, $\langle \overline{\text{tIC1}} \rangle \sim 0.8$, and twisted states, $\langle \overline{\text{tIC1}} \rangle \sim -1.1$, illustrated with MSM states 0 and 4 medoid structures. The columns illustrate the shift in backbone hydrogen bond registry from B24 to B26, the rotation of the B1 helix pseudo-dihedral angle from 103 to 74°, the changes in sidechain contacts at the dimer interface Blue: B24F, Orange: B26Y; Gray: B16Y; Light gray: B12V; Magenta: B13E; Yellow: B9S, and the change in turn structure for the B19C-B23G residues.

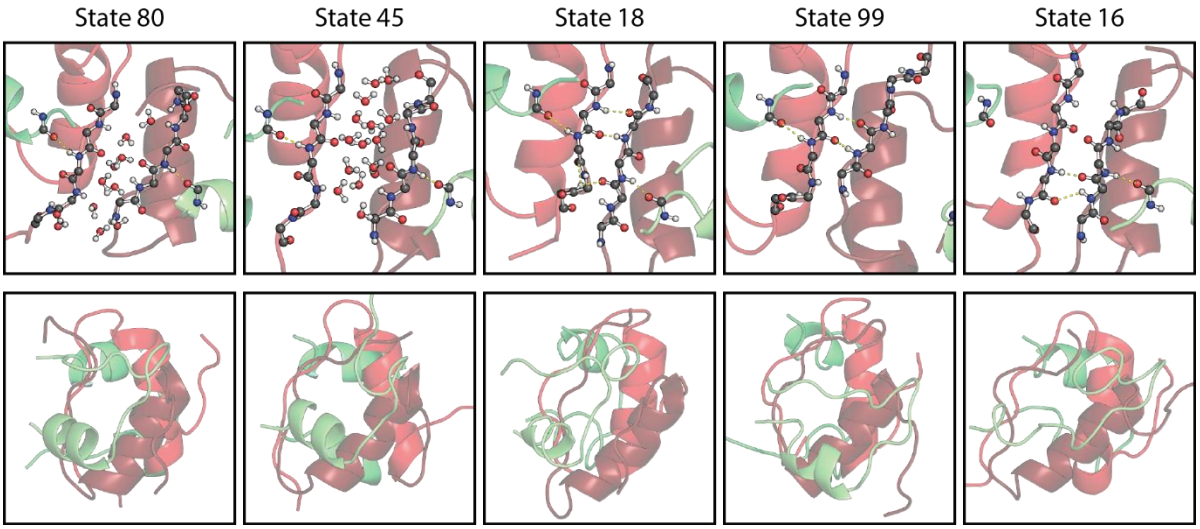

Figure 8.3: Structure of intermediate Markov states, showing a representative frame of each state with backbone atoms from A19Y and B23G–B27T, and a rotated side view illustrating the helix dihedral.

8.4.2 Kinetics

The remaining five states (16, 18, 45, 80, and 99) with intermediate tIC1 values are structurally diverse and account for 7% of the population (Fig. 8.3). Analysis of the top 20 tICs indicates that 11 tICs describe slow kinetics involving the transfer of population between one of these five intermediate Markov states and the rest, suggesting that these intermediate states are kinetic traps. From Fig. 8.1, we also infer that these intermediate states may play an important role in the transitions between the native and twisted states. To further investigate these intermediate states, we reduced the full MSM state space by lumping native and twisted states together and reduced the transition matrix for the resulting seven states using the method of Hummer and Szabo.[93] The network plot for this seven-state lumping is shown in Fig. 8.4a. One can see that states 80 and 45 act as on-pathway intermediates for the conversion of native and twisted forms, whereas 18, 99, and 16, which are identified in tICs 2–7, appear to be off-pathway, kinetic traps in this exchange process.

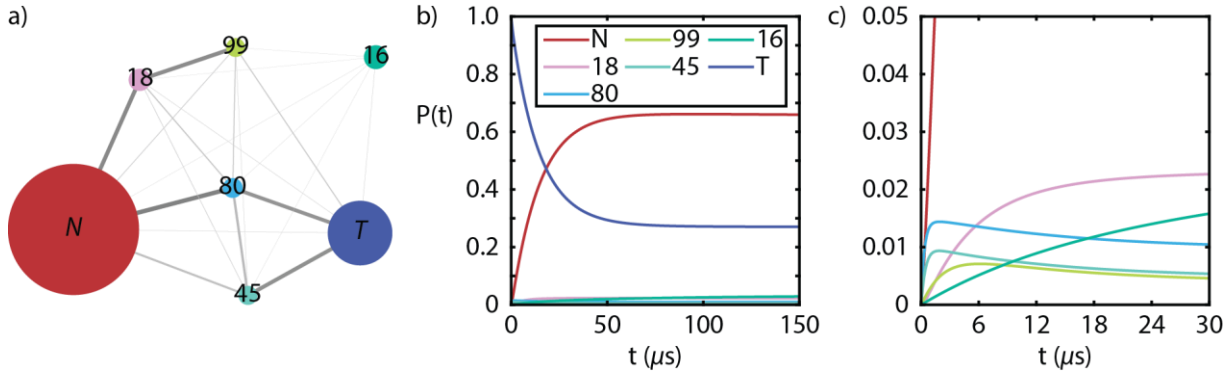

Figure 8.4: Seven-state lumping of native and twisted states with five intermediates. (a) Network plot for the new states and transition matrix. (b) Calculated equilibration kinetics tracking the exchange between native and twisted states when the population is initially in the twisted state.

With this seven-state lumping we also investigated the kinetics of dimer twisting with the reduced rate matrix. Fig. 8.4b shows the time dependent population changes in the 7 states when the system is initiated entirely in the twisted state. Native and twisted populations exchange with a time constant of 14 μs. Further reduction of the MSM to 3 states in which 80 and 45 are lumped as intermediates leads to a small increase in the observed kinetics to 21 μs for the exchange between *N* and *T*. In Fig. 8.4c we also see that population in states 80 and 45 rise with a time-scale of 460 ns and then re-equilibrate with a 21 μs decay as expected of on-path intermediates, whereas states 18 and 16 simply rise slowly to their equilibrium value. State 99's behavior lies between the other indicating that it also plays a non-negligible role in the exchange of native and twisted forms.

Representative structure for the intermediate states are illustrated in Fig. 8.3. States 80 and 45 have water molecules that penetrate the two β-strands such that all of the native HBs present in the β-sheet are replaced by water molecules. This suggests that the primary mechanism of dimer twisting involves water disrupting the specific interactions of the β-sheet without significant disruption to the hydrophobic core. The remaining weak contacts between hydrophobic sidechains of the B-chain helix provide the orientational flexibility to reconfigure the β-strand sidechains and contacts in its new configuration.

A closer look at the makeup of the native and twisted states reveals that both contain conformational substates, which correspond to clustered groups in our network plot. Based on a reduction of the full MSM state space using a Robust Perron Cluster Cluster Analysis (PCCA+) of the first 20 eigenvectors of the transition matrix,[94] we identify four conformational substates which have native dimer contacts, but differ in their fold away from the dimer interface. For instance, two low RMSD native clusters mirror the crystal structure of insulin (N_0, N_1), whereas

the another varies by the unfolding of the A1 helix of one insulin monomer (N_2). The third native substate (N_3) retains the intermolecular α and β contacts of the crystal structure, but have one or both A1 helices unfolded in both monomers, with considerable conformational disorder for the termini of all chains. The kinetics reveal that most conversion between native and twisted passes through the N_3 state. PCCA+ also reveals that the twisted state is primarily one block of well-folded configurations (T_1), with two minor substates (T_2 and T_3) that retain the twisted dimer interface, but vary in the structure and disorder of the A chains and chain termini. These states are discussed further below.

8.4.3 Computational Spectroscopy

The MSM predicts the presence of two dominant conformational states for insulin dimer, one of which corresponds to the well-known dimer structure, in addition to an observed rate constant of 14 μs for the interconversion between native and twisted forms. The prediction of these large-scale conformational changes that have not been previously observed, perhaps due to the μs interconversion timescale, raises the question of how such conformational changes could be observed experimentally. For this purpose, we investigated how local and global conformational changes of insulin dimer could be characterized with amide I infrared spectroscopy.

The amide I vibrational frequency shifts in proportion to the local electric field experienced by the carbonyl, and different carbonyl oscillators can couple to one another by through-bond (mechanical) and through-space (dipole-dipole) couplings.[29] The patterns of characteristic CO frequency shifts and couplings for different secondary structures gives rise to characteristic frequencies and band shapes for α-helices and β-sheets. Most importantly, it is now possible to computationally model the protein amide I spectrum on the basis of atomistic structures drawn

from MD simulation with quantitative accuracy. This tool has been used to characterize and refine conformational ensembles in peptides and small proteins.[14-15, 17, 86, 89, 95]

Amide I spectroscopy can be performed in different manners when combined with site-specific isotope labeling strategies. In the absence of labels, the relatively small frequency variations among the different CO vibrations (σ~10 cm^{-1}) and similar coupling strength (V~0–10 cm^{-1}) means that vibrations spectrally overlap. This leads to broad absorption bands which are insensitive to local structural variation but can be used to quantify secondary structure content. We refer to these as unlabeled (UL) spectra. Alternatively, an isotope-labeled carbonyl—here $^{13}C^{18}O$—can be used to shift the vibrational frequency well outside the band (≈ –60 cm^{-1}), which both spectrally isolates and vibrationally decouples it from other amides in order to identify site-specific contacts.[96-97] In addition to the single-label experiments, dual-label experiments which insert pair of specific isotope labeled carbonyls selected to interact strongly when in close proximity and alignment, are particularly effective for characterizing hydrogen bonding contacts between two residues of the main chain.

With computational IR spectroscopy, one can both compute a spectrum from structures as an interpretive tool, and also predict which isotope labels will be most informative for revealing specific changes in conformation, solvation environment, and hydrogen-bond contacts. We used this strategy to computationally study the IR spectra associated with all 100 Markov states to identify isotope labeling strategies for investigation of structural heterogeneity in insulin dimer.

To begin we calculated the UL amide I absorption spectrum for all 100 Markov states. These spectra are shown in Fig. 8.5a, ordered by the state's tIC1 value. We observe that all spectra are featureless asymmetric absorption bands, similar to experimental IR absorption spectra for the dimer,[49-50] but with little variation in the lineshape between states. Although the lineshape

variations are nearly imperceptible within the *N* and *T* MSM states, the tIC1 value is found to correlate well with the frequency of the absorption maximum (ρ = –0.89). Population-weighted average spectra over the *N* and *T* states reveals a predicted 3 cm^{-1} band shift between states, from $\overline{\langle \omega_N \rangle}$ ~ 1650 cm^{-1} to $\overline{\langle \omega_T \rangle}$ ~ 1647 cm^{-1} (Fig. 8.5b). The asymmetry of the native spectrum can be explained in terms of the two transitions expected from anti-parallel β-sheets,[42, 49-50] including a weak $\nu_{\parallel}$ vibrational transition at 1680 cm^{-1} and a stronger $\nu_{\perp}$ band at 1635 cm^{-1}. Second derivative spectra only marginally improve the spectral differences, so although there are predictable differences, it appears difficult to distinguish *N* and *T* configurations from UL spectra in practice.

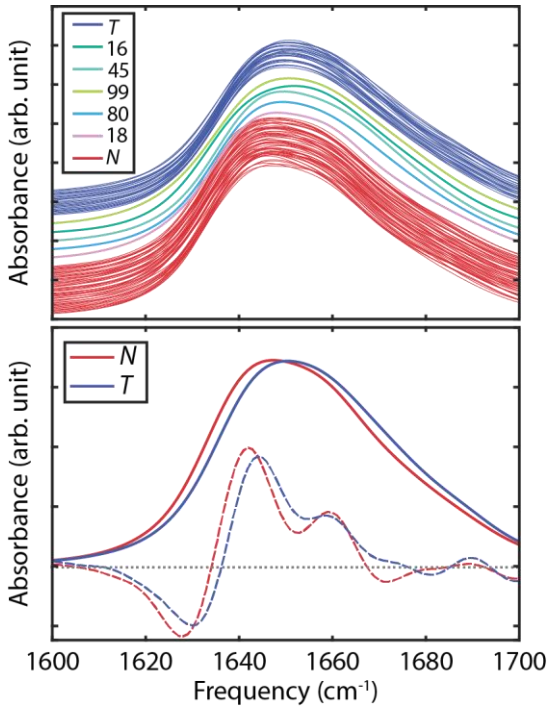

Figure 8.5: (a) Simulated UL FTIR spectra for all 100 Markov states ordered by increasing tIC 1 from top (twisted) to bottom (native). Spectra are vertically displaced for presentation purposes, and colored by their assignment to seven coarse-grained states (b) Comparison of population-weighted average IR spectra (solid line) and second-derivative spectra (dashed line) for the native and twisted states.

Since the *N* and *T* states vary most with the change of β-strand hydrogen bonds and other contacts at the dimer interface, changes in IR spectra between these states are more likely to be observed in spectra from isotope labels placed to interrogate these contacts.[96-97] To identify the most promising candidates, we performed calculations of isotope-edited IR spectra for all 100 states of the MSM starting with single site-specific labels for all 49 amide linkages of the monomer peptide backbone. These single labels shift in frequency depending on the local electric field experienced by the amide carbonyl, but is qualitatively best understood as being sensitive to the number and strength of hydrogen bonds to the carbonyl oxygen. Note that a single isotope label in this homodimer will result in two labels that can couple with one another depending on their proximity. The resulting spectra were analyzed individually or averaged by MSM population over all *N* and *T* coarse-grained states. Additionally, we performed calculations on 16 additional dual labels selected to isolate particular intra- and intermolecular contacts between the two amide groups.

To illustrate isotope labeling IR spectroscopy, Fig. 8.6 shows simulated IR spectra of UL insulin (black curve) and B24B25 double-labeled insulin for one MSM state with a native configuration (red curve). Upon $^{13}C^{18}O$ isotopic substitution on both B24 and B25 amide units on the β-sheet, there are additional isotope-edited features appearing between 1550–1620 cm^{-1}. To highlight vibrational features associated with labeling, we calculate the difference spectrum between B24B25-labeled insulin and UL insulin (dashed curve). From this isotope-labelled difference spectrum, one can see the positive absorption change with the peak frequency of 1576 cm^{-1} corresponding to the labeled amide I vibrations, as well as negative features at 1635 and 1691 cm^{-1} that arise from the loss of those unlabeled residues.

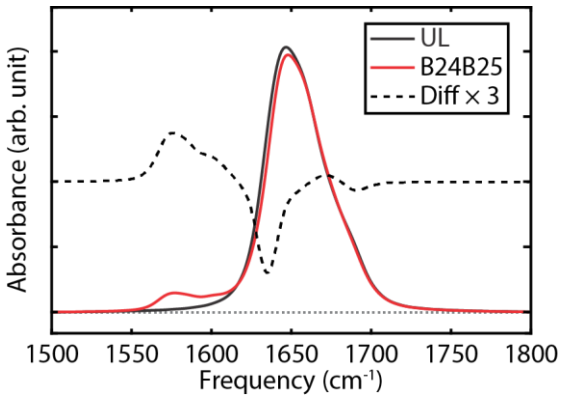

Figure 8.6: Simulated FTIR spectra of the native state 0 of the MSM. Simulated UL FTIR spectrum (black), B24B25 labeled FTIR spectrum (red) and difference spectrum, ΔA, between labeled spectrum and unlabeled spectrum (black dashed). Difference spectrum has been vertically displaced for presentation purpose.

Examples of the calculated isotope difference spectra for residues showing the largest frequency shifts between *N* and *T* states (>5 cm^{-1}) are shown in Fig. 8.7a. Overall 18% of the labels provided a spectrally resolvable distinction between *N* and *T* states. The largest spectral differences are observed for the B24, B25, and B26 single labels that form the *β*-sheet in the native state, as well as double-labels that include a label on one of these sites. Residues at the N-terminus of the A chain also report on a significant conformational change in the A2 helix relative to the B chain between *N* and *T* states, and the A4 and B11 labels report on a change in amide hydrogen bonding strength within the A1 and B helices, respectively. In all cases the frequency shifts observed are less than the linewidth, indicating that quantifying *N* and *T* populations in a mixture will be challenging with only one label; however, the pattern of spectral variation among multiple labels can be used for a more accurate determination.

Looking closer at labels involving the dimer *β*-sheet residues, we now focus on the B24 single label, B24B25 dual label, and A19B24 dual label. The B24 and B24B25 labels are expected to probe intermolecular hydrogen-bond contacts within the *β*-sheet, whereas the A19B24 dual label

should be sensitive to the intramolecular hydrogen bond between the A19 and B24 amide units in the native dimer, as illustrated in Fig. 8.7. The resulting isotope-labeled difference spectra for *N* and *T* states are shown in Fig. 8.7.

Overall, all three difference spectra exhibit common features, including an increase of absorption due to the isotope labelled residues between 1560–1620 cm^{-1}, and the loss of intensity from the unlabeled band in the frequency range above 1620 cm^{-1}. Labeled difference spectra of the native states (blue curves) share common loss features at 1635 cm^{-1} and ~1690 cm^{-1} corresponding to the $\nu_{\perp}$ and $\nu_{\parallel}$ modes of the β-sheet, respectively. For each label, the $\nu_{\perp}$ loss peak for the *T* state is suppressed by about half from the *N* state and blue-shifted to 1642 cm^{-1}, and the $\nu_{\perp}$ loss peak blue-shifts to 1693 cm^{-1}. We observe this pattern also in the A19, B25, and B23B26 label spectra.

Structurally, the B24-isotope label is sensitive to the register shift of the two β-strands between *N* and *T* configurations, partially due to a decrease in hydrogen bonding, but also because of the through-space coupling between B24 labels on each monomer changes significantly (–3 cm^{-1} to 5 cm^{-1}). The calculated B24 difference spectra predicts that the *N* state has an asymmetric isotope-labeled amide I band peaked at 1602 cm^{-1}, whereas the *T* state exhibits a frequency down-shift to 1596 cm^{-1}, a decrease in intensity and change in spectral lineshape.

The B24B25 dual-label of each monomer introduces four labels total into the dimer, which effectively isolates the β-sheet of the native state. In Fig. 8.7, the B24B25 difference spectrum of the *N* state has a peak frequency at 1580 cm^{-1} and a shoulder at 1605 cm^{-1}, which we attribute to the isotope-edited $\nu_{\perp}$ mode and $\nu_{\parallel}$ β-sheet modes.[42, 98] The *T* state, in contrast, shows a symmetric labeled band with a peak frequency of 1596 cm^{-1}.

The A19B24 dual label probes the intramolecular H-bond contact between the A19 and B24 amide units away from dimer interface. The labeled difference spectrum of *N* shows a peak at 1589 cm^{-1} and a pronounced shoulder at ~1610 cm^{-1}, whereas the *T* state is observed to have a more symmetric peak with about the same peak frequency. This reflects the changes in the number of H-bonds between the B24 N–H and A19 C=O, from an average of 0.86 for the *T* state to 0.59 for the *N* state.

These calculations establish that there are labeling strategies available to distinguish *N* and *T* configurations, however, with some of the labels investigated we also found patterns of spectral variation within with *N* and *T* states with slight variation in spectral lineshape corresponding to clusters within our network plot. In Fig. 8.8, we illustrate these shifts with the B24B25 dual label and compare how spectra for all native and twisted MSM states maps onto coarse-grained substates that were lumped with assistance of PCCA+. Although the *N* and *T* states share the general features described above, a closer look within 7 *N* and *T* coarse-grained substates reveals that the individual MSM states also have a different frequency, linewidth and lineshape to their label transition and loss features (Fig. 8.8a). For instance, we find the subset of states N_0, which correspond to the kinetically clustered lowest RMSD states of the MSM, have the lowest frequency labelled bands (<1575 cm^{-1}) compared to the N_1, N_2, and N_3 substates (>1575 cm^{-1}). This is illustrated by the averaged spectra in Fig. 8.8. This comparison also reveals that there is very little spectral variation for the B24B25 label within the individual or coarse-grained twisted states. Only one *T* state (77) is clearly distinguishable by its low frequency resonance. Overall 16% of the labels calculated showed such spectral variation within substates.

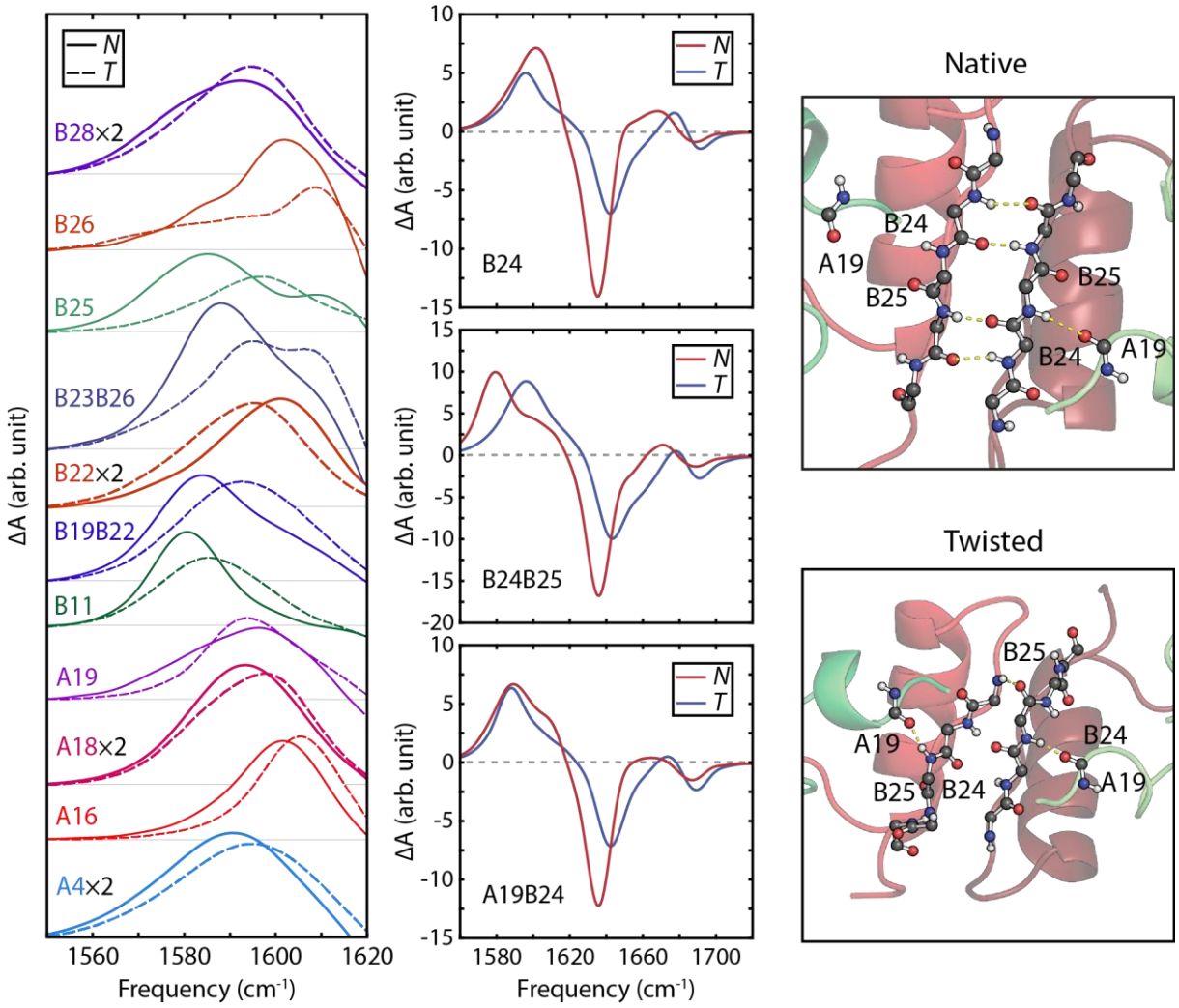

Figure 8.7: Left: Simulated isotope labeled IR difference spectra for several labels illustrating patterns of frequency shifts between the native and twisted states. Center: Simulated isotope difference spectra for B24, B24B25, and A19B24 labels including gain and loss features for *N* and *T* states. Right: Representative structures of both *N* and *T* states indicating structural differences in the A19, B24, and B25 carbonyls.

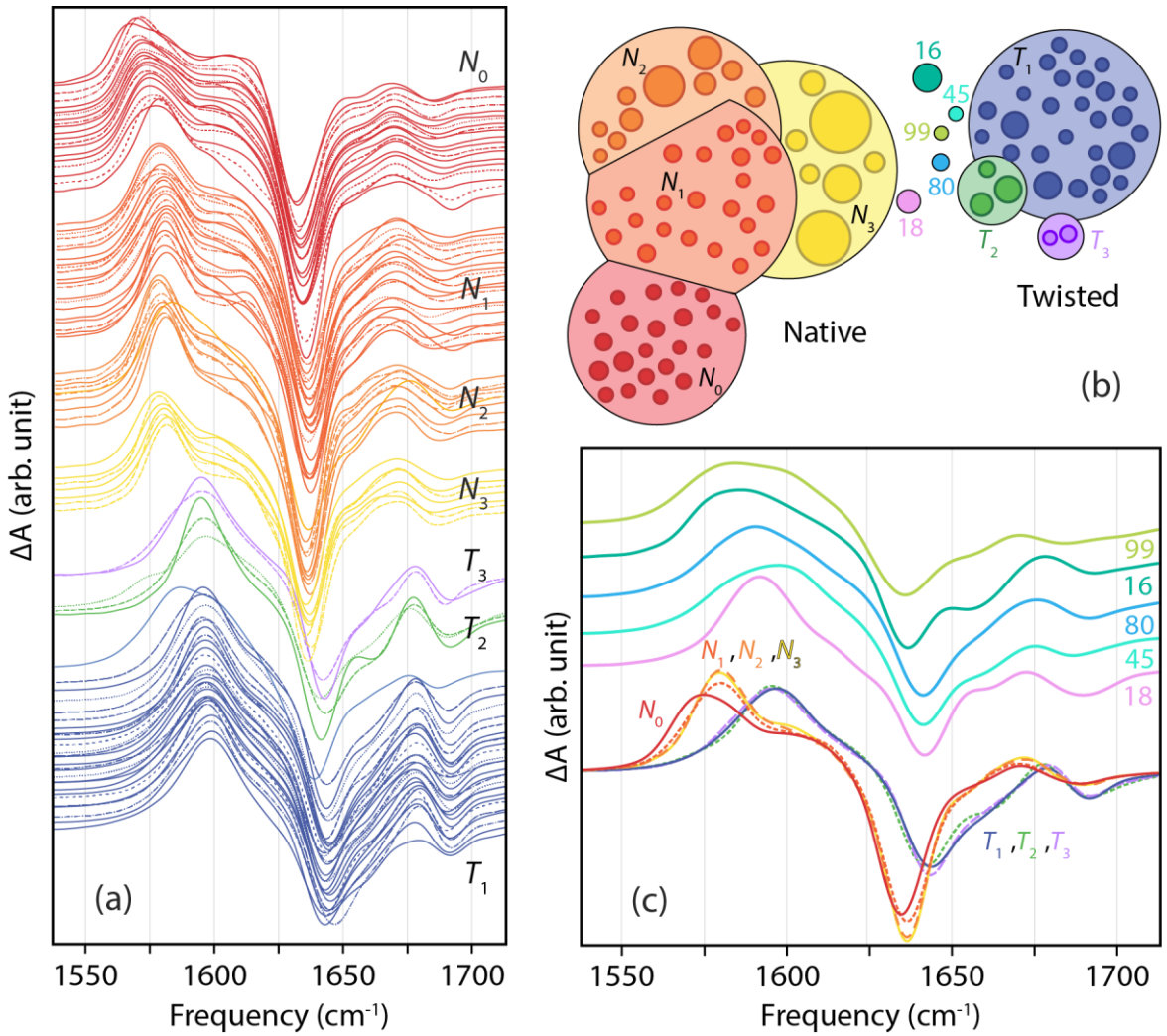

Figure 8.8: Spectral variation of the B24B25 label difference spectra among the 100 Markov states. (a) Individual spectra for native and twisted states ordered by peak transition frequency within the coarse-grained states obtained by PCCA+. (b) Corresponding color-coded native and twisted substates and intermediate states in 12-state coarse graining. (c) Comparison of population-weighted spectra for the four native and three twisted substates and the spectra of intermediate states.

Fig. 8.8c also compares the *N* and *T* spectra with those calculated for the five intermediate states. We observe that the intermediate states have significant spectral variation and are distinct from the native and twisted states. The spectra for these intermediate states do not share clear similarities with either *N* or *T* substates, and are much broader and featureless. Only state 18 has a clearly identifiable sharp resonance in its spectrum. These observations suggest that it may also be

possible to distinguish the populations of intermediate states in IR kinetics measurements of structural interconversion of insulin dimer.

8.5 Discussion and Conclusions

Our investigation of the structural variation of insulin dimer using extensive all-atom simulations, Markov state modeling, and computational amide I spectroscopy predicts the presence of two dominant conformations to insulin dimer, and illustrate how they can be experimentally resolved through isotope-edited IR spectroscopy. The native and twisted dimer conformations primarily differ by the change of contacts at the interface between the two bound monomers, and in the backbone hydrogen bonding and sidechain packing of the B-chain β-strand residues that form the intermolecular β sheet in the native configuration. The MSM kinetics indicate that the exchange of native and twisted populations occurs on a 21 μs time-scale.

The twisted dimer conformation has, to our knowledge, not been experimentally observed. This is not surprising, given that it is a higher energy state than the native form and the predicted exchange kinetics are rapid; however, if present, confirming the presence of a twisted structure could have several consequences. On a fundamental level its presence should be considered in several factors varying from its influence on biological processes to interpretation of NMR experiments. It would influence our understanding of the role of many common B chain mutants found in insulin medications on the monomer-dimer equilibrium, and other principles for drug design. More generally, it provides evidence of the structural rearrangements and dynamical processes that can occur at protein-protein binding interfaces of complexes that are thought to bind in a unique site-specific manner.

From a dynamics perspective, it remains unknown what role a twisted dimer could play in the dimer dissociation and association processes, perhaps as an intermediate. Recent simulations of insulin dimer dissociation free energies described a broad distribution of possible energetically favorable dissociation pathways, bounded by two limiting cases: (1) a sequential process of disrupting B-chain α-helix contacts prior to B-chain β contacts (the α path), or (2) β contacts prior to α contacts (β path).[52] To investigate connections between MSM intermediate states and on-path structures during the course of dissociation in that study, we projected the dimer MSM onto collective variables (CVs) describing the dissociation (see Figure 8A.3). As one might expect, the well-solvated β strands of states 45 and 80 lie along the β path when observing CVs involving β contacts, but the α pseudo-dihedral rotation lies closer to the α path in a high free energy region rarely visited in the sampling of dimer dissociation. It is possible that the dimer MSM contains structures not sampled for the biased sampling of dissociation free energy landscapes,[51-52] but it is also possible that variations in the side-chain protonation state or force field contribute to this discrepancy. On the other hand, projecting the dimer MSM onto the free energy landscape constructed by Bagchi and coworkers showed that the native state and the twisted state lie in the same free energy basin (Table 8.1 and Figs. 3–4 in Ref. 51). As a result, the role of the twisted dimer in insulin dimer dissociation remains unclear at this time.

The simulation conditions we used were selected with IR spectroscopy in mind, since protonated sidechains (low pH) improves insulin solubility, destabilizes insulin hexamer, and reduces IR background absorptions from asymmetric COO^- vibrations in the same region as the labels. It is possible that these conditions favored the stabilization of the twisted configuration; however, a clear rationale is not apparent. The protonatable sidechains are away from the dimer

binding interface, except for the case of B21E which may influence the conformation of the B19C–B23G turn.

To search for the presence of the twisted state, we find that IR spectroscopy targeting the B24, B25, and B26 residues in single and pairwise $^{13}C^{18}O$ isotope labelling provides the best strategy for spectroscopically distinguishing the *N* and *T* configurations. Temperature and pH dependent studies could be used to influence the equilibrium between *N* and *T* states. Such IR spectra could also be used to track the exchange kinetics between *N* and *T* states when used as the probe of a temperature-jump experiment. Separately, we did investigate the variation of computed UV circular dichroism spectra[99] between *N* and *T* states, and found no significant change in the spectral shape, but a decrease in the magnitude of the molar ellipticity in the *T* state.

While a few labels result in large spectral differences between *N* and *T* states, most of the expected spectral changes for any particular label are predicted to show a small up-shift or down-shift in vibrational frequency, often less than the linewidth of the transition. Therefore, a robust strategy for studying insulin dimer is best performed with multiple labels whose pattern of spectral peaks can act as a type of "bar code" to identify the presence of the twisted dimer state. Indeed, we believe that the calculated isotope-labeled difference spectra for all 100 Markov states form a unique basis set for structural ensemble refinement from experiments using maximum entropy or Bayesian refinement tools.[17, 89] Finally, we note that the amide I spectral simulation tools presented here are equally applicable to 2D IR spectroscopy, which has improved capabilities for resolving isotope peak positions and distributions of spectra encoding structural variation. These observations set the stage for IR experimental studies to study insulin dimer structure and the dissociation/association equilibrium between the dimer(s) and monomer, and the kinetics of the coupled dimer conformational change and the dimer dissociation processes.

8.6 Acknowledgments

I thank Anton Sinitskiy in the Pande group for constructing this dimer MSM. I thank Adam Antoszewski, and Bodhi Vani in the Dinner group for fruitful discussions on structural descriptions of the dimer MSM. I also thank Brennan Ashwood, Luis Busto de Moner, and Paul Sanstead for carefully reading through the previous enormous manuscript that leads to part of this chapter.

8.7 References

1. Tompa, P., Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* **2012,** *37* (12), 509-16.
2. Papoian, G. A., Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proc Natl Acad Sci U S A* **2008,** *105* (38), 14237-8.
3. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M., Classification of intrinsically disordered regions and proteins. *Chem Rev* **2014,** *114* (13), 6589-631.
4. Rieping, W.; Habeck, M.; Nilges, M., Inferential structure determination. *Science* **2005,** *309* (5732), 303-6.
5. Buchenberg, S.; Schaudinnus, N.; Stock, G., Hierarchical Biomolecular Dynamics: Picosecond Hydrogen Bonding Regulates Microsecond Conformational Transitions. *J Chem Theory Comput* **2015,** *11* (3), 1330-6.
6. Fleming, G. R.; Wolynes, P. G., Chemical Dynamics in Solution. *Physics Today* **1990,** *43* (5), 36-43.
7. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450* (7172), 964-72.
8. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M., Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **2017,** *42*, 106-116.
9. Markwick, P. R.; Malliavin, T.; Nilges, M., Structural biology by NMR: structure, dynamics, and interactions. *PLoS Comput Biol* **2008,** *4* (9), e1000168.
10. Bryant, R. G., The NMR time scale. *Journal of Chemical Education* **1983,** *60* (11), 933.
11. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press: 2011.
12. Baiz, C. R.; Reppert, M.; Tokmakoff, A., An Introduction to Protein 2D IR Spectroscopy. *Ultrafast Infrared Vibrational Spectroscopy* **2013**, 361-403.
13. Woutersen, S.; Hamm, P., Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J Phys Chem B* **2000,** *104* (47), 11316-11320.

14. Smith, A. W.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A.; Roy, S.; Jansen, T. L.; Knoester, J., Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J Phys Chem B* **2010,** *114* (34), 10913-24.
15. Baiz, C. R.; Tokmakoff, A., Structural disorder of folded proteins: isotope-edited 2D IR spectroscopy and Markov state modeling. *Biophys J* **2015,** *108* (7), 1747-1757.
16. Feng, Y.; Huang, J.; Kim, S.; Shim, J. H.; MacKerell, A. D., Jr.; Ge, N. H., Structure of Penta-Alanine Investigated by Two-Dimensional Infrared Spectroscopy and Molecular Dynamics Simulation. *J Phys Chem B* **2016,** *120* (24), 5325-39.
17. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
18. Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D.; Skinner, J. L.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016,** *353* (6303), 1040-1044.
19. Ghosh, A.; Ostrander, J. S.; Zanni, M. T., Watching Proteins Wiggle: Mapping Structures with Two-Dimensional Infrared Spectroscopy. *Chem Rev* **2017,** *117* (16), 10726-10759.
20. Buchanan, L. E.; Dunkelberger, E. B.; Tran, H. Q.; Cheng, P. N.; Chiu, C. C.; Cao, P.; Raleigh, D. P.; de Pablo, J. J.; Nowick, J. S.; Zanni, M. T., Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. *Proc Natl Acad Sci U S A* **2013,** *110* (48), 19285-90.
21. Lomont, J. P.; Ostrander, J. S.; Ho, J. J.; Petti, M. K.; Zanni, M. T., Not All beta-Sheets Are the Same: Amyloid Infrared Spectra, Transition Dipole Strengths, and Couplings Investigated by 2D IR Spectroscopy. *J Phys Chem B* **2017,** *121* (38), 8935-8945.
22. Huang, J.; MacKerell, A. D., Jr., Force field development and simulations of intrinsically disordered proteins. *Curr Opin Struct Biol* **2018,** *48*, 40-48.
23. Molgedey, L.; Schuster, H. G., Separation of a mixture of independent signals using time delayed correlations. *Phys Rev Lett* **1994,** *72* (23), 3634-3637.
24. Naritomi, Y.; Fuchigami, S., Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J Chem Phys* **2011,** *134* (6), 065101.
25. Schwantes, C. R.; Pande, V. S., Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* **2013,** *9* (4), 2000-2009.
26. Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F., Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* **2013,** *139* (1), 015102.
27. Husic, B. E.; Pande, V. S., Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018,** *140* (7), 2386-2396.
28. Noe, F.; Rosta, E., Markov Models of Molecular Kinetics. *J Chem Phys* **2019,** *151* (19), 190401.
29. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
30. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.

31. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.
32. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
33. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
34. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.
35. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
36. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
37. Torii, H., Amide I Vibrational Properties Affected by Hydrogen Bonding Out-of-Plane of the Peptide Group. *J Phys Chem Lett* **2015,** *6* (4), 727-33.
38. Torii, H.; Tasumi, M., Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *Journal of Raman Spectroscopy* **1998,** *29* (1), 81-86.
39. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
40. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
41. Hayashi, T.; Mukamel, S., Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides. *J Phys Chem B* **2007,** *111* (37), 11032-46.
42. Cheatum, C. M.; Tokmakoff, A.; Knoester, J., Signatures of beta-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J Chem Phys* **2004,** *120* (17), 8201-15.
43. Sengupta, N.; Maekawa, H.; Zhuang, W.; Toniolo, C.; Mukamel, S.; Tobias, D. J.; Ge, N. H., Sensitivity of 2D IR spectra to peptide helicity: a concerted experimental and simulation study of an octapeptide. *J Phys Chem B* **2009,** *113* (35), 12037-49.
44. Woys, A. M.; Almeida, A. M.; Wang, L.; Chiu, C. C.; McGovern, M.; de Pablo, J. J.; Skinner, J. L.; Gellman, S. H.; Zanni, M. T., Parallel beta-sheet vibrational couplings revealed by 2D IR spectroscopy of an isotopically labeled macrocycle: quantitative benchmark for the interpretation of amyloid and protein infrared spectra. *J Am Chem Soc* **2012,** *134* (46), 19118-28.
45. Wang, L.; Middleton, C. T.; Singh, S.; Reddy, A. S.; Woys, A. M.; Strasfeld, D. B.; Marek, P.; Raleigh, D. P.; de Pablo, J. J.; Zanni, M. T.; Skinner, J. L., 2DIR spectroscopy of human amylin fibrils reflects stable beta-sheet structure. *J Am Chem Soc* **2011,** *133* (40), 16062-71.
46. Mukherjee, P.; Kass, I.; Arkin, I. T.; Zanni, M. T., Structural disorder of the CD3zeta transmembrane domain studied with 2D IR spectroscopy and molecular dynamics simulations. *J Phys Chem B* **2006,** *110* (48), 24740-9.
47. Stevenson, P.; Gotz, C.; Baiz, C. R.; Akerboom, J.; Tokmakoff, A.; Vaziri, A., Visualizing KcsA conformational changes upon ion binding by infrared spectroscopy and atomistic modeling. *J Phys Chem B* **2015,** *119* (18), 5824-31.

48. Kratochvil, H. T.; Maj, M.; Matulef, K.; Annen, A. W.; Ostmeyer, J.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Probing the Effects of Gating on the Ion Occupancy of the K(+) Channel Selectivity Filter Using Two-Dimensional Infrared Spectroscopy. *J Am Chem Soc* **2017,** *139* (26), 8837-8845.
49. Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy. *Phys Chem Chem Phys* **2010,** *12* (14), 3579-88.
50. Zhang, X. X.; Jones, K. C.; Fitzpatrick, A.; Peng, C. S.; Feng, C. J.; Baiz, C. R.; Tokmakoff, A., Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J Phys Chem B* **2016,** *120* (23), 5134-45.
51. Banerjee, P.; Mondal, S.; Bagchi, B., Insulin dimer dissociation in aqueous solution: A computational study of free energy landscape and evolving microscopic structure along the reaction pathway. *J Chem Phys* **2018,** *149* (11), 114902.
52. Antoszewski, A.; Feng, C. J.; Vani, B. P.; Thiede, E. H.; Hong, L.; Weare, J.; Tokmakoff, A.; Dinner, A. R., Insulin Dissociates by Diverse Mechanisms of Coupled Unfolding and Unbinding. *J Phys Chem B* **2020**, 5571-87.
53. Desmond, J. L.; Koner, D.; Meuwly, M., Probing the Differential Dynamics of the Monomeric and Dimeric Insulin from Amide-I IR Spectroscopy. *J Phys Chem B* **2019,** *123* (30), 6588-6598.
54. Hua, Q. X.; Shoelson, S. E.; Kochoyan, M.; Weiss, M. A., Receptor binding redefined by a structural switch in a mutant human insulin. *Nature* **1991,** *354* (6350), 238-41.
55. Ludvigsen, S.; Roy, M.; Thogersen, H.; Kaarsholm, N. C., High-resolution structure of an engineered biologically potent insulin monomer, B16 Tyr-->His, as determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **1994,** *33* (26), 7998-8006.
56. Olsen, H. B.; Ludvigsen, S.; Kaarsholm, N. C., Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry* **1996,** *35* (27), 8836-45.
57. Keller, D.; Clausen, R.; Josefsen, K.; Led, J. J., Flexibility and bioactivity of insulin: an NMR investigation of the solution structure and folding of an unusually flexible human insulin mutant with increased biological activity. *Biochemistry* **2001,** *40* (35), 10732-40.
58. Ludvigsen, S.; Olsen, H. B.; Kaarsholm, N. C., A structural switch in a mutant insulin exposes key residues for receptor binding. *J Mol Biol* **1998,** *279* (1), 1-7.
59. Kosinova, L.; Veverka, V.; Novotna, P.; Collinsova, M.; Urbanova, M.; Moody, N. R.; Turkenburg, J. P.; Jiracek, J.; Brzozowski, A. M.; Zakova, L., Insight into the structural and biological relevance of the T/R transition of the N-terminus of the B-chain in human insulin. *Biochemistry* **2014,** *53* (21), 3392-402.
60. Bocian, W.; Sitkowski, J.; Bednarek, E.; Tarnowska, A.; Kawecki, R.; Kozerski, L., Structure of human insulin monomer in water/acetonitrile solution. *J Biomol NMR* **2008,** *40* (1), 55-64.
61. Zoete, V.; Meuwly, M.; Karplus, M., A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. *J Mol Biol* **2004,** *342* (3), 913-29.
62. Baker, E. N.; Blundell, T. L.; Cutfield, J. F.; Cutfield, S. M.; Dodson, E. J.; Dodson, G. G.; Hodgkin, D. M.; Hubbard, R. E.; Isaacs, N. W.; Reynolds, C. D.; et al., The structure of 2Zn pig insulin crystals at 1.5 A resolution. *Philos Trans R Soc Lond B Biol Sci* **1988,** *319* (1195), 369-456.

63. Jørgensen, A. M. M.; Kristensen, S. M.; Led, J. J.; Balschmidt, P., Three-dimensional solution structure of an insulin dimer. *Journal of Molecular Biology* **1992,** *227* (4), 1146-1163.
64. Menting, J. G.; Yang, Y.; Chan, S. J.; Phillips, N. B.; Smith, B. J.; Whittaker, J.; Wickramasinghe, N. P.; Whittaker, L. J.; Pandyarajan, V.; Wan, Z. L.; Yadav, S. P.; Carroll, J. M.; Strokes, N.; Roberts, C. T., Jr.; Ismail-Beigi, F.; Milewski, W.; Steiner, D. F.; Chauhan, V. S.; Ward, C. W.; Weiss, M. A.; Lawrence, M. C., Protective hinge in insulin opens to enable its receptor engagement. *Proc Natl Acad Sci U S A* **2014,** *111* (33), E3395-404.
65. Croll, T. I.; Smith, B. J.; Margetts, M. B.; Whittaker, J.; Weiss, M. A.; Ward, C. W.; Lawrence, M. C., Higher-Resolution Structure of the Human Insulin Receptor Ectodomain: Multi-Modal Inclusion of the Insert Domain. *Structure* **2016,** *24* (3), 469-76.
66. Gutmann, T.; Kim, K. H.; Grzybek, M.; Walz, T.; Coskun, U., Visualization of ligand-induced transmembrane signaling in the full-length human insulin receptor. *J Cell Biol* **2018,** *217* (5), 1643-1649.
67. Weis, F.; Menting, J. G.; Margetts, M. B.; Chan, S. J.; Xu, Y.; Tennagels, N.; Wohlfart, P.; Langer, T.; Muller, C. W.; Dreyer, M. K.; Lawrence, M. C., The signalling conformation of the insulin receptor ectodomain. *Nat Commun* **2018,** *9* (1), 4420.
68. Raghunathan, S.; El Hage, K.; Desmond, J. L.; Zhang, L.; Meuwly, M., The Role of Water in the Stability of Wild-type and Mutant Insulin Dimers. *J Phys Chem B* **2018,** *122* (28), 7038-7048.
69. Wicky, B. I. M.; Shammas, S. L.; Clarke, J., Affinity of IDPs to their targets is modulated by ion-specific changes in kinetics and residual structure. *Proc Natl Acad Sci U S A* **2017,** *114* (37), 9882-9887.
70. Boreikaite, V.; Wicky, B. I. M.; Watt, I. N.; Clarke, J.; Walker, J. E., Extrinsic conditions influence the self-association and structure of IF1, the regulatory protein of mitochondrial ATP synthase. *Proc Natl Acad Sci U S A* **2019,** *116* (21), 10354-10359.
71. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-8.
72. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983,** *79* (2), 926-935.
73. Shirts, M.; Pande, V. S., COMPUTING: Screen Savers of the World Unite! *Science* **2000,** *290* (5498), 1903-4.
74. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **2017,** *13* (7), e1005659.
75. Harrigan, M. P.; Sultan, M. M.; Hernandez, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S., MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys J* **2017,** *112* (1), 10-15.
76. Keller, B.; Daura, X.; van Gunsteren, W. F., Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J Chem Phys* **2010,** *132* (7), 074110.
77. Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S., MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J Chem Theory Comput* **2011,** *7* (10), 3412-3419.

78. Bastian, M.; Heymann, S.; Jacomy, M. In *Gephi: an open source software for exploring and manipulating networks*, Third international AAAI conference on weblogs and social media, 2009.
79. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-54.
80. Boonstra, S.; Onck, P. R.; Giessen, E., CHARMM TIP3P Water Model Suppresses Peptide Folding by Solvating the Unfolded State. *J Phys Chem B* **2016,** *120* (15), 3692-8.
81. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
82. Nosé, S., A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods. *Journal of Chemical Physics* **1984,** *81* (1), 511-519.
83. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A Gen Phys* **1985,** *31* (3), 1695-1697.
84. Reppert, M., g_amide. **2017**.
85. Reppert, M.; Feng, C.-J., g_spec. **2017**.
86. Feng, C. J.; Tokmakoff, A., The dynamics of peptide-water interactions in dialanine: An ultrafast amide I 2D IR and computational spectroscopy study. *J Chem Phys* **2017,** *147* (8), 085101.
87. Torii, H., Effects of intermolecular vibrational coupling and liquid dynamics on the polarized Raman and two-dimensional infrared spectral profiles of liquid N,N-dimethylformamide analyzed with a time-domain computational method. *J Phys Chem A* **2006,** *110* (14), 4822-32.
88. Liang, C.; Jansen, T. L., An Efficient N(3)-Scaling Propagation Scheme for Simulating Two-Dimensional Infrared and Visible Spectra. *J Chem Theory Comput* **2012,** *8* (5), 1706-13.
89. Feng, C. J.; Dhayalan, B.; Tokmakoff, A., Refinement of Peptide Conformational Ensembles by 2D IR Spectroscopy: Application to Ala–Ala–Ala. *Biophys J* **2018,** *114* (12), 2820-2832.
90. Torres, J.; Kukol, A.; Goodman, J. M.; Arkin, I. T., Site-specific examination of secondary structure and orientation determination in membrane proteins: The peptidic13C-18O group as a novel infrared probe. *Biopolymers* **2001,** *59* (6), 396-401.
91. Decatur, S. M., Elucidation of residue-level structure and dynamics of polypeptides via isotope-edited infrared spectroscopy. *Acc Chem Res* **2006,** *39* (3), 169-75.
92. Zoete, V.; Meuwly, M.; Karplus, M., Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* **2005,** *61* (1), 79-93.
93. Hummer, G.; Szabo, A., Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models. *J Phys Chem B* **2015,** *119* (29), 9029-37.
94. Deuflhard, P.; Weber, M., Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications* **2005,** *398*, 161-184.
95. Baiz, C. R.; Lin, Y. S.; Peng, C. S.; Beauchamp, K. A.; Voelz, V. A.; Pande, V. S.; Tokmakoff, A., A molecular interpretation of 2D IR protein folding experiments with Markov state models. *Biophys J* **2014,** *106* (6), 1359-70.
96. Dong, J.; Wan, Z. L.; Chu, Y. C.; Nakagawa, S. N.; Katsoyannis, P. G.; Weiss, M. A.; Carey, P. R., Isotope-edited Raman spectroscopy of proteins: a general strategy to probe individual peptide bonds with application to insulin. *J Am Chem Soc* **2001,** *123* (32), 7919-20.

97. Dhayalan, B.; Fitzpatrick, A.; Mandal, K.; Whittaker, J.; Weiss, M. A.; Tokmakoff, A.; Kent, S. B., Efficient Total Chemical Synthesis of (13) C=(18) O Isotopomers of Human Insulin for Isotope-Edited FTIR. *Chembiochem* **2016,** *17* (5), 415-20.
98. Miyazawa, T.; Blout, E. R., The Infrared Spectra of Polypeptides in Various Conformations: Amide I and II Bands1. *Journal of the American Chemical Society* **1961,** *83* (3), 712-719.
99. Mavridis, L.; Janes, R. W., PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates. *Bioinformatics* **2017,** *33* (1), 56-63.
100. Banerjee, P.; Mondal, S.; Bagchi, B., Effect of ethanol on insulin dimer dissociation. *J Chem Phys* **2019,** *150* (8), 084902.
101. Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; Capelli, R.; Carloni, P.; Ceriotti, M.; Cesari, A.; Chen, H.; Chen, W.; Colizzi, F.; De, S.; De La Pierre, M.; Donadio, D.; Drobot, V.; Ensing, B.; Ferguson, A. L.; Filizola, M.; Fraser, J. S.; Fu, H.; Gasparotto, P.; Gervasio, F. L.; Giberti, F.; Gil-Ley, A.; Giorgino, T.; Heller, G. T.; Hocky, G. M.; Iannuzzi, M.; Invernizzi, M.; Jelfs, K. E.; Jussupow, A.; Kirilin, E.; Laio, A.; Limongelli, V.; Lindorff-Larsen, K.; Löhr, T. M. F.; Martin-Samos, L.; Masetti, M.; Meyer, R.; Michaelides, A.; Molteni, C.; Morishita, T.; Nava, M.; Paissoni, C.; Papaleo, E.; Parrinello, M.; Pfaendtner, J.; Piaggi, P.; Piccini, G.; Pietropaolo, A.; Pietrucci, F.; Pipolo, S.; Provasi, D.; Quigley, D.; Raiteri, P.; Raniolo, S.; Rydzewski, J.; Salvalaglio, M.; Sosso, G. C.; Spiwok, V.; Šponer, J.; Swenson, D. W. H.; Tiwary, P.; Valsson, O.; Vendruscolo, M.; Voth, G. A.; White, A., Promoting transparency and reproducibility in enhanced molecular simulations. *Nat Methods* **2019,** *16* (8), 670-673.
102. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **2014,** *185* (2), 604-613.

Appendix 8A: Characterization of Structural Collective Variables for Markov States

To quantitatively analyze native contacts in the dimer state, number of contacts were calculated by defining atom pairs of interest, and using a rational switching function given below to determine if the pairs were in contact.

$$s\left(r_{ij}\right)=\begin{cases}1, \forall r_{ij} \le d_0 \\ 1-\left(\frac{r_{ij}}{r_0}\right)^n \Big/ 1-\left(\frac{r_{ij}}{r_0}\right)^m, r_{ij} > d_0\end{cases} \tag{8.1}$$

To calculate the number of α contacts (n_α) and β contacts (n_β), the average over $s(r_{ij})$ is calculated for all pairs of C_α atoms summarized in the Table 8A.1. To compare with results from Bahchi and co-workers, we also computed the number of inter-monomer contacts using every single possible pair of C_α atoms between monomer 1 (M1) and monomer 2 (M2), resulting in 51 × 51 = 2601 pairs.[100] r_0 used in the rational switching function was set to 7 Å to match with Bagchi and co-workers calculations, as well as n = 6 and m = 12. For each Markov state consisting of 100 structures, the average number of the contacts was computed over the contacts of those 100 structures within the same Markov state with equal weight. All of the contacts were computed using the open-source, community-developed PLUMED library,[101] version 2.5.2.[102]

The number of HBs (n_{HB}) along the β-strand were also analyzed to illustrate the local structural changes of the residues involved in the β-sheet (Figure 8A.1). HB is counted using geometric criteria, meaning that the distance between O on C=O and H on N–H is less than or equal to 3.5 Å, and the angle between $\mathrm{H\cdots C=O}$, $\angle_{\mathrm{HOC}}$, is greater or equal to 150°, and the calculation was done using GROMACS utility. The average number of HBs of each Markov state, $\langle n_{HB} \rangle$, was computed over 100 structures within the same Markov state with equal weight.

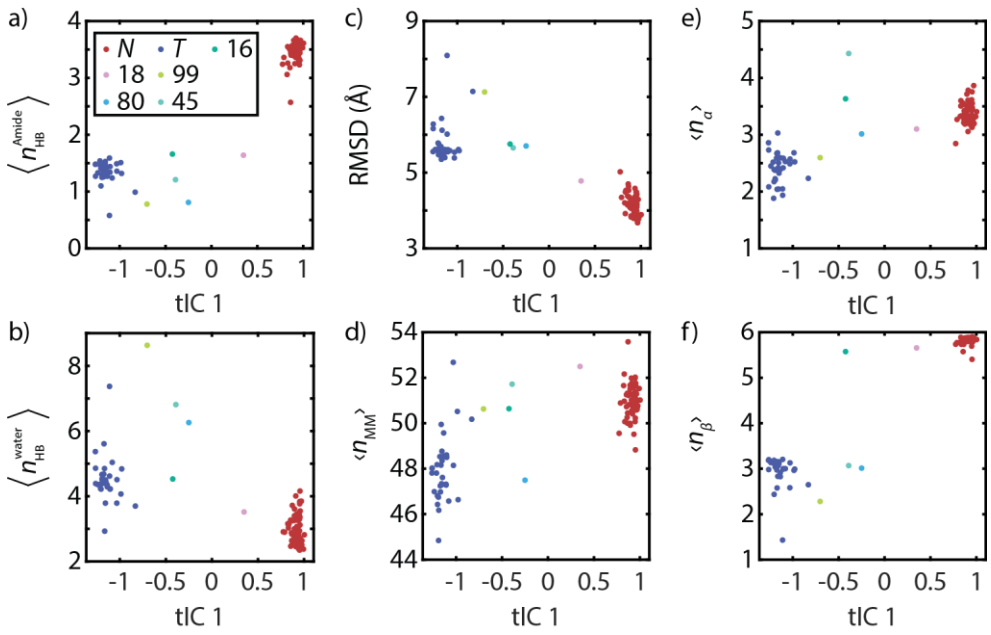

Figure 8A.1: Average contact variables of Markov states along tIC1, including (a) average number of amide hydrogen bonds from B23 to B26 $\langle n_{HB}^{amide} \rangle$, (b) average number of water-amide hydrogen bonds from B23 to B26 $\langle n_{HB}^{water} \rangle$, (c) average RMSD of the heavy atoms with respect to the crystal structure, (d) average number of inter-monomer contacts $\langle n_{MM} \rangle$, (e) average number of α-contact $\langle n_\alpha \rangle$, and (f) number of β-contact $\langle n_\beta \rangle$. The average was computed over the structures within the same Markov state with equal weight.

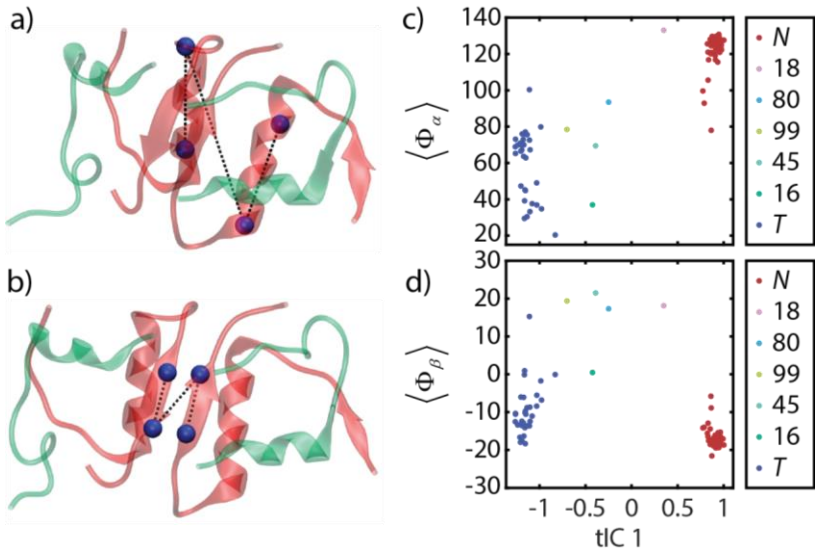

Figure 8A.2: Structural characterization of the MSM including pseudo-dihedral angles of α-helices Φ_α and β-sheet Φ_β. (a) Illustration of the α pseudo-dihedral angle. The blue spheres represent the centers of mass (COMs) used for defining the pseudo dihedral angle. (b) Illustration of the β pseudo-dihedral angle. (c–d) Distribution of pseudo-dihedral angles along tIC1.

α contact	β contact
B9 (M1) – B13 (M2)	B24 (M1) – B24 (M2)
B9 (M1) – B16 (M2)	B24 (M1) – B25 (M2)
B12 (M1) – B16 (M2)	B24 (M1) – B26 (M2)
B13 (M1) – B9 (M2)	B25 (M1) – B24 (M2)
B13 (M1) – B13 (M2)	B25 (M1) – B25 (M2)
B16 (M1) – B9 (M2)	B25 (M1) – B26 (M2)
B16 (M1)– B12 (M2)	B26 (M1) – B24 (M2)
	B26 (M1) – B25 (M2)
	B26 (M1)– B26 (M2)

Table 8A.1: Pairs of C_α atoms for calculating number of contacts. M1 and M2 represents the monomer 1 and 2 in the dimer structure.

To illustrate relevant global changes of different dimer structures, we also defined a set of pseudo dihedral angles shown in Figure 8A.2, including α pseudo-dihedral angle Φ_α, and β pseudo-dihedral angle Φ_β. These dihedral angles were defined using centers of mass (COMs). α pseudo-dihedral angle was computed using COMs of the backbone atoms in B9–B11 of M1, the backbone atoms in B17–B19 of M1, the backbone atoms in B17–B19 of M2, and the backbone atoms in B9–

B11 of M2. Similarly, β-pseudo dihedral angle was calculated using COMs of the backbone atoms of B24 backbone atoms in M1, B26 backbone atoms in M1, B26 backbone atoms in M2, and B24 backbone atoms in M2.

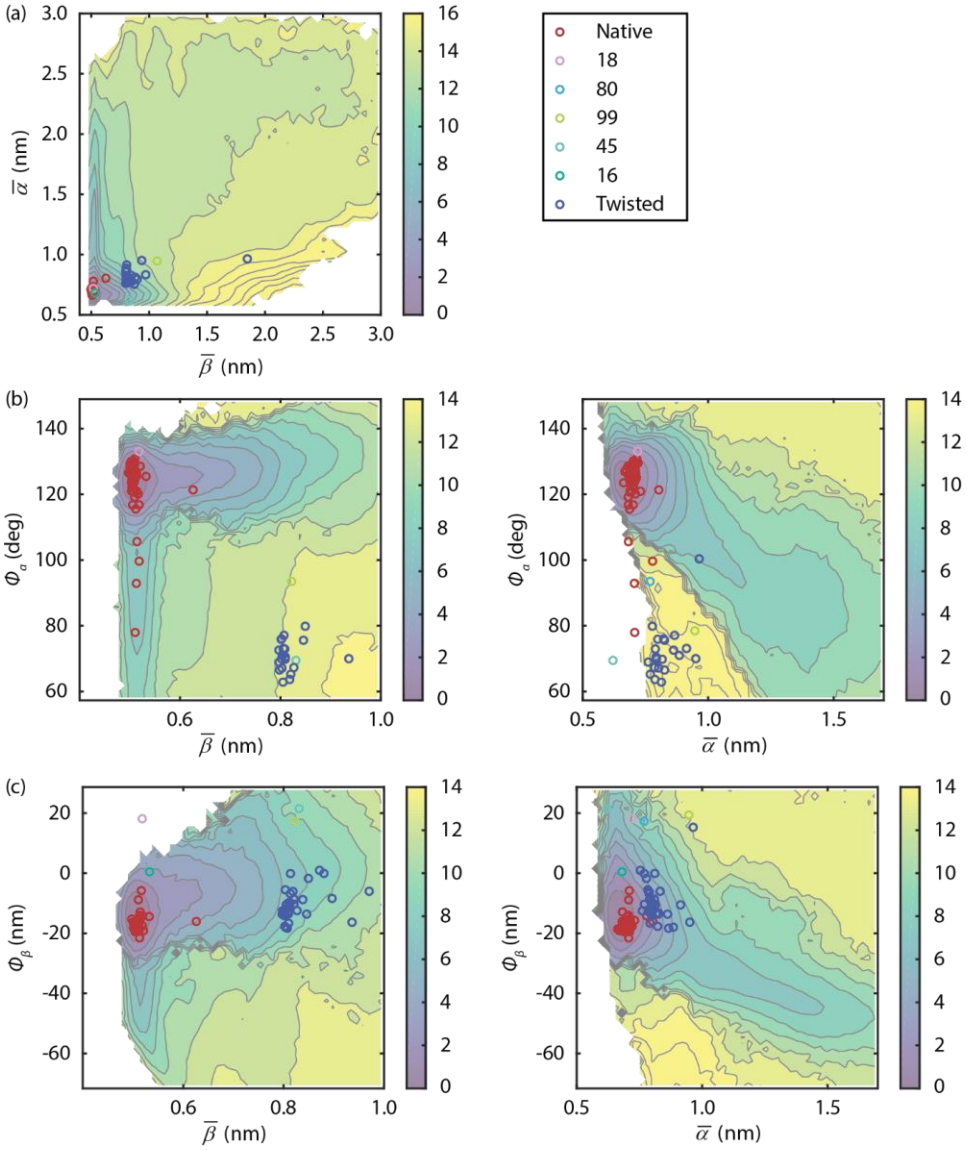

Figure 8A.3: Projection of dimer MSM onto potentials of mean force (PMFs) generated from sampling of dimer dissociation. (a) PMF as a function of average distance of α and β contacts. (b) PMF as function of α pseudo-dihedral angle, average β distance (left), and average α distance (right). (c) PMF as function of β pseudo-dihedral angle, average β distance (left), and average α distance (right).

Appendix 8B: Twelve-State Lumping of Dimer MSM

In addition to identifying hub or intermediate states, the top 20 tICs and eigenvectors of the transition matrix indicate that slower kinetics are associated with subgroups of the native and twisted states which correspond to clustered groups in our network plot. As a result, we investigated a reduction of the full MSM state space using a Robust Perron Cluster Cluster Analysis (PCCA+) of the first 20 eigenvectors of the transition matrix.[94] We concluded that both structural, spectral, and kinetic variations can be usefully lumped into 12 states outlined in Figure 8B.1, including 4 states which lie within the native state (N_0, N_1, N_2, N_3), three in the twisted state (T_1, T_2, T_3), and 5 single Markov states from the 100-state MSM. The corresponding network plot for the 12-state lumping is shown in Figure 8B.2.

The N_1 state has 16% of the population and with a backbone that closely resembles the crystal structure and well-folded secondary structural elements, whereas the N_2 states (21%) have the A1 helix unfolded in one insulin monomer, and are a relatively isolated block of the transition matrix. N_3 are the most populated (29%) and these are states that are typically traversed in converting between native and twisted states. Almost all N_3 conformations retain the intermolecular α and β contacts of the crystal structure, but have one or both A1 helices unfolded in both monomers, with considerable conformational disorder for the termini of all chains. The T_1 state represents most of the twisted state population (20%) and most configurations are well-folded low disorder structures, whereas the T_2 and T_3 states are kinetically separated and generally have more disordered A chains. T_3 retains the twisted dimer structures but the configuration of the A chain differs from T_1 and is disordered.

The network plot readily identifies the intermediate states 80 and 45 as on-pathway intermediates in the conversion between native and twisted forms, with most of the flux passing

between N_3 and T_1. The states 99, 16, 18, and T_3 are observed to be off-pathway intermediates or kinetic traps. Since states 80 and 45 have water intercalated between β-strands on each monomer, we conclude that the primary mechanism of dimer twisting involves water disrupting the specific interactions of the β-sheet without significant disruption to the hydrophobic core. The remaining weak contacts between hydrophobic sidechains of the B chain helix provide the orientational flexibility to reconfigure the β-strand sidechains and contacts in its new configuration.

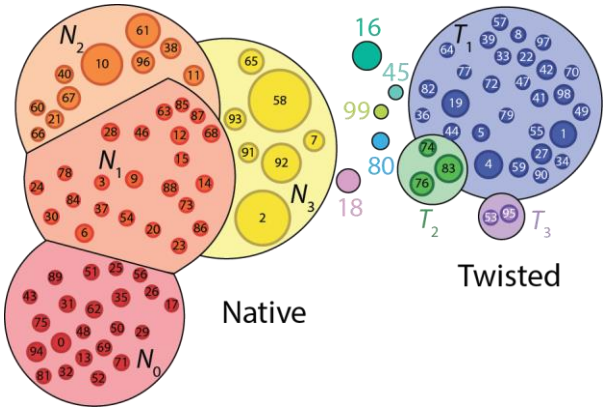

Figure 8B.1: Markov State Model network plot of insulin dimer with state index.

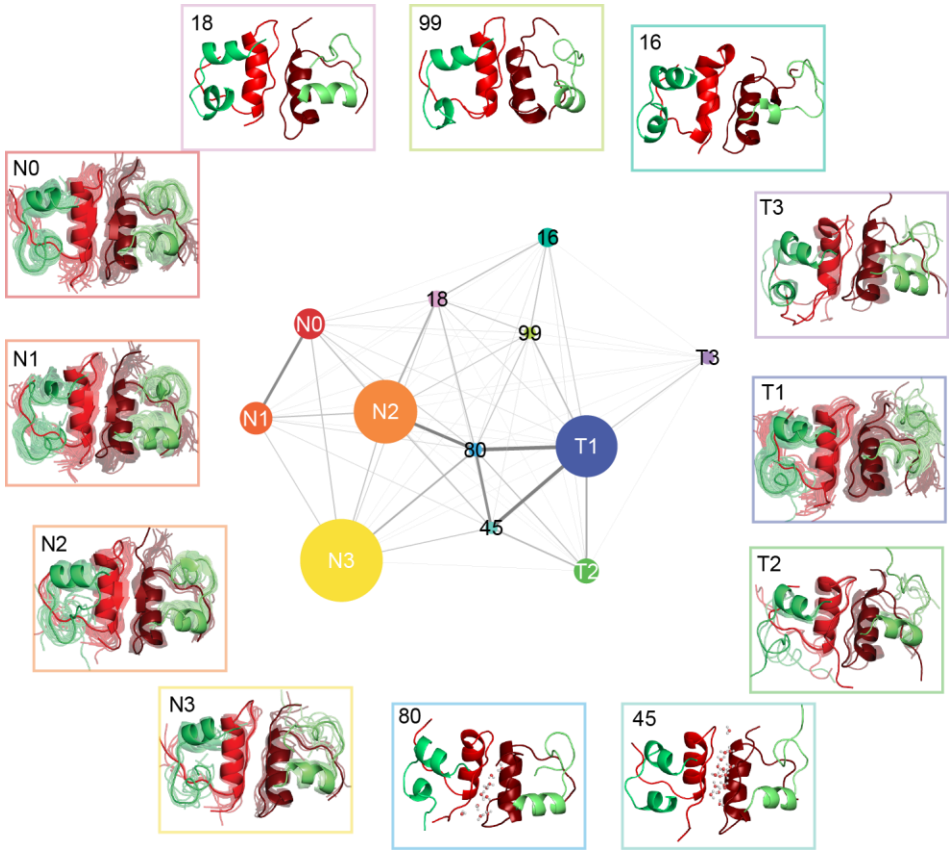

Figure 8B.2: Network plot of reduced 12-state model. Medoids of each reduced state are shown on the side.

Chapter 9

Structural Heterogeneity of Insulin Dimer in Aqueous Solution Probed by Isotope-Edited IR Spectroscopy and Computational Spectroscopy

9.1 Abstract

In the previous chapter, Markov State Model (MSM) of insulin dimer predicted two conformational states of the native dimer and the twisted dimer, which show distinct spectroscopic response with site-specific $^{13}C^{18}O$ labels on the dimer interface. Here, we present an experimental study on conformational characterization of human insulin dimer using site-specific isotope-edited IR and 2D IR spectroscopy. From $^{13}C^{18}O$ labels on the inter-monomer β-sheet, we find an additional spectroscopic species other than the response from the native β-sheet at high ionic strength. Also, the population of these species can be tuned by changing the ionic strength. Character of the cation in solution seems to perturb the equilibrium at high ionic strength, but the mechanism is unclear. The additional spectroscopic species is consistent with the predictions from the twisted dimer in the dimer MSM, which breaks the native hydrogen-bond contacts of the β-sheet. The presence of the twisted conformation suggests a potential kinetic intermediate along the homodimer association and/or multi-pathway nature of the association. This study provides

additional insight on the conformational distribution of dimer and establishes a refined molecular picture of describing coupled folding and binding process in insulin monomer association.

9.2 Introduction

Protein recognition, association, and dissociation are part of the essential processes in biology. Even though the simplest possible protein association events involve little structural rearrangement and conformational changes of each protein during the binding, proteins often exhibit a variety of distinct conformers or conformational disorder, allowing themselves to interconvert between free energy basins on a complex energy landscape or to fluctuate between thermally accessible conformations.[1-3] In the context of protein associations, such conformational heterogeneity can lead to coupled folding and binding processes, which are dynamical interplays between folding to the functional conformation, desolvation around the binding interface, and formation of transient nonnative contacts and eventual native contacts between proteins.[4-10]

It is experimentally challenging to characterize dynamics of coupled folding and binding processes since protein conformational dynamics is often multi-pathway in nature and inherently requires ensemble descriptions for both equilibrium protein conformations and association paths with high structural resolution.[11-16] The coupled dynamic nature between fast conformational fluctuations and protein association also means that the associated time scales span over many decades, which is prohibitive for traditional structural and kinetic tools.[17-19] For instance, structural determination using chemical shifts and J-couplings in NMR spectroscopy is limited by the coalescence timescale of ms such that faster conformational dynamics and association events are averaged.[20] Infrared (IR) and two-dimensional (2D) IR spectroscopies probe molecular vibrations sensitive to underlying structures with fs–ps time resolution, which has been applied for conformational characterizations on peptides and proteins.[21-28] In addition, conformational

dynamics over the course of dissociation and unfolding can be probed using transient IR spectroscopy synchronized with optically triggered temperature-jump (T-jump) laser ranging from ns to ms.[29-35]

Insulin dimerization is one of the protein to study coupled folding and binding processes with distinct secondary structural changes. Insulin monomer is a 51-residue peptide with 21 residues on chain A and 30 residues on chain B. It has α-helical segments and a β-turn located from B20Gly to B23Gly, and the B-chain C-terminus ranging from B24Phe–B30Thr is known to be disordered with the extent depending on mutations and solution environment.[36-43] This intrinsically disordered region folds and binds into a well-defined inter-monomer anti-parallel β-sheet as in the crystal dimer structure (Fig. 9.1a), which includes the aromatic triplet of B24Phe, B25Phe, and B26Tyr.[43-46] This B-chain C-terminus is also involved in the insulin receptor binding, exhibiting detachment of the B-chain C-terminus, a significant dihedral rotation on B24Phe, and hinging motion upon recognition with the insulin receptor,[47-50] which may share similar conformational transitions as in the dimerization.[51]

Recent experimental advances also suggest that insulin dimer dissociation follows coupled unfolding and unbinding dynamics. Amide I spectroscopy of bovine insulin, which probes C=O stretch vibration of the backbone amide group, showed that it is spectroscopically sensitive to thermal dissociation of β-sheet and therefore dimer-monomer transition.[46] T-jump 2D IR spectroscopy measuring thermal dissociation showed an exponential kinetics of β-sheet dissociation between 250 and 1000 μs, and additional non-exponential kinetics on the order 5–150 μs, which was attributed to conformational disordering of the monomers in the dimeric state.[52] Complimentary T-jump X-ray solution scattering spectroscopy of insulin revealed additional kinetic processes assigned to intermediate dimer states with conserved secondary structures prior

to the β-sheet dissociation.[53] These experiments provide qualitative pictures of coupled unfolding and unbinding dynamics, but limited structural resolution due to short vibrational lifetime, the nature of coupled vibrations, and broad scattering profile inhibits detailed structural interpretation of the dissociation process.

The issue of limited structural information can be mitigated from isotope-labeling on amide groups. Introducing site-specific isotope-labeling such as $^{13}C^{18}O$ on the amide group provides additional frequency shift of ~65 cm^{-1} and isolates the corresponding C=O vibration from a congested spectrum. Recently, an effective chemical synthesis utilizing chemical ligation has been developed to incorporate site-specific $^{13}C^{18}O$ label in human insulin, and isotope-edited IR spectrum of B24Phe-labeled insulin demonstrated the sensitivity of β-sheet dissociation and dimer-monomer transition.[54-55] Also, Raman spectroscopy of ^{13}C-label on both B24Phe and B25Phe amide units has been measured to probe site-specific spectroscopic feature on the β-sheet.[56] Site-specific isotope-labeling on human insulin is now feasible and provides the capability of dissecting dissociation kinetics with structural specificity.

Computational studies on insulin dimerization offer a detailed atomistic picture and serves as physical basis and proposals of understanding the dimer dissociation. Free energy simulations between dimer and monomer states identified specific residues for stabilizing the dimerization, including the aromatic triplet, B16Tyr on the interfacial α-helix, B23Gly on the β-turn and B28Pro interacting with the β-turn.[57-58] Metadynamics simulation from Bagchi and co-workers proposed a series of dimeric intermediates along the minimum energy path on the computed free energy surface, and these structures display loss of native inter-monomer contacts before the two monomers separate apart.[59-60] Dinner and co-workers employed multiple enhanced sampling methods to characterize the dimer association using interfacial contacts, and proposed multi-

pathway dissociation processes.[51] One limiting pathway (α-pathway) of the dissociation involves solvation on the interfacial α-helices and the subsequent detachment of the B-chain C-terminus, which is also appearing in the receptor binding process, whereas the other limiting pathway (β-pathway) solvates the interfacial β-strands and exhibits little detachment. The solvation of these interfacial residues were proposed to correlate with spectroscopic changes on isotope-edited IR spectroscopy, serving as structure-based models that can be experimentally tested.[51, 61]

Direct comparison of protein structures with IR experiments is now possible using computational amide I spectroscopy. This method can be used to predict linear IR and 2D IR spectra for simulated conformational distributions drawn from MD trajectories or Markov State Models (MSMs), providing a route to investigate structure-spectrum correlation.[62] Specifically, amide I spectroscopic maps predict amide I vibrational frequencies to high accuracy using local electrostatics calculated from MD simulations such as electrostatic potential or electric field at the site of interest.[63-70] Maps for vibrational coupling between different amide I vibrations are used to calculate the interaction of multiple backbone amide groups.[71-74] These maps have reached the point of predicting amide I spectroscopic observables to a high level of accuracy with 2 cm^{-1} frequency uncertainty and provided a direct way to structurally interpret experimental IR spectra.[69] Also, the idea of interpreting isotope-edited IR spectroscopy using computational spectroscopy with MD simulations and MSMs has been tested and applied on various proteins, such as characterizing equilibrium conformational distribution of TrpZip2,[24] structural disorder of NTL9,[25] investigating permeation mechanism of potassium ion in KcsA.[28]

Insulin dimerization studies have focused almost entirely on the conformational characterization in the monomeric state, thermodynamics of dimer dissociation, and increasingly on the dynamics of dimer dissociation. The current viewpoint is that the starting dimer structure

resembles crystal structure of the dimer (Fig. 9.1), which may be justified from the solution structure of the B9Asp mutant.[45] However, recent simulation study on a mutant dimer, showed that mutation of B24Phe to Gly resulted in additional dimer conformations including strongly interacting dimer and weakly interacting dimer, which involves conformational change between B10His and B13Glu, and increased solvation of the dimer interface.[75] The detailed dimer conformational distribution still requires investigation, in particular accounting for changes of solvation environment such as pH and ionic strength that can mediate dimer conformational changes and in turn dynamics of dimer dissociation and association.[76-77]

In Chapter 8, we presented a detailed study on the conformational characterization of insulin dimer using atomistic Markov State Model (MSM), and predicted amide I spectra of various $^{13}C^{18}O$ site-specific isotope labels using computational spectroscopy. The dimer MSM shows two dominant conformations: A native (*N*) dimer resembling experimental structures of the dimer, and a twisted (*T*) dimer exhibiting ~55° rotation of the native dimer state. Such rotation is coupled to the register-shift of the β-sheet, resulting in breakage of native hydrogen-bond contacts in the β-sheet, and the reorganization of the side-chain contacts. Computational spectroscopy of the MSM indicates that the unlabeled (UL) spectrum may not be the most sensitive IR measurement to distinguish these two conformations. Instead, native dimer and twisted dimer can be effectively identified using a series of site-specific $^{13}C^{18}O$ labels on the dimer interface, including B24F single label, B24FB25F dual labels, B23GB26Y dual labels, *etc*. Important questions regarding the twisted dimer includes: (1) if the twisted dimer can be observed experimentally, (2) what experimental variables are responsible for the dimer conformational transition, and (3) what the role of the twisted dimer is on the homodimer dissociation.

In this study, we present a joint experimental-computational study to characterize conformational ensemble of insulin dimer in aqueous solution using site-specific isotope-edited IR spectroscopy, Markov State Models (MSMs), and computational amide I spectroscopy. Isotope-edited IR spectroscopy probes local structural details around the β-sheet residues, allowing us to characterize conformational changes around the β-sheet. MSM of insulin dimer provides a high-resolution structural basis with associated interconversion rates, which can be directly compared with experimental spectra in conjunction with computational amide I spectroscopy. These tools help us investigate the conformational ensemble of insulin dimer, identify potential experimental variables perturbing the conformational equilibrium, and refine our pictures of dimer dissociation dynamics.

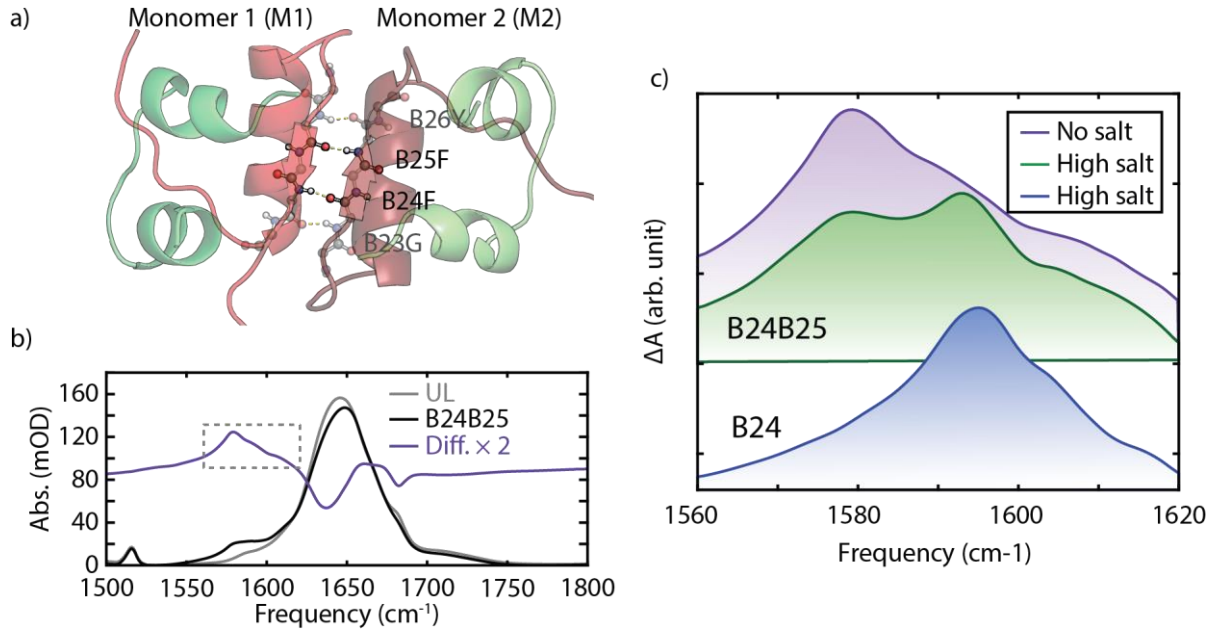

Figure 9.1: (a) Crystal structure of human insulin dimer (PDB: 3W7Y). Two chains A and B of each monomer are colored by green and red, respectively. Inter-monomer β-sheet (red arrows) in the structure is formed by 3–4 hydrogen bonds (HBs) from B24 to B26, indicated by yellow-dashed lines. Site-specific $^{13}C^{18}O$ isotope labels on the amide groups in this study are labeled black, including B24 single label and B24B25 dual labels. (b) FTIR spectra of unlabeled (UL) human insulin (gray), B24B25 labeled human insulin (black), and difference spectrum between B24B25-labeled insulin and UL insulin (purple). The spectra are taken at 3 °C in the no-salt condition. (c) Isotope-edited difference spectra of B24B25 dual labels in the no-salt condition (purple), the high-salt condition (green), and B24 single label in the high-salt condition (blue). The spectra are also taken at 3 °C.

9.3 Materials and Methods

In the previous chapters, Chapter 4 describes in detail about the sample characterization and the sample preparation, and Chapter 8 presents the structural descriptions and the spectroscopic predictions from the dimer MSM, which will be used in the following sections. For brevity, here I only describe the experimental sample preparation.

9.3.1 Unlabeled Human Insulin Samples and Synthesis of Isotope-Edited Human Insulin

Unlabeled (UL) zinc-free human insulin can be purchased from Sigma-Aldrich (91077C) The potency of the sample is ≥ 27.5 insulin unit (IU)/mg. Each IU corresponds to 0.0347 mg of human insulin based on WHO, which translates the potency to ≥ 95.4% weight percentage purity. The impurity has ≤ 1.0 endotoxin unit/mL endotoxin.

For site-specific $^{13}C^{18}O$ isotopic substitution of WT human insulin, 1-^{13}C Phenylanaline (99% ^{13}C-enrichment) was purchased from Cambridge Isotope Laboratories, and isotopically exchanged with ^{18}O-water in the presence of dry hydrochloric acid at 100 °C up to 97% ^{18}O enrichment. Isotopically labelled insulins were synthesized following our reported protocol.[54] The strategy to make the insulin involves one-pot sequential native chemical ligation of three unprotected peptides to get the full-length ester-insulin polypeptide,[78] folding and saponification as the key steps. Ester insulin - a Glu^{A4}-Thr^{B30} side-chain linked single chain insulin molecule was used as the key intermediate in generating native insulin.[79] More details of the synthesis are described in the Chapter 4.

9.3.2 Sample Preparation and Spectroscopic Measurements

To avoid spectral overlap between amide I vibration and water bend vibration, H–D exchange was performed by dissolving insulin in 10 mM DCl/D_2O solution and heating the solution at 60 °C for 30 minutes. The concentration of insulin for H–D exchange was prepared at ~1 mg/mL and checked on a Nanodrop UV/Vis spectrometer (Thermo Scientific). We found concentration values determined based on absorption at 276 nm and at 280 nm are consistent with each other within our experimental precision.[80] The heated solution was then lyophilized to remove excess DCl and D_2O. For IR measurements, low pH solution was required to protonate the COO^-

group to shift its carbonyl vibration to ~1720 cm^{-1}. The lyophilized insulin samples were then dissolved to the final concentration of 10 ± 0.3 mg/mL, or equivalently ~1.7 ± 0.1 mM. In this study, two primary solution conditions are used. First condition is 10 mM DCl/D_2O solution (no-salt condition, pH* = 1.7), which was chosen based on the objective of mitigating irreversible fibril formation due to high ionic strength[81] while maintaining at low pH to avoiding spectral overlap between $^{13}C^{18}O$ amide I vibrations and COO^- asymmetric stretch from sidechains.[82] The other condition of 270 mM DCl/100 NaCl/D_2O (high-salt condition, pH* = 0.5) was chosen to compare with previous studies on bovine insulin.[46, 52-53] To study ionic strength dependence on thermodynamics, the H–D exchanged insulin samples were dissolved in series of 0, 10, 20, 50, 100, 200, 360 mM NaCl in 10 mM DCl/D_2O solution. More practical details, setup of Temperature-dependent IR spectroscopy and 2D IR spectroscopy are described in Chapter 4.

9.4 Results

9.4.1 Infrared Spectroscopy of Human Insulin and Spectral Assignment

To characterize the conformational changes accompanying the dimer-monomer transition, IR spectra were acquired on various isotopologues of wild-type (WT) human insulin, including the natural unlabeled (UL) form of insulin, a singly $^{13}C^{18}O$-labeled insulin on the B24 amide unit (connecting the B24 carbonyl and B25 N–H) and a doubly $^{13}C^{18}O$-labeled insulin on both B24 and B25 amide units (Fig. 9.1a). These isotope labels probe specific local structural contacts within the β-sheet residues at the dimer interface, and are sensitive to the dimer-monomer transition.[54, 56] Also, the predictions from the dimer MSM show that these labels are capable of distinguishing the dimer conformations if any in the experiment. As an illustration, Fig. 9.1b shows FTIR spectra of UL insulin and B24B25-labeled insulin in the no-salt condition (10 mM DCl/D_2O) that favors the dimer structure. The amide I vibrations of UL insulin dimer between 1600–1700 cm^{-1} have a peak frequency of 1646 cm^{-1} and a shoulder at 1682 cm^{-1} that are assigned to the $\nu_{\parallel}$ mode of the dimer's β-sheet arising from dipolar interactions between its hydrogen-bonded amide I vibrations.[46, 52, 83] Other features in this spectrum result from side-chain vibrations. Upon isotopic substitution, additional isotope-edited features appear between 1550–1620 cm^{-1}. The difference spectrum between B24B25-labeled insulin and UL insulin (Fig. 9.1b) shows that the isotope-labeled vibrations have a peak frequency at 1578 cm^{-1}, whereas the intensity losses are observed at 1637 cm^{-1} and 1682 cm^{-1}, which are assigned to the $\nu_{\perp}$ and $\nu_{\parallel}$ modes of the β-sheet.[46, 52, 83] This demonstrates how isotopic substitution helps dissect spectral features and isolate specific contacts within the dimer. Measuring difference spectra between $^{13}C^{18}O$-labeled and UL samples has the additional advantage of removing interference that may arise from Arg vibrations in the 1580-1600 cm^{-1} region.

Fig. 9.1c shows isotope-edited difference spectra of B24B25 dual labels in both 10 mM DCl and 100 mM NaCl/270 mM DCl/D_2O at low temperature (3 °C). The B24B25-labeled spectrum in 10 mM DCl shows a peak centered at 1578 cm^{-1} and a subtle shoulder centered around 1610 cm^{-1}, corresponding to $\nu_{\perp}$ and $\nu_{\parallel}$ vibrations of the isotopically-isolated β-sheet,[83-85] provided that the $^{13}C^{18}O$ isotope label gives additional 60–70 cm^{-1} red shift.[22, 86] There is an additional shoulder appearing around 1590 cm^{-1}. In contrast, the B24B25-labeled spectrum in 100 mM NaCl/270 mM DCl has suppressed intensity of the peaks at 1578 cm^{-1} and 1610 cm^{-1} and an additional peak centered at 1593 cm^{-1}. By changing solution condition such as ionic strength or pH, the isotope-edited spectral responses on both B24 and B25 amide units become distinctly different, suggesting that there is an additional spectroscopic species appearing in 100 mM NaCl/270 mM DCl. Singly B24 $^{13}C^{18}O$-labeled spectrum in the high-salt condition exhibits a peak frequency centered at 1597 cm^{-1}, and a subtle shoulder around 1605 cm^{-1}, consistent with the previous isotope-edited IR spectrum in 20% ethanol at pD 2, which is known to have dominated dimer species in this condition.[54] The existence of the shoulder around 1605 cm^{-1} is also consistent with the presence of additional spectroscopic species that suggests a heterogeneous conformational ensemble instead of a single dominant native structure.

Distinct spectral changes of B24B25-labeled spectra between both conditions at low temperature suggest possibilities of either perturbing dimer-monomer equilibrium due to ionic strength/character of ions/pH, having additional conformations in the dimer or the monomer state, or both of the above. Fig. 4.9 shows that two-state thermodynamics is unaffected between these two conditions within experimental resolution, and Fig. 5.5 shows that dimer is the dominant species at low temperature. Additionally, there are no titratable groups around pH* 1.7,[87] also shown in Fig. 4.7. Therefore, the perturbing experimental variables are most likely related to ions,

including ionic strength and possibly character of ions. For convenience, the condition of 10 mM DCl is called the no-salt condition while the condition of 100 mM NaCl/270 mM DCl is called the high-salt condition. Since the monomer population is almost negligible, the additional spectroscopic species should be dimer-specific, suggesting the high-salt condition leads to a mixture of dimer conformations.

The additional dimer species may be related to the twisted state in the dimer MSM presented in Chapter 8. MSM of the dimer reveals the conformational heterogeneity, including the native (*N*) dimer appearing to be similar to the crystal structure as in Fig. 9.1, and the twisted (*T*) state exhibiting a relative rotation of the monomers by 55 °, reorganization of the side-chain packing, and a register shift along the β-strands that disrupts the native hydrogen bond contacts of the β-sheet (Fig. 8.2). Conformation transition from the *N* state to the *T* state is predicted to change the isotope-edited IR spectra. B24 singly-labeled spectrum shows a frequency red-shift from 1602 cm^{-1} to 1596 cm^{-1}, and B24B25 dual labels appear to have loss of the peak frequency at 1580 cm^{-1} and a shoulder at 1605 cm^{-1}, with instead additional symmetric peak centered at 1596 cm^{-1} (Fig. 8.7).

Spectral predictions from the dimer MSM can help us identify if the additional experimental dimer species originates from the twisted state. In Fig. 9.2, we performed head-to-head comparisons between experimental spectra in the high-salt condition and simulated spectra from the coarse-grained *N* and *T* state in the dimer MSM. Clearly, in the B24B25-labeled spectrum, the β-sheet modes at 1580 cm^{-1} and 1605 cm^{-1} matches the native dimer spectrum, and the additional peak centered at 1593 cm^{-1} can be explained by the twisted dimer spectrum albeit 3 cm^{-1} frequency discrepancy, comparable to the frequency uncertainty of 2.25 cm^{-1}.[69] Also, the peak intensity suggests that the twisted dimer is the dominant species in the high-salt condition, whereas

the B24B25-labeled spectrum in the no-salt condition matches the native dimer spectrum, meaning that the native dimer is the dominant species in this condition. Comparing the B24-labeled spectra shows that the native dimer accounts for the 1602 cm^{-1} shoulder while the twisted dimer matches the peak centered at 1596 cm^{-1}, consistent with the observations found in the experimental B24B25-labeled spectrum. The native dimer and the twisted dimer in the dimer MSM can consistently describe both labeled spectra in the high-salt condition at low temperature, and the twisted dimer appears to be the dominant species.

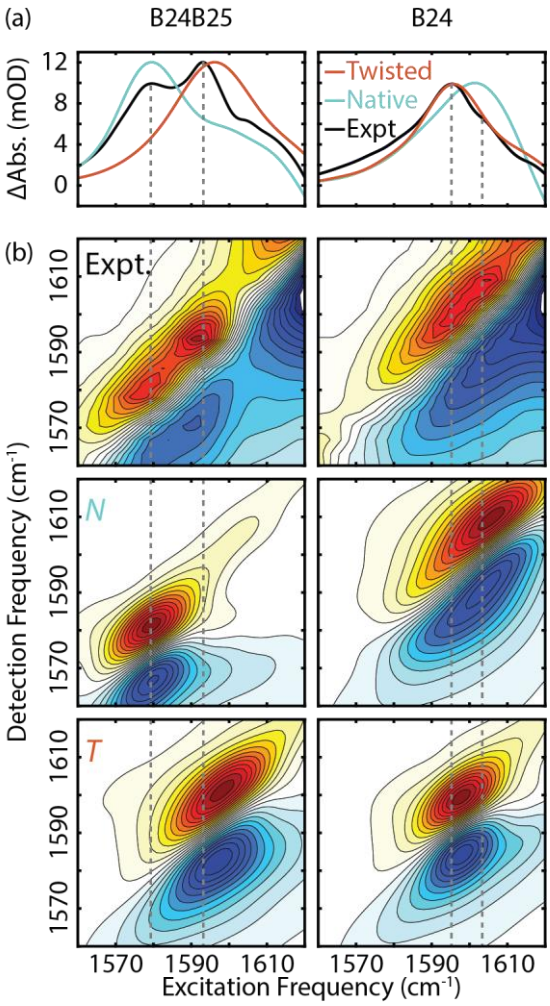

Figure 9.2: (a) Isotope-edited FTIR spectra of B24B25 dual labels and B24 single label from experiments under 100 mM NaCl/270 mM DCl (high-salt condition, black), the native state (*N*, cyan), and the twisted state (*T*, orange). (b) Parallel-polarized isotope-labeled 2D IR spectra from the experiment (top), the *N* state (middle), and the *T* state (bottom).

9.4.2 Thermodynamic Characterization of Human Insulin in the No-Salt Condition

Insulin in the high-salt condition shows a conformational mixture between the native dimer and the dominant twisted dimer. Previously in Section 5.3, we found the thermodynamics of insulin dimer dissociation in the no-salt condition can be treated as a simple two-state model between dimer and monomer.[46, 52, 88] Thermodynamically, the appearance of the twisted dimer

presents a couple questions: (1) How does the twisted dimer affect the thermodynamics of dimer dissociation? (2) How does the experimental variable such as ionic strength perturb the dimer equilibrium? As the first step to fully characterize the thermodynamics of dimer dissociation including the twisted dimer, thermodynamics characterization in the no-salt condition is performed as our reference, in which the native dimer is the dominant species.

Fig. 9.3a shows temperature-dependent FTIR spectra of UL insulin in the no-salt condition across the entire accessible temperature. Increasing temperature results in a blue shift of the amide I peak frequency from 1646 cm^{-1} to 1651 cm^{-1}, disappearance of the shoulder at 1682 cm^{-1}, and slight broadening of the amide I feature. These spectral changes are associated with the β-sheet vibrational modes (Fig. 9.1), and consistently seen in previous IR studies on bovine insulin with both increasing temperature and decreasing insulin concentration,[46, 52] which increases the monomer population. These spectral changes report primarily on the dimer-monomer transition.

Isotope-edited spectroscopy can provide additional experimental evidence of probing the dimer-monomer equilibria. Temperature-dependent spectroscopic changes in B24B25-labeled spectra in both conditions (Fig. 9.3b) show an overall loss of intensity. High temperature spectra around 70 °C in both conditions show very similar spectra, having a broad vibrational profile with two peaks centered around 1582 cm^{-1} and 1589 cm^{-1}, suggesting that this isotope label is either reporting the same monomer state regardless of the conditions investigated, or insensitive to conformational variation in the monomer state upon changing the condition. The growth of intensity around 1620 cm^{-1} in the high-salt solution indicates irreversible insulin aggregation,[89] which is easier to happen at higher ionic strength described in Section 4.3. The twisted dimer peak at 1593 cm^{-1} disappears at high temperature, indicating that the additional spectroscopic species is also distinct from the high temperature monomeric species. In B24-labeled spectra, changes from

low temperature to high temperature also show an overall loss of intensity, and slight broadening, consistent with increasing solvent exposure from dimer state to monomer state.[54]

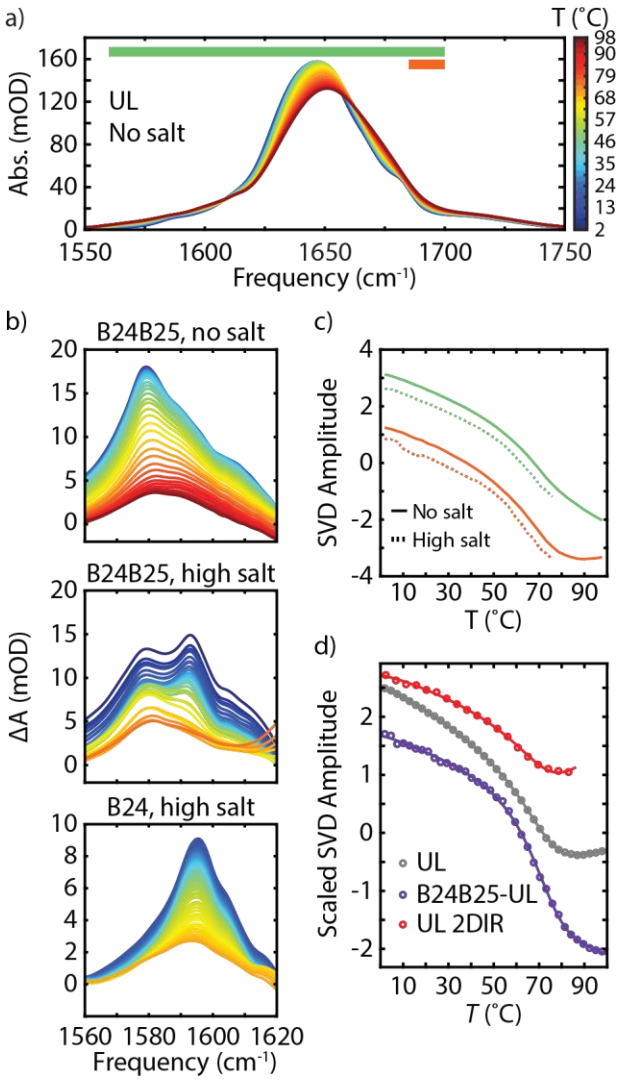

Figure 9.3: (a) Temperature-dependent IR spectra of UL insulin in the no-salt condition from 2 °C to 98 °C. (b) Temperature-dependent isotope-edited difference spectra of B24B25 in the no-salt condition (top), the high-salt condition (middle), and B24 in the high-salt condition (bottom). (c) Second SVD temperature component of UL spectra using the frequency range of 1560–1700 cm^{-1} (green), and 1670–1700 cm^{-1} (orange). Solid curve and dashed curve represent the components from the no-salt condition and the high-salt condition, respectively. (d) SVD temperature components of UL FTIR spectra (gray), difference spectra between B24B25 and UL, and UL 2D IR spectra (red).

To further characterize thermodynamics of the dimer-monomer equilibrium, dissociation curves are extracted from temperature-dependent FTIR spectra and shown in Fig. 9.3c. The second SVD component over the amide I frequency range as a proxy to optical dissociation curve captures the dimer-monomer transition, as demonstrated in Section 5.3, which can also be evaluated based on site-specific $^{13}C^{18}O$ labeled spectra (Fig. 9.1). As a comparison to previous studies on bovine insulin,[46, 52-53] we also included the curve extracted from UL insulin in the high-salt condition (dashed curves), with the spectra shown in Figure 9A.1. The dissociation curves were extracted using the frequency ranges of 1560–1700 cm^{-1} and 1670–1700 cm^{-1} to capture the whole spectral response in the amide I range and to probe more specific changes on the β-sheet that lies around 1682 cm^{-1}, respectively. The dissociation curves in both conditions can be overlaid perfectly regardless of chosen frequency ranges by tuning the vertical scaling factor and vertical offset, which would not change the shape of the curves. This indicates that the underlying two-state dissociation thermodynamics of both solution conditions are the same within our experimental resolution, which is also consistently shown by the ionic strength dependence in Fig. 4.9. Note that the accessible temperature range in the high-salt condition is narrower than in the no-salt condition because insulin is more prone to irreversible aggregation at high temperature with increasing ionic strength,[81] also confirmed by investigating ionic strength dependence in Fig. 4.9.

Fig. 9.3d shows the second SVD components in the no-salt condition from different measurements including the UL spectra, difference spectra between B24B25-labeled insulin and UL insulin, and parallel-polarized UL 2D IR spectra. Despite some difference on the sloping baselines at low temperature and high temperature, they all share similar inflection points in temperature, suggesting that all of the data together can be used to extract the thermodynamics.

A fairly common practice for obtaining the thermodynamics of dimer dissociation is to fit a thermodynamic dissociation curve from the second SVD component to a two-state model:

$$\mathrm{D} \overset{K_\mathrm{d}}{\rightleftharpoons} 2\mathrm{M} \quad K_\mathrm{d} = \frac{[\mathrm{M}]^2}{[\mathrm{D}]} \tag{9.1}$$

In practice, however, there are always undetermined baselines reflecting changes of solvation structures or other temperature-dependent processes such that one has to apply many-parameter fit to a single dissociation curve as seen in Fig. 9.3d, which may lead to over-fitting.[90] To mitigate the over-fitting issue due to additional baseline parameters, we instead performed a global fitting across multiple data at the same no-salt condition including UL FTIR spectra, B24B25-labeled FTIR spectra, B24B25–UL FTIR spectra, UL parallel-polarized 2D IR spectra, and UL perpendicular-polarized 2D IR spectra, with details described in the Appendix 9B. In this two-state global fitting, thermodynamic parameters are globally fit across all data set including mid-point dissociation temperature of the transition T_d, dissociation enthalpy at dissociation temperature $\Delta H°(T_\mathrm{d})$, and constant change of heat capacity $\Delta C_P°$, whereas the baseline parameters are fit in a case-by-case manner. The result of the global fit is shown in Fig. 9.4, with corresponding thermodynamic parameters summarized in Table 9.1. The resulting fit shows good agreement with the experimental data, and is insensitive to initial guesses including baseline parameters, suggesting a global minimum in the parameter space is reached.

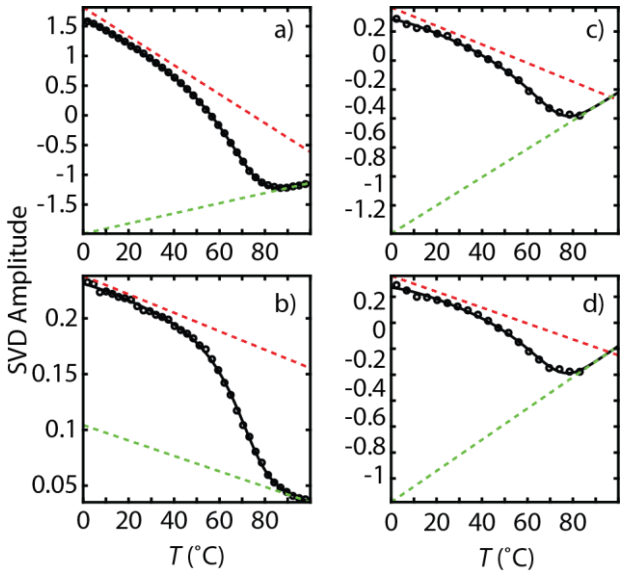

Figure 9.4: Second SVD temperature components from (a) UL FTIR spectra in Fig. 9.3a, (b) B24B25–UL FTIR spectra in Fig. 9.3b, (c) UL parallel-polarized 2D IR spectra in the no-salt condition from Figure 9A.2, (d) UL perpendicular-polarized 2D IR spectra in the no-salt condition solution from Figure 9A.2. Data points (black circles) are fit to the two-state model (black solid curve) globally across the whole data set. Dashed lines indicate dimer baseline (red) and monomer baseline (green) for illustration purpose. Black circles: Data points obtained by subtracting the baselines in (a). Black solid curve: Model prediction.

No-Salt Condition	Global Fit	Literature
T_d (°C)	67.6 (0.03)	68.7/67.6[91]
$\Delta H°(T_d)$ (kJ/mol)	184.9 (0.7)	206.7 (9.8)[91]
$\Delta C_P°$ (kJ/mol.K)	3.5 (0.03)	
θ_D(12 °C)	0.96 (0.02)	
θ_D(25 °C)	0.95 (0.02)	
θ_D(37 °C)	0.93 (0.02)	
K_d(25 °C) (μM)	7.7 (5.6)	8.81 (1.05)[92]
$\Delta G°$(300 K) (kJ/mol)	28.9 (1.7)	23.3[59]

Table 9.1: Thermodynamic two-state model parameters of dimer dissociation based on the global fit in Fig. 9.4, including dissociation temperature T_d, dissociation enthalpy $\Delta H°(T_d)$, change of heat capacity between dimer and monomer $\Delta C_P°$, dimer fraction θ_D at 12 °C, 25 °C and 37 °C, and dissociation constant at 25 °C. Standard deviation σ is indicated in the parenthesis, estimated by error propagation described in Subsection 5.2.4.

The two-state thermodynamic model is the simplest possible model, with the benefit of comparing to literature results. In short, reasonable agreements with literature are found. For

instance, dissociation temperature T_d that defines mid-point of dimer-monomer dissociation (θ_D = 0.5) is estimated to be 67.6 °C, consistent with the mid-point of 68.7 °C or 67.7 °C from differential scanning calorimetry (DSC).[91] Although dissociation-unfolding thermodynamic model has to be used to get good agreement in the DSC model fitting, our isotope-edited IR spectroscopy at high temperature does not seem to be sensitive to unfolding of monomers, which may reduce the dissociation-unfolding model back to the two-state dissociation. Also, two-state dissociation constant K_d at 25 °C is estimated to be 7.6 μM (Table 9.1 and Fig. 9.5a), close to the measured value of 8.81 μM from isothermal titration calorimetry experiment.[92] From this two-state model, we can estimate the temperature-dependent dimer fraction $\theta_D(T)$, which represents population of insulin in the thermodynamic dimer state. Fig. 9.5b and Table 9.1 show that population of the dimer remains over 90% across the temperature range from 12 °C to 37 °C, indicating that monomer cannot contribute to the drastic change of the B24B25-labeled spectra between both conditions at low temperature, which is consistent with the assignment that the additional species originates from the twisted dimer.

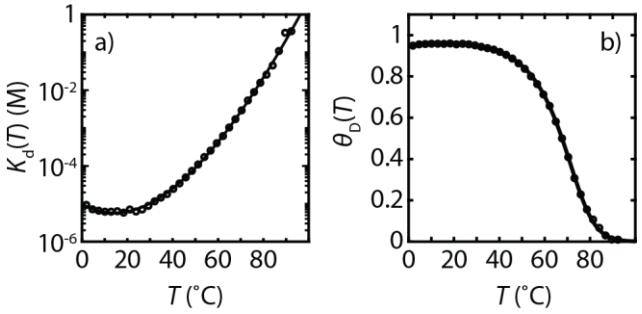

Figure 9.5: (a) Estimated dissociation constant as a function of temperature $K_d(T)$ from the two-state model. (f) Estimated dimer fraction as a function of temperature $\theta_D(T)$. Black circles represent the back-calculated values from the data points whereas the curves represent the model prediction.

9.4.3 Thermodynamics of Dimer Dissociation with Two Dimer Conformations

The additional spectroscopic species can be explained by the twisted dimer, which raises a question of how this additional dimer state influences the thermodynamics of dimer dissociation. The two simplest thermodynamic scheme to account for the additional dimer state is either the following sequential three-state model:

$$\mathrm{D}_1 \overset{K_\mathrm{u}}{\rightleftharpoons} \mathrm{D}_2 \overset{K_{\mathrm{d}2}}{\rightleftharpoons} 2\mathrm{M} \quad (\text{Sequential}) \tag{9.2}$$

or a parallel model that allows each dimer species dissociates into monomer given below

$$\begin{aligned} &\mathrm{D}_1 \overset{K_\mathrm{u}}{\rightleftharpoons} \mathrm{D}_2 \\ &\mathrm{D}_1 \overset{K_{\mathrm{d}1}}{\rightleftharpoons} 2\mathrm{M} \quad (\text{Parallel}) \\ &\mathrm{D}_2 \overset{K_{\mathrm{d}2}}{\rightleftharpoons} 2\mathrm{M} \end{aligned} \tag{9.3}$$

Both models have distinct processes of dimer unfolding and dimer dissociation, with the detailed descriptions in Section 5.2. Each process requires three thermodynamic parameters for the fitting including reference temperature, change of enthalpy at the reference temperature, and constant change of heat capacity. However, thermodynamics of the additional dissociation process from D_1 to 2M is not independent with the rest of the two processes. As a result, both models consist of six thermodynamic parameters against estimated fractions.

Thermodynamic characterization of insulin in the high-salt condition is challenging due to the irreversible aggregation discussed in Section 4.3, and a subtle 3 cm^{-1} blueshift of the UL spectra upon conformational changes from the *N* state to the *T* state (Fig. 8.5). Also, collecting high-quality temperature-dependent 2D IR spectra is challenging due to the same issues. To mitigate the issue

of aggregation and still get clean feature of the twisted dimer, we instead performed fast T-ramp FTIR experiment (Subsection 4.4.2) on B24B25-labeled insulin shown in Fig. 9.7a, which measures IR spectra on the fly of heating from 3 °C to 93 °C to characterize the thermodynamics of dimer dissociation with full temperature range. The temperature-dependent FTIR spectra of the B24B25 label shows qualitatively the same behavior as in Fig. 9.3b. Please note that because the data was taken on the fly of heating with fewer spectral averages (4 averages instead of 32 averages), it is extremely challenging to obtain a set of high-quality temperature-dependent difference spectra between B24B25 label and the UL and extract the quantitative fractions of all possible species.

To obtain thermodynamic fraction of each dimer and monomer species, we applied three-state maximum Entropy (MaxEnt) reconstruction on the temperature-dependent B24B25-labeled IR spectra shown in Fig. 9.6a using the first three SVD components present in these spectra, with the detailed description of MaxEnt reconstruction in Section 5.3. Note that there is an additional constraint (Eqn. (5.87)) that D_1 component resembles the low temperature spectrum in the no-salt condition, in which the native dimer is the dominant species. This MaxEnt reconstruction successfully isolates additional D_2 component that has a peak frequency of 1593 cm^{-1}, corresponding to the additional twisted dimer species appeared in Fig. 9.2, and a monomer component resembling the high temperature spectrum. Therefore, this 3-state MaxEnt reconstruction should be a reasonable approach for investigating the underlying thermodynamics of dimer unfolding and dissociation.

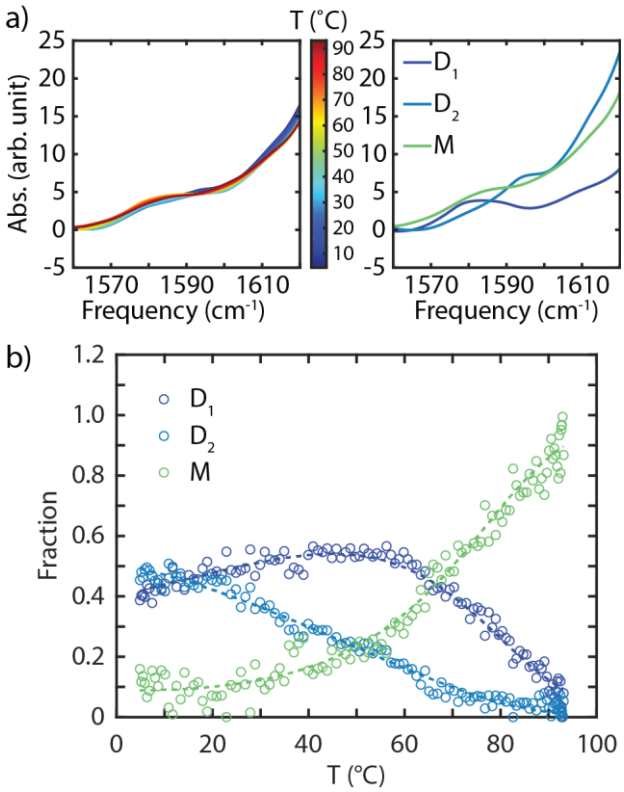

Figure 9.6: (a) Left: Fast T-ramp FTIR spectra of B24B25-labeled insulin in the high-salt condition. Right: MaxEnt spectral components indicating dimer components D_1, D_2, and monomer component M. (b) MaxEnt faction of D_1, D_2, and M as a function of temperature.

The MaxEnt reconstruction has an additional benefit of estimating the fraction of each species without complication of slanted baselines, allowing a direct comparison of population profiles and applying thermodynamic models. The corresponding fraction of each species is shown in Fig. 9.6b. At low temperature below ~15 °C, the twisted dimer (D_2) is indeed more populated than the native dimer (D_1), consistent with head-to-head comparisons in Fig. 9.2. Elevating temperature increases the population of D_1 while decreasing the population of D_2, indicating that the twisted dimer is more stable at lower temperature in this condition. Dissociation from D_1 or D_2 to M also shows distinct thermodynamic behaviors. Dissociation of D_1 has a mid-point at ~65 °C, consistent with the previous two-state thermodynamic model while dissociation of D_2 has

significantly lower mid-point temperature of ~50 °C, which is not resolved in the previous second SVD components.

The fitting results of both models are summarized in Table 9.2. One observation from these fit results is that the mid-point temperature between D_1 and 2M (T_{d1}) matches the dissociation temperature in the previous two-state model, meaning that these three-state models are still consistent with the simplified two-state model while providing more thermodynamic details regarding the additional dimer species. Furthermore, in the limit of negligible D_2 population as in the no-salt solution, both models should simply reduce to the two-state model such that the dissociation temperature should be the same across all of these models. The observation also suggests that the previous second SVD component is the most sensitive to the native dimer dissociation, which has the most spectral changes. The unfolding process gives the unfolding temperature of ~14 $^{\circ}$C, negative enthalpy and heat capacity changes, indicating that increasing temperature actually favors the folded (native) dimer species given this condition. On the other hand, thermodynamics alone cannot distinguish the underlying kinetic schemes since both models give almost identical fits within uncertainties. Additional kinetic experiments are required to resolve the kinetic interconversion between these species.

	Sequential Fit	Parallel Fit	Two-State Global Fit
T_{d1} (°C)	67.8 (N.A.)	67.9 (0.6)	67.6 (0.03)
$\Delta H_{d1}°(T_u)$ (kJ/mol)	97.1 (N.A.)	97.5 (3.0)	184.9 (0.7)
$\Delta C_{P,d1}°$ (kJ/mol.K)	1.7 (0.5)	1.7 (0.1)	3.5 (0.03)
T_{d2} (°C)	56.9 (1.3)	56.9 (1.3)	N.A.
$\Delta H_{d2}°(T_d)$ (kJ/mol)	103.6 (7.6)	104.9 (8.1)	N.A.
$\Delta C_{P,d2}°$ (kJ/mol.K)	1.9 (0.3)	2.0 (0.4)	N.A.
T_u (°C)	13.9 (5.6)	13.9 (N.A.)	N.A.
$\Delta H_u°(T_u)$ (kJ/mol)	-13.3 (9.1)	-12.7 (N.A.)	N.A.
$\Delta C_{P,u}°$ (kJ/mol.K)	-0.3 (0.4)	-0.3 (0.4)	N.A.

Table 9.2: Thermodynamic three-state model parameters of dimer unfolding and dissociation from the fit in Fig. 9.6, including midpoint temperature, change of enthalpy, and change of heat capacity. Subscripts of d1, d2, and u correspond to the dissociation from D_1 to 2M, the dissociation from D_2 to 2M, and the unfolding from D_1 to D_2, respectively.

9.4.4 Ionic Strength Dependence of the Dimer Conformational Equilibrium

Thermodynamic equilibria extracted from site-specific isotope-labeled spectroscopy can be perturbed using different solutions, in particular changing the ionic strength in this study. However, it is challenging to perform exhaustive thermodynamic characterization on precious isotope-edited samples. Rather, the experimental characterization is performed using UL insulin. Experimental UL FTIR spectra in both conditions show that the underlying dissociation curves (Fig. 9.3c) are sensitive to dissociation of the native dimer rather than the dimer equilibrium, indicating that the spectral change from dimer population shift is subtle in UL insulin. Model prediction of the UL spectra also only exhibits a 3 cm^{-1} blueshift from the *N* state to the *T* state (Fig. 8.5). Such a subtle spectral change is better seen using second derivative shown in Fig. 9.7a, that compares second derivatives of the UL FTIR spectra and B24B25 FTIR spectra under both conditions. The UL FTIR spectra have distinct spectral feature in the frequency range from 1630 cm^{-1} to 1650 cm^{-1} in response to the change of sample condition from the no-salt condition to the high-salt condition. This frequency range also exhibits a loss of intensity upon isotopic substitution on B24 and B25 amide units, suggesting that second derivatives in the UL

FTIR spectra within this range can be served as a proxy to underlying dimer conformational changes.

To verify if this frequency range can be used to probe underlying thermodynamics of dimer conformational change, we investigated the second derivative of UL insulin as a function of ionic strength in 10 mM DCl (Fig. 9.7b). With increasing ionic strength from 10 mM to 370 mM, the second derivatives have a clear trend of decreasing intensity at 1639 cm^{-1}, consistent with intensity loss of β-sheet $\nu_{\perp}$ mode, as well as increasing intensity at 1646 cm^{-1}. These two frequency slices exhibit a strong linear correlation with each other (R>0.9), suggesting that the slices can be used to extract populations of the native dimer and the twisted dimer. Interestingly, at the ionic strength of 370 mM, the intensities of both frequency slices are different between 360 mM NaCl/10 mM DCl and 100 mM NaCl/270 mM DCl (high-salt condition), suggesting that there is an additional cation dependence on the underlying thermodynamics rather than simply ionic strength without accounting for character of ions.

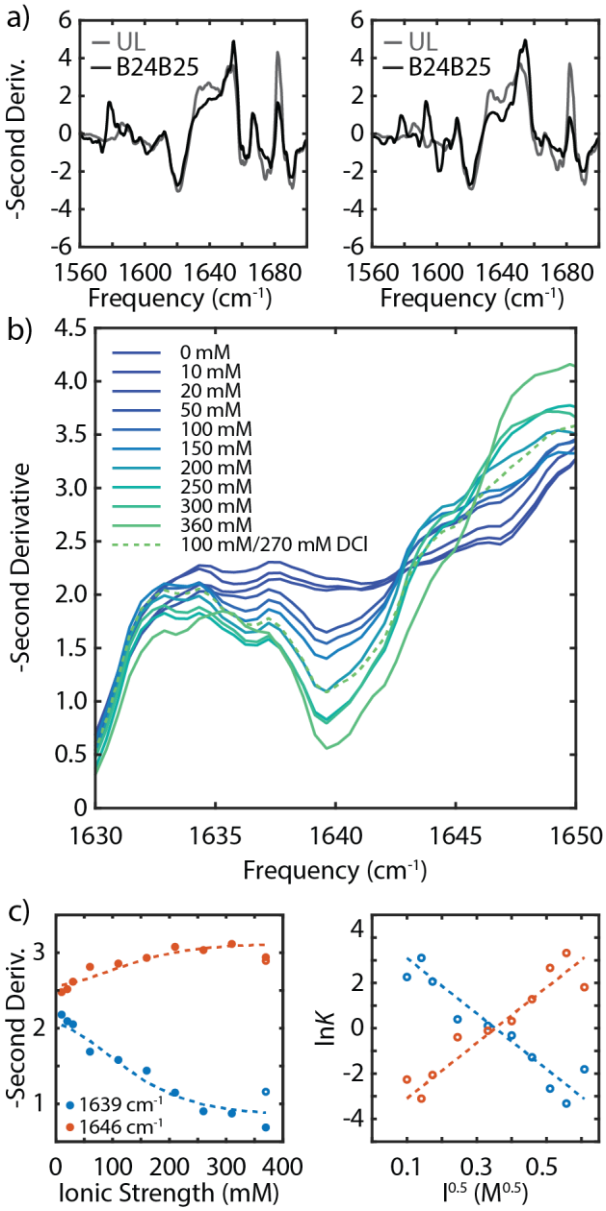

Figure 9.7: (a) Left: Second derivative of the FTIR spectra of UL insulin (gray) and B24B25-labeled insulin (black) in the no-salt condition. Right: Second derivative of the FTIR spectra of UL insulin (gray) and B24B25-labeled insulin (black) in the high-salt condition. (b) Second derivative of the UL FTIR spectra as a function of additional NaCl including 0, 10, 20, 50, 100, 150, 200, 250, 300, and 360 mM NaCl in 10 mM DCl. Dashed curve: Second derivative of the UL FTIR spectrum in the high-salt condition. (c) Left: Frequency slice of the second derivative as a function of ionic strength. Right: Calculated natural log of the equilibrium constant as a function of squared root of the ionic strength. Blue and orange represents the unfolding equilibrium constant and the folding constant, respectively. Open circles represent the amplitude of those slices in the high-salt condition.

To characterize the thermodynamics of dimer unfolding upon change of ion parameters, two-state dimer unfolding models were applied accounting for ionic strength *I*:

$$\mathrm{D}_1 \overset{K_\mathrm{u}}{\rightleftharpoons} D_2 \tag{9.4}$$

To apply model fit to the second derivative, we assume that the frequency slice at 1639 cm^{-1} is proportional to the population of native dimer whereas the slice at 1646 cm^{-1} is proportional to the population of twisted dimer. There are at least two types of thermodynamic models accounting for the ionic strength dependence, which exhibit different relationships between ionic strength and unfolding free energy. One is treating ions around proteins as effective electrostatic screening that shows the square root dependence of ionic strength:[93-95]

$$\Delta G_\mathrm{u}^\circ(I) = \Delta G_\mathrm{u}^\circ(I=0) + m\sqrt{I} \equiv \Delta G_{\mathrm{u},0}^\circ + m\sqrt{I} \tag{9.5}$$

while the other is treating ions analogous to chemical denaturants like urea, resulting in linear dependence of ionic strength:[96]

$$\Delta G_\mathrm{u}^\circ(I) = \Delta G_\mathrm{u}^\circ(I=0) + mI \equiv \Delta G_{\mathrm{u},0}^\circ + mI \tag{9.6}$$

The fitting parameters in both models consist of the free energy change at zero ionic strength $\Delta G_{\mathrm{u},0}°$ and the slope between unfolding free energy and square root of ionic strength/ionic strength, *m*. Details of both models are described in Section 5.2. There are 4 additional baseline parameters to translate from second derivative slices to the underlying population, but there is inter-dependence of these baseline parameters, resulting in only 1 indeterminate baseline parameter for the fit (See the Appendix 9C). Applying the fitting only requires 2 thermodynamic parameters and 1 baseline parameter. Examining the ionic strength dependence on dimer unfolding should help us identify the dominant factor perturbing the equilibrium of dimer conformational changes.

Fit result of the slices is shown in Fig. 9.7c. We found that the extracted free energy change is linearly correlated with square root of ionic strength instead of ionic strength, indicating that the dominant factor perturbing the dimer equilibrium is electrostatic screening instead of chemically denaturing the protein conformation that require specific ion-protein interactions Note that even both conditions of 360 mM NaCl/10 mM DCl and 100 mM NaCl/270 mM DCl have the same ionic strength, the frequency slices at 1639 cm^{-1} and at 1646 cm-1 shows that the population of both dimer species are different, which suggest that the dimer populations is not influenced simply by the ionic strength. Instead, the character of cations such as Na^+ and proton in this situation can also play a role on perturbing the dimer equilibrium. Charactering the full ion-dependence would require more detailed investigations on different cations, but qualitatively a key factor seems to be the ionic strength as electrostatic screening.

Table 9.3 summarizes the corresponding model parameters. The free energy change of unfolding at zero ionic strength is around 10 kJ/mol, indicating that native dimer is the predominant species. Given the value of *m* around -33 kJ/mol.$M^{0.5}$, increasing ionic strength favors the unfolding species or twisted dimer in this case, resulting in the mid-point of equal population around physiological ionic strength of 100 mM, suggesting that the twisted dimer may also be present at physiological condition.

	Two-State
$\Delta G_{u,0}°$ (kJ/mol)	10.1 (0.7)
m (kJ/mol.$M^{0.5}$)	-33.1 (1.9)
Mid-point *I* (mM)	92.1

Table 9.3: Model parameters of dimer unfolding based on the fit in Fig. 9.7, including the free energy change of unfolding at zero ionic strength, and the slope to the square root of ionic strength.

The additional thermodynamic behavior associated with twisted dimer was extracted using the MaxEnt reconstruction with 3 SVD components in B24B25-labeled FTIR spectra, indicating that the conformational changes are indeed subtle on the spectroscopy, and the site-specific isotope-labeling is able to extract this subtle information with distinct spectral changes. Another consistent rationale for having almost identical thermodynamic behaviors across a range of ionic strength (Fig. S4) is that the perturbation due to ionic strength does not affect the overall free energy difference between the dimer state and the monomer state in the two-state picture, which can be supported from our previous analysis on the MSM. The number of contacts across different dimer conformational states suggests that the change of inter-monomer contacts is minimal, and primarily from the change around the β-sheet. Also, the major contribution to the insulin dimerization comes from hydrophobic interactions.[43] So in principle the change on the free energy would not be large. Additionally, comparing the number of inter-monomer contacts to the same metric used in the study of Bagchi and co-workers,[59] we found that the conformational transition between the native dimer and the twisted dimer turns out to be within the same free energy basin in their study. Note that their simulation is performed at low ionic strength, which has consistently the predominant native conformation. Another interesting comparison on thermodynamics is that their prediction of free energy difference between dimer (state A) and monomer (state F) gives 23.3 kJ/mol, and the estimation from the thermodynamic two-state model gives 28.9. kJ/mol drawn from the no-salt condition, which is not drastically different given the chemical accuracy around 4 kJ/mol. Given these pieces together, a consistent physical picture can be proposed that perturbation of ionic strength tilts the local free energy landscape between the two dimer states toward the twisted dimer conformation, but the free energy difference between dimer and

monomer at the high-salt condition does not change much to be observed in the unlabeled experiments.

9.5 Discussions

We systematically investigated the conformational ensemble of insulin dimer using isotope-edited IR spectroscopy, all-atom MSM, and computational amide I spectroscopy. Temperature-dependent isotope-edited spectroscopy reveals the existence of additional spectroscopic species in the dimer state in the high-salt condition, and this conformational equilibrium can be perturbed using experimental variables associated with ions such as ionic strength. MSM of insulin dimer and computational spectroscopy provides a molecular interpretation of the twisted conformational state of dimer that contributes to the additional spectroscopic feature other than the β-sheet vibrations present in the native conformation. Integrating site-specific isotope-edited IR spectroscopy with high resolution structure-based model and computational spectroscopy provides detailed molecular insights on the conformational ensemble of insulin dimer.

This particular twisted dimer conformation has not been observed in previous experiments, and the existence of the twisted dimer identifies several avenues for interpretations on structural dynamics of coupled folding and binding process in insulin association, including the twisted conformation as a potential intermediate during association and dissociation, that the twisted dimer can appear on additional pathways such that the coupled folding and binding process in insulin dimerization exhibits a multi-pathway nature,[51] or that the twisted dimer is an off-pathway intermediate irrelevant to the dissociation of dimer into monomer and vice versa. Also, structural interpretations in this study rely heavily on accuracy of the structure-based model like MSM and

spectroscopic maps. We describe in the following our analysis on these aspects related to the conformational distribution of insulin dimer.

9.5.1 Discrepancy between Model and Experiments

From the comparison between structure-based model and experimental isotope-edited spectroscopy presented in Fig. 9.2, we found the MSM over-emphasizes population of the native conformation, which may be rooted to limited sampling and accuracy of the force fields (FFs). Assuming the sampling is reasonable, which can be justified with ~1.7 ms sampling and sufficient statistics along the conformational exchange between the *N* state and the *T* state, under-estimated population of the twisted dimer originates from a combination of protein FF and in particular ion FFs. Ionic strength is the experimental parameter controlling the appearance of additional spectroscopic species while the current FF for ions including Na^+ and Cl^- in this study is essentially point charges, which is not sufficient to describe proper intermolecular interaction strengths under a condition of somewhat higher ionic strength. As a result, even though relevant conformations may have been sampled, the underlying population distribution is most likely not accurate. However, it is difficult to identify exact issues regarding the origin of discrepancy, which requires higher-level simulations like polarizable FFs for ions and potentially proteins.

Another observation regarding accuracy of the model is 3 cm^{-1} frequency error on the peak frequency in the twisted dimer state between the simulation and experiment of site-specific isotope labeling on B24B25. Even though it is site-specific isotope labeling on B24 and B25 amide carbonyls, there are still in total of 4 labeled amide units spatially close to each other and separated only by a C_α atom. Therefore, frequency map, through-bond coupling map, and through-space coupling map all contribute to the observed frequency error, which is also difficult to pinpoint the

exact issue, not to mention that there are little experimental standards to address uncertainty of the coupling maps. Hence, further development of amide I coupling map against experimental standards may help clarify the origin of frequency error.[97-98]

9.5.2 The Ion Effect on Dimer Conformational Change

The equilibrium unfolding of the insulin dimer exhibits the apparent Debye-Hückel electrostatic screening such that the free energy change of unfolding depends on the square root of the ionic strength. This origin of such apparent behavior remains unclear. Also, the frequency slices in Fig. 9.7b shows different values between the condition of 360 mM NaCl/10 mM DCl and the condition of 100 mM NaCl/270 mM DCl, meaning that ionic strength is not the only factor accounting for the dimer equilibrium and that the character of cations can also influence the equilibrium. Experimentally, exploring thermodynamic behavior with different ions across the Hofmeister series may help identify if the ionic strength dependence is universal regardless of the ion character.

9.5.3 Implication on Dynamics of Insulin Dimer Dissociation

In the scope of coupled folding and binding processes, the presence of this additional spectroscopically observable state like twisted conformations raised a question if this state is involved in the dimerization process as an intermediate, or actually it is an irrelevant off-pathway conformational state. In the previous three-state thermodynamic analyses, we found both models can consistently account for the observed thermodynamic transitions, meaning that thermodynamic data alone cannot distinguish the underlying dynamics. To verify if the addition conformations is on-path intermediates, transient experiments are required, such as T-jump IR

experiments using B24B25-labeled insulin, and this study provides a foundation of physically interpreting isotope-edited T-jump experiments.

9.6 Conclusions

In this study, we investigated the conformational ensemble of human insulin dimer using isotope-edited 2D IR spectroscopy, all-atom MSM of the dimer, and amide I computational spectroscopy. Isotope-edited IR spectroscopy shows that the conformational distribution of insulin dimer is heterogeneous instead of a single dominant native conformation as in crystal structure, and that the conformational heterogeneity can be perturbed by ionic strength without changing the apparent thermodynamics of dimer-monomer transition. MSM of the dimer and computational spectroscopy reveal additional twisted conformational state that is consistent with experimental spectra. Thermodynamic analysis with an additional dimer state cannot definitely identify if the additional twisted dimer state is a kinetic intermediate, which can be further verified by performing transient experiments such as isotope-edited temperature-jump IR experiment, and the twisting motions may be seen from dimer dissociation/monomer association pathways in molecular simulations.

9.7 Acknowledgments

I thank Balamurugan Dhayalan for kindly synthesizing batches of $^{13}C^{18}O$ isotope-edited human insulin and providing the synthesis protocol. I thank Nicholas Lewis and Paul Sanstead for discussions of potential experiments to characterize ionic strength dependence. I also thank

Brennan Ashwood, Luis Busto de Moner, and Paul Sanstead for carefully reading through the previous enormous manuscript that leads to part of this chapter.

9.8 References

1. Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M., Controlling entropy to tune the functions of intrinsically disordered regions. *Curr Opin Struct Biol* **2014,** *26*, 62-72.
2. Burger, V.; Gurry, T.; Stultz, C., Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **2014,** *6* (10), 2684-2719.
3. Papoian, G. A., Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proc Natl Acad Sci U S A* **2008,** *105* (38), 14237-8.
4. Camacho, C. J.; Weng, Z.; Vajda, S.; DeLisi, C., Free Energy Landscapes of Encounter Complexes in Protein-Protein Association. *Biophysical Journal* **1999,** *76* (3), 1166-1178.
5. Cheung, M. S.; Garcia, A. E.; Onuchic, J. N., Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci U S A* **2002,** *99* (2), 685-90.
6. Clementi, C.; Plotkin, S. S., The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci* **2004,** *13* (7), 1750-66.
7. Tang, C.; Iwahara, J.; Clore, G. M., Visualization of transient encounter complexes in protein-protein association. *Nature* **2006,** *444* (7117), 383-6.
8. Sugase, K.; Dyson, H. J.; Wright, P. E., Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **2007,** *447* (7147), 1021-5.
9. Best, R. B.; Hummer, G.; Eaton, W. A., Native contacts determine protein folding mechanisms in atomistic simulations. *Proc Natl Acad Sci U S A* **2013,** *110* (44), 17874-9.
10. Chen, T.; Song, J.; Chan, H. S., Theoretical perspectives on nonnative interactions and intrinsic disorder in protein folding and binding. *Curr Opin Struct Biol* **2015,** *30*, 32-42.
11. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995,** *21* (3), 167-95.
12. Guinn, E. J.; Jagannathan, B.; Marqusee, S., Single-molecule chemo-mechanical unfolding reveals multiple transition state barriers in a small single-domain protein. *Nat Commun* **2015,** *6*, 6861.
13. Gianni, S.; Dogan, J.; Jemth, P., Coupled binding and folding of intrinsically disordered proteins: what can we learn from kinetics? *Curr Opin Struct Biol* **2016,** *36*, 18-24.
14. Neupane, K.; Foster, D. A.; Dee, D. R.; Yu, H.; Wang, F.; Woodside, M. T., Direct observation of transition paths during the folding of proteins and nucleic acids. *Science* **2016,** *352* (6282), 239-42.
15. Best, R. B.; Hummer, G., Microscopic interpretation of folding varphi-values using the transition path ensemble. *Proc Natl Acad Sci U S A* **2016,** *113* (12), 3263-8.
16. Kim, J. Y.; Chung, H. S., Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science* **2020,** *368* (6496), 1253-1257.
17. Fleming, G. R.; Wolynes, P. G., Chemical Dynamics in Solution. *Physics Today* **1990,** *43* (5), 36-43.

18. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450* (7172), 964-72.
19. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M., Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **2017,** *42*, 106-116.
20. Bryant, R. G., The NMR time scale. *Journal of Chemical Education* **1983,** *60* (11), 933.
21. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press: 2011.
22. Baiz, C. R.; Reppert, M.; Tokmakoff, A., An Introduction to Protein 2D IR Spectroscopy. *Ultrafast Infrared Vibrational Spectroscopy* **2013**, 361-403.
23. Woutersen, S.; Hamm, P., Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J Phys Chem B* **2000,** *104* (47), 11316-11320.
24. Smith, A. W.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A.; Roy, S.; Jansen, T. L.; Knoester, J., Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J Phys Chem B* **2010,** *114* (34), 10913-24.
25. Baiz, C. R.; Tokmakoff, A., Structural disorder of folded proteins: isotope-edited 2D IR spectroscopy and Markov state modeling. *Biophys J* **2015,** *108* (7), 1747-1757.
26. Feng, Y.; Huang, J.; Kim, S.; Shim, J. H.; MacKerell, A. D., Jr.; Ge, N. H., Structure of Penta-Alanine Investigated by Two-Dimensional Infrared Spectroscopy and Molecular Dynamics Simulation. *J Phys Chem B* **2016,** *120* (24), 5325-39.
27. Reppert, M.; Roy, A. R.; Tempkin, J. O.; Dinner, A. R.; Tokmakoff, A., Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy: Application to Elastin-Like Peptides. *J Phys Chem B* **2016,** *120* (44), 11395-11404.
28. Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D.; Skinner, J. L.; Perozo, E.; Roux, B.; Valiyaveetil, F. I.; Zanni, M. T., Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016,** *353* (6303), 1040-1044.
29. Gruebele, M.; Sabelko, J.; Ballew, R.; Ervin, J., Laser Temperature Jump Induced Protein Refolding. *Accounts of Chemical Research* **1998,** *31* (11), 699-707.
30. Dyer, R. B.; Gai, F.; Woodruff, W. H.; Gilmanshin, R.; Callender, R. H., Infrared Studies of Fast Events in Protein Folding. *Accounts of Chemical Research* **1998,** *31* (11), 709-716.
31. Chung, H. S.; Tokmakoff, A., Temperature-dependent downhill unfolding of ubiquitin. I. Nanosecond-to-millisecond resolved nonlinear infrared spectroscopy. *Proteins* **2008,** *72* (1), 474-87.
32. Hauser, K.; Krejtschi, C.; Huang, R.; Wu, L.; Keiderling, T. A., Site-specific relaxation kinetics of a tryptophan zipper hairpin peptide using temperature-jump IR spectroscopy and isotopic labeling. *J Am Chem Soc* **2008,** *130* (10), 2984-92.
33. Jones, K. C.; Peng, C. S.; Tokmakoff, A., Folding of a heterogeneous beta-hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc Natl Acad Sci U S A* **2013,** *110* (8), 2828-33.
34. Scheerer, D.; Chi, H.; McElheny, D.; Keiderling, T. A.; Hauser, K., Isotopically Site-Selected Dynamics of a Three-Stranded beta-Sheet Peptide Detected with Temperature-Jump Infrared-Spectroscopy. *J Phys Chem B* **2018,** *122* (46), 10445-10454.
35. Minnes, L.; Greetham, G. M.; Shaw, D. J.; Clark, I. P.; Fritzsch, R.; Towrie, M.; Parker, A. W.; Henry, A. J.; Taylor, R. J.; Hunt, N. T., Uncovering the Early Stages of Domain Melting in Calmodulin with Ultrafast Temperature-Jump Infrared Spectroscopy. *J Phys Chem B* **2019,** *123* (41), 8733-8739.

36. Hua, Q. X.; Shoelson, S. E.; Kochoyan, M.; Weiss, M. A., Receptor binding redefined by a structural switch in a mutant human insulin. *Nature* **1991,** *354* (6350), 238-41.
37. Ludvigsen, S.; Roy, M.; Thogersen, H.; Kaarsholm, N. C., High-resolution structure of an engineered biologically potent insulin monomer, B16 Tyr-->His, as determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **1994,** *33* (26), 7998-8006.
38. Olsen, H. B.; Ludvigsen, S.; Kaarsholm, N. C., Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry* **1996,** *35* (27), 8836-45.
39. Keller, D.; Clausen, R.; Josefsen, K.; Led, J. J., Flexibility and bioactivity of insulin: an NMR investigation of the solution structure and folding of an unusually flexible human insulin mutant with increased biological activity. *Biochemistry* **2001,** *40* (35), 10732-40.
40. Ludvigsen, S.; Olsen, H. B.; Kaarsholm, N. C., A structural switch in a mutant insulin exposes key residues for receptor binding. *J Mol Biol* **1998,** *279* (1), 1-7.
41. Kosinova, L.; Veverka, V.; Novotna, P.; Collinsova, M.; Urbanova, M.; Moody, N. R.; Turkenburg, J. P.; Jiracek, J.; Brzozowski, A. M.; Zakova, L., Insight into the structural and biological relevance of the T/R transition of the N-terminus of the B-chain in human insulin. *Biochemistry* **2014,** *53* (21), 3392-402.
42. Bocian, W.; Sitkowski, J.; Bednarek, E.; Tarnowska, A.; Kawecki, R.; Kozerski, L., Structure of human insulin monomer in water/acetonitrile solution. *J Biomol NMR* **2008,** *40* (1), 55-64.
43. Zoete, V.; Meuwly, M.; Karplus, M., A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. *J Mol Biol* **2004,** *342* (3), 913-29.
44. Baker, E. N.; Blundell, T. L.; Cutfield, J. F.; Cutfield, S. M.; Dodson, E. J.; Dodson, G. G.; Hodgkin, D. M.; Hubbard, R. E.; Isaacs, N. W.; Reynolds, C. D.; et al., The structure of 2Zn pig insulin crystals at 1.5 A resolution. *Philos Trans R Soc Lond B Biol Sci* **1988,** *319* (1195), 369-456.
45. Jørgensen, A. M. M.; Kristensen, S. M.; Led, J. J.; Balschmidt, P., Three-dimensional solution structure of an insulin dimer. *Journal of Molecular Biology* **1992,** *227* (4), 1146-1163.
46. Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy. *Phys Chem Chem Phys* **2010,** *12* (14), 3579-88.
47. Menting, J. G.; Yang, Y.; Chan, S. J.; Phillips, N. B.; Smith, B. J.; Whittaker, J.; Wickramasinghe, N. P.; Whittaker, L. J.; Pandyarajan, V.; Wan, Z. L.; Yadav, S. P.; Carroll, J. M.; Strokes, N.; Roberts, C. T., Jr.; Ismail-Beigi, F.; Milewski, W.; Steiner, D. F.; Chauhan, V. S.; Ward, C. W.; Weiss, M. A.; Lawrence, M. C., Protective hinge in insulin opens to enable its receptor engagement. *Proc Natl Acad Sci U S A* **2014,** *111* (33), E3395-404.
48. Croll, T. I.; Smith, B. J.; Margetts, M. B.; Whittaker, J.; Weiss, M. A.; Ward, C. W.; Lawrence, M. C., Higher-Resolution Structure of the Human Insulin Receptor Ectodomain: Multi-Modal Inclusion of the Insert Domain. *Structure* **2016,** *24* (3), 469-76.
49. Gutmann, T.; Kim, K. H.; Grzybek, M.; Walz, T.; Coskun, U., Visualization of ligand-induced transmembrane signaling in the full-length human insulin receptor. *J Cell Biol* **2018,** *217* (5), 1643-1649.
50. Weis, F.; Menting, J. G.; Margetts, M. B.; Chan, S. J.; Xu, Y.; Tennagels, N.; Wohlfart, P.; Langer, T.; Muller, C. W.; Dreyer, M. K.; Lawrence, M. C., The signalling conformation of the insulin receptor ectodomain. *Nat Commun* **2018,** *9* (1), 4420.

51. Antoszewski, A.; Feng, C. J.; Vani, B. P.; Thiede, E. H.; Hong, L.; Weare, J.; Tokmakoff, A.; Dinner, A. R., Insulin Dissociates by Diverse Mechanisms of Coupled Unfolding and Unbinding. *J Phys Chem B* **2020**, 5571-87.
52. Zhang, X. X.; Jones, K. C.; Fitzpatrick, A.; Peng, C. S.; Feng, C. J.; Baiz, C. R.; Tokmakoff, A., Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J Phys Chem B* **2016,** *120* (23), 5134-45.
53. Rimmerman, D.; Leshchev, D.; Hsu, D. J.; Hong, J.; Kosheleva, I.; Chen, L. X., Direct Observation of Insulin Association Dynamics with Time-Resolved X-ray Scattering. *J Phys Chem Lett* **2017,** *8* (18), 4413-4418.
54. Dhayalan, B.; Fitzpatrick, A.; Mandal, K.; Whittaker, J.; Weiss, M. A.; Tokmakoff, A.; Kent, S. B., Efficient Total Chemical Synthesis of (13) C=(18) O Isotopomers of Human Insulin for Isotope-Edited FTIR. *Chembiochem* **2016,** *17* (5), 415-20.
55. Mandal, K.; Dhayalan, B.; Avital-Shmilovici, M.; Tokmakoff, A.; Kent, S. B., Crystallization of Enantiomerically Pure Proteins from Quasi-Racemic Mixtures: Structure Determination by X-Ray Diffraction of Isotope-Labeled Ester Insulin and Human Insulin. *Chembiochem* **2016,** *17* (5), 421-5.
56. Dong, J.; Wan, Z. L.; Chu, Y. C.; Nakagawa, S. N.; Katsoyannis, P. G.; Weiss, M. A.; Carey, P. R., Isotope-edited Raman spectroscopy of proteins: a general strategy to probe individual peptide bonds with application to insulin. *J Am Chem Soc* **2001,** *123* (32), 7919-20.
57. Zoete, V.; Meuwly, M.; Karplus, M., Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* **2005,** *61* (1), 79-93.
58. Gong, Q.; Zhang, H.; Zhang, H.; Chen, C., Calculating the absolute binding free energy of the insulin dimer in an explicit solvent. *RSC Advances* **2020,** *10* (2), 790-800.
59. Banerjee, P.; Mondal, S.; Bagchi, B., Insulin dimer dissociation in aqueous solution: A computational study of free energy landscape and evolving microscopic structure along the reaction pathway. *J Chem Phys* **2018,** *149* (11), 114902.
60. Banerjee, P.; Bagchi, B., Dynamical control by water at a molecular level in protein dimer association and dissociation. *Proc Natl Acad Sci U S A* **2020,** *117* (5), 2302-2308.
61. Desmond, J. L.; Koner, D.; Meuwly, M., Probing the Differential Dynamics of the Monomeric and Dimeric Insulin from Amide-I IR Spectroscopy. *J Phys Chem B* **2019,** *123* (30), 6588-6598.
62. Reppert, M.; Tokmakoff, A., Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu Rev Phys Chem* **2016,** *67*, 359-86.
63. Bouř, P.; Keiderling, T. A., Empirical modeling of the peptide amide I band IR intensity in water solution. *The Journal of Chemical Physics* **2003,** *119* (21), 11253-11262.
64. Ham, S.; Kim, J.-H.; Lee, H.; Cho, M., Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA–nD2O complexes. *The Journal of Chemical Physics* **2003,** *118* (8), 3491-3498.
65. Hayashi, T.; Zhuang, W.; Mukamel, S., Electrostatic DFT map for the complete vibrational amide band of NMA. *J Phys Chem A* **2005,** *109* (43), 9747-59.
66. la Cour Jansen, T.; Knoester, J., A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J Chem Phys* **2006,** *124* (4), 044502.
67. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and validation of transferable amide I vibrational frequency maps for peptides. *J Phys Chem B* **2011,** *115* (13), 3713-24.

68. Reppert, M.; Tokmakoff, A., Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J Chem Phys* **2013,** *138* (13), 134116.
69. Reppert, M.; Tokmakoff, A., Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy. *J Chem Phys* **2015,** *143* (6), 061102.
70. Torii, H., Amide I Vibrational Properties Affected by Hydrogen Bonding Out-of-Plane of the Peptide Group. *J Phys Chem Lett* **2015,** *6* (4), 727-33.
71. Torii, H.; Tasumi, M., Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *Journal of Raman Spectroscopy* **1998,** *29* (1), 81-86.
72. Ham, S.; Cha, S.; Choi, J.-H.; Cho, M., Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *The Journal of Chemical Physics* **2003,** *119* (3), 1451-1461.
73. la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J., Modeling the amide I bands of small peptides. *J Chem Phys* **2006,** *125* (4), 44312.
74. Hayashi, T.; Mukamel, S., Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides. *J Phys Chem B* **2007,** *111* (37), 11032-46.
75. Raghunathan, S.; El Hage, K.; Desmond, J. L.; Zhang, L.; Meuwly, M., The Role of Water in the Stability of Wild-type and Mutant Insulin Dimers. *J Phys Chem B* **2018,** *122* (28), 7038-7048.
76. Wicky, B. I. M.; Shammas, S. L.; Clarke, J., Affinity of IDPs to their targets is modulated by ion-specific changes in kinetics and residual structure. *Proc Natl Acad Sci U S A* **2017,** *114* (37), 9882-9887.
77. Boreikaite, V.; Wicky, B. I. M.; Watt, I. N.; Clarke, J.; Walker, J. E., Extrinsic conditions influence the self-association and structure of IF1, the regulatory protein of mitochondrial ATP synthase. *Proc Natl Acad Sci U S A* **2019,** *116* (21), 10354-10359.
78. Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B., Synthesis of proteins by native chemical ligation. *Science* **1994,** *266* (5186), 776-9.
79. Sohma, Y.; Hua, Q. X.; Whittaker, J.; Weiss, M. A.; Kent, S. B., Design and folding of [GluA4(ObetaThrB30)]insulin ("ester insulin"): a minimal proinsulin surrogate that can be chemically converted into human insulin. *Angew Chem Int Ed Engl* **2010,** *49* (32), 5489-93.
80. Harrison, D. M.; Garratt, C. J., The accurate measurement of insulin molarity. *Biochem J* **1969,** *113* (4), 733-4.
81. Nielsen, L.; Frokjaer, S.; Brange, J.; Uversky, V. N.; Fink, A. L., Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **2001,** *40* (28), 8397-409.
82. Barth, A., The infrared absorption of amino acid side chains. *Prog Biophys Mol Biol* **2000,** *74* (3-5), 141-73.
83. Cheatum, C. M.; Tokmakoff, A.; Knoester, J., Signatures of beta-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J Chem Phys* **2004,** *120* (17), 8201-15.
84. Miyazawa, T., Perturbation Treatment of the Characteristic Vibrations of Polypeptide Chains in Various Configurations. *Journal of Chemical Physics* **1960,** *32* (6), 1647-1652.
85. Miyazawa, T.; Blout, E. R., The Infrared Spectra of Polypeptides in Various Conformations: Amide I and II Bands1. *Journal of the American Chemical Society* **1961,** *83* (3), 712-719.
86. Torres, J.; Kukol, A.; Goodman, J. M.; Arkin, I. T., Site-specific examination of secondary structure and orientation determination in membrane proteins: The peptidic13C?18O group as a novel infrared probe. *Biopolymers* **2001,** *59* (6), 396-401.

87. Tanford, C.; Epstein, J., The Physical Chemistry of Insulin. I. Hydrogen Ion Titration Curve of Zinc-free Insulin1-3. *Journal of the American Chemical Society* **1954,** *76* (8), 2163-2169.
88. Sanstead, P. J.; Stevenson, P.; Tokmakoff, A., Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J Am Chem Soc* **2016,** *138* (36), 11792-801.
89. Dzwolak, W.; Ravindra, R.; Lendermann, J.; Winter, R., Aggregation of bovine insulin probed by DSC/PPC calorimetry and FTIR spectroscopy. *Biochemistry* **2003,** *42* (38), 11347-55.
90. Dyson, F., A meeting with Enrico Fermi. *Nature* **2004,** *427* (6972), 297.
91. Huus, K.; Havelund, S.; Olsen, H. B.; van de Weert, M.; Frokjaer, S., Thermal dissociation and unfolding of insulin. *Biochemistry* **2005,** *44* (33), 11171-7.
92. Antolikova, E.; Zakova, L.; Turkenburg, J. P.; Watson, C. J.; Hanclova, I.; Sanda, M.; Cooper, A.; Kraus, T.; Brzozowski, A. M.; Jiracek, J., Non-equivalent role of inter- and intramolecular hydrogen bonds in the insulin dimer interface. *J Biol Chem* **2011,** *286* (42), 36968-77.
93. de Los Rios, M. A.; Plaxco, K. W., Apparent Debye-Huckel electrostatic effects in the folding of a simple, single domain protein. *Biochemistry* **2005,** *44* (4), 1243-50.
94. Religa, T. L.; Markson, J. S.; Mayor, U.; Freund, S. M.; Fersht, A. R., Solution structure of a protein denatured state and folding intermediate. *Nature* **2005,** *437* (7061), 1053-6.
95. Bolel, P.; Datta, S.; Mahapatra, N.; Halder, M., Spectroscopic investigation of the effect of salt on binding of tartrazine with two homologous serum albumins: quantification by use of the Debye-Huckel limiting law and observation of enthalpy-entropy compensation. *J Phys Chem B* **2012,** *116* (34), 10195-204.
96. Myers, J. K.; Pace, C. N.; Scholtz, J. M., Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **1995,** *4* (10), 2138-48.
97. Karjalainen, E. L.; Ersmark, T.; Barth, A., Optimization of model parameters for describing the amide I spectrum of a large set of proteins. *J Phys Chem B* **2012,** *116* (16), 4831-42.
98. Reppert, M.; Roy, A. R.; Tokmakoff, A., Isotope-enriched protein standards for computational amide I spectroscopy. *J Chem Phys* **2015,** *142* (12), 125104.
99. Gomez, J.; Hilser, V. J.; Xie, D.; Freire, E., The heat capacity of proteins. *Proteins* **1995,** *22* (4), 404-12.

Appendix 9A: Additional IR Spectra of Unlabeled Human Insulin

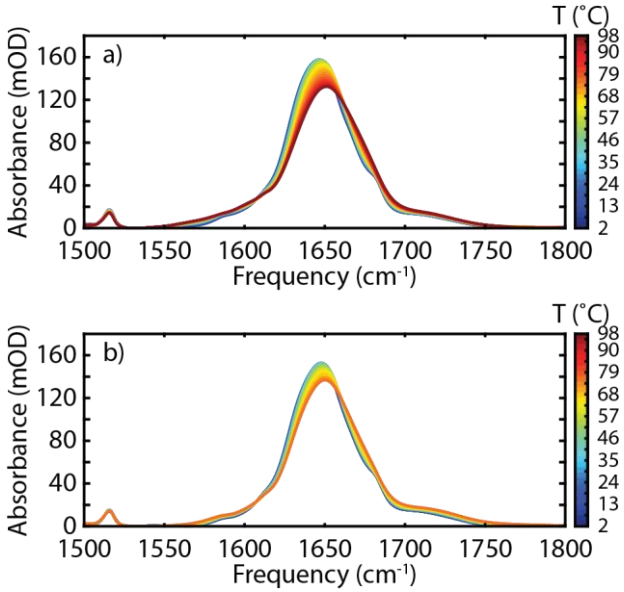

Figure 9A.1. (a) Temperature-dependent FTIR spectra of UL insulin in the no-salt condition. (b) Temperature-dependent FTIR spectra of UL insulin in the high-salt condition.

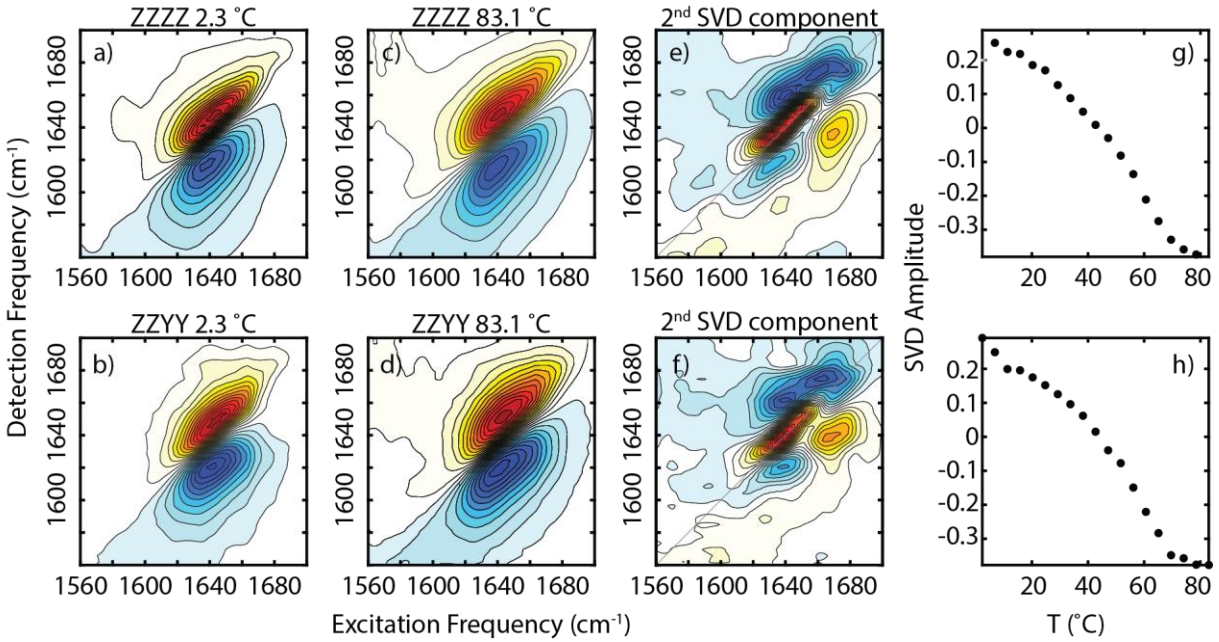

Figure 9A.2. Temperature-dependent 2D IR spectra of UL insulin. (a) Parallel-polarized 2D spectrum at 2.3 °C (b) Perpendicular-polarized 2D spectrum at 2.3 °C (c) Parallel-polarized 2D spectrum at 83.1 °C (d) Perpendicular-polarized 2D spectrum at 83.1 °C Second SVD spectral component of (e) parallel-polarized 2D spectra and (f) perpendicular-polarized 2D spectra. (g) SVD dissociation curve from parallel-polarized 2D spectra. (h) SVD dissociation curve from perpendicular-polarized 2D spectra.

Appendix 9B: Two-State Thermodynamic Model for Global Fit

The detailed model description is provided in Chapter 5.2. In brief, thermodynamic two-state dissociation is given as

$$\mathrm{D} \overset{K_\mathrm{d}}{\rightleftharpoons} 2\mathrm{M} \tag{9.7}$$

In a solution with the total concentration, $C_{\mathrm{tot}} = 2[\mathrm{D}]+[\mathrm{M}]$, we can define fractions of dimer and monomer as θ_D and θ_M such that $\theta_\mathrm{D}+\theta_\mathrm{M}=1$, which will be useful to know the population of each species at a given temperature. Each species is typically assumed to be distinguishable spectroscopically in some way so that measuring spectral change as a function of temperature can

be used as a proxy to the underlying population of some species. The expression of these fractions are given as

$$\begin{aligned}\theta_{\mathrm{D}} &= \frac{2[\mathrm{D}]}{C_{\mathrm{tot}}} \\ \theta_{\mathrm{M}} &= 1-\theta_{\mathrm{D}} = \frac{[\mathrm{M}]}{C_{\mathrm{tot}}}\end{aligned} \tag{9.8}$$

and the dimer fraction and the monomer fraction in terms of K_{d}:

$$\theta_{\mathrm{D}} = \frac{1}{4C_{\mathrm{tot}}}\left[4C_{\mathrm{tot}} + K_{\mathrm{d}} - \sqrt{K_{\mathrm{d}}^2 + 8C_{\mathrm{tot}}K_{\mathrm{d}}}\right] \tag{9.9}$$

It is well-known that there is non-negligible change of heat capacity in protein association/dissociation, and protein folding/unfolding processes.[99] Assuming constant change of heat capacity, ΔC_P°, the expressions of enthalpy and entropy at the standard state are

$$\Delta H_{\mathrm{d}}^\circ(T) = \Delta H_{\mathrm{d}}^\circ(T_{\mathrm{d}}) + \Delta C_P^\circ\left(T - T_{\mathrm{d}}\right) \tag{9.10}$$

$$\Delta S_{\mathrm{d}}^\circ\left(T\right) = \Delta S_{\mathrm{d}}^\circ\left(T_{\mathrm{d}}\right) + \Delta C_P^\circ \ln\left(\frac{T}{T_{\mathrm{d}}}\right) = \frac{\Delta H_{\mathrm{d}}^\circ(T_{\mathrm{d}})}{T_{\mathrm{d}}} + \Delta C_P^\circ \ln\left(\frac{T}{T_{\mathrm{d}}}\right) \tag{9.11}$$

In this thermodynamic two-state model, T_{d} is the dissociation temperature defined at $\theta_{\mathrm{D}} = 0.5$, or equivalently, $K_{\mathrm{d}} = C_{\mathrm{tot}}$. Then, the dissociation constant as a function of temperature can be easily written as

$$\begin{aligned}K_{\mathrm{d}}\left(T\right) &= \exp\left(-\frac{\Delta G_{\mathrm{d}}^\circ\left(T\right)}{RT}\right) \\ \Delta G_{\mathrm{d}}^\circ\left(T\right) &= \Delta H_{\mathrm{d}}^\circ(T_{\mathrm{d}})\left[1-\frac{T}{T_{\mathrm{d}}}\right] - \Delta C_P^\circ\left[T - T_{\mathrm{d}} - T\ln\left(\frac{T}{T_{\mathrm{d}}}\right)\right] + RT\ln\left(C_{\mathrm{tot}}\right)\end{aligned} \tag{9.12}$$

When the dimer fraction is 0.5, or $K_{\mathrm{d}} = C_{\mathrm{tot}}$, the corresponding free energy of dissociation is $\Delta G_{\mathrm{d}}^\circ\left(T_{\mathrm{d}}\right) = -RT_{\mathrm{d}}\ln\left(C_{\mathrm{tot}}\right)$. In other words, if we try to find a temperature where the free energy is

zero in Eqn. (9.12), the temperature corresponds to the dissociation temperature T_d at the total concentration C_{tot}. One can in principle use the Eqns. (9.9)–(9.12) with experimentally measurable C_{tot} to determine unknown thermodynamic parameters including T_d, $\Delta H_d^\circ(T_d)$, ΔC_P°. In practice, the optical dissociation curve is usually obtained by picking a frequency slice at a function of temperature or by performing singular value decomposition (SVD) over a frequency range and taking the second component, which reports the most dominant changes across the temperature. This dissociation curve is, however, most likely convolved with additional spectral response not related to the real dimer-monomer transition. One example is change of solvation environment due to temperature change. As a result, this solvation change will induce slanted baseline(s), and most likely the solvation changes are different between dimer state and monomer state. So in a fitting to a single dissociation curve $I(T)$, one may take the following form for a two-state model fit:

$$I(T) = \left[\left(m_D T - b_D\right) - \left(m_M T - b_M\right)\right]\theta_D(T) + m_M T - b_M \tag{9.13}$$

It can be seen that there are 4 additional slanted baseline parameters including slope, m_i, of the species i, and the intercept b_i, in total of 7 fit parameters to a single curve, which is extremely prone to over-fitting.[90]

To mitigate the issue, we performed a global fitting across multiple data at the same solution condition, including unlabeled (UL) FTIR spectra, B24B25-labeled FTIR spectra, B24B25–UL FTIR spectra, UL parallel-polarized 2D IR spectra, and UL perpendicular-polarized 2D IR spectra. For the global fitting, optical dissociation curves from these experimental spectra are obtained by selecting a specific range of frequency and temperature summarized in Table 9B.1. The frequency ranges are selected to specifically probe the changes around the β-sheet vibrations, which primarily report the dimer dissociation. In this global fitting, thermodynamic parameters T_d, $\Delta H_d^\circ(T_d)$, ΔC_P° are shared among all dissociation curves while the baseline parameters are private

to each dissociation curve, meaning that fitting to n dissociation curves requires $3 + 4n$ parameters. Since the sample condition is exactly the same, the underlying thermodynamics between different dissociation curves should also be exactly the same assuming the spectral changes report the same physical processes. In the limit of infinite number of dissociation curves, one can expect it to be reduced to 4-parameter fit to baselines for each curve without changing the shape of the dissociation curve. By increasing the number of dissociation curves, one can reduce the risk of over-fitting and potentially decrease uncertainty of the fit as well as dependence on initial guesses.

	Frequency (cm^{-1})	Excitation Frequency (cm^{-1})	Detection Frequency (cm^{-1})	Temperature (°C)
UL FTIR	1670–1700	N.A.	N.A.	1.9–97.8
B24B25 FTIR	1560–1620	N.A.	N.A.	1.9–97.8
B24B25 – UL FTIR	1560–1620	N.A.	N.A.	1.9–97.8
UL parallel-polarized 2D IR	N.A.	1560–1700	1560–1700	2.3–83.0
UL perpendicular-polarized 2D IR	N.A.	1560–1700	1560–1700	2.3–83.0

Table 9B.1. Frequency and temperature ranges used for extracting the dissociation curves for global fitting.

Appendix 9C: Ionic Strength Dependence on Thermodynamics of Dimer Unfolding

Knowing that there is an additional dimer conformational state involving in the dimer dissociation (Fig. 9.2), and that the populations of the two dimer states can be perturbed by ionic strength and identity of cations (Fig. 9.7), it is informative to characterize the ionic strength dependence on the thermodynamics of dimer unfolding. Based on Fig. 9.7 and the MSM, the frequency slices of 1639 cm^{-1} and 1646 cm^{-1} can be used as proxies to populations of the native dimer state and the twisted dimer state, respectively. Also, increasing ionic strength decreases the

population of the native dimer while the population of twisted dimer increases. Discussing thermodynamic behaviors of the ionic strength dependence will help us refine the thermodynamic picture of the complex interplay between dimer unfolding and the dimer dissociation. The detailed discussion is provided in Section 5.2.

In brief for a two-state unfolding of the dimer:

$$\mathrm{D}_1 \overset{K_\mathrm{u}}{\rightleftharpoons} D_2 \tag{9.14}$$

The simplest possible effect accounting for the ionic strength dependence is the electrostatic screening from Debye-Hückel theory. In the limit of Debye-Hückel regime, electrostatic interactions are effectively screened by surrounding distribution of ions, which as a result depend on the squared root of ionic strength $\sqrt{I}$ originated from the difference of activity coefficients between different dimer states. Debye-Hückel limiting law shows that the natural log of activity coefficient γ is proportional to the square root of the ionic strength. The free energy change of unfolding can be written as[93-95]

$$\Delta G_\mathrm{u}^\circ(I) = \Delta G_\mathrm{u}^\circ(I=0) + m\sqrt{I} \equiv \Delta G_{\mathrm{u},0}^\circ + m\sqrt{I} \tag{9.15}$$

In Eqn. (9.15), $\Delta G_{\mathrm{u},0}^\circ$ corresponds to the free energy change at zero ionic strength in kJ/mol, and m is the slope to the square root of ionic strength in unit of kJ/mol.M$^{0.5}$. One can find the relationship between the square root of ionic strength and the fractions of dimer as

$$\theta_{\mathrm{D}_1}(I) = \frac{1}{1+K_\mathrm{u}} = \frac{1}{1+\exp\left(-\Delta G_\mathrm{u}^\circ(I)/RT\right)} \tag{9.16}$$

$$\theta_{\mathrm{D}_2}(I) = \frac{K_\mathrm{u}}{1+K_\mathrm{u}} = \frac{\exp\left(-\Delta G_\mathrm{u}^\circ(I)/RT\right)}{1+\exp\left(-\Delta G_\mathrm{u}^\circ(I)/RT\right)} \tag{9.17}$$

These equations indicate that the larger the *m* value is, the steeper the two-state transition is across the ionic strength. Rearranging Eqns. (9.16) and (9.17) gives

$$\ln K_{\mathrm{u}} = \ln\left(\frac{1-\theta_{\mathrm{D}_1}(I)}{\theta_{\mathrm{D}_1}(I)}\right) = \frac{-\Delta G^{\circ}_{\mathrm{u},0}}{RT} - \frac{m\sqrt{I}}{RT} \tag{9.18}$$

$$\ln\left(\frac{1}{K_{\mathrm{u}}}\right) = \ln\left(\frac{1-\theta_{\mathrm{D}_2}(I)}{\theta_{\mathrm{D}_2}(I)}\right) = \frac{\Delta G^{\circ}_{\mathrm{u},0}}{RT} + \frac{m\sqrt{I}}{RT} \tag{9.19}$$

These provide a set of linearized plots relating dimer fractions with the intercept related to the free energy change at zero ionic strength, and the slope associated with the squared root of the ionic strength. These two parameters $\Delta G^{\circ}_{\mathrm{u},0}$ and *m* can be used for fitting to a two-state equilibrium as a function of ionic strength.

In the main text, we assume that frequency slices are proportional to the fractions of folded dimer and unfolded dimer.

$$\begin{aligned} V_1(I) &= a_1\theta_{D_1}(I) + b_1 \\ V_2(I) &= a_2\theta_{D_2}(I) + b_2 \end{aligned} \tag{9.20}$$

There are 4 additional baseline parameters. However, there are three additional constraints that can help reduce the number of fit parameters summarized below. First, conservation of the dimer populations has to hold, meaning the sum of dimer fractions is equal to 1, or $\theta_{\mathrm{D}_1} + \theta_{\mathrm{D}_2} = 1$. Second, there is a strong linear correlation between $V_1(I)$ and $V_2(I)$, such that one can write $V_1(I) = \alpha V_2(I) + \beta$, with α and β obtained from a linear regression between $V_1(I)$ and $V_2(I)$. Third, one can use the estimated fractions from MaxEnt reconstruction of the high-salt condition as an internal standard. Combining these constraints, one can rewrite the Eqn. (9.20) in terms of a_1 as

$$
\begin{aligned}
b_1 &= V_1^{'} - a_1\theta_{D_1}^{'} \\
a_2 &= -a_1/\alpha \\
b_2 &= \left(b_1 - \beta\right)/\alpha - a_2
\end{aligned}
\tag{9.21}
$$

So the fit can be reduced into a 3-parameter fit to both frequency slices including a_1, $\Delta G_{u,0}^{\circ}$, and m. The quality of the fit can be evaluated by looking at the linearized plot from the Eqns. (9.18) and (9.19).

Another potential model to account for the ionic strength dependence is to treat the ionic strength as a chemical denaturant analogous to urea or GdmCl. Then the free energy change can be written as[96]

$$
\Delta G_u^{\circ}\left(I\right) = \Delta G_u^{\circ}\left(I = 0\right) + mI \equiv \Delta G_{u,0}^{\circ} + mI \tag{9.22}
$$

Using the same treatment above, one can get

$$
\begin{aligned}
\ln K_u &= \ln\left(\frac{1-\theta_{D_1}\left(I\right)}{\theta_{D_1}\left(I\right)}\right) = \frac{-\Delta G_{u,0}^{\circ}}{RT} - \frac{mI}{RT} \\
\ln\left(\frac{1}{K_u}\right) &= \ln\left(\frac{1-\theta_{D_2}\left(I\right)}{\theta_{D_2}\left(I\right)}\right) = \frac{\Delta G_{u,0}^{\circ}}{RT} + \frac{mI}{RT}
\end{aligned}
\tag{9.23}
$$

Fitting of this model will have exactly the same parameters as in the model fit with electrostatic screening, but the dependence of ionic strength is linear instead of the squared root. Based on the linearized plot, we found that the linear correlation is stronger with the squared root of the ionic strength than with the ionic strength, shown in Figure 9C.1, indicating that electrostatic screening is a better descriptor than treating the ions as a chemical denaturant. However, note that at high ionic strength, the dimer equilibrium does depend on the identity of ions, shown in Fig. 9.7, suggesting that there is a complex interplay between the electrostatic screening and characteristics of the ions instead of a single dominant electrostatic screening.

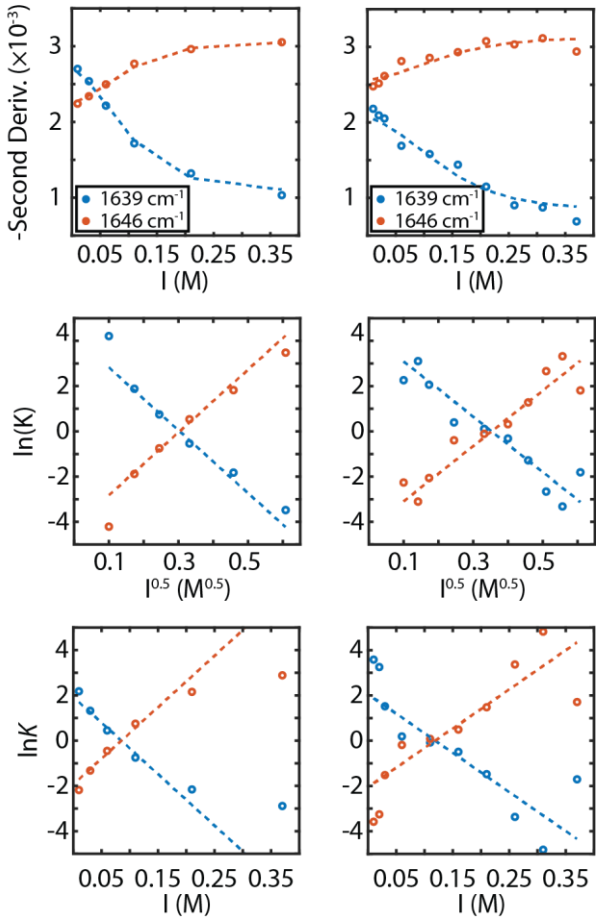

Figure 9C.1 Top: Frequency slice of the second derivative as a function of ionic strength. Left and right correspond to the two independent repeats. Dashed line represents to the Debye-Hückel fit. Middle: Calculated natural log of the equilibrium constant as a function of squared root of the ionic strength, derived from the Debye-Hückel fit. Blue and orange represents the unfolding equilibrium constant and the folding constant, respectively. Bottom: Calculated natural log of the equilibrium constant as a function of the ionic strength, derived from the fit treating ionic strength as a chemical denaturant.