

Vibrational and Scaling Cascades in Conformational Dynamics of (Alanine-Leucine)_n-Peptides

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

Am Fachbereich Physik
der Freien Universität Berlin

vorgelegt von

Irtaza Hassan

Berlin, 2022

Erstgutachterin: Prof. Dr. Petra Imhof
Zweitgutachterin: Prof. Dr. Bettina G. Keller

Tag der Disputation: 25.05.2022

Selbstständigkeitserklärung

Name: Hassan

Vorname: Irtaza

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Datum: _____, Unterschrift: _____

Contents

Abstract	1
Zusammenfassung	5
1 Introduction	9
2 Theoretical Background	13
2.1 The Conformational Dynamics of Peptides/Proteins	13
2.2 Classical Theory	16
2.2.1 Single Particle Dynamics	16
2.2.2 Classical Harmonic Oscillator	17
2.2.3 Force and Potential Energy	18
2.2.4 Force-Fields	19
2.3 Quantum Theory	22
2.3.1 Single Particle Dynamics	23
2.3.2 Quantum Harmonic Oscillator	23
2.3.3 The Born-Oppenheimer Approximation	24
2.3.4 Density Functional Theory	25
2.4 Statistical Theory	30
2.5 Molecular Dynamics	32
2.5.1 Verlet Algorithm	32
2.5.2 Classical MD	33
2.5.3 First-principles MD	35
2.6 Markov State Modelling	36
2.6.1 Perron-cluster cluster analysis	37
2.6.2 Time-lagged Independent Component Analysis	37
2.7 Normal modes	38
2.7.1 Normal Modes of a Linear Triatomic Molecule	39
2.7.2 Quantum Chemical Calculation of Normal Modes Spectra	40
2.8 Theory of Infrared Spectra	41
2.8.1 Schemes to Estimate Instantaneous Molecular Dipole Moments	42
2.9 Theory of Wavelet Analysis	43
3 Methods	45

4	A Combined Approach	47
4.1	Introduction	47
4.2	Methods	48
4.2.1	Classical Molecular Dynamics Simulations	48
4.2.2	Markov State Model	48
4.2.3	Normal mode calculations	50
4.2.4	First-Principles Molecular Dynamics Simulations	50
4.2.5	Experimental Setup	51
4.3	Results	52
4.3.1	Markov State Model	52
4.3.2	Normal Modes	58
4.3.3	Infrared Spectrum	60
4.3.4	Comparison of Sampled Conformations and Resulting Spectra	61
4.4	Discussion	63
4.5	Conclusions	64
5	Hydration Shell Effect	67
5.1	Introduction	67
5.2	Methods	68
5.2.1	Molecular Mechanics Simulations	68
5.2.2	First-Principles Molecular Dynamics Simulations	69
5.2.3	Normal Modes	70
5.2.4	Interaction Energies	71
5.3	Results	71
5.3.1	Conformational Analysis	71
5.3.2	Vibrational Analysis	74
5.3.3	Normal Modes of ALAL–Water Clusters	75
5.3.4	Analysis of the Hydration Shell	76
5.3.5	Instantaneous Frequencies	86
5.3.6	Simulations with Constrained Water Molecules	88
5.4	Discussion	89
5.5	Conclusions	92
6	Metastable-Conformations vs. Hydration Shell	95
6.1	Introduction	95
6.2	Methods	95
6.2.1	Classical Molecular Dynamics simulations	95
6.2.2	Markov State Modelling	96
6.2.3	First-Principles Molecular Dynamics Simulations	97
6.2.4	Normal Modes	98
6.2.5	Hydrogen Bonds and Water Topology	98
6.3	Results	98
6.3.1	Markov State Modelling	98

6.3.2	Normal Modes Analysis	102
6.3.3	First-Principles MD simulations	106
6.4	Discussion	125
6.5	Conclusions	127
7	Effect of Peptide Length	129
7.1	Introduction	129
7.2	Methods	130
7.2.1	Classical Molecular Dynamics simulations	130
7.2.2	Markov State Modeling	131
7.2.3	First-Principles Molecular Dynamics Simulations	131
7.3	Results	132
7.3.1	Markov State Modeling	132
7.3.2	Constrained Classical Simulations	147
7.3.3	First-principles MD Simulations	149
7.4	Discussion and Conclusions	151
8	Discussion	153
9	Conclusions and Outlook	157
A	Supplementary material for: A Combined Approach	159
A.1	Conformational analyses	159
A.1.1	Time series of torsion angles and micro-states in the first-principles MD simulations	159
A.1.2	Hydrogen bond analysis	161
A.1.3	Torsion angle distributions, IR and power spectra from first principles MD simulations	170
B	Supplementary material for: Hydration Shell Effect	175
C	Supplementary material for: Metastable-Conformations vs. Hydration Shell	183
C.1	Markov state modelling	183
C.2	Normal modes	188
C.3	Torsion angle distributions	189
C.4	Root mean square fluctuations	197
C.5	Radial distribution functions	198
C.6	Angular distribution functions	200
C.7	Combined distribution functions	203
C.8	Hydrogen bonds	207
C.9	Power spectra	208
C.10	Density-based vs Wannier centers-based IR spectra	214

D Supplementary material for: Effect of Peptide Length	217
D.1 Markov State Modeling	217
D.1.1 Alanine-Leucine-Alanine	219
D.1.2 Alanine-Leucine-Alanine-Leucine-Alanine	220
D.1.3 Alanine-Leucine-Alanine-Leucine-Alanine-Leucine	222
D.2 Constrained Classical Simulations	225
D.2.1 Dihedral Distributions	225
 Acknowledgments	 231
 Bibliography	 233

Abstract

This thesis aims for understanding the scaling cascades in the vibrational and conformational dynamics of peptides, exemplified on model (Ala-Leu)_n based peptides.

By sampling (classical MD simulations), modeling (MSM's) and vibrational spectroscopy (calculated using first-principles MD simulations and compared to IR experiments), we succeeded in 1) interpreting the measured vibrational spectrum of AlaLeu, 2) understanding the effect of hydration shell dynamics on the vibrational spectrum of Ala-Leu-Ala-Leu, 3) connecting the slow, functionally relevant metastable conformations of Ala-Leu-Ala-Leu and the fast solvent dynamics/local atomic fluctuations, 4) elucidating the effect of change in the length of the (Ala-Leu)_n peptide on the timescales of metastable conformation, hydration shell geometry and the vibrational spectra.

In the first part, we have combined infrared (IR) experiments with molecular dynamics (MD) simulations in solution at finite temperature to analyse the vibrational signature of the small floppy peptide Alanine-Leucine. IR spectra computed from first-principles MD simulations exhibit no distinct differences between conformational clusters of α -helix or β -sheet-like folds with different orientations of the bulky leucine side chain. All computed spectra show two prominent bands, in good agreement with the experiment, that are assigned to the stretch vibrations of the carbonyl and carboxyl group, respectively. Variations in band widths and exact maxima are likely due to small fluctuations in the backbone torsion angles.

In the second part, we analysed the hydration shell around the Ala-Leu-Ala-Leu peptide and, in particular, the central C₂=O₂ carbonyl group, as it is the least affected carbonyl group by the charged termini. The factors that influence the vibrational frequency, and thus the spectroscopic signature, in the amide I region were inspected. We observed that the frequency differences between the three distinct carbonyl groups are due to their interactions with the surrounding water. The probabilities of groups forming hydrogen bonds with water are consistent with the observed shifts in computed stretching frequencies. The one, two or mixed hydrogen bonded states of the central carbonyl group exhibited a clear trend of the red-shift of the C₂=O₂ vibrational frequencies with the averaged number of hydrogen-bonded water molecules. The analysis of interaction energies between the closest water molecules and the peptide fragment showed that the amount by which the frequencies are lowered is reflected in the strengths of

the interaction energies containing the $C_2=O_2$ group. Interestingly, it has been observed that the interaction of the second water molecule determines the amount of the (additional) red-shift, as the first water molecule almost always interact strongly. Moreover, the geometric definition of a hydrogen bond by distance and angle criteria, typically used in the MD community and also in our work, is justified by the distribution of interaction energies between water molecules and the $C_2=O_2$ group of the Ala-Leu-Ala-Leu peptide over water-carbonyl distances and angles.

In the third part, we studied how the underlying conformation affects the vibrational signatures and how the vibrational signatures are impacted by the dynamics of water surrounding exposed polar residues of each conformation. The metastable-conformations of the Ala-Leu-Ala-Leu peptide are estimated using a MSM and classical MD simulations. The β -sheet like conformation is the most probable one, and the L_α -like conformation is a highly unlikely conformation. The structural difference between the conformations, notably the presence of intramolecular hydrogen bonds for α and L_α -like conformations, result in different normal mode frequencies of the carbonyl groups. The flexibility of the peptide bonds varies among the conformations and with it the frequency of changes in the hydrogen-bonded situations. The level of hydration, lifetimes of hydrogen bonds between individual polar groups and water, and the hydration shell topology are also different for different conformations. The central carbonyl group ($C_2 = O_2$), however, has in all conformations a hydrogen bond probability intermediate to the other two groups, but shows a clear change in hydrogen bond probability with a change in simulated conformation. Being least influenced by the termini the $C_2 = O_2$ group is the best representative of a carbonyl group in a longer peptide or protein. The power spectra of each polar group, particularly, the carbonyl group, are affected by the conformation-dependent different interactions with the water molecules in the first hydration shell. There is a strong correlation between the solvation probabilities and peak frequencies of $C_2 = O_2$ of metastable-conformations. The underlying conformation and conformation-dependent variability of hydration are indeed manifested in the entire Amide-I region of IR spectra as different vibrational signatures for different metastable-conformations.

In the fourth part, we investigated the effect of peptide length on the timescales of slow conformation transitions and Amide I spectra using Alanine-Leucine (AL)peptides of different lengths, i.e., AL, ALA, ALAL, ALALA, ALALAL. The MSM's of peptide of different lengths show that the most probable conformation of all peptides is the β -sheet like and the timescale of the slow transition increases with the peptide length (from AL to ALALAL) except for the ALA-peptide. ALA tends to be in a more closed shape as evidenced by the constituent conformations i.e., ($L_\alpha, L_\alpha, \beta/\alpha, \beta/\alpha$)-conformations, of the second most probable metastable set for this peptide. This might be due to the presence of the fast moving, two bulky leucine sidechains in such a short peptide. On the other hand, a pure L_α -like

conformation is not identified in any of the nine metastable sets of ALALAL, likely requiring even longer classical MD simulations. Our data clearly shows that the larger the peptide, the longer the classical MD simulations required to explore the entire configurational space for the construction of a MSM to estimate all possible metastable conformations. Furthermore, the construction of a MSM involves many steps (from the selection of essential internal coordinates to the spectral clustering) and the success of the kinetic model is dependent on the modeler. Thus, a modern framework such as VAMPnets that combines the whole data processing pipeline in a single end-to-end framework and provides an easily interpretable few-state kinetic model should be employed. Moreover, we observed that with the increase in the length of the peptide the properties of the hydration shell improve i.e., the average number of hydrogen bonds per carboxyl group and the well-defined second and third hydration shells (which are a better representative of a bulk system). Lastly, the power spectra of the carbonyl groups calculated from the first-principle simulations of the peptides in explicit solvent show the typical trend, i.e., a higher number of carbonyl groups gives rise to a broader Amide-I band and the increased interaction of water molecules enhances the intensity features of the power spectra.

Zusammenfassung

Diese Dissertation hat zum Ziel, die Skalenkaskaden der Schwingungs- und Konformationsdynamik von Peptiden anhand von $(\text{Ala-Leu})_n$ -basierten Modellpeptiden zu verstehen.

Mittels Sampling (klassische Molekulardynamik (MD)-Simulationen), Modellierung (Markov State Modelle (MSM) und Schwingungsspektroskopie (berechnet durch first-principles MD Simulationen und Vergleich mit Infrarot (IR)-Experimenten), ist es gelungen, 1) das gemessene Schwingungsspektrum von Ala-Leu zu interpretieren, 2) den Einfluss der Dynamik in der Hydratationshülle auf das Schwingungsspektrum von Ala-Leu-Ala-Leu zu verstehen, 3) die langsamen, funktionell relevanten metastabilen Konformationen von Ala-Leu-Ala-Leu und die schnelle Dynamik des Solvens sowie die lokalen atomaren Fluktuationen in Beziehung zu setzen, 4) den Einfluss der Länge der $(\text{Ala-Leu})_n$ -Peptide auf die Zeitskalen der metastabilen Konformationen, der Geometrie der Hydrathülle und des Schwingungsspektrums zu beleuchten.

Im ersten Teil, wurden IR-Experimente mit MD-Simulationen in Lösung bei endlicher Temperatur kombiniert, um die Schwingungssignatur des kleinen, beweglichen Peptids Alanin-Leucin zu analysieren. Mittels first-principles MD-Simulationen berechnete IR-Spektren weisen keine auffälligen Unterschiede zwischen Gruppen von α -Helix oder β -Faltblatt-artigen Faltungen mit verschiedenen Anordnungen der raumerfüllenden Leucin-Seitenkette auf. Alle berechneten Spektren zeigen, in guter Übereinstimmung mit dem Experiment, zwei prominente Banden, die der Streckschwingung der Carbonyl- bzw. Carboxylgruppe zugeordnet werden. Variationen in der Breite der Banden und der exakten Lage der Maxima gehen vermutlich auf kleinere Fluktuationen in den Torsionswinkeln des Peptidrückgrats zurück.

Im zweiten Teil, wurde die Hydratationshülle um das Ala-Leu-Ala-Leu-Peptide analysiert. Hierbei wurde der zentralen, $\text{C}_2=\text{O}_2$, Carbonylgruppe spezielle Aufmerksamkeit gewidmet, da diese die am wenigsten durch die geladenen Termini beeinflusste Carbonylgruppe ist. Die Faktoren, welche die Schwingungsfrequenz und damit die spektroskopische signatur in der Amide I-Region beeinflussen wurden untersucht. Es konnten klare Unterschiede zwischen den drei verschiedenen Carbonylgruppen beobachtet werden, die auf die Interaktionen mit dem umgebenden Wasser zurückgeführt werden können. Die Wahrscheinlichkeiten, mit welchen die Carbonylgruppen Wasserstoffbrücken zu Wasser bilden sind konsistent mit den beobachteten Verschiebungen in den berechneten Streckschwingungsfrequenzen. Die ein-, zwei-, oder gemischten Wasserstoffbrückenbindungszustände der zentralen Carbonylgruppe weisen einen deutlichen Trend in der Rotverschiebung der $\text{C}_2=\text{O}_2$ Schwingungsfrequenzen mit der mittleren Anzahl an wasserstoffbrückengebundenen Wassermolekülen auf. Die Analyse der Interaktionsenergien zwischen dem Peptidfragment und den nächstliegen-

den Wassermolekülen zeigt, dass der Betrag, um welchen die Frequenzen erniedrigt sind, in der Stärke der Wechselwirkung mit der $C_2=O_2$ -Gruppe widergespiegelt wird. Interessanterweise ist es die Wechselwirkung mit dem zweitnächsten Wassermolekül, welche die (zusätzliche) Rotverschiebung bestimmt, da das erste (nächst benachbarte) Wassermolekül fast ausschließlich stark wechselwirkt. Weiter konnte durch die Verteilung der Wechselwirkungsenergien zwischen Wassermolekülen und der $C_2=O_2$ -Gruppe des Ala-Leu-Ala-Leu-Peptids über entsprechende Abstände und Winkel gezeigt werden, dass die geometrische Definition einer Wasserstoffbrückenbindung durch ein Distanz- und ein Winkelkriterium, wie es zur Analyse von MD-Daten typischerweise und auch in dieser Arbeit verwendet wird, gerechtfertigt ist.

Im dritten Teil, wurde untersucht, wie die zugrundeliegende Konformation des Peptids und die Dynamik des Wassers um die lösemittel-exponierten polaren Gruppen der verschiedenen Konformationen deren Schwingungssignatur beeinflusst. Die metastabilen Konformationen des Ala-Leu-Ala-Leu-Peptids wurden mit Hilfe eines MSM und klassischen MD-Simulationen ermittelt. Die β -Faltblatt-artige Konformation ist die wahrscheinlichste, und die L_α -artige Konformation ist eine sehr unwahrscheinliche Konformation. Der strukturelle Unterschied zwischen den Konformationen, insbesondere das Vorliegen einer intramolekularen Wasserstoffbrücke in den α und L_α -artigen Konformationen, hat unterschiedliche Normalmodenfrequenzen der Carbonylgruppen zur Folge. Die Flexibilität der Peptidbindungen variiert zwischen den Konformationen und mit der Häufigkeit von Wechseln des wasserstoffbrückengebundenen Zustands. Der Grad der Hydratisierung, Lebensdauern der Wasserstoffbrücken zwischen einzelnen polaren Gruppen und wasser und die Topologie der Hydratationshülle sind ebenfalls verschieden für unterschiedliche Konformationen. Die zentrale Carbonylgruppe ($C_2=O_2$), hingegen, hat in allen Konformationen eine Wasserstoffbrückenbindungswahrscheinlichkeit, die zwischen der der anderen beiden Carbonylgruppen liegt. Diese Carbonylgruppe zeigt aber deutliche Unterschiede in der Wasserstoffbrückenbindungswahrscheinlichkeit beim Wechsel zwischen den Konformationen. Als die am wenigsten von den Termini beeinflusste Carbonylgruppe ist die $C_2=O_2$ -Gruppe am repräsentativsten für die Carbonylgruppe in einem längeren Peptid oder Protein. Die Powerspektren der einzelnen polaren Gruppen, insbesondere der Carbonylgruppen, sind durch die konformationbedingt verschiedenen Wechselwirkungen mit den Wassermolekülen der ersten Hydrathülle beeinflusst. Es gibt eine starke Korrelation zwischen der Solvationswahrscheinlichkeit und der $C_2=O_2$ -Peak-Frequenz der metastabilen Konformationen. Die zugrundeliegenden Konformationen und konformationsabhängigen Variabilitäten der Hydratation können in der Tat in der gesamten Amide-I-Region der IR-Spektren als unterschiedliche Schwingungssignaturen verschiedener metastabiler Konformationen wahrgenommen werden.

Im vierten Teil, wurde der Einfluss der Peptidlänge auf die Zeitskalen der langsamen Konformationsänderungen und der Amide-I-Spektre von Alanin-Leucin

(AL)- Peptiden unterschiedlicher Länge, d.h. AL, ALA, ALAL, ALALA, ALALAL, untersucht. Die MSM's der Peptide unterschiedlicher Länge zeigen, dass die wahrscheinlichste Konformation in allen diesen Peptiden eine β -Faltblattartige ist und dass die Zeitskalen der langsamsten Übergänge mit der Peptidlänge (von AL zu ALALAL) ansteigt, mit Ausnahme des ALA-Peptids. ALA tendiert zu einer mehr geschlossenen Form, wie sich in den die zweitwahrscheinlichste metastabile Menge bildenden Konformationen (L_α , L_α , β/α , β/α) zeigt. Dies mag an der sich schnell bewegenden grossen Leucin-Seitenkette in diesem kurzen Peptid liegen. Für ALALAL ist hingegen keine L_α -artige Konformation beobachtet worden, was daran liegen kann, dass hierfür noch wesentlich längere klassische MD-Simulationen erforderlich sind. Die vorliegenden Daten zeigen deutlich, dass, um den gesamten Konformationsraum abzutasten, der zur Konstruktion eines MSM und zur Ermittlung aller möglichen metastabilen Konformationen benötigt wird, umso längere MD-Simulationen gebraucht werden, je länger das betrachtete Peptid ist. Die Konstruktion eines MSMs umfasst viele Schritte (von der Auswahl der essentiellen inneren Koordinaten hin zum spektralen Clustern) und der Erfolg des kinetischen Modells hängt vom Modellierer ab. Folglich ist es wichtig, ein modernes framework wie VAMPnets zu verwenden, welches die gesamte Abfolge der Datenverarbeitung in einem einzigen framework vereint und ein einfaches, leicht zu interpretierendes kinetisches Modell mit wenigen Zuständen liefert.

Desweiteren wurde beobachtet, dass mit der Länge der Peptide die Eigenschaften der Hydratationshülle insofern „besser“ werden als dass die mittlere Anzahl an Wasserstoffbrücken per Carbonylgruppe und eine wohl definierte zweite und dritte Hydrathülle einem Bulksystem näher kommen. Schliesslich konnte in den aus den first-principles MD-Simulationen berechneten Powerspektren der Carbonylgruppen ein typischer Trend herausgelesen werden: eine grössere Anzahl an Carbonylgruppen hat eine breitere Amid-I-Bande zur Folge und die vermehrten Wechselwirkungen mit den Wassermolekülen verstärken die Intensitätsmuster in den Powerspektren.

Chapter 1

Introduction

Proteins are vital for life because they perform a wide variety of functions in living organisms. For instance, the operation and survival of neurons are dependent on several critical functions of the related proteins, such as communication, metabolism, and repair [BER78]. Moreover, the unique tertiary structure of proteins determines their specific function, and the conformational changes on multiple time and length scales are common during the protein function. Motor proteins, which convert chemical energy into directed motion, are perhaps the most obvious example [SH18]. Numerous experimental and computational techniques have been used to study cytoskeletal motors at multiple spatial and temporal resolutions (e.g., angstrom to millimetre and microseconds to seconds) [GCC⁺21]. Therefore, being flexible and dynamic molecules proteins inherit conformational diversity, a crucial element for their function in the aqueous environment [HWK07]. This structural heterogeneity of proteins causes interconversion of their thermally accessible conformations on a complex free energy landscape [BNW09]. The synergy between non-covalent intramolecular interactions of proteins and protein-water interactions is considered the driving element of such conformational transitions [BSS15, NK13]. The conformational dynamics of proteins, water fluctuations, and their coupling play a crucial role in many biological functions, such as ligand binding [BLS⁺13, DHS⁺19] or protein-protein interactions [AGGH11, SHZ09], to name a few. As an example, the role of water in amyloid aggregation/misfolding [SS19, SKM⁺21] and its importance in designing new drugs has been reviewed [CMNF21, SAB⁺17].

Thus, it is crucial to study the function and dynamics of proteins in order to understand the underlying biochemical processes. Which requires 1) detailed *information of the molecular structures* of relevant proteins, 2) and the ability to access/analyse their *conformations and associated dynamics at the longer timescales*.

Multiple experimental spectroscopic techniques (such as X-ray crystallography, nuclear magnetic resonance (NMR), vibrational spectroscopy etc) are capable of providing structural information of proteins. Furthermore, with the recent advancements in NMR [AK21], single-molecule techniques [MZS⁺17], ultrafast 2D infrared (IR) techniques [Hun09, KC13] to determine the structure of the protein, to probe heterogeneity of free energy states across a molecular population, and to

observe molecular dynamics on the picosecond timescale, respectively, it is possible to accurately determine the molecular structure and the molecular dynamics up to some extent.

However, the both sluggish and fast conformational changes of proteins are not easily accessible simultaneously in experiments and the interpretation of experimental results almost always required some calculations to help assigning the measured spectra and to understand the underlying features precisely [TBF⁺13, Gai21]. As an example, the experimentally measured IR spectra are assigned with the help of static calculations i.e., quantum chemical geometry optimization followed by the harmonic normal mode analysis of one or few conformations of the respective molecule in vacuum or by using implicit solvent models. The calculated spectra of different conformations are then matched with experimentally measured spectrum to search which conformation is responsible for the measured vibrational signatures.

Molecular dynamics (MD) simulations (Section 2.5) offers atomistic descriptions of protein/peptides structures and motions, and protein-water interactions. However, also with simulations a combined approach is required: whereas first-principles MD simulations (Section 2.5.3) can accurately analyze the dynamics but only small timescales are accessible, large timescales (hundreds of nanoseconds up to microseconds) can only be explored with classical forcefield based MD simulations (Section 2.5.2).

Peptides are ideal candidates for such combined studies as they can be sampled extensively using classical force-field MD simulations as well as computationally expensive first-principles MD simulations are possible for the calculation of the property of interest. They are often used as small, tractable model systems for proteins, in order to study their conformational dynamics and the dynamics of the surrounding water molecules, which play a key role in protein function [FL14]. Typically, protein or peptide dynamics take place at longer timescales while the timescales of water dynamics are around picoseconds, mainly due to high mobility of water molecules and frequent changes in hydrogen bonding state [RMH02, Bag05, NDK⁺07]. Therefore, to understand both protein dynamics and water dynamics and the interplay between them and to complement experimental results, it is essential to study small model peptides in explicit solvent using MD simulations.

In this regard, many different combined approaches exist to understanding the interplay between conformational dynamics and water dynamics in the literature. A combined experimental and computational approach on dialanine in water found that the spectral diffusion of the Amide-I vibration is dictated by water solvation dynamics [FT17a]. The combination of classical MD simulations with several data science algorithms showed the significance of water bridges around the peptide, trialanine [JH18]. Most recently, the computational study of insulin dimer in aqueous solution using Markov state models (MSM's) (Section 2.6) and

computational Amide-I spectroscopy predicts how conformational substates can be observed in experiments [FSPT21]. Also, purely using MD simulations, [XB01] showed that the kinetics of breaking and forming water–water hydrogen bonds is slower in the first solvation shell than in bulk water. [MS02] revealed that a specific conformation's stabilization depends on the interplay between enthalpy and entropy. [KNG04] studied the role of the water network during the formation of β -turns. [KNS10] and [JGH17] showed the importance of the formation of critical water bridges that define the peptide's structure.

For our combined approach, the first step is an extensive sampling of conformational space of the solvated model peptide in order to explore the possible conceivable conformations that require longer timescales to be observed. The empirical forcefields based MD simulations are pretty successful in this regard. Further, if combined with MSM's, they (the MD + MSM's) allow identifying the metastable-conformations and estimation of their corresponding timescales [CSP⁺07, CN14, PWS⁺11]. Further, to understand how the underlying conformation affects the vibrational signatures and how the vibrational signatures are impacted by the dynamics of water surrounding exposed polar residues of each conformation the next step is the calculation of accurate vibrational signatures of each metastable conformation in explicit solvent, which requires more sophisticated first-principles MD simulations. Density functional theory (DFT) MD simulations have been successfully applied, as it allows on the fly calculation of electronic potential energy and nuclear dynamics. Using such simulations, the computation of vibrational spectra of small systems in the explicit solvent is also achievable [MGD⁺06, GMV07, Gai10b, TBF⁺13]. The DFT-MD simulations provide reliable infrared (IR) spectra of metastable-conformations in explicit solvent that account for finite temperature, anharmonic effects, and the dynamic average of the starting coordinates.

The spectrum is obtained by Fourier transformation of the time correlation function of dipole moment trajectory. The calculation of IR spectra using the first principles MD simulation is computationally very demanding. Depending upon the size of the system, it is only possible to perform these simulations typically up to a few hundred picoseconds. For smaller peptides, 20 – 25 ps long trajectories are enough for fast processes such as hydrogen bond breaking [GS03] and rearrangements of water molecules around exposed polar residues to average out [HFI21]. The DFT-MD based IR spectra give accurate insights into the structure of molecules and the intra- and intermolecular interactions. Using IR fingerprints, one can also distinguish between multiple metastable-conformations.

The DFT-MD based IR can be compared directly with the experimentally measured IR spectra of the conformational state of a protein in the so-called amide region. As the vibrational fingerprints of this region are due to the motions of the groups involved in the peptide bond, that is carbonyl, $C = O$, and $N - H$ group stretching and bending motions. Depending on the backbone conformation of a peptide, e.g. a fully formed α -helix or a β -sheet (Section 2.1), the characteristic fre-

quencies of the bands in this region differ, in principle, allowing an assignment of the observed conformations. Moreover, for proteins/peptides, as many possible conformations are conceivable, and it is not a priori clear which one dominates and whether and how these interchange. With the help of distinguishable calculated spectra of estimated metastable-conformations, it is possible to decipher the measured IR spectrum into spectra of its constituent metastable-conformers.

Apart from the metastable conformations and associated timescales, the hydration shell directly influences the vibrational properties/strength of the individual polar bonds, impacting the calculated/measured IR spectra. Moreover, the hydration properties are also interconnected with the underlying conformation's spatial organization, flexibility, and characteristic dynamics. Furthermore, length and time scales are greatly influenced by the change in length of the protein, as well as the hydrating water molecules. Therefore, it is valuable to study model systems such as $\text{Ala}_n\text{-Leu}_n$ peptides in explicit solvent using combined classical and first-principles MD simulations plus MSM's. This enables to dissect/quantify each factor (metastable conformations, hydration shell effect on vibrational properties of individual conformation, impact of change in conformation on the hydration shell, etc.) separately.

Therefore, as a first step in this thesis, a combined approach was established (Chapter 4). The effect of the hydration shell on vibrational spectra was then examined (Chapter 5). The interplay between vibrational signatures of the metastable conformations and the hydration shell is then analysed (Chapter 6). Finally, the effect of peptide length on slow timescales and hydration properties is studied (Chapter 7).

Chapter 2

Theoretical Background

2.1 The Conformational Dynamics of Peptides/Proteins

There are only 20 different amino acids which are the molecular building blocks of natural proteins. Figure 2.1 shows two of them i.e., Alanine (Ala) and Leucine (Leu). Each of the amino acids is composed of a central carbon atom (C_{α}) bonded to a hydrogen atom, an amino group (NH_2), a carboxyl group ($COOH$), and a sidechain (see Figure 2.2 a)). The sidechain is the distinguishing factor among the 20 natural amino acids. Any two amino acids can combine end-to-end such that the carboxyl group of one amino acid condenses with the amino group of the other to eliminate water resulting in the formation of a peptide bond (see Figure 2.2 b)). This procedure is repeated to combine any number of amino acids. Whether two amino acids or more are combined, the amino group of the first amino acid and the carboxyl group of the last amino acid remain intact. The formation of a series of peptide bonds results in the formation of a “mainchain” or “backbone” from which the various side chains project. Thus, the linear polymers formed by the head-to-tail linkage of amino acids via peptide bonds are called **peptides**. The backbone of the peptides is composed of carbon atoms (C_{α}), an N-H group and a carbonyl group ($C=O$). Peptide bonds connect the C atom of one residue to the nitrogen atom of the next, forming a polypeptide. The linear polypeptide chains of particular sequences of amino acids form the so-called *primary structure* of proteins, followed by the formation of the local three-dimensional structure of the backbone of the polypeptide, known as the *secondary structure* of proteins, due to the non-bonded interactions between the atoms of the polypeptide chains. The local secondary structure of the proteins is defined by the term “**conformation**”. The most common conformations are β -sheet like, α -helices, and $L\alpha$ -helices. These conformations act as building blocks for the overall three-dimensional shape (called the *tertiary structure* of proteins), which is stabilized by the interactions of the secondary structure elements. The unique tertiary structure of protein is associated to its specific function, and the conformational changes on multiple time and length scales are common during

the protein function.

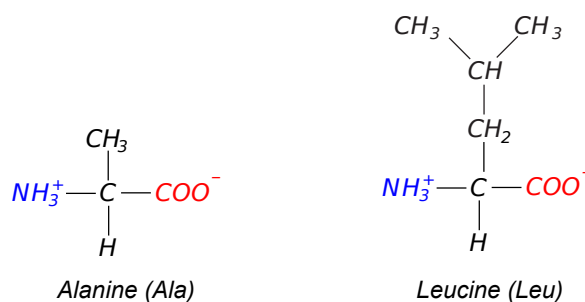


Figure 2.1: Schematic of two natural amino acids, Alanine (Ala) and Leucine (Leu).

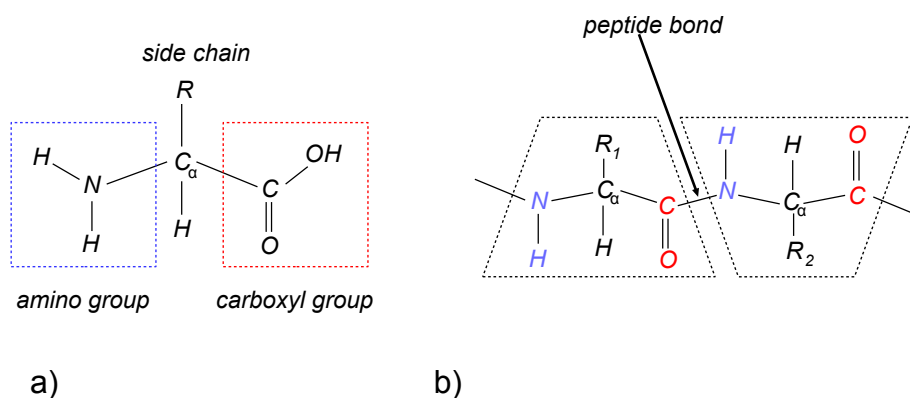


Figure 2.2: a) Schematic of an amino acid, b) Schematic of formation of the peptide bond.

The stiff peptide units of the proteins are connected by covalent bonds at the C_α atoms. Each unit can only rotate around $\text{N} - \text{C}_\alpha$ and $\text{C}_\alpha - \text{C}$ bonds. The angle of rotation around the $\text{N} - \text{C}_\alpha$ bond is called phi (ϕ dihedral angle), while the angle around the same $\text{C}_\alpha - \text{C}$ bond is called psi (ψ dihedral angle), as shown in Figure 2.3. Each amino acid residue is thus assigned two conformational angles, ϕ , and ψ . Typically, the angle pairs ϕ and ψ are plotted against one another in a diagram known as a **Ramachandran plot** [RRS63]. Due to the steric hindrance between mainchain and sidechain atoms, only certain combinations of these angles are allowed. Most likely and sterically “allowed” $[\phi, \psi]$ -combinations are associated with β -sheets and α/L_α -helices like conformations as can be seen in Figure 2.4 [LDAI+03]. For more details on the structure of proteins, see [BT12].

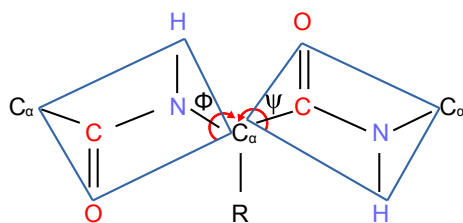


Figure 2.3: Definition of a dihedral angle ϕ and ψ .

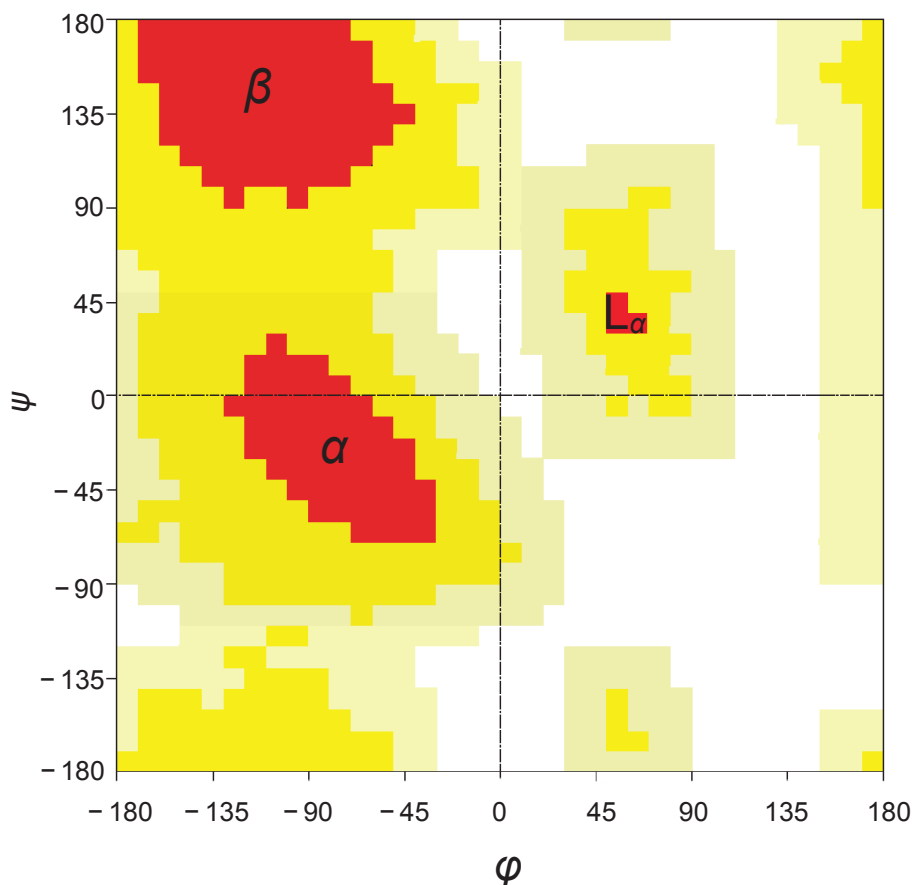


Figure 2.4: Ramachandran plot for all 20 amino acids in 163 experimental structures resolved to 2 Å. The probability distribution takes its largest values in the regions shaded red and decreases through the yellow regions into the least probable $[\phi, \psi]$ -combinations, which are shaded white. The three distinct maxima are associated with β -sheets, and α -helices, and L_α -helices which correspond to elements of secondary structure [Wal04].

Due to the conformational flexibility of proteins/peptides, their structure and dynamics are essentially manifestations of the underlying potential energy surface

(PES). The PES is a function of all relevant atomic coordinates that represents the potential energy of a given system such as proteins.

The potential energy, U , of a three-dimensional system containing N atoms is a function of $3N$ spatial coordinates, expressed as the components of a three-dimensional vector, X . Thus, the PES, $U(X)$, is a three-dimensional object embedded in a $(3N + 1)$ -dimensional space, with the additional dimension corresponding to the value of the potential energy function. It is not difficult to realize how the PES determines the structure and dynamics of the protein. The local minima of $U(X)$ correspond to mechanically stable conformations where, $\nabla U = 0$, and all the forces vanish. On the other hand, the non-vanishing forces everywhere else on the PES, determine the protein's dynamics and other possible conceivable conformations [Wal04].

There are two theories to describe the PES (also known as energy landscape) of proteins/peptides i.e., classical theory (Section 2.2) and the quantum (Section 2.3) theory which are discussed in the following sections.

2.2 Classical Theory

For a detailed introduction to the theory of classical mechanics the reader is referred to the books [GPS02, Cra13].

2.2.1 Single Particle Dynamics

Consider a particle of mass m constrained to move in a particular coordinate direction q , subject to force $F(q, t)$. The general scheme of classical mechanics is to determine the position of the particle at any time i.e., $q(t)$. Once known, the other dynamical variables of interest can also be determined. For example, the velocity ($v = \frac{dq}{dt}$), the momentum ($p = mv$), the kinetic energy ($\frac{1}{2}mv^2$), etc, and the state of the particle can be characterized.

The $q(t)$ is determined by the Newton's second law

$$F = ma \tag{2.1}$$

And, the classical relationship between force F in a particular coordinate direction q and potential energy U is as follows

$$F = -\frac{\partial U}{\partial q} \tag{2.2}$$

Comparing Equations 2.1 and 2.2

$$m \frac{d^2 q}{dt^2} = -\frac{\partial U}{\partial q}, \quad a = \frac{d^2 q}{dt^2} \quad (2.3)$$

By solving above equation with appropriate initial conditions, $q(t)$ can be determined.

2.2.2 Classical Harmonic Oscillator

The classical harmonic oscillator (mass-spring system) is the model system. It is a model for many other systems in nature that behave as harmonic oscillators, see Figure 2.5. As, typically, we are interested in determining the motion of the system in its ground state or other low energy states.

According to the Hooke's law the force F is a restoring force and is proportional to the displacement q of the mass from equilibrium

$$F_q = -kq \quad (2.4)$$

By equating Equations 2.1 and 2.4

$$-kq = m \frac{d^2 q}{dt^2}$$

And defining

$$\omega = \sqrt{\frac{k}{m}} \quad (2.5)$$

The equation of motion for the classical harmonic oscillator is written as

$$\frac{d^2 q}{dt^2} = -\omega^2 q(t) \quad (2.6)$$

And, the solution of the above equation is

$$q(t) = A \cos(\omega t + \phi) \quad (2.7)$$

where the amplitude A and phase constant ϕ are determined by the initial state of the motion of the system (i.e., initial conditions). The motion is characterized by a single angular frequency, ω (i.e., a single harmonic).

Hence it is possible to determine the complete set of all harmonic oscillator **trajectories** which fill the corresponding two-dimensional space (known as **phase space**).

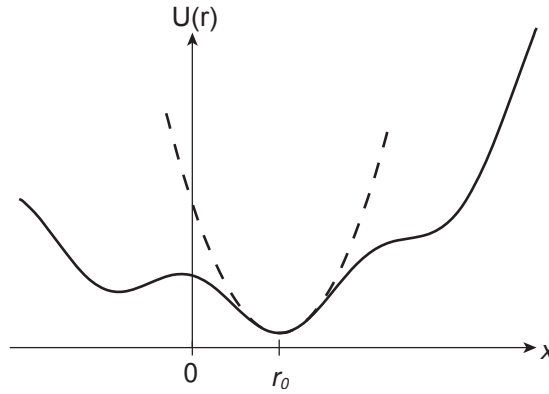


Figure 2.5: A general potential energy function (solid) is approximated by a quadratic harmonic potential (dashed) in the vicinity of the potential minimum.

2.2.3 Force and Potential Energy

Consider any two particles i and j , and the bond length between them r (Diatomic-Spring model). In order to write the potential energy function in an abstract form it should be continuously differentiable and if the dissociation energy for the bond is positive. It can be defined that if the minimum of the function have a potential energy of zero then the bond length between the particles at minimum is r_0 .

By taking the Taylor expansion about r_0

$$U(r_{ij}) = U(r_0) + \left. \frac{dU}{dr} \right|_{r=r_0} (r - r_0) + \frac{1}{2!} \left. \frac{d^2U}{dr^2} \right|_{r=r_0} (r - r_0)^2 + \frac{1}{3!} \left. \frac{d^3U}{dr^3} \right|_{r=r_0} (r - r_0)^3 + \dots \quad (2.8)$$

The first two terms of the above equation can be set to zero. The first term $U(r_0)$ because a constant potential energy does not affect the motion and second $\left. \frac{dU}{dr} \right|_{r=r_0} (r - r_0)$ by virtue of r_0 being the minimum, as the potential derivative (i.e., slope) is zero at the minimum.

So, by truncating after the first non-zero term, the pair-potential energy function U as a function of r can be written

$$U(r_{ij}) = \frac{1}{2}k_{ij}(r_{ij} - r_0)^2 \quad (2.9)$$

where, $k_{ij} = \left. \frac{d^2U}{dr^2} \right|_{r=r_0}$ is the force constant according to the Hooke's law.

For such pair potentials, according to the relation (Equation 2.2), if the slope of the energy curve with respect to the 'bond-length' coordinate is zero, this implies that there is no force. The potential energy remains zero as the two particles approach one another.

Moreover, the following properties of the harmonic potential can be realized

1. resistance to compression, implying that the interaction is repulsive at close range i.e., ($F > 0$ indicates that a positive force acts along the $+r$ direction to separate the particles)
2. particles must attract each other across a range of separations i.e., ($F < 0$ indicates that a positive force acts along the $-r$ direction to bring the particles together)
3. no interaction if particles are far apart i.e., ($U \rightarrow 0$ and $F \rightarrow 0$ as $r \rightarrow \infty$)

2.2.4 Force-Fields

Generally, in the context of force-fields [MT13], the molecules are described in terms of "bonded atoms", which have been distorted from some idealized ("equilibrium") geometry due to non-bonded van der Waals and Coulombic interactions.

The classical mathematical model for the potential energy of the molecular system is defined as

$$U_{total} = \underbrace{U_{bond} + U_{angle} + U_{dihedral}}_{bonded} + \underbrace{U_{van\ der\ Waals} + U_{coulomb}}_{non-bonded} \quad (2.10)$$

Bonded Interactions: From Equation 2.9:

$$U_{bond} = \frac{1}{2}k_{ij}(r_{ij} - r_0)^2 \quad (2.11)$$

Likewise, assuming three atoms i, j and k and the angle θ between the bonds $i - j$ and $j - k$, the energy function for the angle is

$$U_{angle} = \frac{1}{2}k_{ijk}(\theta_{ijk} - \theta_0)^2 \quad (2.12)$$

and for the dihedral angle (see Figure 2.3 for definition)

$$U_{dihedral} = k_{ijkl}(1 + \cos(\omega_{ijkl} - \omega_0)) \quad (2.13)$$

Note that the dihedral angle (torsion) is periodic, and so the torsional potential energy. It is therefore modeled as an expansion of periodic functions, e.g., a Fourier series.

Non-Bonded Interactions: The simplest functional form to represent the combination of dispersion and repulsion energies is

$$U(r_{ij}) = \frac{a_{ij}}{r_{ij}^{12}} - \frac{b_{ij}}{r_{ij}^6} \quad (2.14)$$

where a and b are constants specific to atoms i and j . This equation define a so-called 'Lennard-Jones' potential [LJ31].

More typical form of the Lennard-Jones potential is written as

$$U(r_{ij}) = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (2.15)$$

Where the constants a and b are replaced by constants ε and σ .

If we differentiate Equation 2.15 w.r.t r_{ij} , we get

$$\frac{dU(r_{ij})}{dr_{ij}} = \frac{4\varepsilon_{ij}}{r_{ij}} \left[-12 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} + 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (2.16)$$

To find the minimum in the Lennard–Jones potential, we set the derivative to zero and by rearranging

$$r_{ij}^* = r_0 = 2^{\frac{1}{6}}\sigma_{ij} \quad (2.17)$$

Where r_{ij}^* is the minimum bond length. By inserting this value for the bond length in Equation 2.15, we get

$$U(r_{ij}) = -\varepsilon_{ij} \quad (2.18)$$

This indicating that the parameter ε is the Lennard–Jones well depth, as shown in Figure 2.6.

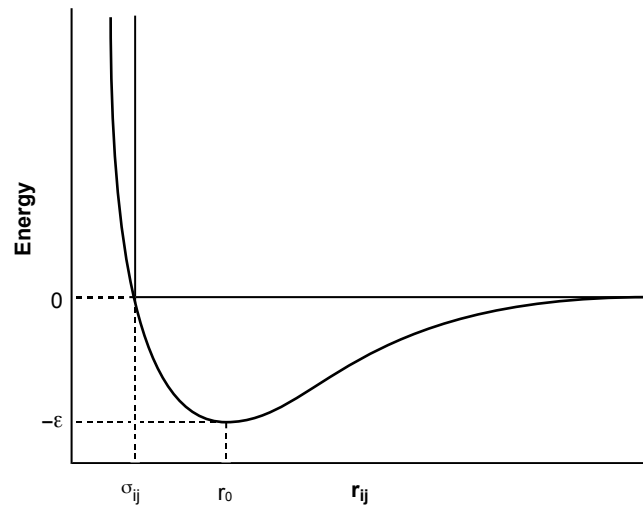


Figure 2.6

Lastly, if atoms are assigned a partial charge, the Coulomb interaction energy between atoms i and j is simply

$$U_{ij} = \frac{q_i q_j}{\varepsilon_{ij} r_{ij}} \quad (2.19)$$

Thereby, the general form of the Equation 2.10 becomes

$$\begin{aligned}
U_{total} = & \sum_{bonds} \frac{1}{2} k_{ij} (r_{ij} - r_0)^2 + \sum_{angles} \frac{1}{2} k_{ijk} (\theta_{ijk} - \theta_0)^2 \\
& + \sum_{dihedrals} k_{ijkl} (1 + \cos(\omega_{ijkl} - \omega_0)) \\
& + \sum_{i=1}^N \sum_{j=i+1}^N \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\varepsilon_{ij} r_{ij}} \right)
\end{aligned}$$

The empirical potential energy functions based on the above equation are known as 'force-fields' and is the classical description of the energy landscapes of protein. There are multiple force-fields available and the exact forms of the equations are force field dependent [WWC⁺04, LLPP⁺10, VHA⁺10, OVMVG04].

Combining the potential energy function (e.g. from a force field) with a kinetic energy function $T(p) = \frac{p^2}{2m}$, provides a Hamiltonian function

$$H(q, p) = T(p) + U(q)$$

So, in general, for N particles of masses $m_1, m_2 \dots m_N$ with position vector $q = (q_1, q_2 \dots q_N)$, momentum vector $p = (p_1, p_2 \dots p_N)$, kinetic energies $\sum_{n=1}^N \frac{p_n^2}{2m_n}$, potential $U(q) = (q_1, q_2 \dots q_N)$, their evolution can be described by using the Hamiltonian equations of motion

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}(q, p) \quad (2.20)$$

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q}(q, p) \quad (2.21)$$

2.3 Quantum Theory

For a detailed introduction to the theory of quantum mechanics, the reader is referred to the books [GS18, SN11, Cra13].

2.3.1 Single Particle Dynamics

In quantum mechanics, the state of the particle at time t is described by the wave function $\Psi(x, t)$, and it is determined by the Schrödinger equation [Sch26]

$$i\hbar \frac{\partial \Psi(x, t)}{\partial t} = H\Psi(x, t) \quad (2.22)$$

Where $i = \sqrt{-1}$, \hbar is the Plank's constant and H is the Hamiltonian operator.

Assuming the potential energy function is time-independent then the SE can be solved by separation of variables, using

$$\Psi(x, t) = \psi(x)\phi(t) \quad (2.23)$$

and solving we get

$$\frac{d\phi}{dt} = -\frac{iE}{\hbar}\phi \Rightarrow \phi(t) = e^{-\frac{iEt}{\hbar}} \quad (2.24)$$

and

$$H\psi = E\psi \quad (2.25)$$

This equation is known as time-independent Schrödinger equation and the resulting wave function $\Psi(x, t) = \psi(x)e^{-\frac{iEt}{\hbar}}$ represent stationary states of definite total energy and time dependence vanishes during the calculation of dynamic variable of interest i.e., expectation values.

2.3.2 Quantum Harmonic Oscillator

The potential energy of the harmonic oscillator (see Section 2.2.3) is

$$U(x) = \frac{1}{2}kx^2 \quad (2.26)$$

Using $k = m\omega^2$, and rewriting the classical Hamiltonian into quantum Hamiltonian for the harmonic oscillator

$$H = \frac{\hat{p}^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2 \quad (2.27)$$

Where \hat{p} and \hat{x} are the momentum and position operators.

So the TISE (Equation 2.25) for the harmonic oscillator become

$$-\frac{\hbar^2}{2m} \frac{d^2\psi(x)}{dx^2} + \frac{1}{2}m\omega^2 x^2 \psi(x) = E\psi(x) \quad (2.28)$$

The wavefunctions that satisfy this equation are Gaussian functions multiplied by Hermite polynomials (shown in Figure 2.7). And the energy eigenvalues are

$$E_n = \left(n + \frac{1}{2}\right) \hbar\omega, \quad n = 0, 1, 2, 3, \dots \quad (2.29)$$

Only discrete energy values with a constant spacing of $\hbar\omega$ are allowed and the ground state energy (so-called zero-point energy) is $E_0 = \frac{1}{2}\hbar\omega$.

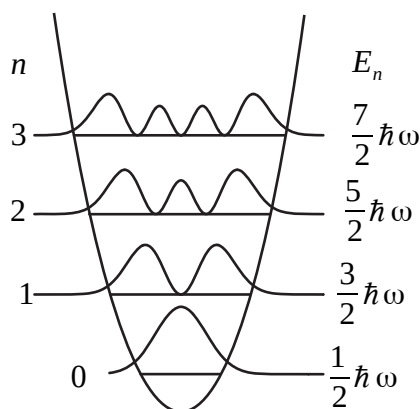


Figure 2.7: The wavefunctions of the quantum harmonic oscillator.

2.3.3 The Born-Oppenheimer Approximation

According to the Born–Oppenheimer approximation, the nuclei of molecular systems move at a much slower rate than the electrons under typical physical conditions implying that the electronic ‘relaxation’ with respect to nuclear motion is

almost instantaneous [BO27]. So, the electronic and nuclei motions can be decoupled and only requirement is the estimation of electronic energies for fixed nuclei positions.

The typical form of the molecular Hamiltonian operator takes into account five contributions to the total energy of a system

$$H = - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 + - \sum_k \frac{\hbar^2}{2m_k} \nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i < j} \frac{e^2}{r_{ij}} + \sum_{k < l} \frac{e^2 Z_k Z_l}{r_{kl}} \quad (2.30)$$

Where i and j run over electrons, k and l run over nuclei, m_e is the mass of electron, m_k is the mass of nuclei k , ∇^2 is the Laplacian operator, e is the charge on electron, Z is the atomic number, and r is the distance between the two particles.

The Equation 2.30 contains terms indicating pairwise attraction and repulsion, implying that no particle moves independently of the others i.e., 'correlation'. The nuclear kinetic energy term (2^{nd}) is taken to be independent of the electrons, correlation in the attractive electron–nuclear potential energy term (3^{rd}) is eliminated, and the repulsive nuclear–nuclear potential energy term (5^{th}) becomes constant for a given system. Hence, the SE becomes

$$E_{elec} \psi_{elec} = (H_{elec} + U_N) \psi_{elec} \quad (2.31)$$

where, E_{elec} is the electronic energy, U_N is the nuclear-nuclear repulsion energy and is constant for a given set of fixed nuclear coordinates. Note that the E_{elec} depends parametrically on the nuclear positions.

2.3.4 Density Functional Theory

For a molecular system, assuming the fixed nuclear positions (i.e., BO approximation), in order to solve the electronic Schrödinger equation, the electronic wave function ψ_{elec} , which depend on one spin and three spatial coordinates for each electron present in the system, is not easy to determine.

A widely used method to solve the ESE is the density functional theory (DFT), a theory of correlated many-body systems [KBP96] based on the Hohenberg–Kohn theorems (Section 2.3.4.1) [HK64, KS65]. It make use of a priori construction of the Hamiltonian operator which depends on the positions, atomic numbers of the nuclei and the total number of electrons. The total number of electrons, N relate

to the physical observable density n as

$$N = \int n(r) dr \quad (2.32)$$

As the nuclei are effectively point charges, clearly their positions correspond to local maxima in the electron density. So, in order to completely specify the Hamiltonian, the assignment of the nuclear atomic numbers is needed. Since for each nucleus i located at the electron density maximum r_i

$$\left. \frac{\partial \bar{n}(r_i)}{\partial r_i} \right|_{r_i=0} = -2Z_i n(r_i) \quad (2.33)$$

where Z is the atomic number of nucleus i , r_i is the radial distance from i , and ρ is the spherically averaged density.

2.3.4.1 The Hohenberg–Kohn Theorems

Existence Theorem The external potential of a uniform electron gas is the uniformly distributed positive charge, whereas the external potential of a molecule is the attraction to the nuclei. The existence theorem proves that the ground-state electron density $n_0(r)$ determines the external potential $U_{ext}(r)$ [HK64].

Proof: Let two different external potentials $U_{ext}^{(1)}(r)$ and $U_{ext}^{(2)}(r)$ which lead to the same ground state density $n(r)$. The two external potentials lead to two different Hamiltonians, $H^{(1)}$, and $H^{(2)}$ which have different ground-state wave functions $\psi^{(1)}$ and $\psi^{(2)}$, which are hypothesized to have same ground-state density $n_0(r)$. Since $\psi^{(2)}$ is not the ground state of $H^{(1)}$, it follows that

$$E^{(1)} = \langle \psi^{(1)} | H^{(1)} | \psi^{(1)} \rangle < \langle \psi^{(2)} | H^{(1)} | \psi^{(2)} \rangle \quad (2.34)$$

The inequality follows if the ground-state is not degenerate. The last term can be written

$$\langle \psi^{(2)} | H^{(1)} | \psi^{(2)} \rangle = \langle \psi^{(2)} | H^{(2)} | \psi^{(2)} \rangle + \langle \psi^{(2)} | H^{(1)} - H^{(2)} | \psi^{(2)} \rangle \quad (2.35)$$

$$= E^{(2)} + \int d^3r [U_{ext}^{(1)}(r) - U_{ext}^{(2)}(r)] n_0(r) \quad (2.36)$$

so that

$$E^{(1)} < E^{(2)} + \int d^3r \left[U_{ext}^{(1)}(r) - U_{ext}^{(2)}(r) \right] n_0(r) \quad (2.37)$$

Similarly, $E^{(2)}$ can be considered in exactly the same way, we will get the same equation with superscripts interchanged.

$$E^{(2)} < E^{(1)} + \int d^3r \left[U_{ext}^{(2)}(r) - U_{ext}^{(1)}(r) \right] n_0(r) \quad (2.38)$$

By adding Equations 2.37 and 2.38

$$E^{(1)} + E^{(2)} < E^{(2)} + E^{(1)} \quad (2.39)$$

Hence, the density uniquely determines the external potential.

Variational Theorem The optimization of the ground-state electron density is needed and the variational theorem proves that the density obeys a variational principle [HK64].

Proof: Since all the properties are uniquely determined if the ground state density $n(r)$ is specified, then each property can be viewed as a functional of $n(r)$, including the total energy functional

$$E_{HK} = T[n] + E_{int}[n] + \int d^3r U_{ext}(r)n(r) + E_{II} \quad (2.40)$$

$$E_{HK} \equiv F_{HK}[n] + \int d^3r U_{ext}(r)n(r) + E_{II} \quad (2.41)$$

where, E_{II} is the interaction energy of the nuclei and the functional $F_{HK}[n] = T[n] + E_{int}[n]$ which being the only functional of density, is universal.

Further, consider a system with a ground state density $n^{(1)}(r)$ corresponding to the external potential $U_{ext}^{(1)}(r)$, the Hohenberg-Kohn functional is equal to the expectation value of the Hamiltonian in the unique ground state, which has wave

function $\psi^{(1)}$

$$E^{(1)} = E_{HK}[n^{(1)}] = \langle \psi^{(1)} | H^{(1)} | \psi^{(1)} \rangle \quad (2.42)$$

A different density $n^{(2)}(r)$ corresponds to a different wave function $\psi^{(2)}$. It follows that the energy of this state $E^{(2)}$ is greater than $E^{(1)}$, since

$$E^{(1)} = E_{HK}[n^{(1)}] = \langle \psi^{(1)} | H^{(1)} | \psi^{(1)} \rangle < \langle \psi^{(2)} | H^{(1)} | \psi^{(2)} \rangle = E^{(2)} \quad (2.43)$$

Hence, the energy given by the Equation 2.41 calculated for the correct ground state density $n_0(r)$ is lower than energy value for any other density $n(r)$.

It follows that if the functional $F_{HK}[n]$ was known, then by minimizing the total energy of the system w.r.t. to variations in the density function, one would find the exact ground state density and energy.

2.3.4.2 Kohn-Sham Equations

Consider a fictitious system of non-interacting electrons with the same overall ground-state density as a real system of interest with electrons that interact i.e., As a starting point, the Hamiltonian operator are the one for a non-interacting system of electrons.

Using a set of orthonormal orbitals as a basis for charge density $n(r)$

$$n(r) = 2 \sum |\psi_i(r)|^2 \quad (2.44)$$

where, $\psi_i(r)$ are Kohn-Sham orbitals that are used to expand the charge density. Typically, the Kohn-Sham orbitals are represented as linear combinations of a finite set of n well-known basis functions

$$\psi_i(r) = \sum_{j=1}^m C_{ij} \phi_j(r) \quad (2.45)$$

The commonly used basis sets are atomic orbital-like Gaussian type orbitals (GTO's), plane waves (PW's), and Gaussian and plane wave methods (GPW's [LPM97, VKM⁺05]) (implemented in CP2K [KIDB⁺20a]), each with its own computational efficiency, system-related advantages and disadvantages. We make use of GPW's,

in which the Kohn–Sham orbitals are expanded in terms of GTO basis functions, but a second representation of the electron density is also made using a PW auxiliary basis. Note that due to high cut-off of the PW's, either pure or mixed implementations, are usually accompanied by the so-called pseudopotentials. We employed the pseudopotentials of Goedecker, Teter, and Hutter (GTH) [GTH96, HGH98, Kra05]. A pseudopotential, in-essence, replaces the interaction potential of the non-valence electrons of an atom and the Coulomb potential of the nucleus with an effective potential. For more detail about the different type of basis set implementations, pseudopotentials, see [Mar20].

The kinetic energy in terms of such orbitals is expressed as

$$T[n(r)] = \sum \frac{-\hbar^2}{2m} \int d^3r \psi_i(r)^* \nabla^2 \psi_i(r) \quad (2.46)$$

Using Kohn-Sham orbitals, the SE like equation can be solved for the effective potential $V_{eff}(r)$

$$\left(\frac{-\hbar^2}{2m} \nabla^2 + U_{eff}(r) \right) \psi_i(r) = E_i \psi_i(r) \quad (2.47)$$

where,

$$U_{eff}(r) = U_{ext}(r) + \int d^3r' \frac{e^2 n(r')}{r - r'} + U_{xc}(r), \quad U_{xc} = \frac{\delta E_{xc}}{\delta n(r)} \quad (2.48)$$

U_{xc} is the exchange-correlation potential and the Kohn-Sham equations (Equation 2.47) are solved self-consistently. However, the exchange-correlation energy functional $E_{xc}[n(r)]$ is unknown and different approximations are required. It takes into account the remaining electronic energy that is not accounted for in the non-interacting kinetic and electrostatic terms.

It is usually split into two parts: exchange and correlation

$$E_{xc}[n(r)] = E_x[n(r)] + E_c[n(r)] \quad (2.49)$$

In order to account for exchange-correlation energy, many exchange-correlation functionals are developed. As an example, a simple local-density approximation (LDA) model which is based on the assumption of slowly varying electron den-

sity so it can be treated as uniform electron gas locally [Blo29, Dir30]. Some functionals are based on the gradient of the electron density called generalized gradient approximation (GGA). We employed so-called BLYP exchange-correlation functional which is the combination of the exchange functional of Becke(B) [Bec88] and the correlation functional of Lee, Yang, and Parr (LYP) [LYP88], GGA-type functionals. The shortcoming of exchange-correlation functionals is their inability to adequately describe dispersion interactions, several approaches exist to include the dispersion interactions such as empirical DFT-D3 [GAEK10].

The solution of the Equation 2.47 for a given effective potential U_{eff} i.e., $U_{eff}^{input} \rightarrow n^{output}$ is computationally very expensive and is considered as a 'black-box' because except for the exact solution the input and output potential and densities are not consistent. In order to reach exact solution a new potential U_{eff}^{new} is defined using the output density n^{output} and this procedure is repeated iteratively i.e., $U_i \rightarrow n_i \rightarrow U_{i+1} \rightarrow n_{i+1} \rightarrow \dots$, until the potential converges as a function of density or vice-versa called self-consistency and thereby the Kohn-Sham equations (Equation 2.47) sometime referred as self-consistent Kohn-Sham equations [KS65].

For the details of DFT, see [Mar20, Cra13].

2.4 Statistical Theory

The Boltzmann distribution is fundamental to statistical mechanics. It is derived by maximizing the entropy of the system subject to the constraints on the system.

For a system of N particles with n_1 particles in energy levels ε_1 , and n_2 particles in the energy level , and so on and so forth, there are W ways to achieve this distribution:

$$W(n_1, n_2, \dots) = \frac{N!}{n_1! n_2! \dots} \quad (2.50)$$

The most favorable distribution correspond to the configuration with just one particle in each energy level, i.e., $W = N!$, and it has the the highest weight. However there are two important constraints on the system.

First, the total energy is fixed

$$\sum_i n_i \varepsilon_i = E \quad (2.51)$$

Second, the total number of particles is fixed

$$\sum_i n_i = N \quad (2.52)$$

The Boltzmann distribution gives the number of particles n_i in each energy level ε_i as

$$\frac{n_i}{N} = \frac{\exp(-\varepsilon_i/k_B T)}{\sum_i \exp(-\varepsilon_i/k_B T)} \quad (2.53)$$

Where k_B is the Boltzmann constant.

The denominator in the above equation is the molecular partition function

$$q = \sum_i \exp(-\varepsilon_i/k_B T) \quad (2.54)$$

The exciting properties of a system in MD simulations are made up of a number of particles, and an *ensemble* is a collection of such a system (see second 2.5.2 for different types of thermodynamic ensembles). The energy of each member of the ensemble is distributed according to the Boltzmann distribution. This leads to the concept of the ensemble partition function Q .

Various thermodynamics properties can be calculated from the partition function such as the most common Helmholtz free energy A , and Gibbs free energy G

$$A = -k_B T \ln Q$$

$$G = -k_B T \ln Q + k_B T V \left(\frac{\partial \ln Q}{\partial V} \right)_T$$

Where T is the temperature and V is the volume.

For a detailed introduction to the theory of statistical mechanics, the reader is referred to the book [Cra13].

2.5 Molecular Dynamics

To explore the energy landscape of proteins/peptides described by the force-fields and DFT (see Sections 2.2.4 and 2.3.4), and is accomplished by the molecular dynamic (MD).

In essence, MD is a simulation method based on the numerical integration of Newton's equation of motion (Equation 2.3) i.e.,

$$m \frac{d^2 r}{dt^2} = - \frac{\partial U}{\partial r} \quad (2.55)$$

and the commonly used integration scheme to solve these equations is known as the Verlet algorithm [Ver67].

2.5.1 Verlet Algorithm

By Taylor expansion of the coordinate of a particle, around time t

$$r(t + dt) = r(t) + v(t)(dt) + \frac{d^2 r}{dt^2} dt^2 + \dots O(dt^4) \quad (2.56)$$

Similarly

$$r(t - dt) = r(t) - v(t)(dt) + \frac{d^2 r}{dt^2} dt^2 + \dots O(dt^4) \quad (2.57)$$

Where the acceleration is calculated using the interatomic forces. By adding Equations 2.56, 2.57 and generalizing

$$r_i(t + dt) = 2r_i(t) - r_i(t - dt) + \frac{d^2 r_i}{dt^2} dt^2 + \dots O(dt^4) \quad (2.58)$$

It estimates the positions of the particles at time $t + dt$ given the positions at the earlier times t and $t - dt$.

Likewise, the velocities are determined by

$$v_i(t) = \frac{r_i(t + dt) - r_i(t - dt)}{2dt} + \dots O(dt^2) \quad (2.59)$$

The velocities calculated using the Verlet algorithm [Ver67] are less accurate than the position.

A more widely used algorithm is however the Velocity Verlet algorithm [SABW82]. The derivation is quite similar, except that velocity is explicitly included in equation. 2.58

$$r_i(t + dt) = r_i(t) + v_i(t)(dt) + \frac{1}{2}a_i(t)dt^2 + \dots O(dt^4) \quad (2.60)$$

$$v_i(t + dt) = v_i(t) + \frac{a_i(t) + a_i(t + dt)}{2}dt + \dots O(dt^4) \quad (2.61)$$

It is simple and straightforward to implement, and the accuracy reaches up to the fourth order.

2.5.2 Classical MD

The classical MD is referred as an empirical method to determine the PES by the parametrization of the potential energy as an analytic function, i.e., force-fields (see Section 2.2.4) without treating electronic degrees of freedom and the time evolution of the system is governed by the classical Newton equations of motion (Equation 2.3).

Numerically, it is solved using the Verlet algorithm (see Section 2.5.1) [Ver67]. The most frequently used approach is to use periodic boundary conditions (by considering the appropriate volume of the system) and to calculate the energy of long-range electrostatic interactions using the particle-mesh Ewald method [Ewa21]. Newtonian dynamics entails that the system maintains constant total energy (microcanonical ensemble i.e., number of particles N , volume V and energy E are constant (NVE)) and progresses in a manner prescribed by its initial conditions. However, actual systems include certain stochastic degrees of freedom due to their connection to an external environment that functions as a heat bath. In this situation, the system's total energy varies within a specified distribution defined by a specific temperature and pressure (canonical ensemble (NVT) or isothermal-isobaric ensemble (NPT)).

Once successfully performed, MD enables the computation of global quantities that typically correspond to the thermodynamics of the system being modeled.

For example, the total kinetic energy of the modeled system

$$K.E = \left\langle \sum_i \frac{1}{2} m v_i^2 \right\rangle \quad (2.62)$$

or the radial distribution function as the output of the MD simulation is the set of positions r_{ij}

$$\rho g(r) = \left\langle \frac{1}{N} \sum_i \neq \delta(r - r_{ij}) \right\rangle \quad (2.63)$$

where the brackets represent the ensemble average and N is the mean number density of the system.

or the vibrational density of states i.e., power spectra

$$g(\omega) = \frac{1}{V} \sum_k \delta(\omega - \omega_k) \quad (2.64)$$

The Fourier transform of the velocity autocorrelation function can be used to estimate the sum over the system's eigenmodes

$$g(\omega) = \int_0^\infty dt \langle v(t)v(0) \rangle \exp(-i\omega t) \quad (2.65)$$

2.5.2.1 Nose-Hoover Thermostat

In this approach, the Hamiltonian of the system is extended by including a heat bath and a friction term in the equations of motion [Nos84a, Nos84c, Hoo85]. The friction force is proportional to the product of the velocity of each particle and a friction coefficient, ζ . It has its own momentum and equation of motion and the time derivative is calculated by subtracting the current kinetic energy from the reference temperature. Equation of motion includes an extra term in this case

$$\frac{d^2 \mathbf{r}_i}{dt^2} = \frac{\mathbf{f}_i}{m_i} - \frac{p \zeta}{Q} \frac{d \mathbf{r}_i}{dt} \quad (2.66)$$

where Q is the coupling constant and the equation of motion for the heat bath is

$$\frac{dp_\xi}{dt} = T - T_0 \quad (2.67)$$

where T_0 is the reference temperature and T is the current temperature.

2.5.3 First-principles MD

As the electronic and nuclei motions can be decoupled by the Born-Oppenheimer approximation (see Section 2.3.3) [BO27]. The first-principles MD relies on the estimation of the PES by the first-principles methods for electronic structure calculations such as DFT [HK64, KS65] (see Section 2.3.4) and nuclei are treated classically and there are no other empirical parameters involved.

For a set of nuclei with an interaction energy $E[\{\mathbf{R}_I\}]$ dependent upon the position of the particles \mathbf{R}_I , as they are treated classically, their motion is propagated by the Newtons equation of motion

$$M_I \frac{\partial^2 \mathbf{R}_I}{\partial t^2} = - \frac{\partial E}{\partial \mathbf{R}_I} = F_I[\{\mathbf{R}_I\}] \quad (2.68)$$

Just like in classical MD, the Verlet algorithm (see Section 2.5.1) is used for numerical integration of the above equation.

Rewriting equation 2.40 as a function of Kohn-Sham orbitals ψ_i and position of the nuclei \mathbf{R}_I

$$E[\{\psi_i\}, \{\mathbf{R}_I\}] \equiv \sum_{i=1}^N \int \psi_i^*(\mathbf{r}) \left(-\frac{1}{2} \nabla^2 \right) \psi_i(\mathbf{r}) d\mathbf{r} + U[n] + E_{II}[\{\mathbf{R}_I\}], \quad (2.69)$$

$$U[n] = \int d\mathbf{r} V_{ext}(\mathbf{r}) n(\mathbf{r}) + \frac{1}{2} \int \int d\mathbf{r} d\mathbf{r}' \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + E_{xc}[n], \quad (2.70)$$

$$n(\mathbf{r}) = 2 \sum_{i=1}^N |\psi_i(\mathbf{r})|^2 \quad (2.71)$$

$$F_I = -\frac{\partial E}{\partial \mathbf{R}_I} \quad (2.72)$$

Where, E_{II} is the interaction energy of the nuclei, $n(\mathbf{r})$ is the electron charge density, $V_{ext}(\mathbf{r})$ is the electron-nuclei interaction, $E_{xc}[n]$ is the exchange-correlation energy, and F_I is the force.

In summary, within the Born-Oppenheimer approximation, the problem becomes one of minimizing the energy of electrons (at fixed nuclei positions \mathbf{R}_I) and solving for the motion of the nuclei simultaneously. For more details on first-principles MD see, [MH09].

2.6 Markov State Modelling

A Markov state model of the long-time conformational dynamics is constructed from discrete partitioning of the conformational space into micro-states¹ (Markov states) [SPS04a, PWS⁺11]. To this end, a transition matrix \mathbf{T} is set up that estimates the underlying stochastic process [SHD01], here transitions between conformational micro-states of the system. The entries of the matrix are

$$T_{ij}(\tau) = P(x_{t+\tau} = j | x_t = i) \quad (2.73)$$

The elements of the transition matrix represent the conditional probabilities of finding the molecule in state j at time $t + \tau$, given it was in state i at time t . This matrix defines a Markov process in which the propagation of the system is entirely determined by knowing its present state x_t and is independent of its past. The dynamic system furthermore fulfils detailed balance, that is in equilibrium all processes are reversible with the number of transitions $i \rightarrow j$ equal to the number of transitions $j \rightarrow i$.

The eigenvalues $\lambda_i(\tau)$ and eigenvectors $\mathbf{r}_i, \mathbf{l}_i^T$ (right and left) of the transition matrix are important ingredients to understand the prominent features of the conformational dynamics

$$\begin{aligned} \mathbf{T}(\tau) \mathbf{r}_i &= \mathbf{r}_i \lambda_i(\tau) \\ \mathbf{l}_i^T \mathbf{T}(\tau) &= \lambda_i(\tau) \mathbf{l}_i^T \end{aligned} \quad (2.74)$$

¹Please note that a micro-state in the context of this work refers to a set of conformations and not to a point in phase space.

The transition matrix is row-stochastic

$$\sum_{j=1}^N T_{ij}(\tau) = 1 \quad \forall i \quad (2.75)$$

and for ergodic dynamics its largest eigenvalue $\lambda_1(\tau) = 1$. The corresponding left eigenvector is the stationary distribution.

The other eigenvalues $|\lambda_{i>1}(\tau)| < 1$ define the implied time-scales which can be understood as molecular relaxation timescales

$$t_i = -\frac{\tau}{\log |\lambda_i(\tau)|} \quad (2.76)$$

The corresponding eigenvectors represent processes e.g., the conformational transitions that occur on those timescales [SPS04a, BH08, PWS⁺11]. The implied time scales and Chapman-Kolmogorov tests [PWS⁺11] can be used to determine the quality of an MSM.

2.6.1 Perron-cluster cluster analysis

A coarse-grained transition matrix can be constructed using the perron-cluster cluster analysis (PCCA+). It make use of the eigenvector structure and tries to finds an optimal linear transform of the eigenvector coordinates into a probability simplex in order to define the metastable-sets of microstates [SFHD99, DW05, RW13].

2.6.2 Time-lagged Independent Component Analysis

The time-lagged independent component analysis (TICA) defines a linear transform of a high dimensional set of input coordinates to a low dimensional set of output coordinates [MS94, PHPG⁺13, SP13].

Suppose the input data $x(t)$, first we compute two covariance matrices i.e, instantaneous $C_{ij}(0)$ and time-lagged $C_{ij}(\tau)$

$$C_{ij}(0) = \langle x_i(t)x_j(t) \rangle \quad (2.77)$$

$$C_{ij}(\tau) = \langle x_i(t)x_j(t + \tau) \rangle \quad (2.78)$$

Where $x_i(t)$ represent the mean-free i^{th} feature at time t .

It becomes the generalized eigenvalue problem

$$C(\tau)U_i = C(0)(\tau)U_i\lambda_i \quad i = 1, \dots, N \quad (2.79)$$

Where U_i is the eigenvector matrix whose columns are the independent components, λ_i is the eigenvalue matrix.

The data can now be projected onto the TICA space as

$$z^\top(t) = r^\top(t)U_i \quad (2.80)$$

So, by selecting only the first m columns of the full-rank U_i , the dimension reduction is performed.

2.7 Normal modes

Normal mode analysis is a technique for describing the flexible states of the molecular system such as protein/peptide. Once the system is perturbed from equilibrium (from the energy minimum conformation) i.e., when the forces acting on a system are equal to zero, such states are accessible.

It is shown in Section 2.2.3 by the Taylor expansion of potential energy minimum

$$U(r_{ij}) = \frac{1}{2}k_{ij}(r_{ij} - r_0)^2 \quad (2.81)$$

In terms of mass-weighted coordinates $q_i = \sqrt{m_i}\Delta\tilde{q}_i$, where \tilde{q}_i denotes the i^{th} coordinate's displacement from the energy minimum and m_i denotes the mass of the corresponding atom. The above equation can be written as

$$U = \frac{1}{2} \sum_{i,j=1}^{3N} \left. \frac{\partial^2 U}{\partial q_i \partial q_j} \right|_0 q_i q_j \quad (2.82)$$

The double derivatives can be written in the form of a matrix, so-called the Hessian matrix \mathbf{H} . The diagonalization of which implies

$$\omega_j^2 A_j = \mathbf{H} A_j \quad (2.83)$$

Where, ω_j^2 is the j^{th} eigenvalue and A_j is the j^{th} eigenvector. Each eigenvector represent a normal coordinate by

$$Q_j = \sum_{i=1}^{3N} A_{ij}q_i \quad (2.84)$$

These normal mode coordinates are shown to oscillate harmonically and independently of one another with the angular frequency ω_j

$$Q_j = W_j \cos(\omega_j t + \phi_j) \quad (2.85)$$

W_j is the amplitude and ϕ_j is the phase of the oscillation.

2.7.1 Normal Modes of a Linear Triatomic Molecule

Consider a model based on a linear symmetrical triatomic molecule. Two atoms of mass M are symmetrically located on either side of an atom of mass m in the molecule's equilibrium configuration. All three atoms are on one straight line, the equilibrium distances between them is k . Using different coordinates for each atom i.e., x_1, x_2, x_3 , the Newtonian equations are

$$\ddot{x}_1 - \frac{k}{M}(x_1 - x_2) \quad (2.86)$$

$$\ddot{x}_2 - \frac{k}{m}(x_2 - x_1) - \frac{k}{m}(x_2 - x_3) \quad (2.87)$$

$$\ddot{x}_3 - \frac{k}{M}(x_3 - x_2) \quad (2.88)$$

Looking for the frequencies ω such that all masses vibrate simultaneously know as normal modes. Suppose

$$x_i = x_i e^{i\omega t}, \quad i = 1, 2, 3.$$

Putting this into Newtonian equations above and writing in a matrix form

$$\begin{pmatrix} \frac{k}{M} & -\frac{k}{M} & 0 \\ -\frac{k}{m} & 2\frac{k}{m} & -\frac{k}{m} \\ 0 & -\frac{k}{M} & \frac{k}{M} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = +\omega^2 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad (2.89)$$

This is a eigenvalue equation with the matrix asymmetric. The secular equation is

$$\begin{vmatrix} \frac{k}{M} - \omega^2 & -\frac{k}{M} & 0 \\ -\frac{k}{m} & 2\frac{k}{m} - \omega^2 & -\frac{k}{m} \\ 0 & -\frac{k}{M} & \frac{k}{M} - \omega^2 \end{vmatrix} = 0 \quad (2.90)$$

This leads to

$$\omega^2 \left(\frac{k}{M} - \omega^2 \right) \left(\omega^2 - \frac{2k}{M} - \frac{k}{M} \right) \quad (2.91)$$

The eigenvalues are

$$\omega^2 = 0, \quad \frac{k}{M}, \quad \frac{k}{M} + \frac{2k}{m}$$

all real. And by evaluating the corresponding eigenvectors for each eigenvalue we get

For $\omega^2 = 0 \rightarrow x_1 = x_2 = x_3$, pure translation motion.

For $\omega^2 = \frac{k}{M} \rightarrow x_1 = x_3, x_2 = 0$, the two outer masses are moving in opposite direction.

For $\omega^2 = \frac{k}{M} + \frac{2k}{m} \rightarrow x_1 = x_3, x_2 = -\frac{2M}{m}x_1$, the two outer masses are moving together and the central mass is moving opposite to the two outer ones.

Such type of vibrations are called normal modes. Hence, it is possible to describe the displacements of three masses as a linear combination of above three type of motions.

2.7.2 Quantum Chemical Calculation of Normal Modes Spectra

Within the Born-Oppenheimer Approximation [BO27] (Section 2.3.3), it is possible to obtain all normal modes of a single molecule and the corresponding eigen-

values (i.e., frequencies) using DFT (Section 2.3.4) based quantum chemical methods. The normal modes analysis involve solving the TISE (Equation 2.25) for the nuclei of a molecule. The energy eigenvalues of the quantum harmonic oscillator (Section 2.3.2) are

$$E_n = \left(n + \frac{1}{2}\right) \hbar\omega, \quad n = 0, 1, 2, 3, \dots \quad (2.92)$$

Since, only the discrete energy levels are allowed, in order to jump from one state to another, electromagnetic absorption or emission of energy equivalent to the energy difference of the two energy levels is needed. For the two adjacent energy levels of the quantum harmonic oscillator, it is equivalent to $\hbar\omega$ (see Figure 2.7). This energy difference for the vibrational energy levels of the nuclei of the molecules lies in the infrared (IR) range of electromagnetic spectrum. These vibrational energy levels correspond to the characteristic peak position in the IR spectra and the intensities of the peaks is calculated by taking the derivatives of the dipole moment along the normal coordinates [NRKH02]

$$A_k = \frac{1}{4\pi\epsilon_0} \frac{N_A\pi}{3c^2} \left(\frac{\partial\bar{\mu}(\mathbf{Q})}{\partial Q_k}\right)_0^2 \quad (2.93)$$

Where $\bar{\mu}(\mathbf{Q})$ is the molecular dipole moment as a function of mass-weighted normal coordinates Q , N_A is the Avogadro constant and c is the speed of light in vacuum. The above equation is derived using the Fermi's golden rule and the evaluation of the transition dipole matrix. For more detail, see [WDC80, NRKH02].

2.8 Theory of Infrared Spectra

According to Fermi's Golden Rule, an infrared spectrum can be calculated through [McQ00]

$$I(\omega) = 3 \sum_i \sum_f \rho_i |\langle f | \vec{\mathcal{E}} \vec{\mu} | i \rangle|^2 \delta(\omega_{fi} - \omega) \quad (2.94)$$

where $\vec{\mathcal{E}}$ is the applied external field vector, $\vec{\mu}$ is the dipole vector of the molecular system. $I(\omega)$ is the intensity as a function of the reciprocal wavenumber of the radiation, ω (in cm^{-1}) and ω_{fi} is the reciprocal wavenumber associated with the transition between the initial and final vibrational states of the system $|i\rangle$

and $\langle f |$, respectively. ρ_i is the density of the molecules in the initial vibrational state $| i \rangle$. Within Linear Response Theory [KTH12], the above equation can be rewritten as the Fourier transform of the dipole moment autocorrelation

$$I(\omega) = \frac{2\pi k_B T \omega^2}{3cV} \int_{-\infty}^{\infty} dt \langle \vec{\mu}(t) \cdot \vec{\mu}(0) \rangle \exp(i\omega t) \quad (2.95)$$

$$f(\omega) = \frac{2\pi k_B T \omega^2}{3cV} \int_{-\infty}^{\infty} dt \langle \vec{r}(t) \cdot \vec{r}(0) \rangle \exp(i\omega t) \quad (2.96)$$

where T is the temperature, k_B the Boltzmann constant, c is the speed of light in vacuum, and V is the volume. The angular brackets represent the statistical average, as sampled by MD simulations, of the autocorrelation of the dipole moment $\vec{\mu}$ of the absorbing molecule. Equation 2.95 yields the complete IR spectrum of the molecular system. For the assignment to vibrations of molecular groups, power spectra are computed from the Fourier transform of the autocorrelation of the velocities (equation 2.96) of individual groups. See also review [Gai10d] on theoretical spectroscopy of floppy peptides at room temperature. For more detail, see [WDC80, NRKH02, HR06].

2.8.1 Schemes to Estimate Instantaneous Molecular Dipole Moments

In MD simulations, generally, a maximally localized Wannier functions scheme [Wan37] is used to estimate the instantaneous molecular dipole vectors. However, recently, another approach based on the radical Voronoi tessellation [GF82], called Voronoi integration (Section 2.8.1.2) has proven successful in achieving the same at a lower computational cost. The former approach tries to find unitary transformation for optimizing the degree of localization of molecular orbitals. In contrast, the latter utilizes a periodic radical Voronoi tessellation in three-dimensional space to the atom positions and finally integrates the total electron density within each molecule's Voronoi cell to obtain a molecular dipole vector. Wannier localization, an iterative process, is computationally expensive, and sometimes does not converge. Voronoi integration, in contrast, is considered cheap [BT21, TBK15], but is still not widely used.

2.8.1.1 Maximally localized Wannier functions scheme

A maximally localized Wannier functions scheme [Wan37] tries to find unitary transformation for optimizing the degree of localization of molecular orbitals. The specific form of this transformation is chosen to minimise a specific spread

functional of Wannier orbitals (see references). The molecular dipole moment is given by

$$\mu_{mol} = -2e \sum_{i=1}^N r_i + e \sum_{I=1}^M Z_I R_I \quad (2.97)$$

Where r_i is the position expectation values of the Wannier orbitals called Wannier function center.

Wannier localization, an iterative process, is computationally expensive, and sometimes does not converge.

2.8.1.2 Voronoi tessellation

Voronoi tessellation [GF82] utilizes a periodic radical Voronoi tessellation in three-dimensional space to the atom positions and finally integrates the total electron density within each molecule's Voronoi cell to obtain a molecular dipole vector. Voronoi integration, is considered cheap [BT21, TBK15], but is still not widely used.

2.9 Theory of Wavelet Analysis

For the calculation of spectrogram and instantaneous stretching frequencies of carbonyl groups, we used wavelet theory. For any signal in time domain $f(t)$, it can be described as

$$W(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} dt f(t) \psi^* \left(\frac{t - \tau}{s} \right) \quad (2.98)$$

$\psi(t)$ is the so-called mother wavelet, which is translated and compressed/dilated by the τ and s parameter, respectively. The τ parameter localizes the frequency in time, and $1/s$ is proportional to frequency. We used the Morlet–Gabor mother wavelet [CHT98], which has been successfully applied in many previous studies to calculate instantaneous stretching frequencies. It is defined as

$$\psi(t) = \pi^{-\frac{1}{4}} e^{i\omega_0 t} e^{-t^2/2\sigma^2} \quad (2.99)$$

where ω_0 represents the main oscillation frequency of the plane wave, and σ represents the width at half-height of the Gaussian time window.

For a mother wavelet to be applicable for wavelet analysis, it should be localized (i.e., have finite energy) and admissible (i.e., have zero mean). In the above-described wavelet, parameter σ controls its locality properties, and it also directly affects the time–frequency resolution of a spectrogram.

The discretized version of the continuous WT is given by

$$W(n, s) = \sum_{n'=0}^{N-1} f(n' \cdot \delta t) \psi \left[\frac{(n' - n) \cdot \delta t}{s} \right] \quad (2.100)$$

The product $n' \delta t$ shows the total time at the n' th time step, which localizes the signal in time, and consequently, WT gives the frequency content of a signal over a Gaussian time window centered at $n \delta t$.

The wavelength for the Morlet–Gabor set of basis functions is defined as

$$\lambda = \frac{s4\pi}{\omega_0 + \sqrt{2 + \omega_0^2}} \quad (2.101)$$

We used $\sigma = 8$, and $\omega_0 = 2\pi$ which yields the value of corresponding effective frequency $\nu = 1.01/s$. For the detailed theory and implementation of wavelet analysis, see [TC98]. The value of s is found such that [it] maximizes the modulus $|W(n, s)|^2$ of the wavelet at a given time step n' . The corresponding value of $1/s$ is taken as the “instantaneous stretching frequency” [MMPCS11]. To calculate the spectrogram and instantaneous frequencies using the wavelet transform, we used the FORTRAN code developed by the group of Pagliai [MMPCS11].

Chapter 3

Methods

The standard methods used in this thesis are briefly described in this Chapter, and the materials/methods specific to the project are provided in the respective Chapter.

Systems: The partially deuterated (Ala-Leu)_n peptides i.e, AL, ALA, ALAL, ALALA, ALALAL, in a cubic simulation box of explicit deuterated water molecules.

Classic MD simulations: As a first step, very long MD simulations are performed based on empirical forcefields (Sections 2.2.4 and 2.5.2), i.e., classical MD simulations (CMD). The standard procedure, which includes energy minimization, system equilibration, and multiple production runs for each system, was carried out in a canonical ensemble (NVT). It enables extensive sampling of the conformational space of the solvated model peptides in order to identify any conceivable conformations that require longer timescales to observe.

Following the same procedure, we also performed short constrained simulations of the most probable conformations of the model peptides in order to get better statistics for the structural analysis of the hydration shell.

Markov state modeling (MSM): Following that, a Markov state model (MSM) of each system was built on the conformational space spanned by the appropriate torsion angles using the trajectories of partly deuterated (Ala-Leu)_n peptides obtained from the classical MD simulations, followed by a perron-cluster cluster analysis. It (CMD + MSM's + PCCA+) enables the identification of metastable conformations and estimation of their corresponding timescales for model peptides (Section 2.6).

Normal modes calculations: For each system (partially deuterated metastable conformations in an implicit solvent), geometry optimisation and subsequent normal mode calculation at the DFT level of theory are performed. Along with power spectra, it aids the assignment of the measured/calculated IR spectra.

First-Principles MD simulations: For each system (partially deuterated metastable conformation in a fully deuterated explicit solvent), initially energy minimisation using a conjugate-gradient algorithm was performed where the positions of the solute atoms were fixed. This allows the solvent molecules to relax around the

solute and find energetically favourable positions, followed by a short NVT equilibration runs from the minimized systems, during which the solute was kept fixed to avoid the transition to an undesired conformation. Subsequently, another long NVT run to sample the independent starting configurations for the production runs was performed. Finally, using the sampled starting configurations, multiple, several picoseconds (20-50 ps) long, NVT/NVE production runs was performed for the desired metastable conformation of the each system (Section 2.5.3). It enables calculation of several static/dynamic/vibrational properties of the peptides.

Following the same procedure, we also performed constrained simulations where certainly number/s of water molecules were constrained to desired hydrogen bond distance/angle.

Molecular dipole trajectories: During the first-principles MD simulations using maximally localized Wannier functions scheme [Wan37], and Voronoi integration based on the radical Voronoi tessellation [GF82] of the electron density is used to estimate the instantaneous molecular dipole vectors. The former approach tries to find unitary transformation for optimizing the degree of localization of molecular orbitals. In contrast, the latter utilizes a periodic radical Voronoi tessellation in three-dimensional space to the atom positions and finally integrates the total electron density within each molecule's Voronoi cell to obtain a molecular dipole vector (Section 2.8).

Infrared (IR) and power spectra: Subsequently, precise vibrational signatures for each metastable conformation are obtained by Fourier transformation of the time correlation function of molecular dipole moment trajectories of the model peptides' molecular dipole moment trajectories and the trajectories of the model peptides' atoms' velocities obtained from the the first-principles MD simulations.

Instantaneous frequencies: Wavelet analysis is used for the calculation of spectrogram and instantaneous stretching frequencies of carbonyl groups using the the first-principles MD trajectories of the desired bond/s (Section 2.9).

Interaction energies: The molecular fragmentation method was employed to calculate the interaction energies of the polar group with the closest water molecules using the data from the obtained from the the first-principles MD simulations.

Geometrical analyses: These analyses include hydrogen bond analysis, radial distribution functions (RDF's), angular distribution functions (ADF's), water bridges, combined distribution functions (CDF's), etc.

Dynamical analyses: These analyses include hydrogen bond life-times, mean square displacement (MSD), etc.

Chapter 4

A Combined Approach

This chapter is based on the publication:

Hassan, I., Donati, L., Stensitzki, T., Keller, B. G., Heyne, K., & Imhof, P. (2018). The vibrational spectrum of the hydrated alanine-leucine peptide in the amide region from IR experiments and first principles calculations. *Chemical Physics Letters*, 698, 227-233.
DOI: <https://doi.org/10.1016/j.cplett.2018.03.026>

4.1 Introduction

To determine the conformations responsible for measured vibrational fingerprints and to investigate the precise conformational dynamics of a small floppy peptide Ala-Leu, we followed a combined approach, i.e., classical MD simulations plus MSM's and theoretical IR spectrum calculations in conjunction with experimental measurements. Representative conformations were estimated using MSM of trajectories obtained from classical MD simulations. The IR spectra using static quantum chemical calculations (harmonic approximation) for each representative conformation are calculated. First-principles MD simulations are performed to move beyond the static approximation and explicitly account for finite temperature effects, anharmonicities, and an explicit solvent. The power spectra is calculated using the Fourier transform of atom velocities autocorrelation. All vibrational frequencies are contained within this power spectrum. Further information about the collective motion of a group of atoms, for example, C=O, can be extracted from such power spectra. The IR spectra whose intensity is determined by the dipole moment's change with respect to the vibrational mode are calculated using the Fourier transform of the dipole moment's autocorrelation. The MD-based spectra contain all frequency and intensity shifts caused by anharmonicity, line broadening due to solute-solvent interactions, and conformational dynamics. A reliable assignment of the spectra to different conformations is possible.

4.2 Methods

4.2.1 Classical Molecular Dynamics Simulations

We performed classical molecular dynamics (MD) simulations of the Ala-Leu peptide in a cubic simulation box of explicit water (844 molecules modelled as TIP3P water [JCM⁺83]) employing the AMBER ff99SB-ILDN [HAO⁺06, LLPP⁺10] force field. We used a minimum distance of 1 nm between the solute and the periodic boundaries of the box, resulting in a total number of atoms of 2564. Water hydrogen atoms and polar hydrogen atoms of the peptide (ND3 and ND) were modelled with the mass of deuterium, mimicking the experimentally measured system. For Lennard-Jones interactions and electrostatic interactions (Particle-Mesh Ewald [DYP93a, EPB⁺95a] with grid spacing of 0.16 and an interpolation order of 4) we used cut-off value of 1 nm.

The system was minimised and equilibrated for 500 ps. Then three MD simulations of 400 ns each were launched which yields a total simulation time of 1.2 μ s. A V-rescale thermostat [BDP07a] was applied to control the temperature at 300 K (NVT ensemble). The positions of the solute atoms were saved to file every 0.25 ps. No constraints were applied and the leap-frog integrator with the time step of 1 fs was employed using the GROMACS simulation package [PPS⁺13].

4.2.2 Markov State Model

Using the trajectory of partially deuterated Ala-Leu (see Figure 4.1) in deuterated water obtained from classical MD simulations, a Markov state model was constructed on the conformational space spanned by the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} (highlighted in Figure 4.1). The distributions of the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} obtained from the classical MD simulations are shown in Figure 4.2. Such torsion coordinates have proven useful to capture the essential dynamics of small peptides such as Ala-Leu [NHSS07, SPS04a, SPS⁺04b, SW08, AON⁺08]. For the projection of the MD trajectory onto the microstates defined by the torsion angles and the PCCA+ analysis we used PyEMMA [STSP⁺15a].

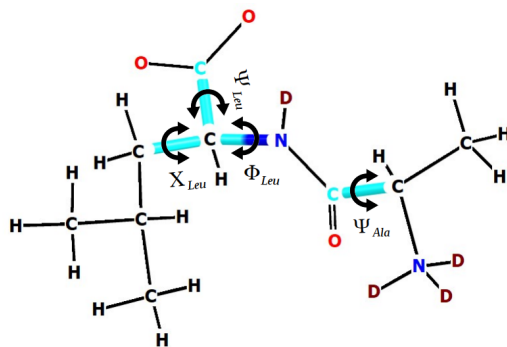


Figure 4.1: Scheme of the Ala-Leu peptide and the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} used for discretisation of the conformational space. Note that ψ_{Leu} is a pseudo-backbone dihedral angle since there is no nitrogen atom of a subsequent amino acid, but a second oxygen atom of the carboxyl terminus. “D” denotes a deuterium atom. A 180° torsion around Ψ_{Leu} , thus does not mean an actual conformational change, but rather an inter-conversion of two chemically equivalent structures.

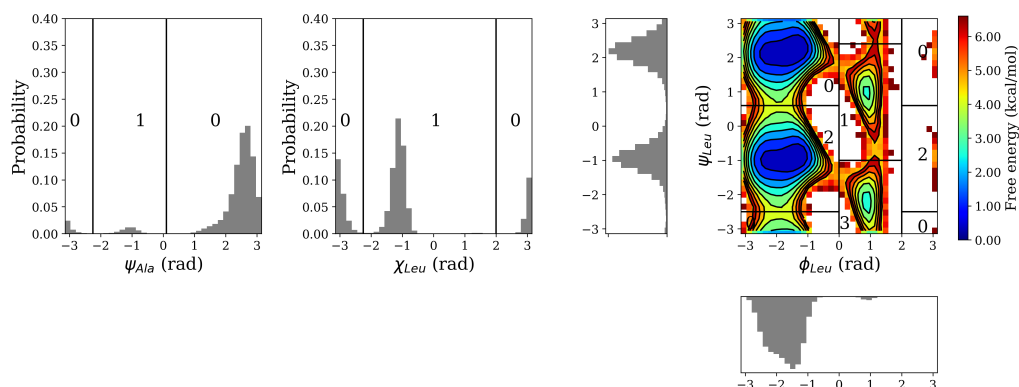


Figure 4.2: Distribution of the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} obtained from the classical MD simulation of Ala-Leu in water.

4.2.3 Normal mode calculations

For each cluster of conformations with a significantly high probability, we performed a geometry optimisation (convergence criterion $3.00\text{E-}04 E_H/\text{\AA}$) and subsequent normal mode calculation in implicit water (modelled by a polarisable continuum model, PCM, with a dielectric of $\epsilon=80$) at the DFT level of theory. The BLYP exchange-correlation density functional and a $6-31\text{G}(d)$ basis set was employed for all these static calculations using the Gaussian programme package [FTS⁺]. The N-terminus and the amino group of the peptide was set deuterated (ND₃ and ND). All normal modes were scaled by 0.992 in frequency. This is the same scaling factor as used for the spectra computed from first-principles simulations but without an extra shift (see below). The spectra obtained from normal mode calculations lack thermal and anharmonic effects. It is therefore reasonable that the exact frequencies in the vibrational spectra computed with the different approaches (normal modes and first-principles MD simulations) do not match exactly. Still, the spectra computed from normal modes facilitate the assignment of the spectra based on the first-principles simulations

4.2.4 First-Principles Molecular Dynamics Simulations

For one representative conformation of each microstate, we performed first-principles molecular dynamics simulations in explicit water. The same deuteration (ND₃ and ND in the peptide and D₂O water) as for the classical force field simulations (see Molecular mechanics simulations in supplementary material) was applied. The cubic simulation boxes had a minimum distance of 0.4 nm minimum between the solute and boundaries of the box and periodic boundary conditions were applied in all three dimensions. To keep the computational cost of the first principles simulations moderate, this box size, and therefore the number of water molecules, is smaller than in the classical simulations (see section "Molecular mechanics simulations") but is still just large enough to avoid interactions between periodic images.

Each Ala-Leu representative conformation was further energy minimised using a conjugate gradient algorithm [PTA⁺92] by keeping the atom positions of the solute fixed and let the water molecules relax around the solute.

The first principle simulations were performed using the CP2K package [HISV14]. We employed the default Gaussian and plane waves (GPW) electronic structure method [LHP99] as implemented in the Quickstep module [VKM⁺05]. We used double zeta valence plus (DZVP) basis set and interaction between valence and core electrons is described by Geodecker-Teter-Hutter (GTH) norm conserving pseudopotentials. A plane wave expansion for the charge density is employed using energy cutoff of 500Ry. BLYP functional is used as exchange correlation functional.

Simulations were performed in an NVT ensemble. The temperature was controlled to be at 300 K using a chain of three Nose-Hoover thermostats [Nos84b] with a time constant for the thermostat chain of 100fs. CP2K default values of other thermostat parameters are used. The time-step for the numerical integration was 0.5 fs.

Atom positions were saved every step and every 5th step Wannier localisation was performed so as to monitor the changes in dipole moment [KSV93, Res94, MV97, SP99b, SP99a, KH04]. For each of the four representative conformations of Ala-Leu three to four individual first-principles simulations were run for 20 – 80 ps (see Figure A.1 for individual runs).

Table 4.1: Details of the first principles simulations

Microstate (see Section 4.3.1)	Length of cubic simulation box (nm)	Number of water molecules	Total number of atoms
0	1.794	160	512
2	1.809	162	518
4	1.657	123	401
6	1.727	151	485

Spectra were calculated using the TRAVIS [BK11] programme. All calculated spectra are scaled by 0.992 in frequency. Such a scaling has been found to correct for the overestimation of high vibrational frequencies such as those in the amide region by BLYP functional [LCW12] and is commonly applied although different approaches to obtain optimal scaling factors exist [HVS01, MMR07, SBG+04]. Since such scaling factors are derived for harmonic mode calculations, i.e. to correct as well for anharmonic effects, different scaling procedures may be necessary for vibrational frequencies calculated from MD simulations. We have introduced an additional shift of 85 cm⁻¹ so as to match best the most intense band in the experimental spectrum and thus simplify assignment.

Conformation and hydrogen-bonds analyses were carried out using MDtraj [MBH+15] and our own Python and Java scripts.

4.2.5 Experimental Setup

Infrared absorption spectra were taken with an Equinox 55 FTIR device (Bruker). Ala-Leu (Sigma-Aldrich, CAS 3303-34-2) was dissolved in D₂O and placed between two CaF₂ windows with a spacer thickness of 0.05 mm. Absorption of D₂O was subtracted in Figure 4.8 to stress the absorption signals of Ala-Leu. Note, in D₂O the exchangeable protons will exchange to deuterons. However, a residual

amount of partially or undeuterated Ala-Leu remains to be present in the sample. The experiments were carried out in the group of Karsten Heyne. group.

4.3 Results

4.3.1 Markov State Model

The conformational space reduced to the four dimensions corresponding to torsion angles was then discretised into microstates (corresponding to conformational clusters), based on the one-dimensional distribution of the torsion angles ψ_{Ala} and χ_{Leu} and the two-dimensional joint distribution (the ramachandran plot) of the torsion angles ϕ_{Leu} and ψ_{Leu} (see Figure 4.2). Two states for the torsion angles ψ_{Ala} and χ_{Leu} , respectively, were found, while the ramachandran plot was divided into four conformational states. All the possible combinations define $2 \times 2 \times 4 = 16$ microstates onto which the MD trajectory was projected.

A transition matrix has been computed for transitions between the microstates in the classical molecular dynamics trajectory with varying the lag time τ up to 500ps. The spectrum of the matrix, calculated for these lag times, indicates a convergence of the implied time scales (see equation 3) of the slowest process at about 50 ps. Using thus a lag time of $\tau = 50$ ps a transition matrix was set up by counting the transitions between the microstates. Based on the transition probabilities between them, microstates were merged into three metastable sets and a coarse-grained transition matrix was constructed by employing a robust Perron Cluster Analysis (PCCA+) [DW05].

The implied time time-scales in Figure 4.3 suggest three slow processes, corresponding to transitions between four metastable sets. The spectral gap of the transition matrix is, however, after the third eigenvalue (see Table in Figure 4.3 b)). Therefore, we tried to perform PCCA+ to identify three and four metastable sets, respectively. The grouping of microstates into the metastable sets is listed in Table 4.2. The fourth partitioning results in a regrouping of formerly separated microstates, denoting that there are several transitions on this time-scale, as already suggested by the eigenvalues. We therefore worked with a MSM of three metastable sets as presented in Figure 3. The first three eigenvectors are given in (see Figure 4.3 b)). The first eigenvector corresponds to the stationary distribution and the other two correspond to the slowest processes which can be understood as transitions between metastable sets of clusters.

4.3 Results

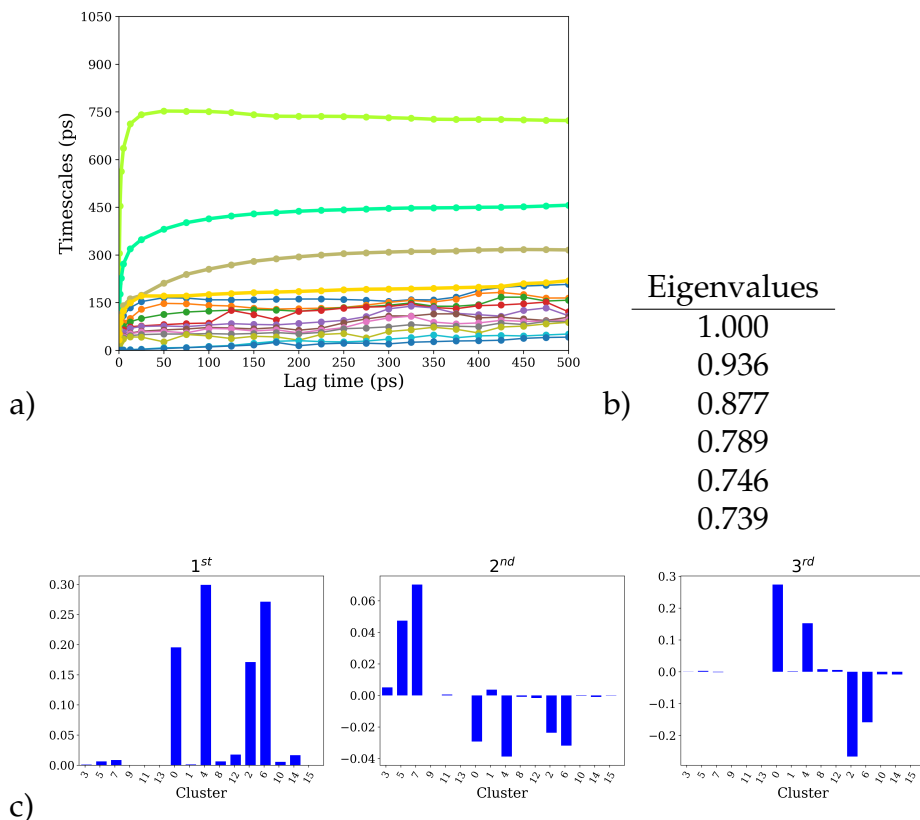


Figure 4.3: a) Implied time scales, b) First eigenvalues of the transitions matrix sampled with lag time $\tau=50\text{ps}$. c) Corresponding first three eigenvectors expressed as contributions of the 16 microstates.

Table 4.2: Meta-stable sets of microstates according to PCCA+

I		II	
[3, 5, 7, 9, 11, 13]		[0, 1, 2, 4, 6, 8, 10, 12, 14, 15]	
I		II	III
[3, 5, 7, 9, 11, 13]		[0, 1, 4, 8, 12]	[2, 6, 10, 14, 15]
I	II	III	IV
[5, 7, 13]	[0, 1, 8]	[4, 6, 12, 14]	[2, 3, 9, 10, 11, 15]

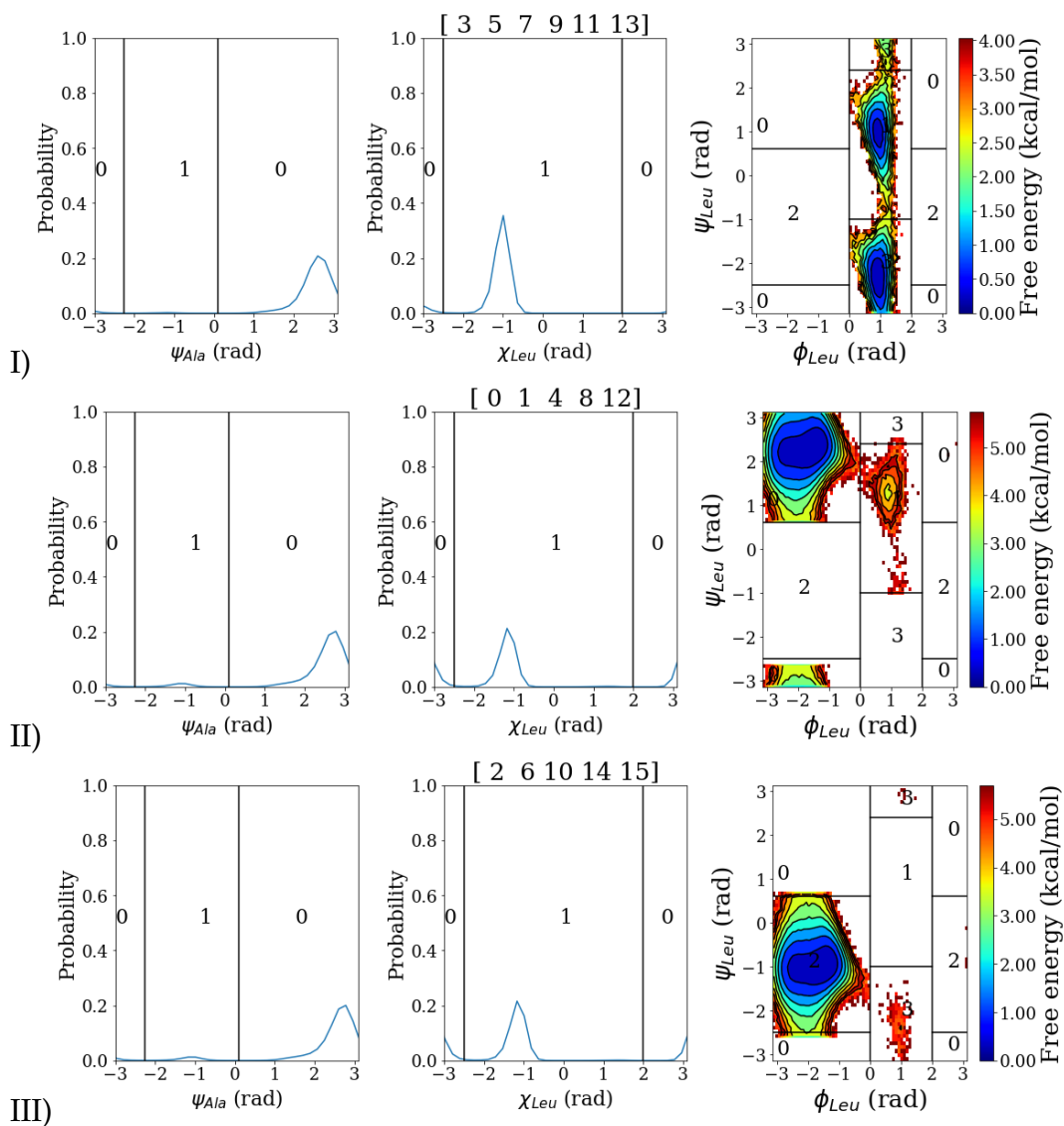


Figure 4.4: Distribution of torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} , and ψ_{Leu} in the conformational clusters of the metastable sets I, II, and III, as obtained from the classical MD simulation of Ala-Leu in water.

Table 4.3: Torsion angles of the representative conformations of the most probable microstates from the classical MD simulations, shown in Figure 4.6 and subjected to further first-principles simulations (except for cluster 5).

Micro-state	ψ_{Ala} (rad)	χ_{Leu} (rad)	ϕ_{Leu} (rad)	ψ_{Leu} (rad)
0	2.48634	-3.00789	-1.26109	2.07314
4	2.39830	-1.24703	-1.54770	2.06933
2	2.31969	-3.09427	-1.35842	-1.17242
6	2.59943	-1.25551	-1.53055	-1.17207
5	2.21018	-1.12151	1.03061	1.40457

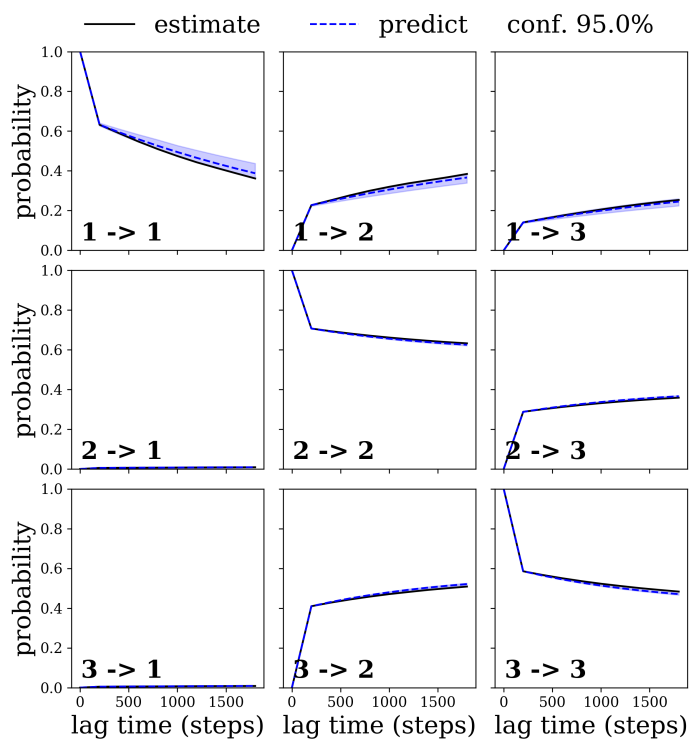


Figure 4.5: Chapman-Kolmogorov test of the Markov state model (MSM) with three states 1,2,3, corresponding to metastable sets I,II,III, respectively. Using the MSM estimated at lag time τ , transition probabilities are predicted for $n \cdot \tau$, and compared to an estimate of the model at lag time $n\tau$.

Figure 4.6 shows the coarse-grained model as a transition graph between metastable sets, together with representative conformations of the most probable microstates in the set. Meta-stable set I consists of microstates with a left-handed helix conformation and has the lowest probability. The transition into this set, corresponding to a torsion around Φ_{Leu} , is the slowest process. The two other metastable sets, II and III, have similar probability and are dominated by microstates, labelled 0 and 4, and 2 and 6, respectively (see Table 4.2 for the complete list of microstates in each metastable set). The transition between the conformations in the two metastable sets II and III corresponds mainly to a torsion around Ψ_{Leu} . Transitions between microstates within the same set, i.e. between 0 and 4, and between 2 and 6, respectively, both correspond to a torsion of the leucine side chain χ_{Leu} .

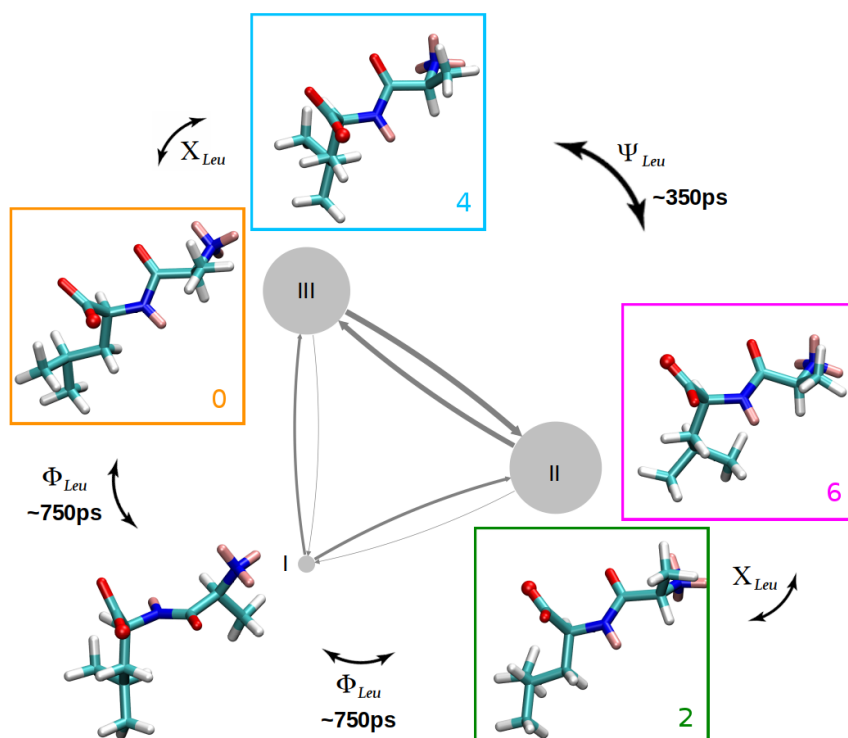


Figure 4.6: Coarse-grained model of the conformational dynamics of Ala-Leu in water. The three metastable sets, I-III, are represented as circles whose size corresponds to the probability of the respective set. The thickness of the arrows between the circles indicates the transition probability between a pair of metastable sets. The molecular structures in the boxes next to the circles are representative conformations of the most probable microstates in the respective set. See Table S1 for the microstates that constitute each set. Arrows between the coloured boxes indicate the coordinate of the conformational transition connecting two microstates. The colour of the boxes correspond to microstate 0 (orange), 2 (green), 4 (blue), and 6 (magenta), respectively. Carbon atoms are shown in cyan, oxygen atoms in red, nitrogen atoms in blue, hydrogen atoms in white and deuterium atoms in pink. One of the two carboxyl oxygen atoms is shown as sphere to illustrate the change in Ψ_{Leu} between metastable set II and III. Note that a Ψ_{Leu} -torsion of 180° inter-converts two chemically equivalent conformations.

4.3.2 Normal Modes

The IR-spectra computed from the normal-modes in implicit water are shown in Figure 4.7, top (dashed lines), together with the optimised geometries. With regard to the backbone conformation, all cluster snapshots converge to the same state which allows the charged COO and the polar ND group, and the CO and the terminal ND₃ group to come close to each other and interact optimally. The optimised structures of conformations 0 and 2 and 4,6 differ, however, in their leucine side chain conformations.

Normal modes computed for “microsolvated” Ala-Leu, that is with a total of seven water molecules added, one of each at each hydrogen bond donor or acceptor in the polar groups, ND₃, CO, ND, and COO, respectively are shown as dotted grey lines in Figure 4.7, top.

The effect of adding one water molecule at a time at each polar group, ND₃, CO, ND, and COO, respectively, is shown in the bottom traces of Figure 4.7. Addition of a water molecule to the C=O group of conformers 0,4, leads to a red-shift of the corresponding stretch vibration. Addition of a water molecule to the C=O group in conformations 2,6 results in a slightly more pronounced red-shift of the CO band. Addition of a water molecule to the carboxyl group has also a similar effect on all conformers, i.e. the appearance of a rather strong COO band at $\sim 1370\text{ cm}^{-1}$. For the frequencies of the N-D stretch vibrations in all conformations upon addition of a water molecule to the respective group, another large red-shift, bringing this band to the amide I region, is observed.

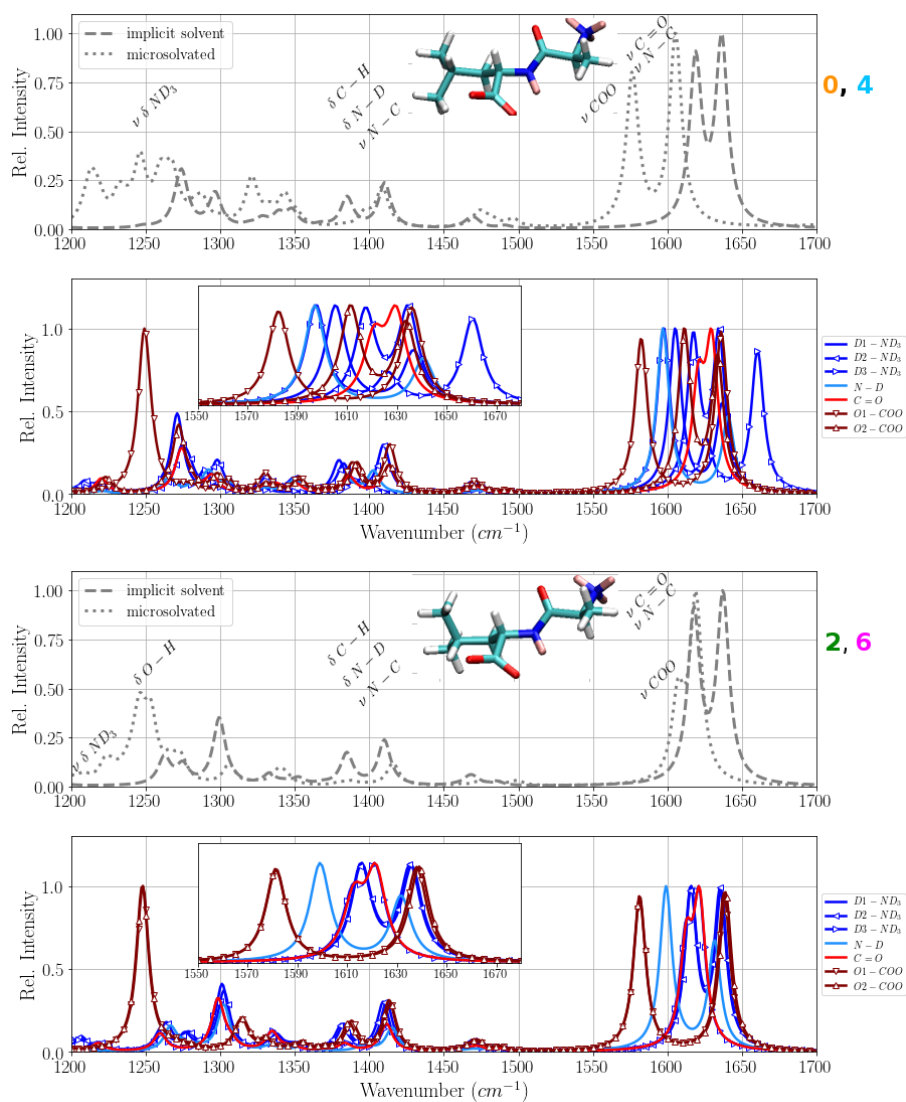


Figure 4.7: IR spectra computed in harmonic approximation for optimised geometries of the four conformers, corresponding to the most probable microstates 0, 2, 4, and 6, respectively. Since conformations 0 and 2 and 4 and 6, respectively, are chemically equivalent only one of them is shown, and labelled as 0, 2 and 4, 6, respectively. Top: normal modes of Ala-Leu in implicit solvent (dashed) and additionally hydrogen-bonded to seven water molecules bound to the polar groups, ND₃, CO, ND, and COO (dotted). Bottom: Normal modes of Ala-Leu in implicit solvent with one water molecule at a time hydrogen-bond to the polar groups, ND₃ (dark blue), CO (red), ND (light blue), and COO (maroon), respectively.

4.3.3 Infrared Spectrum

In the course of the individual first-principles simulations initiated from conformations representing microstates 0, 2, 4, and 6, respectively, the system occasionally undergoes changes in the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} that correspond to transitions between microstates. Hence, an individual simulation can be composed of parts that belong to different microstates, e.g., 0 and 2. In most simulations, there are only a few jumps between microstates. In order to analyse the vibrational fingerprint for individual microstates, we have partitioned the first-principle trajectories by the microstates 0, 2, 4, and 6 and computed spectra from the respective parts of the trajectories. For the time series of torsion angles and the assigned microstate (see Appendix A Figure A.1).

The experimental infrared spectrum of Ala-Leu in water is presented in Figure 4.8 together with the spectra computed from the first-principles MD simulations. The assignment (given as labels in Figure 4.8) is based on the computed power spectra with further aid from normal mode calculations (see supplementary material). The most prominent band is the stretch vibration of the carboxyl group (νCOO) at $\sim 1590\text{ cm}^{-1}$. Note that the most intense band of the νCOO vibration has been used to normalise intensities and therefore shows a relative intensity of one for all spectra. The other band in the amide I region at $\sim 1660\text{ cm}^{-1}$ contains components of the carbonyl group ($\nu\text{C=O}$) and the peptide bond ($\nu\text{N-C}$). The intensity ratio of the two bands, $\nu\text{C=O}$ and νCOO , is well reproduced by the computed spectra. The $\nu\text{C=O}$ band is actually composed of two contributions with varying intensity ratios as can be seen from the spectra computed from the individual microstates (Figure 4.8b)) and also indicated in the considerable error in the composed, weighted spectrum. (Figure 4.8 a) middle).

The amide II bands at $\sim 1450\text{ cm}^{-1}$ and $\sim 1480\text{ cm}^{-1}$, assigned to bend ($\delta\text{N-D}$ and $\delta\text{C-H}$) and stretch ($\nu\text{N-C}$), with some contribution from the CO group, are slightly less well reproduced; the higher frequency band is calculated at too high frequency ($\sim 1540\text{ cm}^{-1}$) with too little intensity. The small shoulder at $\sim 1550\text{ cm}^{-1}$ in the experimental spectrum of Ala-Leu is likely due to remains of undeuterated Ala-Leu in the sample. Experiments on N-methylacetamide [MSM58] report the $\delta\text{N-H}$ bend vibration of the undeuterated species at this frequency ($\sim 1570\text{ cm}^{-1}$) and the $\delta\text{N-D}$ at $\sim 1450\text{ cm}^{-1}$ as in our spectrum of Ala-Leu).

The three smaller bands in the amid III region at $\sim 1360\text{ cm}^{-1}$, $\sim 1390\text{ cm}^{-1}$, and $\sim 1410\text{ cm}^{-1}$ in the experimental spectrum are computed as one broad band at $\sim 1380\text{ cm}^{-1}$ due to the averaging of several simulations, with also a significant variance in the intensities. The spectra computed individually for the microstates (Figure 4.8b)) give rise to two shoulders, albeit with some error, which can be interpreted as corresponding to the lower and higher frequency bands resolved in the experimental spectrum. The main vibrational contribution stems from the terminal ND_3 -group and $\delta\text{N-D}$ and $\nu\text{N-C}$, but there are also COO contributions in varying amounts, depending on the individual simulation. The C=O group

does not contribute to the bands in this frequency range. According to Krimm and Bandekar, both amide II and amide III bands are linear combinations of the same group movements, i.e. $\delta\text{N-D}$ and $\nu\text{N-C}$ [KB86]. In our simulations these bands show different intensity ratios for different simulation runs. As can be seen in the computed power spectra (Appendix A, Figure A.1.3) not only do the $\delta\text{N-D}$ and $\nu\text{N-C}$ contributions fluctuate, there are additional contributions by the C=O and COO group to the bands in the amide II and amide III region, respectively, that vary considerably, explaining the different intensities computed for those bands.

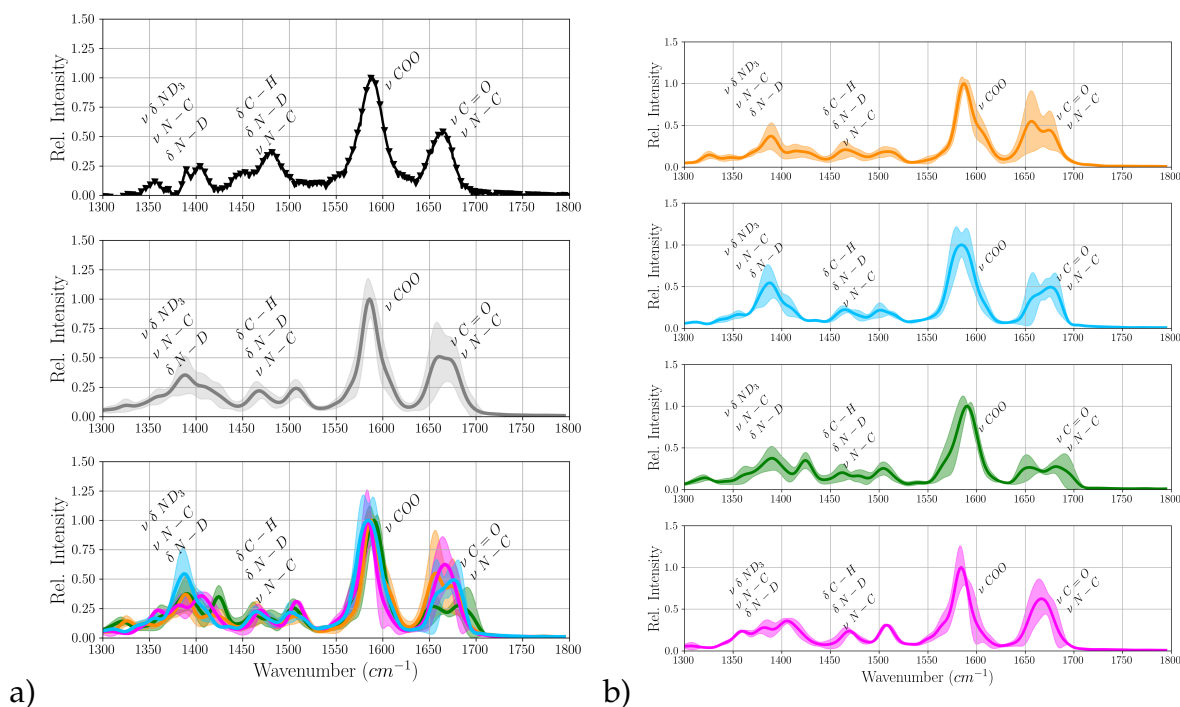


Figure 4.8: a) Experimental (top) and computed (middle) Infrared spectrum of Ala-Leu in water combined from the Boltzmann-weighted average of spectra computed for different microstates (bottom). b) Individual computed spectra of microstates 0 (orange), 2 (green), 4 (blue), and 6 (magenta). The shaded areas indicate the error as computed from the standard deviation from the mean.

4.3.4 Comparison of Sampled Conformations and Resulting Spectra

The distribution of the torsion angles within the partitions of trajectories belonging to microstate 0, 4, 2, or 6, respectively, are shown in Appendix A Figure A.1. Within the range of the torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} , assigned to the respective microstates, there are different fluctuations within the individual runs, resulting in wider, narrower, and occasionally almost bi-modal distributions.

The computed IR spectra show variations in the band assigned to the $\nu\text{C}=\text{O}/\nu\text{N}-\text{C}$ stretch vibration, both in intensity and the exact location of the maximum. In many of the simulations, the computed $\nu\text{C}=\text{O}/\nu\text{N}-\text{C}$ band has a shoulder, others show broadening, and in extreme cases (0-run4, 2-run2 in (see Appendix A Figure A.1), a second band can be observed. Comparison with the distributions of ϕ_{Leu} shows that these correspond to the shape of the $\nu\text{C}=\text{O}/\nu\text{N}-\text{C}$ band in the IR spectrum in as much as narrow bands are only observed for narrow distributions, and broad distributions, with shoulders, correspond to a broadening or splitting in the $\nu\text{C}=\text{O}/\nu\text{N}-\text{C}$ peak. This effect is generally more pronounced in the power spectra than in the IR spectra, indicating that the dipole moments are less affected by the variance in torsion angle than the velocities of the C=O and N-C groups.

A similar observation can be made for the distributions of ψ_{Leu} and the shapes of the νCOO bands. Narrow distributions correspond to narrow bands, shoulders in the ψ_{Leu} distribution correspond to shoulders or broadening in the νCOO band.

The two bands in the amide II region are present in almost all computed spectra, albeit with some fluctuation in their intensities. The $\delta\text{C}-\text{H}$ contribution, mainly of the Leu side chain varies in exact frequency, not always matching the computed IR-band at $\sim 1540\text{ cm}^{-1}$, suggesting only a minor contribution of $\delta\text{C}-\text{H}$ to the IR intensity.

The amide III region exhibits considerable variation of the band intensities. Neither amide II nor amide III region show any relation to the torsion angle distributions, likely because the bands in these regions are composed of motions by several groups, N-C, ND, and ND_3 with contributions of the COO group.

Analysis of the distribution of number of hydrogen bonds from the same parts of the first-principles trajectories (see Appendix A Figure A.5) shows that all polar groups are almost always engaged in at least one hydrogen-bond with a water molecule. In a few cases, (2-run1, 4-run3, and 6-run1) a second hydrogen bond between C=O and water is counted with $> 30\%$ probability. These are the runs of the respective microstate with the red-most $\nu\text{C}=\text{O}$ band, indicating a slightly higher probability of further weakening of the C=O bond by an additional hydrogen-bond in these cases. Observed and computed broad or even split $\nu\text{C}=\text{O}/\nu\text{N}-\text{C}$ bands can thus be explained by the CO group being in different hydrogen-bonded states, resulting in different amounts of red-shift. For the simulation of microstate 0 with the highest probability of a second hydrogen-bond to C=O, the $\nu\text{C}=\text{O}$ band is on the low frequency side, too. This is, however, also the case for other simulations of microstate 0 with no particular relation to the hydrogen-bond probabilities. More detailed analysis of the hydrogen-bond interactions reveals differences in the hydrogen-bond distances and the distributions of donor-hydrogen-acceptor angles between the different simulations of the microstates (see Figures S10 and S11). The impact of water molecules within hydrogen-bond distance on the frequency of the $\nu\text{C}=\text{O}$ band, such as a red-shift, is thus further modulated by the orientation within the hydrogen-bond and thus

the strength of that interaction.

4.4 Discussion

The probability distribution of peptide conformations obtained from the classical molecular dynamics simulations suggest a left-handed helix to be very improbable in the Ala-Leu peptide whereas conformations that correspond to a right-handed α -helix or β -sheet dominate. PolyProline (pPII) conformations, which have been suggested to coexist with β - conformers for the dialanine peptide [Gai10c], are observed in the classical simulations of Ala-Leu only transiently. Accordingly, such conformations have not been taken into account explicitly in the conformation-specific first-principles simulations. In one first-principle simulation, backbone angles that correspond to pPII-conformations have been observed as transition states between two conformational states (4 and 6) and thus only for short life-times (see Figure A.1 4- run1).

All transitions between metastable sets as well as those between the two most probable microstates in the same metastable set ($0 \longleftrightarrow 4$ and $2 \longleftrightarrow 6$, respectively), equilibrate on time scales that are not accessible in the first-principles dynamics. Still, a few conformational transitions between different microstates are observed in the course of some of the first-principles simulations. We have therefore dissected the first-principle trajectories into parts that sample only the same microstate and used these parts for the computation of spectra.

The computed IR spectra show a considerable variance in the band intensities for simulations of the same microstate. In contrast, there are no (further) differences between spectra computed for the different microstates. This is to be anticipated for spectra of microstates 0 and 2, and 4 and 6, respectively, since these correspond to chemically equivalent conformations. The conformational difference between microstates 0 and 4, and between 2 and 6, is the orientation of the leucine side chain as defined by torsion angle χ_{Leu} . The effect on the amide region, if any, is smaller than or comparable to the variances between spectra from different runs initiated from the same conformation.

The experimental and the computed spectra are dominated by the bands assigned to the carbonyl ($\nu_{C=O}$) and carboxyl (ν_{COO}) stretch vibration, respectively. IR spectra of several other small peptides (N-acetyl-Gly-N-methylamide, N-acetyl-Ala-N-methylamide) with capped termini all show only a broad band assigned to the $\nu_{C=O}$ stretch vibration [ECNSS02]. Computations of the spectra reveal that the $\nu_{C=O}$ of Alanine-dipeptide absorbs at almost the same frequency in either the α -helix or polyproline/ β -sheet conformation but with a width that corresponds to the experimentally observed spectrum [Gai10c, VDD⁺15]. Slightly longer peptides, that can form intramolecular hydrogen bonds and thereby stabilise folds such as turns, Ac-Ph-Pro or trialanine, are reported with $\nu_{C=O}$ frequencies that

differ by $\sim 20 - 30 \text{ cm}^{-1}$ [MJRG15, MCG07, ZAH01], giving rise to a shoulder or a double peak. Two-dimensional IR-experiments have furthermore revealed that the two peaks are due to coupled C=O dipoles rather than two conformations [WH00]. Spectra of (Ala)₅, unlabelled and isotope-labelled with $^{13}\text{C}=\text{O}$ and $^{13}\text{C}=\text{O}$ and ^{18}O at individual C=O groups to shift their vibrational frequencies, show dual bands, separated by $\sim 20 \text{ cm}^{-1}$, for both, the isotope-shifted and the unshifted groups [FHK⁺16]. Conformational analysis of classical MD simulations, combined with models for transition dipole coupling, reveals the coupling strength, and hence the detailed band shape, to depend on the conformation [FHK⁺16].

In the short peptide Ala-Leu studied in this work, there is only one C=O group. Any coupling would therefore have to be with the COO group. The νCOO band is observed at 1590 cm^{-1} , at the same position reported for IR spectra of tripeptides ((Ala)₃, (Ser)₃, (Val)₃) [ECNSS02]. In Ala-Leu, the two groups, C=O and COO, exhibit two well-separated ($\sim 50 \text{ cm}^{-1}$) vibrational signals of rather different intensity, suggesting no or only very weak coupling. The $\nu\text{C}=\text{O}$ band maximum differs by $\sim 20 - 30 \text{ cm}^{-1}$ between simulations, some of which indicate a dual peak also within the same simulation. The width of the frequency fluctuations for the computed $\nu\text{C}=\text{O}$ band indicates a relation with the width of the distribution in torsion angle ϕ_{Leu} sampled in that particular simulation. The small differences in this torsion, and similarly of the ψ_{Leu} torsion, give rise to fluctuations in the relative orientation of the C=O and the COO groups (and their corresponding dipoles). Likely, this also leads to fluctuations in the mutual impact of the two groups. Whether and how much the two groups are indeed coupled has to be revealed by future 2D-IR experiments.

4.5 Conclusions

The slow conformational dynamics of Ala-Leu in water are dominated by torsions around backbone angles ϕ_{Leu} and ψ_{Leu} . The slowest process can be attributed to changes in the ϕ_{Leu} torsion angles that lead to transitions to the least probable conformation, a left-handed helix. The most probable part of the conformational space can be formally assigned to the α -helix and β -sheet regions (as assigned by a discretisation of the relevant torsion angles). The inter-conversion of these two regions along ψ_{Leu} is the second slowest process. In the uncapped peptide, these two conformations are, however, actually chemically equivalent and correspondingly exhibit the same spectral signature. Another subdivision of conformations can be made by the orientation of the leucine side chain, corresponding to a torsion around χ_{Leu} .

The IR spectra computed from the first-principles MD simulations reproduce the experimental spectrum of Ala-Leu in reasonable agreement. In accordance with the chemical equivalence of the conformers with the same absolute ψ_{Leu} value,

their spectra are very similar. The orientation of the leucine side chain is not reflected in the amide region of the vibrational spectrum of Ala-Leu as can be seen from comparison of the spectra computed for individual conformations.

The amide I region is very well reproduced by the simulations. The two prominent bands are assigned to the stretch vibrations of the carboxyl group, COO, and the carbonyl group, C=O, respectively. Fluctuations in the backbone torsion angle ψ_{Leu} result in a broadening of the ν COO band. The simulations furthermore reveal the ν C=O band to be composed of (at least) two frequency components. The variance in the exact frequency of this band can be attributed to mainly variations in the backbone torsion angle ϕ_{Leu} within the same area of the peptide fold. These small fluctuations occur on short time-scales and are therefore averaged out in the experimental spectrum, explaining the observation of only one broad ν C=O band.

Chapter 5

Hydration Shell Effect

This chapter is based on the publication:

Hassan, I., Ferraro, F., & Imhof, P. (2021). Effect of the Hydration Shell on the Carbonyl Vibration in the Ala-Leu-Ala-Leu Peptide. *Molecules*, 26(8), 2148.
DOI: <https://doi.org/10.3390/molecules26082148>

5.1 Introduction

The vibrational frequencies of polar groups such as $C = O$, $N - H$, charged termini etc., are sensitive to their interaction with the surrounding water, classified by e.g. their hydrogen bonding states [TT02, Bar07]. The fluctuations in molecular motions of solvent molecules give rise to fluctuations in the vibrational frequencies of polar groups. Similarly, the vibrational frequencies of individual polar groups are influenced by the presence of the surrounding polar groups, either due to direct or indirect vibrational couplings [WCW⁺03], or water-mediated intramolecular interactions [Buc58]. The amide I vibration is depicted by a prominent band in the IR spectra, and is governed by the motion of carbonyl groups. This band is also sensitive to the hydrogen bonding state of the peptide, and due to its intensity in the IR spectrum, a popular marker for the peptide's conformation. It is therefore of interest to study variations in the characteristic amide I frequency, due to changing interactions with the solvent.

In order to obtain both, time and frequency information an analysis of the instantaneous frequencies is required. The localization of the frequency of an input signal in time can be achieved by another integral transform approach called wavelet transform [Dau90, CHT98, TC98] that has recently gained popularity in the molecular dynamics community [MSC08a, MSC08b, PMMC⁺10, MMPCS11, HRGM⁺16, OKK18].

Many experimental and computational efforts have been made to better understand the solute-solvent interaction and the consequences on the amide I region, mainly on short peptides such as N-methyl amide (NMA), di- or tri-alanine [FT17a, FDT18] and other small model peptides [BBB⁺20, ZBC⁺10, KC03, KLAH09].

And several approaches have been used to quantitatively determine how the hydration induced shift on the amide I vibrational band is related to the intermolecular interactions between solute and solvent. Such interactions can easily be computed between individual molecules but, unless for empirical potentials, a dissection of interaction with groups of atoms within one molecule is more involved. To this end energy decomposition schemes based on quantum mechanical calculations and linear scaling techniques to take into account electrostatics, polarization, and charge transfer terms have been successfully applied to NMA [FRLB⁺15]. An alternative way is to implement a molecular fragmentation method and using quantum mechanical methods to calculate these interactions. In the approach used in our study we take out a small fragment of the full molecule to describe the impact of intermolecular interactions on the amide I mode. In another study [FBBB15], a computational protocol (ONIOM) aimed at the quantitative reproduction of the spectra of bio-organic and hybrid organic/inorganic molecular systems with a proper account of the variety of intra- and intermolecular interactions, was applied. By static density functional theory calculations of NMA and NMA–water complexes the impact of hydrogen bonding on the $C = O$ and $N - H$ as well as the amide bond geometry and on the amide I, amide II, and amide III vibrations has been studied [MAA08].

In this work, we investigate the vibrational signature of the small peptide Alanine-Leucine-Alanine-Leucine (ALAL) and the effect of the fluctuations of solute molecules and hydrogen bonding states on the amide I frequencies by employing a combination of first-principles MD simulations, fragmentation methods to quantify interaction energies, and geometrical analyses.

5.2 Methods

5.2.1 Molecular Mechanics Simulations

We performed classical MD simulations of the Ala-Leu-Ala-Leu (ALAL) peptide in a cubic simulation box of explicit water (1477 molecules modeled as TIP3P [JCM⁺83] water) employing the AMBER 99SB-ILDN [HAO⁺06, LLPP⁺10] force field. We used a minimum distance of 1 nm between the solute and the box's periodic boundaries, resulting in side length of 3.61 nm and a total number of atoms of 4492. Water hydrogen atoms and polar hydrogen atoms of the peptide (ND_3 , $N - D$) were modeled with deuterium mass. For Lennard-Jones interactions and electrostatic interactions (Particle-Mesh Ewald [DYP93b, EPB⁺95b] with a grid spacing of 0.16 an interpolation order of 4), we used a cutoff value of 1 nm. The system was minimized and equilibrated for 500 ps. We ran six 2.5 μ s-long MD simulations, which result in a total simulation time of 15 μ s. A V-rescale [BDP07b] thermostat was applied to control the temperature at 300 K (NVT ensemble). The positions of the solute atoms were saved to file every 0.25 ps. No

constraints were applied, and the leap-frog integrator with a time step of 1 fs was employed using the GROMACS simulation package [PPS⁺13].

Free energy distributions are calculated as

$$F = -k_B T \ln Z$$

from the two-dimensional histogram of the ψ and ϕ angles, where k_B is Boltzmann's constant, T is the temperature and $Z = \frac{H_r}{H_0}$ is the count in the histogram, relative to the state with maximal counts.

5.2.2 First-Principles Molecular Dynamics Simulations

The first-principle MD simulations were performed using the CP2K package [HISV14, KIDB⁺20b]. We employed the default Gaussian and plane waves (GPW) electronic structure method [LHP99] as implemented in the Quickstep module [VKM⁺05]. We used a double zeta valence plus (DZVP) basis set, and interaction between valence and core electrons is described by Geodecker–Teter–Hutter (GTH) norm-conserving pseudopotentials. A plane wave expansion for the charge density is employed using an energy cutoff of 500 Ry. BLYP with Grimme's D3 dispersion correction is used as exchange–correlation functional [Bec88, LYP88, GAEK10]. It provides a robust electronic representation for dynamical spectroscopy of hydrated peptides [GS03, GVS05]. During the NVT equilibration, the temperature was controlled to be at 300 K using a chain of three Nose–Hoover thermostats [Nos84a] with a time constant for the thermostat chain of 100 fs. CP2K default values of other thermostat parameters are used. To keep the computational cost of the first-principles simulations moderate, the box size and the number of water molecules are smaller than in the classical simulations but still large enough to avoid interactions of the periodic images. The cubic simulation box had a minimum distance of 0.4 nm between the solute and the box's boundaries, and periodic boundary conditions were applied in all three dimensions. This results in cubic box of side lengths 22.2 Å and 328 solvent molecules. Albeit using a small minimum distance between the solute and the box's boundaries, the total number of atoms jumps to 1045. For such a large number of atoms, the computational cost to perform first-principles MD simulations is already high. Like in the classical MD simulations, water hydrogen atoms and polar hydrogen atoms of the peptide (ND_3 , $N - D$) were modeled with deuterium mass.

First, the system was energy minimized using a conjugate-gradient algorithm where the positions of the ALAL atoms were fixed. This allows the solvent molecules to relax around the peptide and find energy favorable positions. We performed a 5 ps NVT equilibration run from the minimized system, during which the solute was kept fixed to avoid the transition to an undesired conformation, followed by the production run of 50 ps in an NVE ensemble. The time-step

for the numerical integration was 0.5 fs. Atom positions were saved every step, and every fifth step Wannier localization was performed to monitor the changes in the dipole moment [KSV93, Res94, MV97, SP99a, SP99b]. The gathered Wannier centers and position data sets are large enough to compute the reliable IR and power spectra of the amide region of ALAL in explicit solvent. Notably, the hydrogen bond breaking and water rearrangement, which occurs at a 2–3 ps timescale [GS03], is adequately sampled to have a prominent effect on the resulting spectra. In TRAVIS, these saving rates together with the correlation resolution of 1024 and 4096, allow a spectral resolution of infrared and power spectra of $\sim 1.63 \text{ cm}^{-1}$ and $\sim 2.04 \text{ cm}^{-1}$, respectively.

5.2.2.1 Constrained Simulations

We performed three constrained simulations, launched from a snapshot from the previous, unconstrained trajectory. The overall atomic position of either a single or two water molecules was fixed.

(1) One D_2O molecule constrained to a distance of 3.0 Å from the $C_2 = O_2$ group and the acceptor–donor hydrogen angle, $\angle ADH$, constrained to $\sim 30^\circ$.

(2) One D_2O molecule constrained to a distance of 3.2 Å from the $C_2 = O_2$ group and the acceptor–donor hydrogen angle, $\angle ADH$, constrained to $\sim 0^\circ$ and another water molecule, only distance constrained at 4.0 Å from the $C_2 = O_2$ group so as to prevent another water molecule to enter the hydration sphere and to construct a single hydrogen bond situation.

(3) Two D_2O molecules constrained to a distance of 2.6 Å from the $C_2 = O_2$ group and the acceptor–donor hydrogen angle, $\angle ADH$, constrained to $\sim 0^\circ$.

We performed another 5 ps NVT equilibration run for each constraint scenario, followed by the production run of 20 ps each in an NVE ensemble. For these constrained simulations, no Wannier localization was performed.

For both, unconstrained and constrained simulation, Fourier-transform-based spectra and the structural analysis were conducted using the TRAVIS program [BK11, BTGK20] and our own TCL, Python, and Java scripts.

5.2.3 Normal Modes

For ALAL bound to different numbers of water molecules, corresponding to topologies of hydrated ALAL observed in the first-principle MD simulations, we carried out a normal mode analysis using the Gaussian program package [FTS+16]. The snapshots were optimized (convergence criterion $3.00E - 04 E_H / \text{Å}$) in implicit water (modeled by a polarisable continuum model, PCM, with a dielectric of $\epsilon = 80$) at the DFT level of theory. On the optimized geometries, a

frequency calculation was performed in which all polar hydrogen atoms are assigned an atomic mass of 2. Like for the first-principles simulations, the BLYP exchange–correlation density functional was used. To also use a basis set of comparable double-zeta quality, a cc-pVDZ basis set was employed for all these static calculations.

5.2.4 Interaction Energies

The molecular fragmentation method was employed to calculate the interaction energies of the central carbonyl group with the closest water molecules for different snapshots of the system, taken at 50 fs intervals. In this study, we consider as fragment of the molecule the $-CONH$ group to preserve the electronic structure of the peptide bond. First, the molecule was fragmented and hydrogen caps were inserted at the broken $C - C$ and $N - C$ bonds to preserve the valency of the fragment. Then, the water molecule of interest was added to the fragment according to its distance from the $C = O$ group (see Figure 5.1). A similar approach was used in [HF19]. Our fragmentation approach assumes that the other polar groups are far enough from the fragment to not affect the interaction energies with the water molecules. Interaction energies were calculated with Gaussian 16 [FTS⁺16] for comparability at the same level as the normal modes, that is, BLYP/cc-pVDZ level of theory.

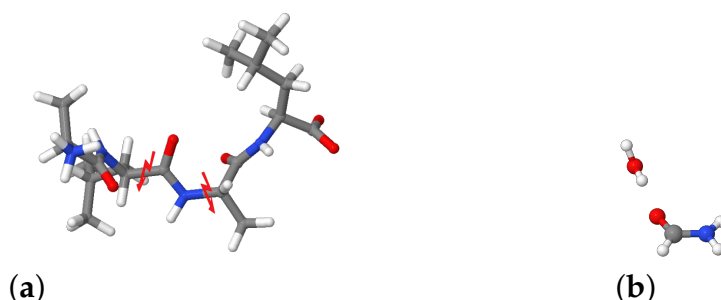


Figure 5.1: Fragmentation of the ALAL peptide for the calculation of interaction energies with individual water molecules. (a) Positions for splitting the fragment (b) Resulting H-saturated $-CONH$ group with one water molecule.

5.3 Results

5.3.1 Conformational Analysis

Analysis of the MD simulations of the ALAL peptide in water with an empirical force field shows that the conformational space of the backbone torsion angles

ψ, ϕ (see Figure 5.2a) for definition) is well sampled and all sterically “allowed” regions in a Ramachandran plot, i.e., around the angles that define α -helix (α : $\phi \approx -57^\circ; \psi \approx -47^\circ$), β -sheet (β : $\phi \approx -130^\circ; \psi \approx +140^\circ$), or left-handed helix (L : $\phi \approx 80^\circ; \psi \approx +70^\circ$) conformations, were visited and regions corresponding to secondary structures such as α -helix, β -sheet, or left-handed helix, show the lowest free energies (see Figure 5.2c). The side chain torsion angles, however, exhibit only significantly populated state in the first-principles simulations, whereas in the longer classical simulations, two well-populated states can be observed for χ_{Leu2} and χ_{Leu4} (Appendix B, Figure B.1). Among the possible backbone conformations, a conformation in which the first and second peptide bonds are in a β -sheet-like conformation, labeled as β, β , clearly dominates (see Figure 5.2b). The second and third most probable conformations, β, α , and α, β , respectively, both also have one of the two peptide bonds in a β -conformation, indicating a preference for a more “stretched” conformation in this ALAL peptide, likely due to the steric demands of the bulky Leu side chains.

This conformation is preserved in the course of the first-principles simulations, launched from the β, β conformation as can be seen in the two-dimensional free energy distributions of the two ψ, ϕ -pairs (see Figure 5.2d). Because of its pre-dominance, we confine our spectroscopic analysis to the β, β conformation and from now on drop the label β, β .

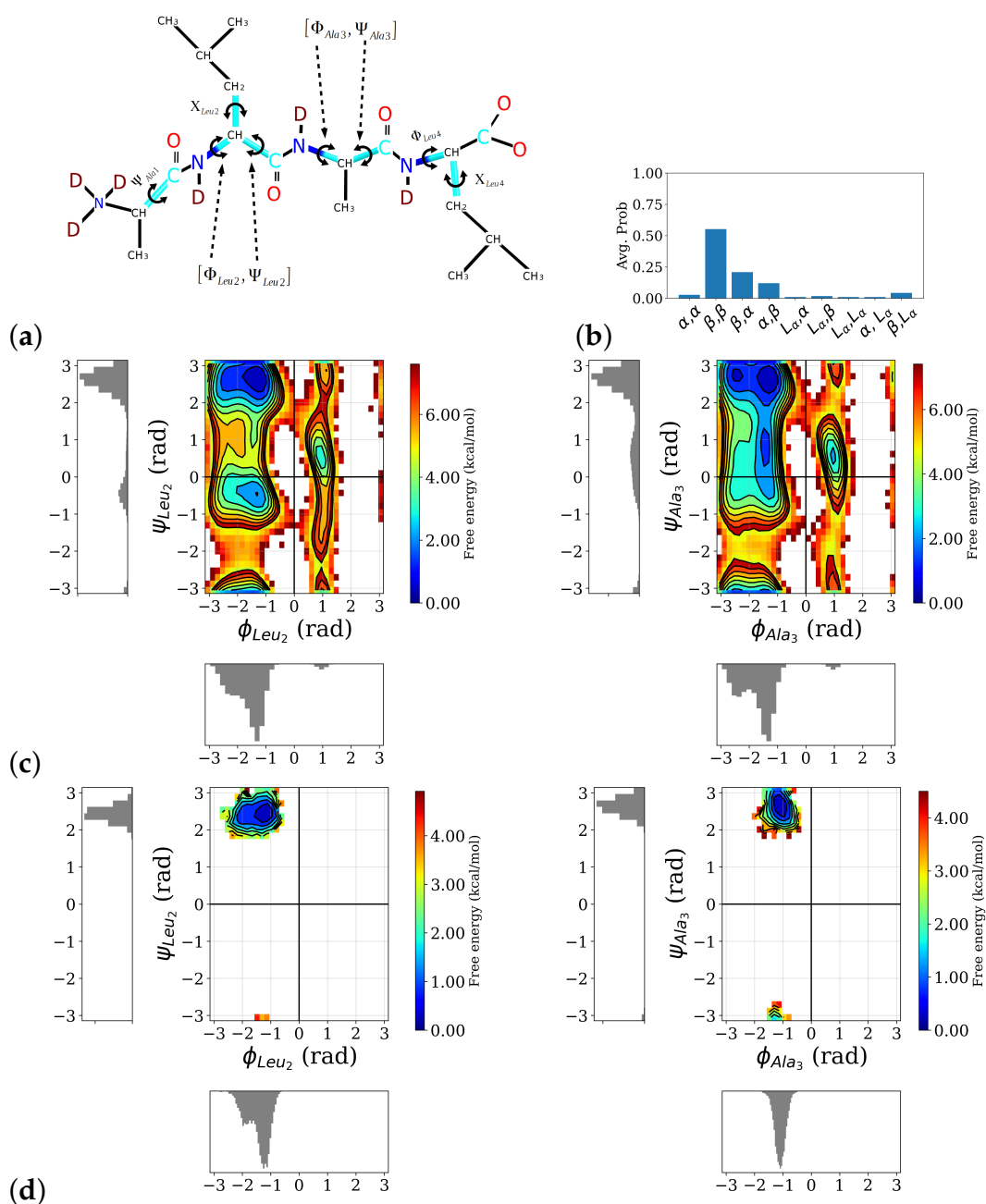


Figure 5.2: Distribution of backbone conformation of the ALAL peptide. (a) Definition of backbone, ψ , ϕ and side chain, χ , torsional angles, (b) probability distribution of different backbone conformations as observed in the classical MD simulations, the first label refers to the first peptide bond, i.e., the ψ_{Leu2}, ϕ_{Leu2} -pair, and the second one to the second peptide bond, i.e., the ψ_{Ala3}, ϕ_{Ala3} -pair, (c) free energy profile and marginal probability distributions of the central ψ , ϕ torsion angles (for the other torsion angles see Appendix B, Figure B.1), (d) two-dimensional free energy distributions of the two ψ , ϕ -pairs, computed from the first-principles simulations, confirming the peptide stays in the β, β conformation.

5.3.2 Vibrational Analysis

Figure 5.3 shows the infrared spectrum computed from first-principle MD simulations of the ALAL peptide in deuterated water. The amide I region shows one intense band centered at $\sim 1600\text{ cm}^{-1}$, which can be attributed to the carbonyl stretch vibration ($\nu C = O$), and another one at $\sim 1528\text{ cm}^{-1}$ which we assign to the stretch vibration of the carboxyl group (νCOO). Also bands in the amide II and amide III region between $\sim 1250\text{--}1440\text{ cm}^{-1}$ and $\sim 1100\text{--}1250\text{ cm}^{-1}$, respectively, are visible. According to the computed power spectrum, these bands correspond to vibrations of the $N - D/N - C$ groups and the N-terminal ND_3 group, respectively. There is also a significant contribution of the carboxyl group to the bands of the amide II region, as can be seen from the power spectrum (Figure 5.3 bottom panel). The motion of both the Ala and the Leu side chain have a large peak at $\sim 1470\text{ cm}^{-1}$ in the power spectrum which, due to the low change in dipole moment of these unpolar groups, translates to only little intensity in the infrared spectrum.

Closer inspection of the motions, i.e., by means of power spectrum, responsible for the most prominent band at $\sim 1600\text{ cm}^{-1}$, the “carbonyl band”, reveals this band to be a superposition of the motion of the three carbonyl groups and a component along the peptide $N - C$ bond. Note, however, that the same C -atom is part of the $N - C$ and the $C = O$ vibration. The frequencies of the three carbonyl groups $C_1 = O_1$, $C_2 = O_2$, $C_3 = O_3$ are $\sim 1606\text{ cm}^{-1}$, $\sim 1592\text{ cm}^{-1}$ and $\sim 1580\text{ cm}^{-1}$ respectively. The three $C = O$ groups in the ALAL peptide hence move with frequencies that differ by $\sim 12\text{--}14\text{ cm}^{-1}$, not enough to be resolved in the $C = O$ band of the IR spectrum, but sufficiently large to wonder why this is the case.

A normal mode calculation of the ALAL peptide in the β -conformation in implicit solvent (see Appendix B, Figure B.2) shows the frequencies of the three carbonyl groups $C_1 = O_1$, $C_2 = O_2$, $C_3 = O_3$ are $\sim 1645\text{ cm}^{-1}$, $\sim 1631\text{ cm}^{-1}$, and $\sim 1619\text{ cm}^{-1}$, respectively. Note that the $C_2 = O_2$ group also contributes to the normal mode at $\sim 1619\text{ cm}^{-1}$ and the $C_3 = O_3$ group to that at $\sim 1631\text{ cm}^{-1}$. The normal-mode-based frequencies differ by the same amount as those computed in explicit solvent. However, the comparatively higher frequency of the normal modes assigned to the carbonyl groups indicate effects not contained in the calculations in implicit solvent, such as temperature or, more likely, explicit interactions with the solvent.

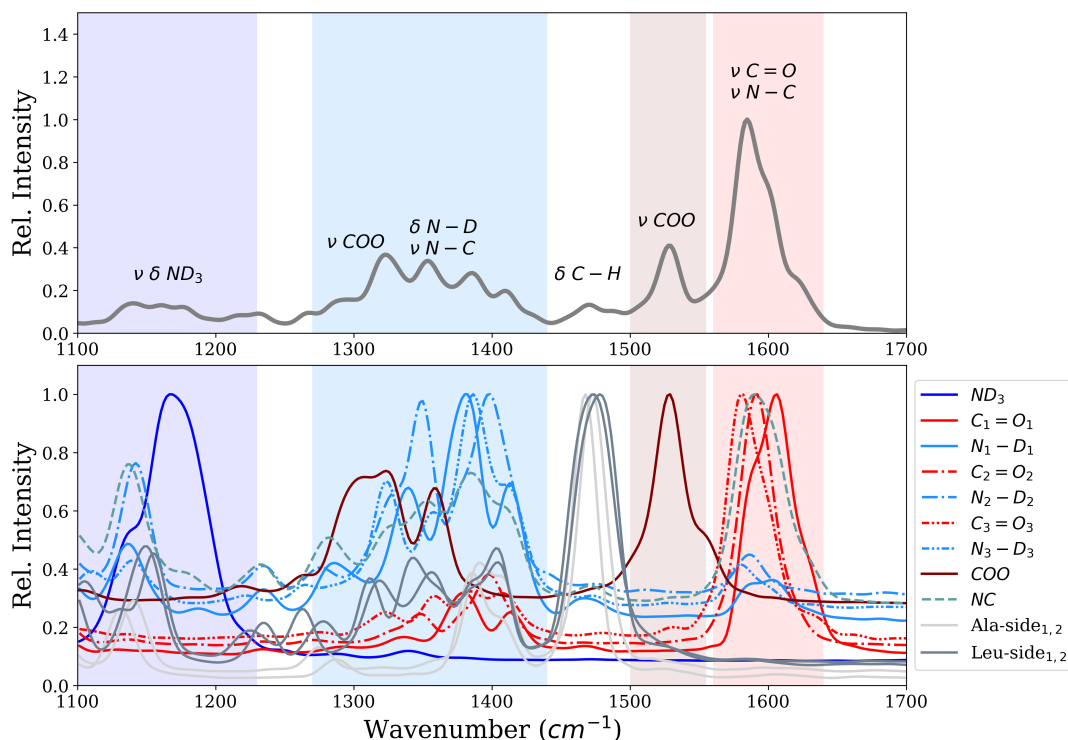


Figure 5.3: Infrared spectrum (**top**) and power spectrum (**bottom**) of the ALAL peptide in water in β, β conformation.

5.3.3 Normal Modes of ALAL–Water Clusters

In order to analyze the impact of a hydrogen-bonded water molecule on the individual $C = O$ groups, we performed normal mode calculations of the ALAL peptide hydrogen-bonded to one or more water molecules.

Each of the $C = O$ groups exhibits, as anticipated, lowered stretching frequencies when a water molecule is hydrogen-bonded to it (see Appendix B, Figure B.2). The red-shift compared to the frequencies of the unbound ALAL peptide is about 29 cm^{-1} for all the carbonyl groups, resulting in an about 10 cm^{-1} higher frequencies than those calculated from the first-principles MD with full explicit solvation.

It is interesting to note that not only the frequency of the normal mode of the carbonyl group that carries the water molecule is affected, but also the other two carbonyl groups show small changes in their stretching frequencies to higher or lower values. With one water molecule hydrogen-bonded at each of the carbonyl groups (three water molecules in total), the red-shifts are slightly different for the three carbonyl groups (25 , 27 , and 31 cm^{-1} , for the $C_1 = O_1$, $C_2 = O_2$, and $C_3 = O_3$ group, respectively). Note that the water molecules are hydrogen bond donors to the respective carbonyl group and at the same time hydrogen bond

acceptors to the neighboring *ND* groups. Adding one more water molecule at the COO^- group hardly affects the frequencies of the $\text{C} = \text{O}$ groups.

An extreme is a water cluster in which all polar groups are involved in at least one hydrogen bond (see Appendix B, Figure B.2). For such a case, the computed normal modes are 1595, 1577, and 1620 cm^{-1} , for the $\text{C}_1 = \text{O}_1$, $\text{C}_2 = \text{O}_2$, and $\text{C}_3 = \text{O}_3$ group, respectively. That is, no change in frequency is observed for the $\text{C}_3 = \text{O}_3$ group compared to the unbound peptide. In contrast, the frequencies of the other two carbonyl groups show significant red-shifts. The $\text{C}_2 = \text{O}_2$ group is involved in more than one hydrogen bond and one might therefore expect a strong red-shift.

Computing the normal modes for the central carbonyl group, $\text{C}_2 = \text{O}_2$, with one or two hydrogen-bonded water molecules at only this group, confirms the idea of stronger red-shifts by more hydrogen-bonded partners, since one hydrogen bond results in a frequency of 1601 cm^{-1} whereas two hydrogen bonds lead to a frequency of 1582 cm^{-1} for the stretching vibration of the $\text{C}_2 = \text{O}_2$ group, but not quite as much as in the model with hydrogen bonded water molecules at all polar groups.

Such “clean” scenarios are, however, not representative of the full hydration in explicit water. As can be seen from the calculated $\text{C}_2 = \text{O}_2$ stretching frequencies, different topologies of hydrogen-bonded water molecules (numbers of water molecules and different connections between the polar groups) around the ALAL peptide have different, and sometimes even opposing, effects. Thus, not only to go beyond the harmonic approximation, but also to obtain a more comprehensive picture of the effect of the water solvation on the carbonyl frequencies of the ALAL peptide, it is important to have a closer look at the first-principles MD simulations.

5.3.4 Analysis of the Hydration Shell

Analysis of the number of water molecules that form a hydrogen bond to the polar groups (see Figure 5.4) shows a trend of higher probability for more hydrogen-bonded water molecules from $\text{C}_1 = \text{O}_1$, over $\text{C}_2 = \text{O}_2$ to $\text{C}_3 = \text{O}_3$. The average number of hydrogen bonds for the three carbonyl groups $\text{C}_1 = \text{O}_1$, $\text{C}_2 = \text{O}_2$, and $\text{C}_3 = \text{O}_3$, are 1.3 ± 0.6 , 1.6 ± 0.5 , and 1.9 ± 0.6 , respectively. This trend is in agreement with the order of the observed $\text{C} = \text{O}$ frequencies in the sense that the most red-shifted $\text{C} = \text{O}$ vibration corresponds to the carbonyl group that has, on average, the highest number of hydrogen-bonded water molecules. In other words, for $\text{C}_3 = \text{O}_3$, the $\text{C} = \text{O}$ bond is weakened most often, for $\text{C}_2 = \text{O}_2$ second, and for $\text{C}_1 = \text{O}_1$ least, by the presence of a hydrogen bond, and hence the vibrational frequency, which is also computed from an average of all the hydrogen-bonded (or not) scenarios, is more, or less, shifted to lower frequencies.

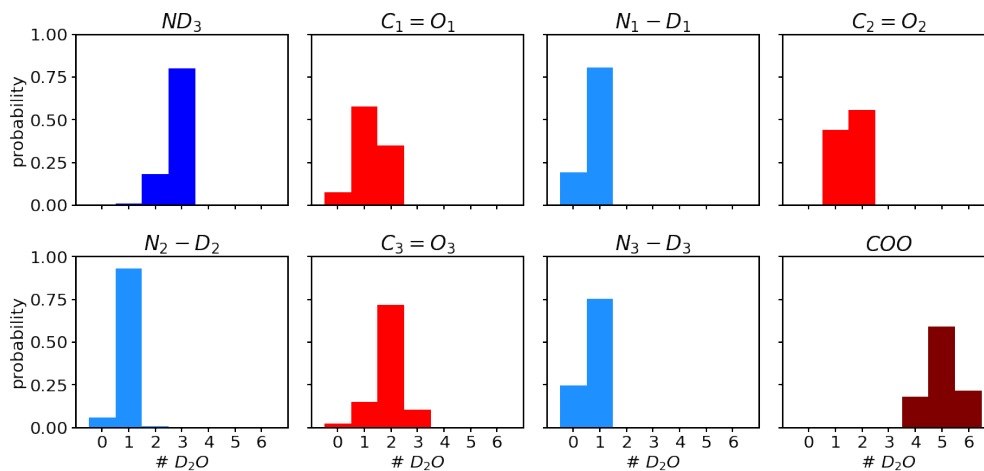


Figure 5.4: Probability distribution of number of D_2O molecules hydrogen-bonded to the polar groups i.e., ND_3 , $C = O$'s, $N - D$'s and COO , of the ALAL peptide.

Closer inspection of the positional distribution of the hydrogen-bonded water molecules in terms of the hydrogen-acceptor distance and the donor–hydrogen–acceptor angles shows a well-defined first solvation region for all three carbonyl groups (see Figure 5.5) and also for the other polar groups (see Appendix B, Figure B.3). The highest density regions are confined to distances of 1.5 to 2.2 Å and angles between 0 and 15 degree deviation from linearity for all polar groups. The integrated number of water molecules that obey at least one hydrogen bond criterion, that is, distance or angle criterion, for the three carbonyl groups are 1.5, 1.6, and 2.0, respectively. Albeit the difference is less pronounced, in particular between the $C_1 = O_1$ and $C_2 = O_2$ group, these numbers follow the same trend as the numbers of water molecules that are within both criteria, i.e., the number of hydrogen-bonded water molecules. Beyond the hydrogen-bonded region, as defined by hydrogen–acceptor distance and hydrogen bond angle, there is still a non-negligible probability for water molecules to be close to, i.e., with a distance below 5 Å, and therefore possibly interacting with the carbonyl groups.

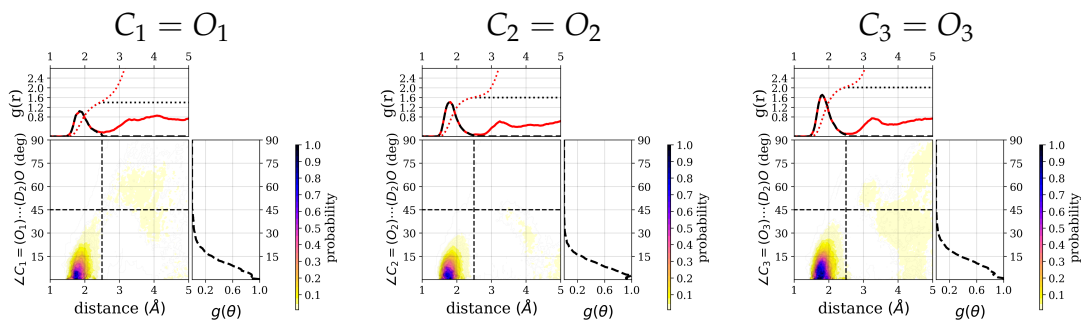


Figure 5.5: Combined radial distribution functions, $g(r)$, and angular distribution functions, $g(\theta)$, of hydrogen-bonded water (D-atoms) around the three carbonyl groups of the ALAL peptide. For the other groups, see Appendix B, Figure B.3. Each top marginal plot shows $g(r)$ and right marginal plot shows $g(\theta)$ for the respective distribution function. A black dashed line is used to show the restriction according to the hydrogen bond criteria. In each $g(r)$ plot, the black and red dotted curves represent the running integration of hydrogen-bonded water molecules and of all water molecules, respectively.

For all three carbonyl groups, the number of hydrogen-bonded water molecules fluctuates by about half a molecule. Since this is an averaged deviation from the mean, the carbonyl groups actually switch between states in which they have one hydrogen-bonded water molecule more or less.

To analyze this further, we have taken a closer look at the water structure around the central carbonyl group, $C_2 = O_2$, whose frequency and number of surrounding/hydrogen-bonded water molecules, is between that of the other two carbonyl groups. Furthermore, this carbonyl group is in a central position, that is least affected by the charged termini, ND_3^+ and COO^- , respectively, and therefore the best representative of a carbonyl group in a longer peptide or protein.

Looking at the time series of water oxygen–carbonyl oxygen distances and wO–H··O angles (see Figure 5.6b) of the water molecules that are closest to the $C_2 = O_2$ group (see (Figure 5.6a)), one can indeed see changes in the local water structure. One water molecule (labeled as resid 118 for the purpose of distinguishing the individual water molecules) stays close (within 3.5 Å) to the carbonyl oxygen atom throughout the simulation time and in an angle within the hydrogen bond criteria most of the simulation time. Between ~8 and 16 ps, this is the only water molecule that qualifies for a hydrogen bond. Before 8 ps and after 16 ps simulation time, another water molecule (resid 141) is close enough to the $C_2 = O_2$ group and at the correct angle to also be counted as hydrogen-bonded, and yet another water molecule (resid 80) transiently comes close enough to be a candidate for a hydrogen bond. Between 16 and 36 ps simulation time, there is again a rather stable state with two water molecules (resid 118 and 141) in hydrogen-bonded position and orientation to the $C_2 = O_2$ group. At about 36 ps

simulation time, the second water molecule (resid 141) leaves again and is replaced by a third water molecule (resid 132) that has been at about 4 – 4.5 distance until then, located in the middle of a three-water bridge to the $C_3 = O_3$ group. As this water molecule moves closer to the $C_2 = O_2$ group, such that it forms a hydrogen bond, the bridge breaks, and is transiently replaced by a two-water bridge. This water molecule stays at the $C_2 = O_2$ group until the end of the simulation (50 ps), rendering this last window again a state with two hydrogen bonded water molecules (see Figure 5.6). Based on geometric criteria, the average numbers of hydrogen bonds between water and the $C_2 = O_2$ group for these three time windows, i.e., 8–16 ps, 16–36 ps, and 40 – 50 ps, are 1.0 ± 0.2 , 1.5 ± 0.5 , and 1.7 ± 0.5 , respectively.

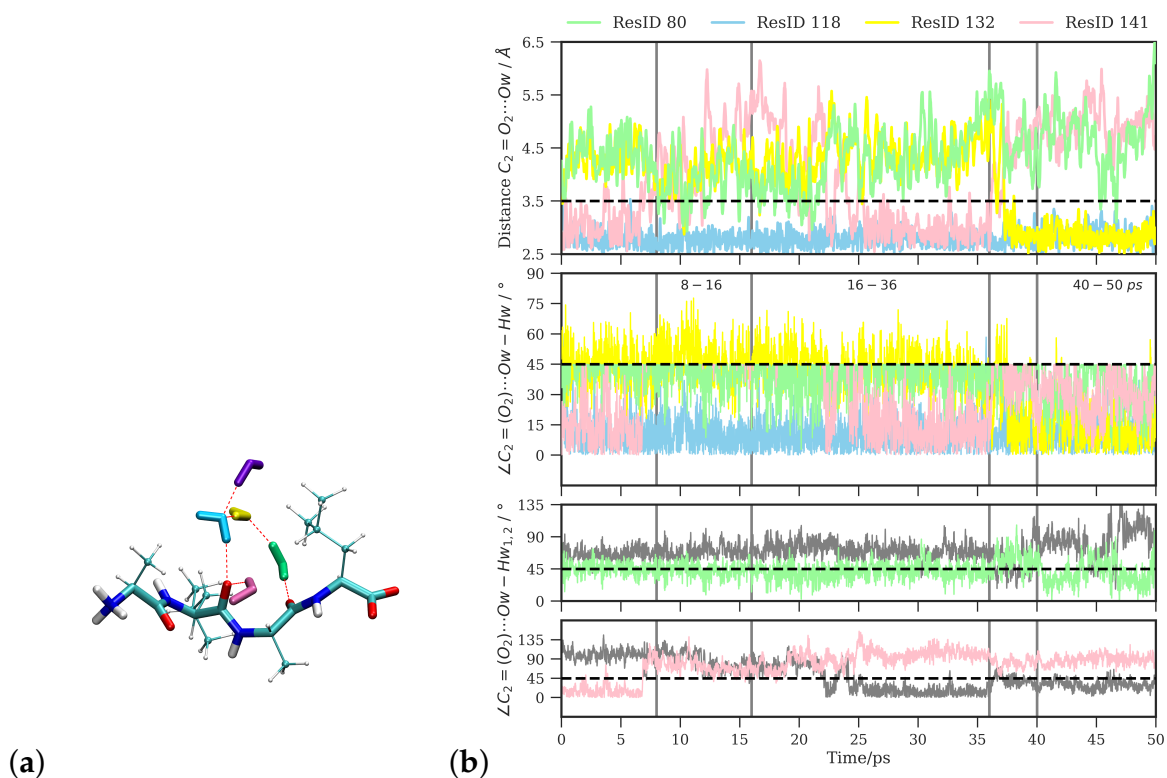


Figure 5.6: (a) Water molecules closest to the central carbonyl group ($C_2 = O_2$). (b) Time series of distances and angles between these water molecules and the $C_2 = O_2$ group. The ResID labels are used solely to distinguish and refer to the individual water molecules.

Correspondingly, the distribution of water molecules around the $C_2 = O_2$ group has a high density in the hydrogen-bonding region (see Figure 5.7) and some additional low density at a distance around 4 Å that is comparatively higher for the 40–50 ps window than for the other two windows. The integrated number of water molecules in the hydrogen-bonded region for the three time windows are 1.1 (8–16 ps), 1.6 (16–36 ps), and 2.0 (40–50 ps), respectively.

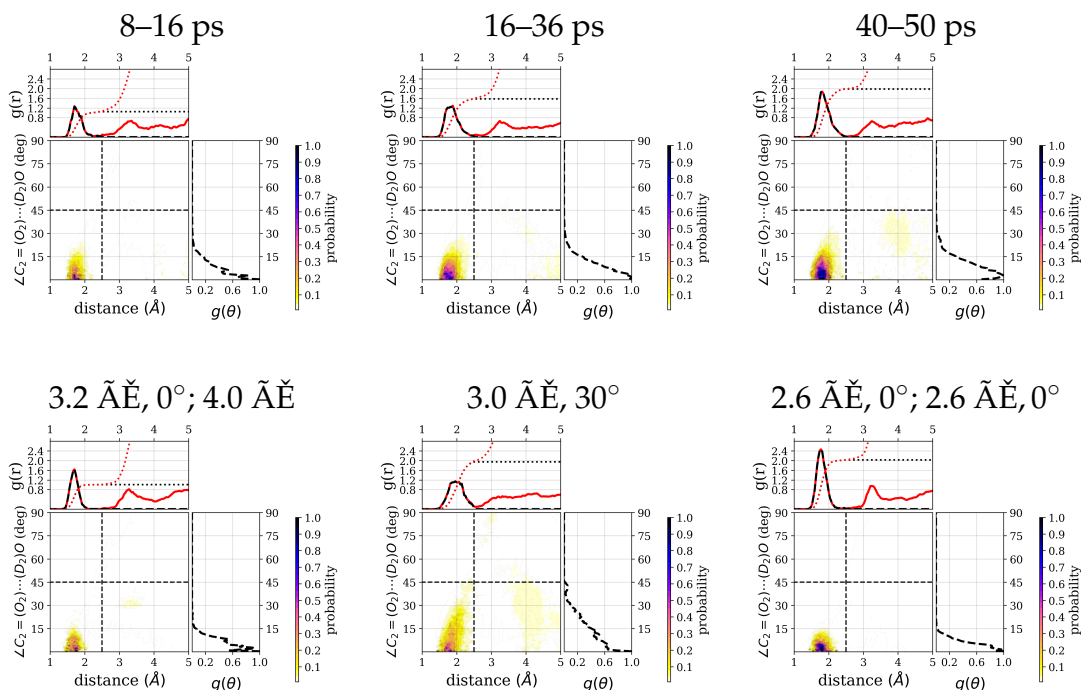


Figure 5.7: Radial distribution functions, $g(r)$, and angular distribution functions, $g(\theta)$, of hydrogen-bonded water (D-atoms) around the $C_2 = O_2$ group of the ALAL peptide, computed for three different time windows: 8–16 ps, 16–36 ps, and 40–50 ps and from constrained simulations (see methods for details). Each top marginal plot shows $g(r)$ and right marginal plot shows $g(\theta)$ for the respective distribution function. A black dashed line is used to show the restriction to hydrogen bond criteria. In each $g(r)$ plot, the black and red dotted curves represent the running integration of hydrogen-bonded water molecules and of all water molecules, respectively.

Power spectra computed from these time windows of this simulation, corresponding to the one-water, one- and two-water, and two-water situations, indeed show different carbonyl frequencies for these different parts of the trajectory (see Table 5.1 and Figure 5.8a). The part that corresponds to a one-water molecule close to the $C_2 = O_2$ group shows a higher frequency (1600 cm^{-1}) than the other two parts (1594 cm^{-1} and 1584 cm^{-1} , respectively). The frequency computed from the middle part with a mixed state of one and two water molecules close to the $C_2 = O_2$ group shows almost the same frequency as the power spectrum computed from entire trajectory. These data confirm the carbonyl frequency to be a result of the averaged interactions of the water molecules with the $C_2 = O_2$ group.

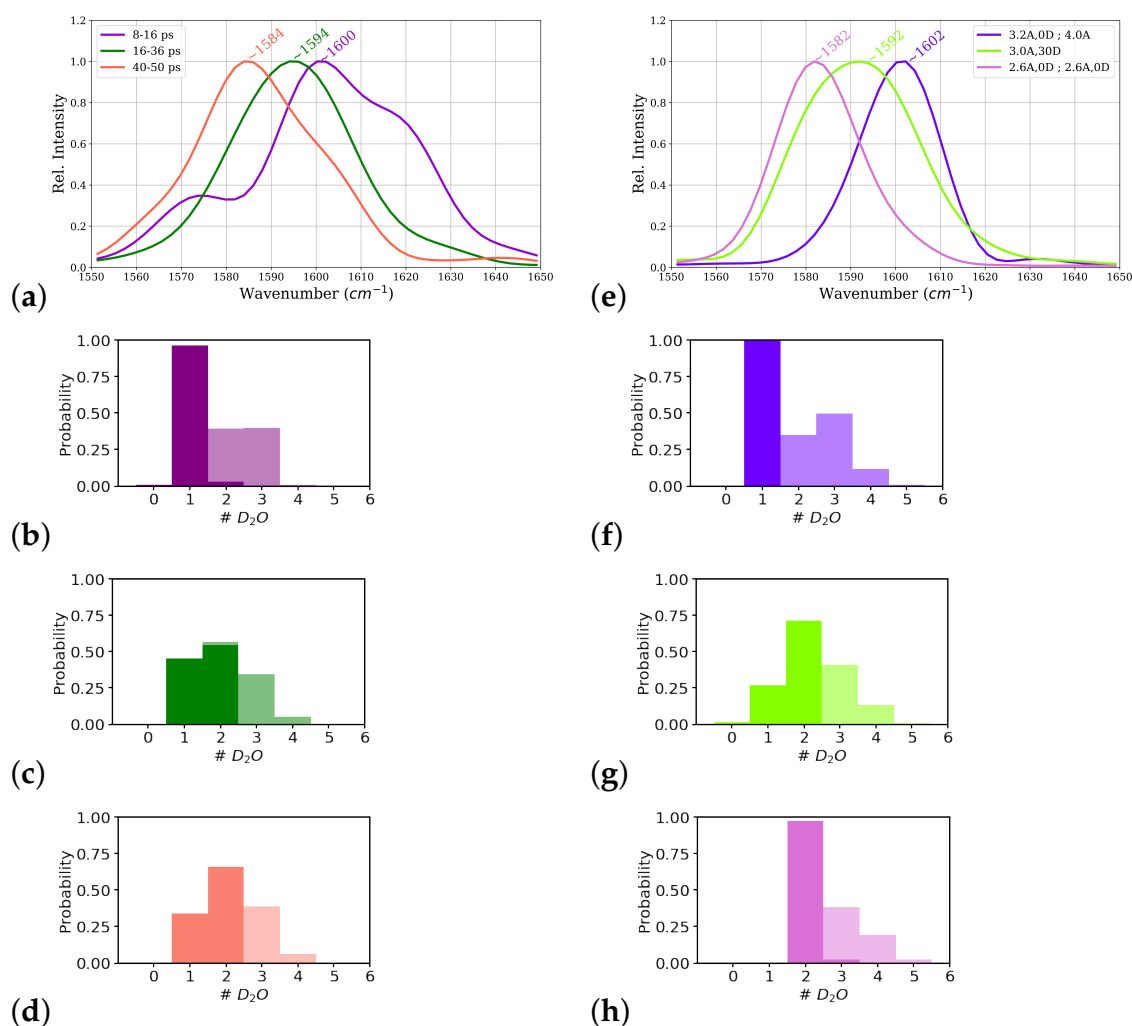


Figure 5.8: Power spectrum of the central carbonyl group, $C_2 = O_2$, of the ALAL peptide (For the full range power spectra see Appendix B, Figure B.5), computed from (a) three time windows of the first-principles simulation: 8–16 ps, 16–36 ps, and 40–50 ps simulation time and (e) constrained simulations. Hydrogen bond probabilities as quantified by number of hydrogen-bonded water molecules in the strong (opaque) and moderately (transparent) interacting zone (see Figure 5.9 for definition), computed from time windows of the unconstrained simulation (b) 8–16 ps, (c) 16 – 36 ps, and (d) 40 – 50 ps) and from constrained simulations (see methods for details) (f) 3.2 Å, 0°; 4.0 Å, (g) 3.0 Å, 30°, and (h) 2.6 Å, 0°; 2.6 Å, 0° .

Table 5.1: Frequencies of the stretching vibration of the second carbonyl group, $C_2 = O_2$, in the ALAL peptide in solution, corresponding to the highest peak of the respective power spectrum (see also Figure 5.8). Interaction energies of the four closest water molecules with the $C_2 = O_2$ group, distances and angles of the closest hydrogen atom to the $C_2 = O_2$ group, and distances of the closest hydrogen atom of the closest water molecules to the $N_2 - D_2$ group. # D_2O denotes the number of water molecules within the ‘strong’, i.e. hydrogen-bonded, and ‘moderately’ interacting zone (see text for definitions). The four water molecules are ranked and (re-)labelled W1, W2, W3, or W4 by their oxygen atom distances to oxygen atom of the $C_2 = O_2$ group at each frame such that e.g. W2 can refer to different individual water molecules.

	Unconstrained				Constrained		
	full (0–50 ps)	8–16 ps	16–36 ps	40–50 ps	3.2 Å, 0°; 4.0 Å	3.0 Å, 30°	2.6 Å, 0°; 2.6 Å, 0°
Frequency (cm ⁻¹)	1592	1600	1594	1584	1602	1592	1582
Energy (kcal·mol ⁻¹)							
W1	-5.9±1.2	-5.8±1.1	-6.2±1.2	-5.5±1.3	-6.5±1.3	-5.2±1.0	-5.9±1.0
W2	-4.3±2.3	-1.8±1.3	-4.6±2.2	-5.6±1.6	-1.7±0.8	-5.1±1.0	-6.1±1.0
W3	-0.7±1.0	-1.0±0.8	-0.7±0.9	-1.0±1.2	-1.2±0.7	-0.5±1.6	-1.1±1.2
W4	-0.6±1.1	-0.7±1.2	-0.6±1.3	-0.6±0.8	-0.8±0.8	-0.8±1.2	-0.5±1.0
# D_2O (strong)	1.6±0.5	1.0±0.2	1.5±0.5	1.7±0.5	1.0±0.1	1.7±0.5	2.0±0.2
# D_2O (moderate)	0.9±0.8	1.3±0.8	0.9±0.8	0.9±0.8	1.9±0.9	1.0±0.8	0.9±0.8
Distance (Å)							
$CO \cdots H - Ow$							
W1	1.8±0.2	1.8±0.2	1.8±0.2	1.8±0.1	1.7±0.1	1.9±0.2	1.7±0.1
W2	2.5±0.6	3.2±0.5	2.5±0.6	2.0±0.2	3.2±0.3	2.2±0.2	1.9±0.1
W3	3.7±0.5	3.6±0.4	3.7±0.4	3.5±0.5	3.5±0.4	3.6±0.6	3.4±0.5
W4	4.1±0.5	4.1±0.6	4.1±0.6	4.0±0.4	3.9±0.5	3.8±0.5	3.8±0.6
Distance (Å)	4.5±0.7	4.5±0.6	4.3±0.8	4.7±0.7	4.7±0.9	4.5±0.8	4.7±0.6
$ND \cdots Ow$							
Angle (°)							
$CO \cdots H - Ow$							
W1	17.3±10.1	17.2± 9.5	17.2±10.1	19.1±10.6	13.6± 7.5	27.8±14.1	12.4± 5.2
W2	38.5±27.2	66.1±21.1	35.5±24.3	22.6±15.0	53.7±15.2	30.9±15.6	8.4± 4.9
W3	118.8±25.8	125.7±17.8	124.4±20.3	121.7±23.0	57.6±19.9	82.5±28.9	73.3±23.2
W4	78.1±29.0	81.7±26.6	75.9±30.2	77.4±29.1	63.7±28.1	70.2±25.6	80.7±28.0

The interaction energies between the *CONH* fragment containing the $C_2 = O_2$ group and individual water molecules vary between almost nothing for the more distant water molecules and ~ 6 kcal/mol, which is about the upper limit for the strength of hydrogen bonds [oPC09]. Figure 5.9 shows the distribution of distances and angles to the $C_2 = O_2$ group of the closest four water molecules and their interaction energies. Within the region that is considered to be hydrogen-bonded, as by hydrogen-acceptor distance and donor-hydrogen-acceptor angle (indicated by dashed lines in Figure 5.9a)), the interaction energies are strongest. Note, however, that toward shorter distances, i.e., at ~ 1.6 Å and below, the interaction energies are less favorable. With larger distances and angles, the interaction energy strength generally decreases. One can, however, group these “outer” region also in different zones, based on the (average) interaction energies observed there. Doing this by *k*-means clustering with three clusters, a “strongly interacting zone” can be recognized that coincides with the hydrogen-bonded region (indicated by dashed lines in Figure 5.9b)), a “moderately interacting zone” at larger distances and angles, but below 4.5 Å and 75° (indicated by dotted lines in Figure 5.9b)), and a “weakly interacting zone” (at even larger distances and angles) can be distinguished. The classification into “strong”, “moderate” and “weak” is based on the average interaction energies within the clusters (see Figure 5.9b)). Note that there are also water molecules that geometrically qualify as “strongly hydrogen-bonded” but are members of the “moderately interacting” cluster. Grouping the data into two clusters mainly results in hydrogen-bonded and not hydrogen-bonded water molecules whereas a grouping into four cluster partitions the hydrogen-bonded zone into two groups of “strong” and “very strong” interactions and a separation into “moderately” and “weakly” interacting (see Appendix B, Figure B.6), comparable to the results for three clusters. In order to use geometric criteria for a qualitative description of interaction strengths, we decided to use three clusters, i.e., hydrogen-bonded corresponds to “strong”, other close water molecules are classified as “moderately” interacting, and the remaining water molecules are considered as weakly or non-interacting.

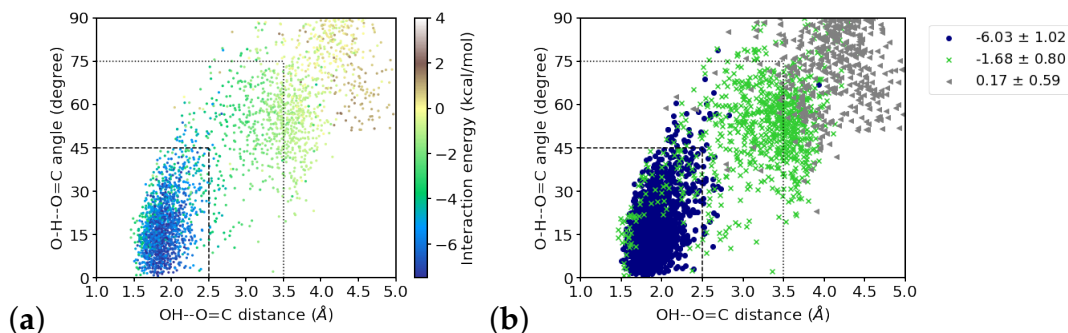


Figure 5.9: Distribution of interaction energies of the four water molecules closest to the $C_2 = O_2$ group over hydrogen-bond distances and hydrogen bond angles. (a) energy values indicated by color, (b) interaction energies clustered into “strong” (navy circles), “moderate” (green crosses) and “weak” (grey triangles) interactions. The dashed lines indicate the used hydrogen bond criteria and mark the “strong interaction zone”. The dotted lines mark the “moderate interaction zone”.

Inspecting the interaction energies, and the resulting number of hydrogen-bonded water molecules in the two interacting zones for the time windows of the simulation, we find the closest water molecule (W1) that is hydrogen-bonded throughout the simulation to be interacting with a strength that varies only slightly, i.e., within the error margin, between the three windows (see Table 5.1). The interaction energies computed for the second closest water molecule (W2), however, differs significantly between the first window (-1.8 ± 1.3 kcal/mol) and the other two windows (-4.6 ± 2.2 and -5.6 ± 1.6 kcal/mol, respectively; see Table 5.1). This is in agreement with the larger average distance ($3.2 \pm 0.5 \text{ \AA}$) of this water molecule in the first window than in the other two ($2.5 \pm 0.6 \text{ \AA}$ and $2.0 \pm 0.2 \text{ \AA}$, respectively) and also the larger angle (66.1 ± 21.1 , 35.5 ± 24.3 , and 22.6 ± 15.0 in the first, second, and third window, respectively; see Table 5.1). Indeed, we find a considerable correlation between the distances, and also the angles, of the second (and third) water molecule and the interaction energies, but only a correlation with the angle for the first water molecule (see Appendix B, Table B.1).

As already noted before, in the first window, there is on average only one (strong) hydrogen bond formed between the $C_2 = O_2$ group and a water molecule, W1, whereas in the other two windows, another water molecule, W2, is (strongly) hydrogen bonded for almost the entire last window (40–50 ps) and partially in the second window (16–36 ps). In those frames where W2 is just outside the (strong) hydrogen-bonded zone, it is still close enough to the $C_2 = O_2$ group to interact (probably more than “moderately”) as manifested by the rather strong interaction energy calculated for this water molecule. Within error, this interaction energy is comparable to that computed for the last window, albeit the fluctuation, i.e., the error, is significantly larger.

Relating these interaction energies with the frequencies computed for the three

windows, the weak interaction in the first window indeed corresponds to the lowest red-shift (to 1600 cm^{-1}). Regarding the other two windows, the red-shift is largest (to 1584 cm^{-1}) in the last window with the highest average number of strongly hydrogen-bonded water molecules and most favorable interaction energies. Both these values are still close, at least within error, to those of the second window for which an intermediate red-shift (to 1594 cm^{-1}) has been computed. One can therefore argue that it is either the combined effect of two close water molecules interacting less strongly with the $C_2 = O_2$ group in the second than in the last window, or, and probably in addition, the larger fluctuations in the second window, which give rise to the lower red-shift.

The third and fourth water molecules are at the edge of the “moderately” interacting zone, as also manifested by the probability distribution of number of water molecules in the two zones (see Figure 5.8b–d), for the first, second, and third time window, respectively). These more distant water molecules, moreover, show only weak interactions with the $C_2 = O_2$ group in all three time windows. For all four closest water molecules considered in the calculation of interaction energies, the distance of the closest hydrogen atom (among the hydrogen atoms of all four water molecules) to the $N_2 - D_2$ group, which is also part of the fragment, is large enough ($\sim 4.5\text{ \AA}$, see Table 5.1) that one can consider the calculated interaction energies to be dominantly with the $C_2 = O_2$ group.

5.3.5 Instantaneous Frequencies

Figure 5.10 shows the instantaneous $C_2 = O_2$ frequency (positions of maxima) as calculated from a wavelet analysis. For the full simulation time, the averaged frequencies calculated by the wavelet analysis is 1594 cm^{-1} which is close to the frequency computed from the Fourier transform of the entire simulation.

The averaged frequency from the wavelet analysis from a first time window of the simulation, 8–16 ps, that corresponds to a situation with one-water hydrogen-bonded to the $C_2 = O_2$ group is 1602 cm^{-1} and for the two windows that correspond to a mixed and a two-hydrogen-bonds state (16–36 ps and 40–50 ps, respectively) are 1594 cm^{-1} and 1588 cm^{-1} , respectively. Again, the $C = O$ frequencies for the individual time windows computed by the direct Fourier transform are well reproduced by the wavelets.

The wavelet spectrum contains a number of sudden “jumps” to very low values ($<1500\text{ cm}^{-1}$) that have to be considered artefacts of the transformation not being able to capture some fluctuations in the $C_2 = O_2$ signal properly. These data points have been omitted and smoothed over for clarity in Figure 5.10. The complete time series of the instantaneous frequencies is shown as Appendix B, Figure B.7, together with the water topology at the $C_2 = O_2$ group, that is the identity of the hydrogen-bonded water molecule(s) and also the hydrogen-bonded connections to other polar groups via hydrogen bonds (water bridges). The artificial

“jumps” occur mainly around times, when water molecules exchange positions and/or a water bridge between polar groups forms/breaks or reforms (see Appendix B, Figure B.7). A leaving or incoming water molecule distorts the electric field around the $C_2 = O_2$ group and has therefore likely an effect on its bond strength and hence instantaneous frequency. Since there will also be some latency, those changes are not confined to a single frame, but may lead to a response in terms of changed frequency also a few frames after the water positions are rearranged.

Note that the switching between discrete states of formed/broken water bridges suggest the water topology to be more labile than it would appear with an overlapping (instead of binary) definition of hydrogen bonds and thus water bridges. In particular, the last time window at 40–50 ps exhibits frequent changes between existing/non-existing water bridges between the $C_2 = O_2$ group and the $C_3 = O_3$ group or the COO^- group of different lengths. The window at 8–16 ps, in contrast has a continuous three-water bridge between the $C_2 = O_2$ group and the $C_3 = O_3$ group and transient formation of a bridge to the COO^- group, corresponding to a water molecule (marked as resid 80) being close to the $C_2 = O_2$ group or not (see Appendix B, Figure B.7).

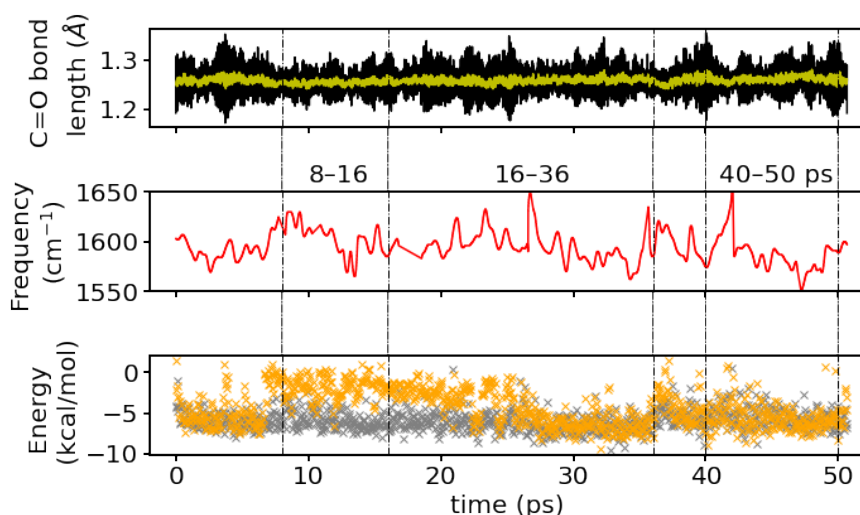


Figure 5.10: **Top:** Time series of the $C_2 = O_2$ bond length (black) and its running average (yellow), **middle:** Instantaneous frequencies from a wavelet analysis, and **bottom:** interaction energies of the closest (grey) and second closest (orange) water molecule with the $C_2 = O_2$ group. The dashed lines indicate the time windows which were analyzed individually.

With one, and the very same water molecule, hydrogen-bonded to the $C_2 = O_2$ group, (instantaneous) changes in the frequency of this group have to be attributed to the other close water molecules. Indeed, the correlation coefficient

of the interaction energy of the second closest water molecule (changing between resid 141 to 132) with the instantaneous frequency is 0.4. For the closest water molecules, this correlation is only 0.2 and below 0.1 for the other water molecules. While a correlation of 0.4 is not striking it is certainly not negligible, suggesting that the interaction energies and the instantaneous, and therefore most probably also the averaged, frequencies are related. From comparison of the time series of interaction energies (see Figure 5.10 middle panel) with the time series of instantaneous frequencies (see Figure 5.10 bottom panel), one can see a tendency for higher frequencies around times with weaker interactions of the second closest water molecule.

5.3.6 Simulations with Constrained Water Molecules

In order to further probe the effect of water molecules within hydrogen bond distance on the frequency of the $C_2 = O_2$ vibration, we have performed additional simulations, in which one or two water molecules are constrained at different distances (2.6 – 3.2 Å or 4.0 Å donor–acceptor distance, see methods for details), such that they are within hydrogen-bonded distance or beyond the cutoff for hydrogen bonds, but still close enough to have an effect.

From the average number of hydrogen bonds, the three constrained simulations correspond to a situation with one (strongly) hydrogen-bonded water molecule (3.2 Å, 0°; 4 Å), two (strongly) hydrogen-bonded water molecules (2.6 Å, 0°; 2.6 Å, 0°), and a mixed situation (3.0 Å, 30°), comparable to the three windows in the unconstrained simulation (see Table 5.1). The integrated number of hydrogen-bonded water molecules of 1.0, 2.0, and 2.0 for the 3.2 Å, 0°; 4 Å constraints, 3.0 Å, 30° constraints, and 2.6 Å, 0°; 2.6 Å, 0° constraints, follow essentially the same trend (see Figure 5.7).

The carbonyl frequencies computed for the constrained simulations are, ~ 1582 , ~ 1592 , and $\sim 1602 \text{ cm}^{-1}$ for the two-water, mixed, and one-water molecule scenarios, respectively (see also Table 5.1 and Figure 5.8e), similar to those observed in the time windows of the corresponding states. The interaction energies between close water molecules and the $C_2 = O_2$ group are similar to those computed for the time windows with comparable scenarios: the closest water molecule interacts strongly in all the constrained simulations, and the second water molecule, again, interacts strongest in the two-water case and weakest in the one-water case. In the latter scenario, this second water molecule is again on the border between the “strong” and “moderately” interacting zone as far as distances and angles are concerned, but its interaction energies rather classifies it as a member of the “moderately interacting” group. As also observed in the unconstrained simulation, the third and fourth water molecule interact only weakly, with a little stronger interactions of the third water molecule in the two-water or mixed scenarios. The distance to the $N_2 - D_2$ group is for all four water molecules large

enough to again consider the interaction energies to be dominated by the interaction with the $C_2 = O_2$ group. There is a probability of ~ 0.3 or more to find a third water molecule in the “moderately interacting zone” in all the cases, that is one-water, two-water or mixed scenarios, modeled in the time windows of the unconstrained simulation (see Figure 5.8b–d) and the different constrained simulations (see Figure 5.8f–h). Hence, it is not possible to tell whether the presence of this third water molecule has a direct impact on the $C_2 = O_2$ frequency or not. It is however, conceivable that this water molecule contributes indirectly by connections to the strongly hydrogen-bonded water molecule(s) and maintaining the water topology, e.g., water bridges to other polar groups, in the hydration shell of the $C_2 = O_2$ group.

5.4 Discussion

Our analyses show that the interactions between water molecules and the $C_2 = O_2$ group clearly have an effect on the stretching frequency of this carbonyl group. From the normal mode analysis of isolated ALAL-water clusters, it becomes apparent that more hydrogen-bonded water molecules lead to a more pronounced red-shift, but also that in water clusters with several water molecules bound to the different polar groups of the peptide, thought to be more representative of an actually solvated peptide, the hydrogen-bonded situation is too complicated to allow for a simple prediction of (possible) red-shifts. MD simulations in explicit water are therefore a great way to sample different, more or less complex, water topologies in the hydration shell of the peptide. The agreement of the trends in red-shift of the carbonyl frequencies and average number of hydrogen-bonded water molecules observed in our first principles MD simulation of the ALAL peptide in water corresponds to the, perhaps idealised, trend in red-shift by hydrogen-bonding one or two water molecules to the $C_2 = O_2$ group. In fact, the frequencies computed by the two approaches, i.e. MD and normal modes, are in a striking agreement: (on average) one hydrogen-bonded water molecule at the $C_2 = O_2$ group leads to a frequency of $\sim 1600 \text{ cm}^{-1}$ and (on average) two water molecules hydrogen-bonded to the $C_2 = O_2$ group result in a frequency of $\sim 1582 \text{ cm}^{-1}$. While one can argue that this shows how well the implicit solvent approach models the average effect of the explicit solvent, being aware of the other approximations in the normal mode analyses, i.e. harmonic model and zero temperature, we consider the almost exact match of the frequencies rather as a coincidence, but appreciate the similarities. They give us some confidence that one-water and two-water states can, to some extent, be mimicked by the respective water clusters.

Our 100 ps long first principles MD simulation is sufficient to sample one-water and two-water states, with full hydration, and also something we call a mixed state, that is a period in which the hydrogen-bonded states change between one

and two. Since these states (luckily) prevailed long enough in our present simulation, we were able to obtain spectra from the different states that still contain a dynamic average of the system. The simulations with water molecules constrained such that one-water, two-water and mixed states prevail by construction, lead to similar results in the carbonyl frequencies, confirming the averages obtained from the (shorter) time windows to be sufficient.

The calculation of interaction energies between the closest water molecules and the $C_2 = O_2$ group (as the CONH fragment) clearly demonstrate the (weakening) effect of strong interactions with the water molecules on the $C_2 = O_2$ bond by the resulting red-shifts of the vibration frequency. Since throughout all the MD simulations one water molecule is strongly interacting with the $C_2 = O_2$ group, differences in the red-shifts have to be attributed to other water molecules, and it turns out that the interaction strength of a second water molecule, and fluctuations therein, can indeed explain this effect. A linear correlation coefficient between the fluctuations of the interaction energies of this second water molecule and the instantaneous frequencies, computed by a wavelet analysis, of 0.4 confirms the relation between water - $C_2 = O_2$ interaction, but also reveals that there are other effects, and/or higher-order correlations, that need to be considered. Such other effects are found, at least qualitatively, as changes in the water topology, i.e. water bridges connecting the $C_2 = O_2$ group and the other polar groups, which occur around times when also the instantaneous frequencies exhibit large jumps. This is again indicative of the higher order hydration shells also affecting the carbonyl frequencies and this has at least to be averaged out, to render the somewhat simple minded one-water or two-water states to be sufficient descriptors.

Relating the computed interaction energies with the water-hydrogen-carbonyl-oxygen distances and water oxygen-hydrogen-carbonyl oxygen angles, that is hydrogen-acceptor distances and donor-hydrogen-acceptor angles that are typically used to geometrically define hydrogen bonds, we find a strong correlation of the interaction energies with both, the distances and the angles for the second water molecule. By clustering the interaction energies we could identify a 'strongly interacting zone' that coincides with the geometric criteria for hydrogen bonds, often used in the MD community: 2.5 Å of maximal hydrogen-acceptor distance and a donor-hydrogen-acceptor angle that deviates by at most 45° from linearity. (When analysing MD simulations with classical force fields the distance criterion is often taken as 3.5 Å of maximal donor-acceptor distance, but with ~1 Å as the typical donor-hydrogen distance these two distance criteria can be considered equivalent.) This is also the hydrogen-bond criterion used for a geometrical definition of a hydrogen bond in the present study. The distribution of interaction energies over the hydrogen-bond distance/hydrogen-bond angle space, however, reveals also that the interactions in the 'strongly interacting', hydrogen-bonded zone are not necessarily strong since in some frames water molecules with a correct hydrogen-bonded position interact only moderately. In turn, also with water molecules positions outside the hydrogen-bonded region, strong in-

interactions with the $C_2 = O_2$ group are occasionally computed. We therefore like to stress that, though the geometric criteria has been confirmed by the averaged interaction energies in the hydrogen-bonded zone, this geometric definition is useful for looking at probabilities for hydrogen bonds, taken from averaging over many water positions. If used in this manner, a simple geometric criterion is indeed a fast and representative metric to describe the hydrogen-bonded state of a system, or at least a carbonyl group surrounded by water molecules.

The different hydrogen-bonded scenarios of the $C_2 = O_2$ group observed in the time windows and constructed by constraints, representing the averaged interactions between the water molecules and the $C_2 = O_2$ group, can thus be used to qualitatively explain the observed red-shifts in the vibrational frequency. We are furthermore confident that the different probabilities of the three carbonyl groups in the ALAL-peptide to form hydrogen bonds with the solvent can also be used to explain the observed differences in their individual vibrational frequencies. The averaged number of hydrogen bonds simply has to be considered not the cause of a red-shift but rather a marker for a hydration situation with water-carbonyl interactions that leads to such a shift.

The definition of a hydrogen bond is something scientists argue about since at least Pauling and the one found in a IUPAC technical report [ADK⁺11] ‘The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.’ may be precise but not directly helpful. Even in the very publication, several ways to provide ‘evidence of bond formation’ are presented, and the ‘attractive interaction’ or ‘nature of physical forces’ are among them.

Different hydrogen-bond criteria have been evaluated in e.g. [PGSR13, CLS⁺11, KSS07, Mat07, OZC14], all having their different merits. But the most reassuring statement is probably ‘The fact that different choices for the relevant geometric variables and a quite different electronic structure approach all lead to quite similar results for both the statics and dynamics of H-bond number fluctuations does perhaps suggest that these ways [i.e. geometric definitions] of considering H bonds in the liquid can be insightful.’ [KSS07]. And another consensus is that the relation between water position, i.e. distance and angles, with respect to the interaction partner, is related to the interaction strength as also found in the present study. The difficulty is rather where, not whether or not, to put the cut-offs, be it on the energy scale or geometrically. Our compromise by clustering the interaction energies and determine the (‘strongly interacting’) hydrogen-bonded zone from the distance/angle distribution of the cluster members reduces some of the arbitrariness but is of course unnecessary if one chooses to work with interaction energies directly.

One also needs to keep in mind that the fragmentation approach used to calculate

the interaction energies introduces errors (due to the capping hydrogen atoms) that can in principle be different for the various frames. Correction methods such as scaling the capping hydrogen atoms [MH16] or using embedded charges as done in another fragmentation method [Col14] exist. Having a decent number of data points though, we are confident that such errors are averaged out.

In recent works, both the distance to a hydrogen-bonded partner as well as the strength of a hydrogen bond (and the amount of charge transfer) have been found to correlate with the *OH* stretch frequency in liquid methanol [YKCC12] or water [CYK⁺12, OKK18] or the *ND* stretch frequency in NMA [BM17]. These correlations have been determined by comparing the instantaneous frequency of the stretching vibrations, as computed by a wavelet analysis of first-principles MD simulations, with the hydrogen-bond distances [CYK⁺12, BM17, MSC08a] or the hydrogen bond strength [OKK18]. In our work, we relate solute-solvent interactions with solvent vibrations, and therefore have significantly fewer data points than available for solvent-solvent interactions and corresponding frequencies. This may be one reason why the correlation between interaction energies and instantaneous frequencies of the $C_2 = O_2$ vibration are less pronounced than in other works in the literature. Still, the approach has proven useful to identify the relations.

Even without a detailed analysis of the instantaneous frequencies, computing vibrational spectra of solvated peptides by first-principles MD simulations, provides significant insight into the dynamics of the molecule in water and the effect of the solvent on the (calculated) vibrational properties [GMV07, TBF⁺13, MGD⁺06, Gai10d]. The carbonyl frequencies computed in this work are in the same range as those computed for other peptides of similar size [MGD⁺06, Gai10a, Gai08]. For example, the calculated frequency of the alanine dipeptide is 1605 cm^{-1} and the experimental value is 1635 cm^{-1} [KWH05].

Our present results are in particular comparable to those of to the Ala-Leu peptide, with measured carbonyl vibration at 1660 cm^{-1} and a calculated frequency at $\sim 1600 \text{ cm}^{-1}$ [HDS⁺18]. This is the same frequency, we find in this work for the central carbonyl group in a one-water state, in agreement with the one hydrogen bond (on average) observed for the carbonyl group in Ala-Leu [HDS⁺18], confirming again the interplay of hydrogen-bonds and vibrational frequencies.

5.5 Conclusions

The present analysis of the hydration shell around the ALAL peptide and, in particular, the central $C_2 = O_2$, carbonyl group, afforded us to closer inspect the factors that influence the vibrational frequency, and thus the spectroscopic signature, in the amide I region. Differences in the frequencies of the three individual carbonyl groups can be attributed to their interactions with the surrounding wa-

ter. The probabilities of the groups to form hydrogen bonds with water is in agreement with the observed shifts in the computed stretching frequencies.

States in which the central carbonyl group has one or two hydrogen-bonded water molecules, or a mixture thereof, (either observed over time windows of an unconstrained simulation or constructed by constraints) exhibit a clear trend of the red-shift of the $C_2 = O_2$ vibrational frequencies with the averaged number of hydrogen-bonded water molecules. The amount by which the frequencies are lowered is reflected in the strengths of the interaction energies between the closest water molecules and the peptide fragment containing the $C_2 = O_2$ group. Since one water molecule is strongly interacting throughout the simulations, it is in particular the second water molecule that is decisive for finding one or two strongly interacting, and also hydrogen-bonded, water molecules. It is this second interaction that determines the amount of the (additional) red-shift.

The geometric definition of a hydrogen bond by distance and angle criteria, typically used in the MD community and also in this work, is justified by the distribution of interaction energies between water molecules and the $C_2 = O_2$ group of the ALAL peptide over water-carbonyl distances and angles. Since, however, strong interactions are also observed for water molecules that are outside the geometric cutoffs and, likewise, weaker interactions for water molecules within the criteria for hydrogen bonds, the geometric definition holds only on average. Still, when used together with ensemble averages or dynamical averages, the average number of (geometrically defined) hydrogen bonds can serve as a qualitative representative of stronger or weaker interactions which, in turn, more strongly or more weakly, impact the strength of, e.g., a carbonyl bond, and thus its vibrational frequency. With first-principles MD simulations, these averages can be computed, providing simultaneous insight into the dynamical interaction of water with the peptide and the vibrational dynamics of the individual groups involved.

Chapter 6

Metastable-Conformations vs. Hydration Shell

6.1 Introduction

The simultaneous study of a range of timescales and the underlying atomistic/molecular origins of protein-water interactions is challenging due to the complex structural and dynamical properties of both protein and water. To understand how the underlying conformation affects the vibrational signatures and how the vibrational signatures are impacted by the dynamics of water surrounding exposed polar residues of each conformation. In this work, using the combined approach, we investigated the conformational diversity of a model peptide Ala-Leu-Ala-Leu (ALAL) in water, using forcefield-based MD simulations, followed by the construction of a MSM. For representative metastable-conformations, individual first-principles MD simulations were performed. The vibrational signatures were then computed from these trajectories by employing both maximally localized Wannier functions scheme and Voronoi integration. Geometrical analyses and dynamic analyses of surrounding water molecules afforded as an understanding the structural changes of the model peptide and the underlying atomistic/molecular origins of peptide-water interactions.

6.2 Methods

6.2.1 Classical Molecular Dynamics simulations

We performed six 2.5 μ s-long classical MD simulations of the Ala-Leu-Ala-Leu (ALAL) peptide in a cubic simulation box of explicit water (1477 molecules modeled as TIP3P [JCM⁺83] water) employing the AMBER 99SB-ILDN [HAO⁺06, LLPP⁺10] force field and the gromacs 5.0.8 programme [PPS⁺13]. The positions of the solute atoms were saved to file every 0.25 ps. Classical MD simulations are set up similarly to those described in the previous chapter.

6.2.2 Markov State Modelling

A Markov state model has been constructed using the trajectory (15 μ s combined) of partially deuterated ALAL in deuterated water obtained from classical MD simulations on the conformational space spanned by the torsion angles ψ_{Ala1} , $[\phi_{Leu2}, \psi_{Leu2}]$, χ_{Leu2} , $[\phi_{Ala3}, \psi_{Ala3}]$, ϕ_{Leu4} and χ_{Leu4} (highlighted in Figure 6.1).

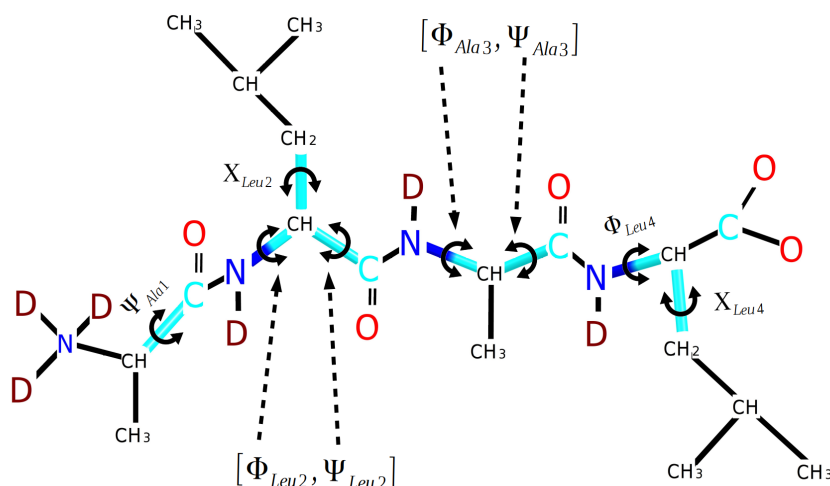


Figure 6.1: Scheme of ALAL and torsion angles ψ_{Ala1} , $[\phi_{Leu2}, \psi_{Leu2}]$, χ_{Leu2} , $[\phi_{Ala3}, \psi_{Ala3}]$, ϕ_{Leu4} and χ_{Leu4} used for Time-lagged independent component analysis (TICA). Carbon atoms are shown in cyan, nitrogen in blue and oxygen in red. "D" denotes a deuterium atom. Note that bonds with hydrogen atoms are not shown. Note that $[\phi_{Leu2}, \psi_{Leu2}]$ -pair refers to the first peptide bond and $[\phi_{Ala3}, \psi_{Ala3}]$ -pair refers to the second peptide bond respectively.

The conformational space reduced to the eight dimensions corresponding to torsion angles (Figure 6.1) is further reduced using the time-lagged independent component analysis (TICA), which is well known to capture the slow reaction coordinates [PHPG⁺13]. The reduced conformational space to major TICA components is then clustered employing a k-means clustering algorithm [HW79]. The number of cluster centroids, 350, was sufficient to discretize the dynamics and results in the convergence of implied time scales (ITS). The free energy surface along the first two TICA components and the distribution of cluster centroids on TICA space are shown in the Appendix C, Figure C.2. A transition matrix has been computed for the transitions between the clusters, found after k-means with varying the lag time, up to 500 ps. The ITS (Appendix C, Figure C.3) graph indicates a convergence of the slowest process at about 100 ps and the existence of

mainly four slow processes. The MSM was then constructed using a lag time of 100 *ps*. Appendix C, Figure C.3 shows the eigenvalue spectrum of the transition matrix sampled with lag time $\tau = 100$ *ps*. There is a clear spectral gap after the first five eigenvalues. A robust Perron Cluster Analysis (PCCA+ [RW13]) was applied to merge the clusters into metastable sets. We performed PCCA+ for four, five, and six metastable-sets. Proper partitioning of the cluster centroids is achieved for four and five metastable-sets, resulting in clean macrostates. In the case of six metastable-sets, there was no further partitioning, leading to mixed macrostates. Therefore, we chose to stick with five metastable-sets; the ITS and eigenvalues spectrum analysis also leads to the same selection. We also validated the MSM using the Chapman-Kolmogorov Test [PWS⁺11] (see Appendix C, Figure C.4). By analyzing eigenvectors of the transition matrix sampled with lag time $\tau = 100$ *ps* and the distributions of dihedral angles of each metastable-set, slow processes are identified. Their corresponding timescales are identified by using eigenvalues and $t_i = -\frac{\tau}{\log|\lambda_i(\tau)|}$. For all the steps involved in the construction and validation of the MSM, we used PyEMMA [STSP⁺15b].

6.2.3 First-Principles Molecular Dynamics Simulations

The details of the first-principles MD simulations setup used for the simulations of β -sheet like conformation of ALAL are given in [HFI21]. The same setup is being used for all other conformations. The default Gaussian and plane waves (GPW [LHP99]) electronic structure method, as implemented in the Quickstep module [VKM⁺05] of the CP2K package [HISV14, KIDB⁺20b], was used. The Geodecker–Teter–Hutter (GTH) norm-conserving pseudopotentials, double zeta valence plus (DZVP) basis set and BLYP with Grimme’s D3 dispersion correction exchange–correlation functional were employed.

Initially, the system was energy minimized using a conjugate-gradient algorithm where the positions of the ALAL atoms were fixed. This allows the solvent molecules to relax around the peptide and find energetically favourable positions, followed by a 5 *ps* NVT equilibration run from the minimized system, during which the solute was kept fixed to avoid the transition to an undesired conformation. Subsequently, another 50 *ps* long NVT run using massive Nose-Hoover chain thermostats to sample the independent starting configurations for the NVE simulations was performed.

Finally, we performed the production runs of each 25 *ps* in an NVE ensemble. We performed five independent runs for each conformation, and runs with non-negligible thermal fluctuations were not included in the analysis. The time-step for the numerical integration was 0.5 *fs*, and atom positions were saved every step.

To calculate electronic density and Wannier centers based IR spectra, Voronoi integration [GF82, BT21, TBK15] of the total electron density and Wannier localiza-

tion [KSV93, Res94, MV97, SP99a, SP99b] was performed every step, respectively. The gathered molecular electromagnetic moments using Voronoi tessellation and Wannier centers are large enough to compute reliable IR spectra of ALAL conformations in explicit solvent.

Fourier-transform-based spectra and the structural analysis were conducted using the TRAVIS program [BK11, BTGK20], and in-house, Python scripts.

All spectra reported are the mean over three simulation runs and errors are estimated as the standard deviation from the mean.

6.2.4 Normal Modes

We have carried out a normal mode analysis using the Gaussian programme package [FTS⁺16]. The representative conformations of ALAL have been optimised (convergence criterion $3.00E - 04E_H / \text{\AA}$) in implicit water (modelled by a polarisable continuum model, PCM, with a dielectric of $\epsilon = 80$) at the DFT level of theory. The same BLYP exchange-correlation density functional as for the first-principles simulations and a comparable basis set, 6-31G(d) were used. On the optimised geometries, a frequency calculation was performed in which all polar hydrogen atoms are assigned an atomic mass of 2.

6.2.5 Hydrogen Bonds and Water Topology

We used a geometric criterion of 2.5\AA as the maximal distance between the hydrogen atom and the acceptor and a donor-hydrogen-acceptor angle of 45° as the maximal deviation from linearity for the analysis of hydrogen bonds and water bridges. As is discussed in our previous paper [HFI21], for fast and efficient evaluation of hydrogen-bonding states, the use of geometric criteria is justified by the calculated interaction energies. Analysis of hydrogen bonds and water bridges is done with MDAnalysis [MADWB11]. All of the other analyses, except where noted, were done with in-house Python scripts. The plots were generated with the help of matplotlib [Hun07].

All values reported are the mean over three simulation runs and errors are estimated as the standard deviation from the mean.

6.3 Results

6.3.1 Markov State Modelling

We named the conformations of ALAL based on the areas populated on the conformational-space (α, β , or L_α region) spanned by the two dihedral angles

pairs distributions, i.e., the $[\phi_{Leu2}, \psi_{Leu2}]$ -pair which refers to the first peptide bond, and the $[\phi_{Ala3}, \psi_{Ala3}]$ -pair which refers to the second peptide bond respectively (see Figure 6.1). For example, (β, β) represents that the sampled values of both dihedral angle pairs populate the β -sheet region of their corresponding 2D distributions. This notation will be used from now onwards.

We have reported in the previous chapter that the probabilities of each conformation extracted directly from the simulations data show that the (β, β) -conformation is the most probable one, followed by the (β, α) and (α, β) conformations. The (L_α, L_α) -conformation is the highly unlikely conformation whereas the (β, L_α) and the (α, α) conformations have a comparatively slightly higher probability, see Figure 6.2.

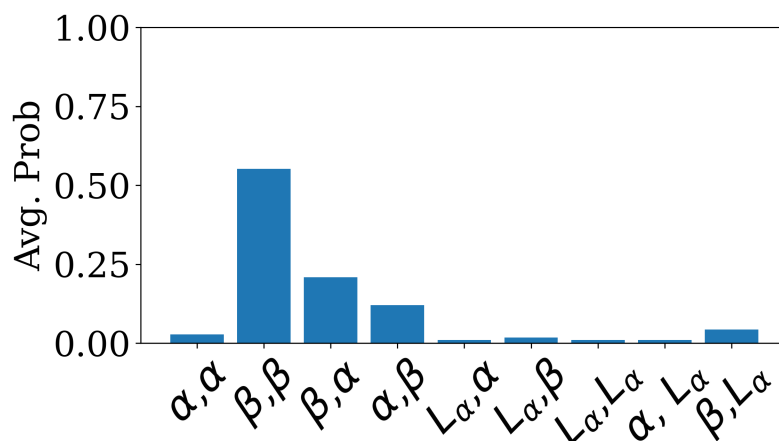


Figure 6.2: Probability distribution of different backbone conformations as observed in the classical MD simulations [HFI21], the first label refers to the first peptide bond, i.e., the ψ_{Leu2}, ϕ_{Leu2} -pair, and the second one to the second peptide bond, i.e., the ψ_{Ala3}, ϕ_{Ala3} -pair (of Figure 6.1).

The transition network plot, along with a representative conformation of each metastable-set and 2D distributions of $[\phi_{Leu2}, \psi_{Leu2}], [\phi_{Ala3}, \psi_{Ala3}]$ -pairs is shown in Figure 6.3. The metastable-sets (found after kinetic clustering), V and III, of ALAL contain more than one type of conformation. The Set-V contains the (β, β) , (β, α) , (α, β) , and the (α, α) conformations, and it has the highest probability due to highly probable (β, β) -conformation and second highly probable (β, α) and (α, β) -conformations present in this set. The possible transitions between the member conformations of Set-V require rotations of either ψ_{Leu2} or ψ_{Ala3} . Further, note that the (α, α) -conformation is less probable than the other member conformations and does not contribute much to the overall probability of this set.

The remaining four sets have significantly less probability compared to Set-V. Looking at Set-IV and Set-I, which contains (β, L_α) -conformation and (α, L_α) -conformation respectively, Set-IV has greater probability compared to Set-I, once again because

of the presence of β -like conformation for $[\varphi_{Leu2} - \psi_{Leu2}]$ dihedral angle pair. Moreover, the transition probability is high from Set-I to Set-IV i.e., α -helix to β -sheet like, involving rotation of ψ_{Leu2} . In contrast, the transition from Set-V to Set-I or Set-IV mainly involves φ_{Ala3} rotation. Like Set-V, the Set-III also contains more than one conformation, i.e., (L_α, β) and (L_α, α) -conformations, with (L_α, β) being more probable one, and it requires φ_{Leu2} rotation to jump from Set-V to this set. Lastly, the Set-II is the one which consists of highly improbable (L_α, L_α) -conformation which requires rotations of $\psi_{Ala3}, \varphi_{Ala3}, \psi_{Leu2}, \varphi_{Leu2}$ dihedral angles, in order to reach this set from the highly desired Set-V. Note that within each set χ_{Leu2} and χ_{Leu4} rotations are possible.

From slowest to fastest, we can specify the slow processes for ALAL in terms of transitions between the metastable-sets and assign the corresponding timescales using the ITS:

1. Transition from Set-V $((\beta, \beta), (\beta, \alpha), (\alpha, \beta), (\alpha, \alpha))$ to Set-II (L_α, L_α) which involves $\psi_{Ala3}, \varphi_{Ala3}, \psi_{Leu2}, \varphi_{Leu2}$ rotations and the timescale for this process is ~ 24 ns.
2. Transition from Set-IV (β, L_α) or Set-I (α, L_α) to Set-II (L_α, L_α) , which involves $\psi_{Leu2}, \varphi_{Leu2}$ rotations, and the timescale for this process is ~ 13 ns .
3. Transition between Set-III $(L_\alpha, \alpha / \beta)$ and Set-II (L_α, L_α) which involves $\psi_{Ala3}, \varphi_{Ala3}$ rotations, and the timescale for this process is ~ 6 ns .
4. Transitions within Set-V, i.e., (between $(\beta, \beta), (\beta, \alpha), (\alpha, \beta), (\alpha, \alpha)$ -conformations), those requires ψ_{Leu2} and ψ_{Ala3} rotations, and their corresponding timescales are ~ 3 ns and ~ 1 ns respectively.

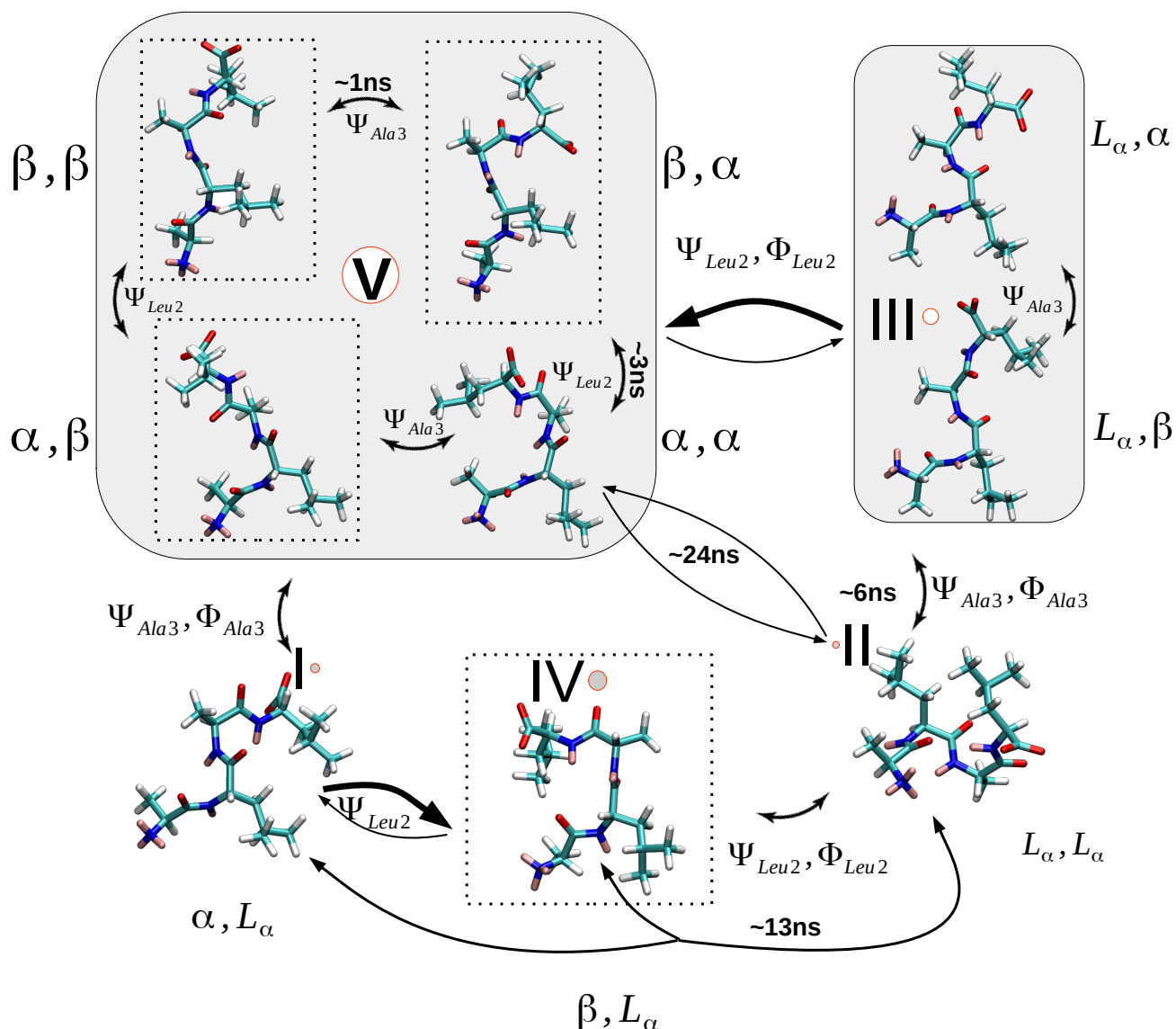


Figure 6.3: Coarse-grained model of the conformational dynamics of ALAL in water. The five metastable sets, I–V, are represented as grey/white circles whose size corresponds to the probability of the respective set. The shaded rectangular region for Set-V and Set-III shows the presence of more than one conformation in these sets. The thickness of the arrows between the circles indicates the transition probability between a pair of metastable sets. The molecular structures near to the circles are representative conformations in the respective set. Arrows between molecular structures indicate the coordinate of the conformational transition connecting two conformations or sets. Carbon atoms are shown in cyan, oxygen atoms in red, nitrogen atoms in blue, hydrogen atoms in white and deuterium atoms in pink.

6.3.2 Normal Modes Analysis

The normal modes spectra of the (β, β) -conformation in implicit water is shown in Appendix C, Figure C.6. In the Amide-I frequency range, the two most prominent bands around $\sim 1629\text{cm}^{-1}$ and $\sim 1648\text{cm}^{-1}$ of spectra in implicit solvent corresponds to stretching vibration of carboxyl (νCOO^-) and carbonyl ($\nu\text{C=O}$) groups, respectively. Moreover, the Amide-II and Amide-III frequency ranges ($\sim 1250\text{cm}^{-1}$ - $\sim 1450\text{cm}^{-1}$ and $\sim 1100\text{cm}^{-1}$ - $\sim 1250\text{cm}^{-1}$) contain the fingerprints of components of peptide bonds, N-C bond stretchings, N-D groups bendings, and ND_3^+ stretching/bendings motions, respectively. The contributions of the COO^- group can also be seen in the Amide-II region, and a gentle peak around $\sim 1270\text{cm}^{-1}$ is assigned to the C-H vibrations of sidechains. During optimization, all conformations tend to converge to a state that provides for close proximity and optimum interaction between the C=O and the ND_3^+ terminal group, as well as between the COO^- terminal and polar N-D groups.

The metastable-conformations of ALAL are structurally different from each other due to the repositioning of the polar groups. The optimised representative conformations of ALAL are shown in Figure 6.4. Upon optimisation, the characteristic metastable-structure of the conformations is maintained. Notice the presence of intra-molecular hydrogen bonds (indicated by the dashed lines). The details of intra-molecular hydrogen bonds and several other geometric parameters (end-to-end distance, radius of gyration, etc) extracted from the optimised conformations are given in Table 6.1. The end-to-end distance is highest for the (β, β) -conformation and lowest for the (L_α, L_α) -conformation. Likewise, other geometric parameters vary between the optimised conformations.

The intrinsic structural properties of ALAL metastable-conformations may cause a marginal shift in the normal mode frequencies of polar groups. In Table 6.2, the normal mode frequencies of individual carbonyl groups of all conformations of ALAL are reported and a conformational dependence of the frequencies of individual carbonyl groups can be seen. Carbonyl frequencies of other conformations are either red-shifted or blue-shifted compared to the (β, β) -conformation due to the difference in the position of carbonyl groups for each conformation, the presence of intramolecular hydrogen bonds, change in the end-to-end termini distances, etc. The frequencies of $\text{C}_1 = \text{O}_1$ range from $\sim 1639\text{cm}^{-1}$ for the (α, α) -conformation to $\sim 1662\text{cm}^{-1}$ for the (β, α) -conformation. Lack of β -sheet like content causes redshift in the frequency of the first carbonyl group of ALAL, and it is apparent by looking at the frequencies of pure α -helix, L_α -helix like conformations. These pure conformations tend to form intramolecular hydrogen bonds; the $\text{N}_1\text{-D}_1 \cdots \text{COO}^-$, $\text{C}_1=\text{O}_1 \cdots \text{N}_3\text{-D}_3$ intra-molecular hydrogen bonds are present for pure α -helix, L_α -helix like conformations, respectively. The frequencies of $\text{C}_2 = \text{O}_2$ range from $\sim 1635\text{cm}^{-1}$ for the (β, β) -conformation to $\sim 1677\text{cm}^{-1}$ for the (β, L_α) -conformation. This carbonyl group is positioned centrally, where it is least influenced by the charged termini, ND_3^+ and COO^- , re-

spectively. The third carbonyl group is located close to the COO^- group and its frequencies range from $\sim 1633\text{cm}^{-1}$ for the (α, β) -conformation to $\sim 1652\text{cm}^{-1}$ for the (α, α) -conformation. The second and third carbonyl groups are not involved in intramolecular hydrogen bonds in the optimised structures used for the calculation of normal modes. And, there is a large redshift in the COO^- group frequency of the (α, α) -conformation compared to the other conformations, due to the presence of two intramolecular hydrogen bonds, $\text{N}_1\text{-D}_1 \cdots \text{COO}^-$ and $\text{Ala}_1\text{-sidechain} \cdots \text{COO}^-$.

Previously, we calculated normal modes of the (β, β) -conformation for several hydrogen bonding situations and showed that the hydrogen bonding situation of polar groups (C=O's, N-D's) is not straightforward, and none of these clean scenarios are representative of the full hydration in explicit water. Furthermore, at least for smaller systems like ALAL, it is essential to look at the components of the most prominent bands, for example, the Amide-I band, which results from the superposition of peaks originating from the individual carbonyl groups. The interaction of each polar group with water also varies depending upon neighboring residues [HFI21].

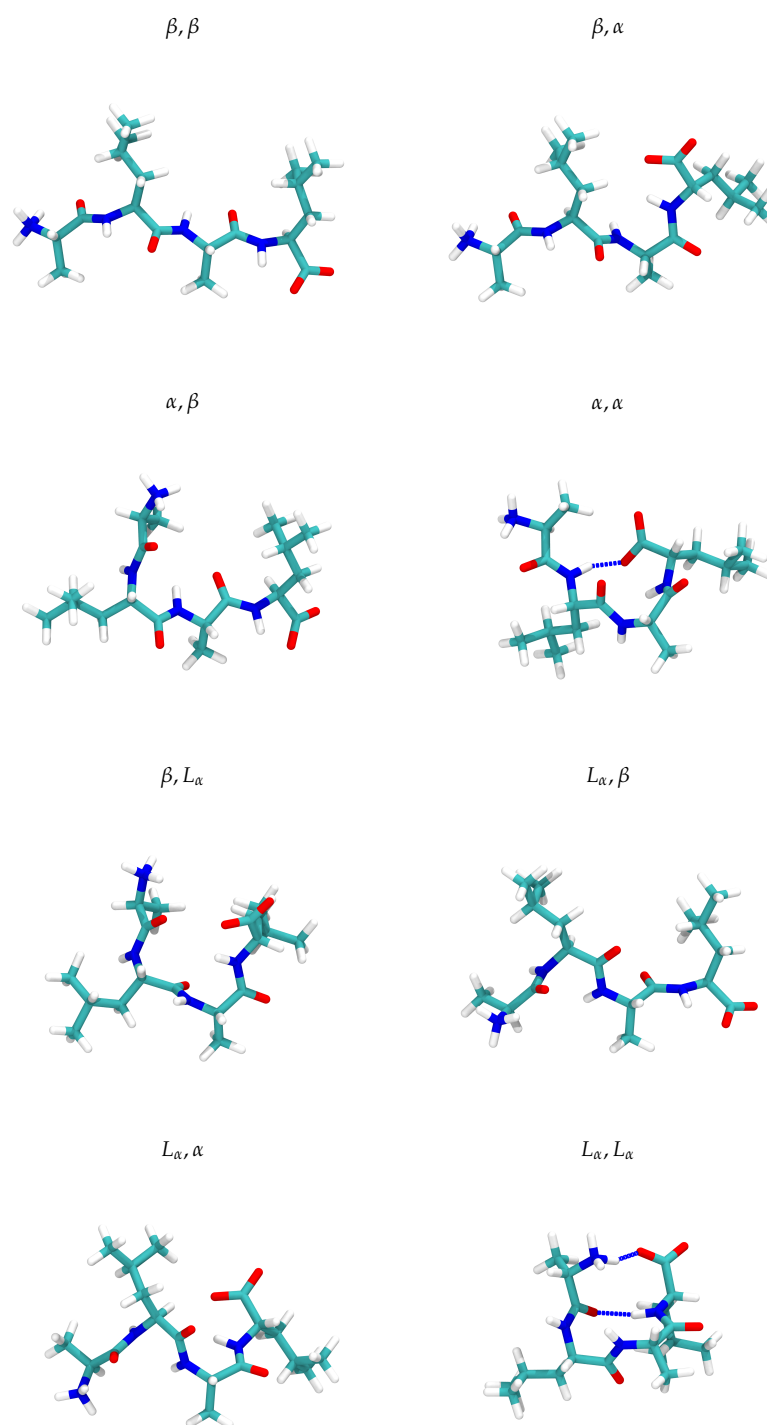


Figure 6.4: Optimised geometries of representative conformations of ALAL used for the calculation of normal modes frequencies.

6.3 Results

Table 6.1: Geometric parameters extracted from the optimised geometries of representative conformations of ALAL.

Conformations	End-to-End distance (Å)	Radius of gyration (Å)	Distance		Intramolecular Hydrogen bonds	O-C-N angle (degree)			N-C-C α angle (degree)		
			between COM of Ala-SC's (Å)	between COM of Leu-SC's (Å)							
β,β	12.8	4.8	6.5	8.3	-	125.6	124.1	125.3	115.5	115.6	114.6
β,α	10.2	4.4	7.3	8.5	-	125.3	124.2	125.3	115.8	115.5	116.6
α,β	9.2	4.4	6.4	10.8	-	124.3	124.2	125.2	116.3	116.6	114.5
α,α	6.0	3.9	8.5	9.2	N ₁ -D ₁ ...COO ⁻	125.7	122.0	125.0	116.0	117.8	115.3
β,L_α	6.3	3.9	7.4	9.6	-	124.5	124.2	124.8	116.2	113.9	115.5
L_α,β	11.1	4.5	7.2	8.2	-	124.4	122.9	125.3	116.7	115.9	113.8
L_α,α	8.8	4.1	7.8	9.6	-	124.7	122.6	125.5	116.2	115.6	115.4
L_α,L_α	3.7	3.7	8.5	7.7	1-ND ₃ ⁺ ...COO ⁻ 2-C ₁ =O ₁ ...N ₃ ⁻ D ₃	123.0	124.3	124.1	116.8	113.8	117.6

Table 6.2: Normal modes frequencies of carbonyl groups and charged C-terminal of each conformation of ALAL in implicit solvent.

Conformations	C ₁ = O ₁ (cm ⁻¹)	C ₂ = O ₂ (cm ⁻¹)	C ₃ = O ₃ (cm ⁻¹)	COO ⁻ (cm ⁻¹)
β,β	1661	1635	1647	1629
β,α	1662	1646	1643	1624
α,β	1652	1658	1633	1627
α,α	1639	1658	1652	1604
β,L_α	1654	1677	1644	1629
L_α,β	1654	1667	1651	1628
L_α,α	1656	1666	1649	1623
L_α,L_α	1649	1673	1635	1629

6.3.3 First-Principles MD simulations

6.3.3.1 Structural Analysis

Metastable-conformations of proteins/peptides exhibit a wide range of flexibility, particularly when compared to the most probable conformation. Appendix C, Figure C.15 shows the root mean square fluctuations of the backbone of pure β -sheet like, pure α -helix and L_α -helix like conformations of ALAL. It can be seen that the first peptide bond exhibits the greatest flexibility in all three conformations, the second peptide bond is less flexible and the third peptide bond has the least spatial fluctuations. The main contributor of flexibility for each peptide bond is the oxygen atom of the polar carbonyl groups (C=O's). Between the conformations, β -sheet like conformation of ALAL is the most flexible and the L_α -helix conformation is the most compact. The distance between polar atoms also changes (and more notably, the end-to-end distance) for each metastable-conformation, and this difference is typically maintained during the first-principle MD simulations. (see Figure 6.5 and 6.6). The end-to-end distance is largest for β -sheet like conformation and shortest for the (β, L_α)-conformation, and L_α -helix like conformation. Similarly, the distances between the polar groups of the first and third peptide bonds differ between the different conformations. Additionally, the spacing between Ala-sidechains and Leu-sidechains varies among the different conformations (see Table 6.3).

6.3 Results

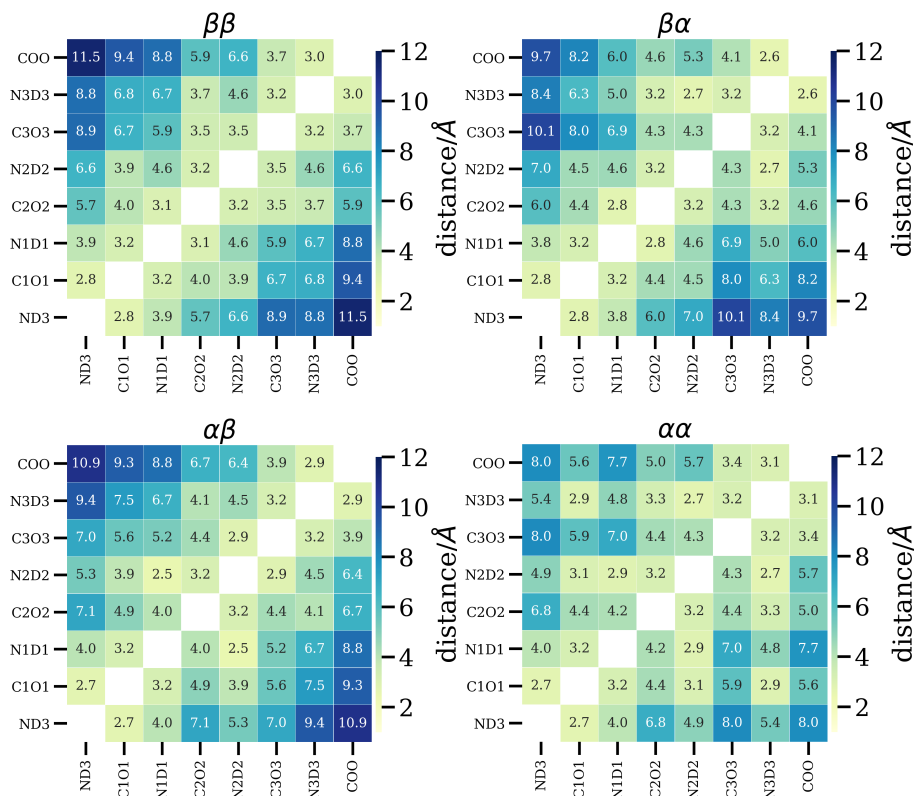


Figure 6.5: Average distance between polar atoms extracted from the first-principle simulations of ALAL conformations of Set-V.

Table 6.3: Geometric parameters extracted from the first-principle MD simulations of representative conformations of ALAL.

Conformations	End-to-End distance (Å)	Radius of gyration (Å)	Distance between COM of SC's (Å)					
			Ala ₁ -Leu ₂	Ala ₁ -Ala ₃	Ala ₁ -Leu ₄	Leu ₂ -Ala ₃	Leu ₂ -Leu ₄	Ala ₃ -Leu ₄
β,β	11.46±0.16	11.09±0.04	6.55±0.11	7.80±0.04	7.41±0.13	6.72±0.06	8.23±0.58	6.57±0.05
β,α	9.72±0.27	11.05±0.01	6.50±0.09	7.99±0.16	11.5±0.34	6.81±0.01	7.84±0.54	5.55±0.20
α,β	10.88±0.12	11.05±0.01	6.22±0.11	5.71±0.41	8.89±0.06	7.05±0.11	9.26±0.21	6.55±0.05
α,α	7.96±0.60	10.98±0.02	6.34±0.24	5.33±0.47	6.15±0.68	7.35±0.08	8.82±0.21	6.09±0.20
β,L_α	6.06±0.46	11.03±0.01	7.14±0.03	7.89±0.15	4.82±0.04	4.89±0.12	8.08±0.13	5.82±0.10
L_α,β	10.22±0.62	11.04±0.01	5.14±0.34	7.04±0.28	11.65±0.1	7.58±0.03	9.15±0.26	6.51±0.05
L_α,α	8.79±0.49	11.01±0.01	5.09±0.30	6.79±0.21	9.85±0.62	7.28±0.10	10.34±0.15	5.11±0.25
L_α,L_α	6.41±0.62	11.11±0.02	5.06±0.09	8.26±0.00	7.09±0.30	7.07±0.05	5.49±0.26	6.41±0.09

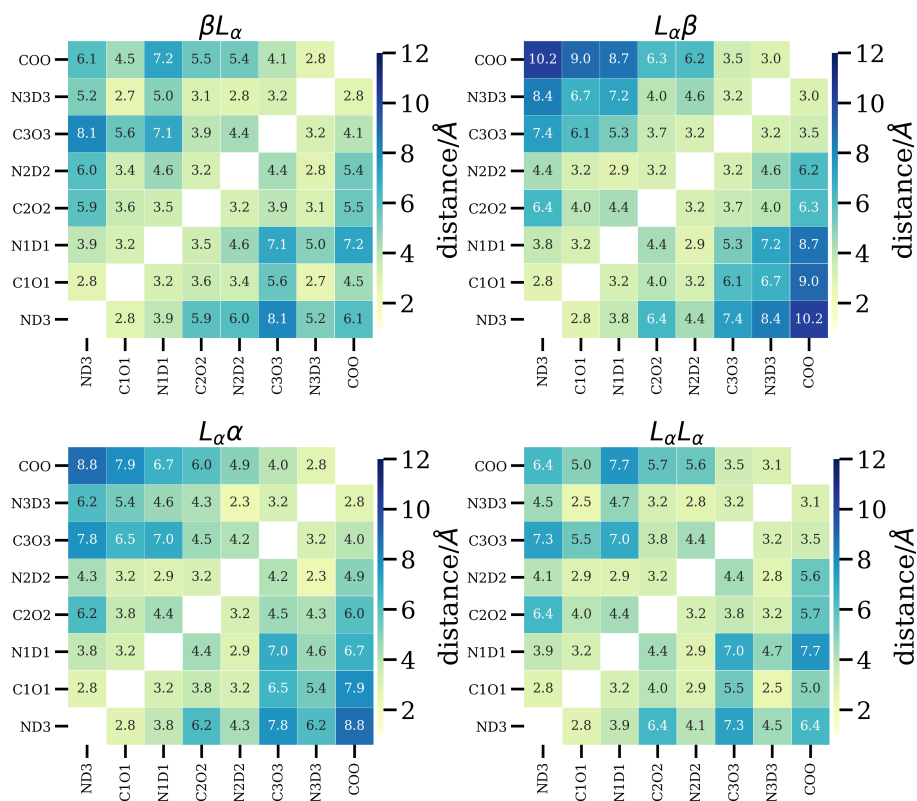


Figure 6.6: Average distance between polar atoms extracted from the first-principle simulations of ALAL conformations of Set II-IV.

6.3.3.2 Distribution Functions

The radial distribution functions (RDF) of carbonyl groups are shown in Figure 6.7 (see Appendix C for the RDF's of other polar groups). The solvation probability of carbonyl groups of the (β, β) -conformation of ALAL decreases from $C_1 = O_1$, over $C_2 = O_2$, to $C_2 = O_3$ [HFI21]. We can see a similar trend for other conformations of ALAL as well. However, for individual carbonyl groups, the different heights of the first RDF peak tells us that the sampling of preferred hydrogen bond distance is conformation dependent and the first minimum depicts the presence of a well-defined first hydration shell. The number of water molecules in the first hydration shell of an individual carbonyl group, i.e., the coordination number (shown as the horizontal black line, also given in the Table 6.4), also varies from conformation to conformation. The coordination number of $C_1 = O_1$ for α -helix, L_α -helix like conformations and (β, L_α) -conformation is lower compared to other conformations. For $C_2 = O_2$, it is lower for (L_α, L_α) , (β, α) and (β, L_α) -conformations and for $C_3 = O_3$, the level of solvation is almost the same for each conformation. Furthermore, the angular distribution function (ADF) analysis of the carbonyl groups indicates the degree of similarity in the hydrogen bond angle preference, irrespective of the position of the carbonyl group of the underlying conformation (see Appendix C for the RDF plots of other polar groups and for the ADF plots).

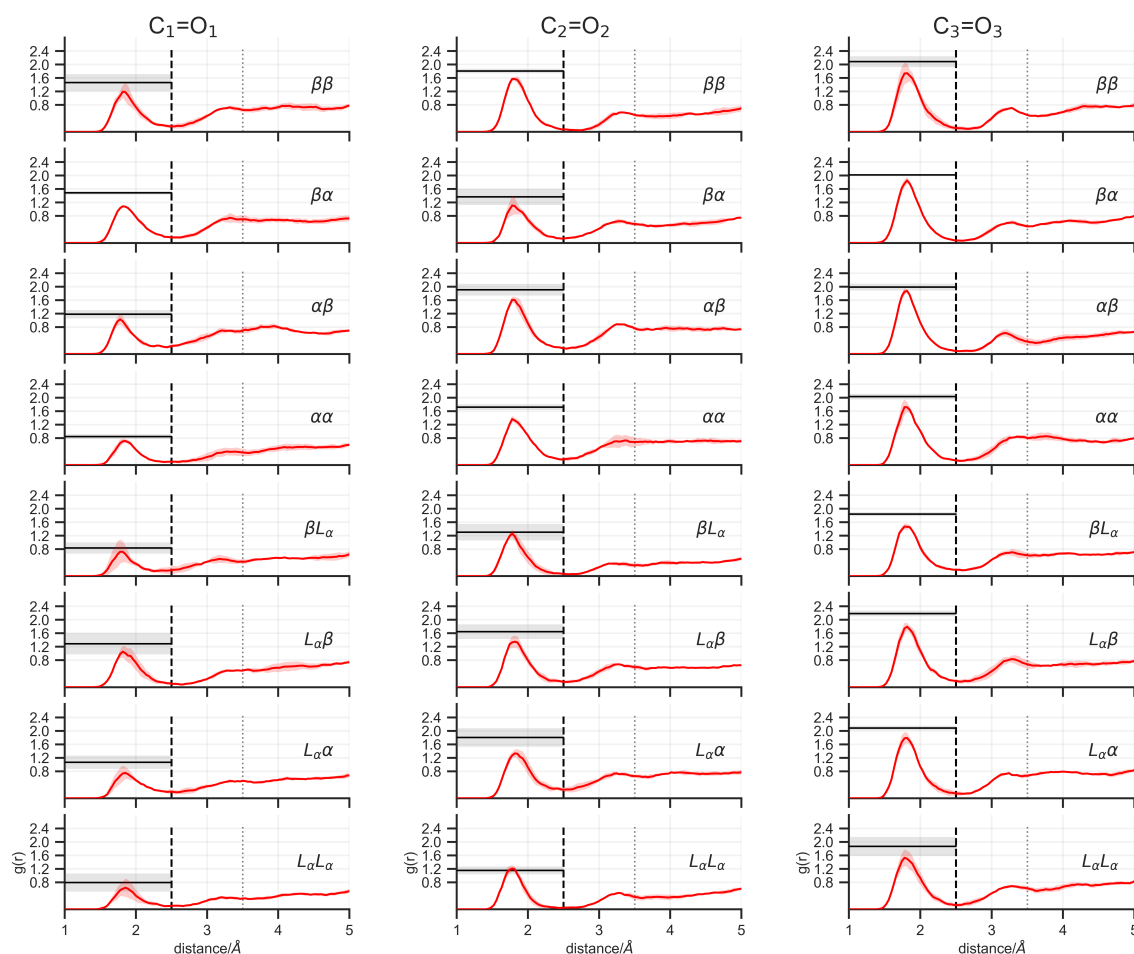


Figure 6.7: Mean RDF plot of all carbonyl groups with shaded region as standard deviation. The vertical dashed line 2.5 Å (black) around first minimum represents the 1st hydration shell i.e., strongly interacting zone [HFI21] (note that this is the geometric criterion for hydrogen bond calculations), the other vertical dotted line at 3.5 Å (gray) represents the so called moderately interacting zone [HFI21]. The horizontal black line represents the mean coordination number with gray shaded region as error band.

The energy of polar group-water interaction is correlated with the distance and angle of hydrogen bonds [HFI21]. In order to examine the spatial distribution of water molecules around the carbonyl groups, the combined distribution function (CDF) of the central carbonyl group, $C_2 = O_2$ of all conformations of ALAL are shown in Figure 6.8 (see Appendix C for the CDF's of other polar groups). Generally, the high-density regions are located in the range, hydrogen-acceptor distance 1.5-2.2 Å and donor-hydrogen-acceptor angles, 0-15° for all carbonyl groups, regardless of their parent conformation. However, the intensity of the distribution in the strongest interacting zone (marked by the dashed lines) varies among the conformations. Looking at the CDF of the first carbonyl group (Ap-

pendix C, Figure C.21), the most noticeable are the distributions of (α, β) , (L_α, α) and (L_α, L_α) -conformations, which are less confined and less intense compared to others, representing the lack of solvation and a poorly defined first solvation shell around the $C_1 = O_1$ of these conformations. The solvation of the $C_1 = O_1$ group is influenced by the relative near location of the charged N-terminal group and possibly due to the greater flexibility of first peptide bond. The moderately interactive zone (the region between the dashed and dotted lines) indicates the likelihood of non-hydrogen-bonded but fairly close water molecules interacting with the first carbonyl group. The only exception is the first carbonyl group of (L_α, β) -conformation that seems to have less space in this zone for free water molecules. The CDF of the central carbonyl group shows that, in case of pure α -helix, L_α -helix like conformations, it is significantly less hydrated. The marginally tilted distribution (in the first zone) towards right and left for α -helix, L_α -helix like conformations respectively, demonstrate a preference for relatively larger and shorter hydrogen bond distance. The $C_2 = O_2$ of (L_α, β) -conformation oddly enough, have a dual likelihood of favorite hydrogen-bond distance and angle values. Finally, if we look at the CDF of the third carbonyl group (Appendix C, Figure C.26) located close to the charged C-terminal group, the first solvation shell is well-defined for all conformations, and there is less probability for free nearby water molecules (moderately interacting zone) except in the case of (L_α, β) -conformation and (L_α, L_α) -conformation. Moreover, there is always a non-zero probability of water being present in the weakly interacting zone. Note that the solvation of N-D groups of ALAL seems to be impacted more with a change in conformation of ALAL (see Appendix C for CDF plots of other polar groups).

Table 6.4: Integrated number of water molecules per polar group with standard deviation.

	ND_3^+	$C_1 = O_1$	$N_1 - D_1$	$C_2 = O_2$	$N_2 - D_2$	$C_3 = O_3$	$N_3 - D_3$	COO^-
β, β	1.02±0.01	1.46±0.26	0.94±0.03	1.80±0.07	0.90±0.01	2.09±0.16	0.90±0.06	2.70±0.14
β, α	0.99±0.02	1.49±0.05	0.93±0.05	1.37±0.24	0.98±0.01	2.02±0.01	0.12±0.10	2.42±0.05
α, β	1.00±0.01	1.18±0.12	0.99±0.01	1.91±0.17	0.39±0.11	1.99±0.10	0.89±0.01	2.68±0.07
α, α	0.98±0.02	0.85±0.07	0.97±0.02	1.72±0.08	0.66±0.06	2.03±0.08	0.11±0.11	2.61±0.09
β, L_α	1.02±0.01	0.83±0.17	0.98±0.01	1.30±0.24	0.97±0.01	1.84±0.06	0.07±0.07	2.59±0.03
L_α, β	1.00±0.01	1.29±0.32	0.98±0.01	1.64±0.22	0.79±0.16	2.19±0.08	0.90±0.06	2.69±0.09
L_α, α	1.01±0.03	1.07±0.20	0.98±0.03	1.80±0.27	0.87±0.08	2.09±0.07	0.78±0.15	2.60±0.12
L_α, L_α	1.02±0.03	0.79±0.27	1.00±0.01	1.15±0.12	0.86±0.01	1.87±0.28	0.08±0.06	2.72±0.10

6.3.3.3 Hydrogen Bonds

The average number of hydrogen bonds of each polar group of each conformation is shown in Figure 6.9. The probabilities to form hydrogen bonds per polar group and intra-molecular hydrogen bonds are provided in Tables S2 and

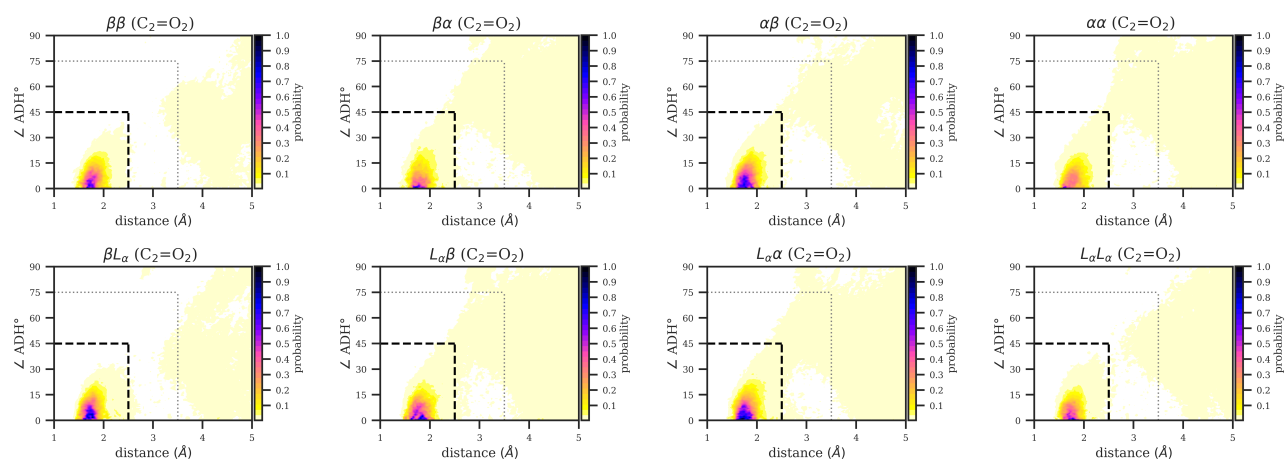


Figure 6.8: CDF of second carbonyl group using data from all three runs of each conformation of ALAL.

6.5. Starting from the charged N-terminal group of ALAL, we can see that for the (β, β) , (β, α) , and (α, β) conformations, the probability of water molecules to form a hydrogen bond with ND_3^+ is higher, whereas, for (α, α) -conformation, it is lower compared to the previous three conformations. The lower hydrogen bond probability for the (α, α) -conformation could be linked to the occurrence of highly probable $\text{C}_1=\text{O}_1 \cdots \text{N}_3=\text{D}_3$, and less likely $\text{C}_1=\text{O}_1 \cdots \text{N}_2=\text{D}_2$ intramolecular hydrogen bonds (see Table 6.5). Moreover the hydrogen bond probability for ND_3^+ of pure left-handed helix, (L_α, L_α) -conformation and the two combinations, $(L_\alpha, \beta), (L_\alpha, \alpha)$ is low compared to (β, L_α) -conformation. This result is also consistent with the presence of an $\text{C}_1=\text{O}_1 \cdots \text{N}_3=\text{D}_3$ intramolecular hydrogen bond in the case of the (β, L_α) -conformation (see Table 6.5). Further, the hydration number of the first carbonyl group ($\text{C}_1 = \text{O}_1$) varies significantly with the conformation and the changes in the hydration number of $\text{N}_1\text{-D}_1$ can also be seen. For more compact (shorter end-to-end distance, as well as a smaller radius of gyration, see Table 6.1 and 6.3) conformations of ALAL, i.e., $(\alpha, \alpha), (\beta, L_\alpha)$ and (L_α, L_α) , there is always a chance that $\text{C}_1 = \text{O}_1$ will form a hydrogen bond with either $\text{N}_2=\text{D}_2$ or $\text{N}_3=\text{D}_3$, with a significantly higher hydrogen bonding probability for $\text{N}_3=\text{D}_3$ (see Table 6.5). The central carbonyl ($\text{C}_2 = \text{O}_2$) being the best representative of a carbonyl group in a longer peptide or protein, as it is least influenced by the termini, shows a clear change in hydrogen bond probability with a change in simulated conformation. The intramolecular hydrogen bonding probability for $\text{C}_2 = \text{O}_2$ is almost zero for each conformation. In the case of $\text{N}_2\text{-D}_2$, the apparent differences are in solvation of $(\alpha, \beta), (\alpha, \alpha)$ -conformations; for (α, β) -conformation the Ala-sidechains are pretty close to each other in comparison to all other conformations, and the $\text{C}_1 = \text{O}_1, \text{C}_2 = \text{O}_2$ groups are aligned against each other, leaving less room for water to interact with $\text{N}_2\text{-D}_2$. For (α, α) -conformation, apart from its compactness, the Leu₂-sidechain and the Ala₃-sidechain are close in proximity,

a possible reason for the lower hydration of N_2-D_2 . Furthermore, the hydration of the third carbonyl group ($C_3 = O_3$) also changes according to the conformation of ALAL, notice the (β, L_α) -conformation interacts with the least number of water molecules. The participation of the N_3-D_3 group in hydrogen bonding also depends on the conformation. For the (β, β) , (α, β) , (L_α, β) , and (L_α, α) conformations, the hydrogen bond probability of N_3-D_3 is close to 1 while for other conformations it is almost zero. Finally, the solvation of the C-terminal group of ALAL also varies slightly with conformation. Note that the hydrogen bond probabilities can also vary slightly between independent simulation runs for the same conformation (see Table S3).

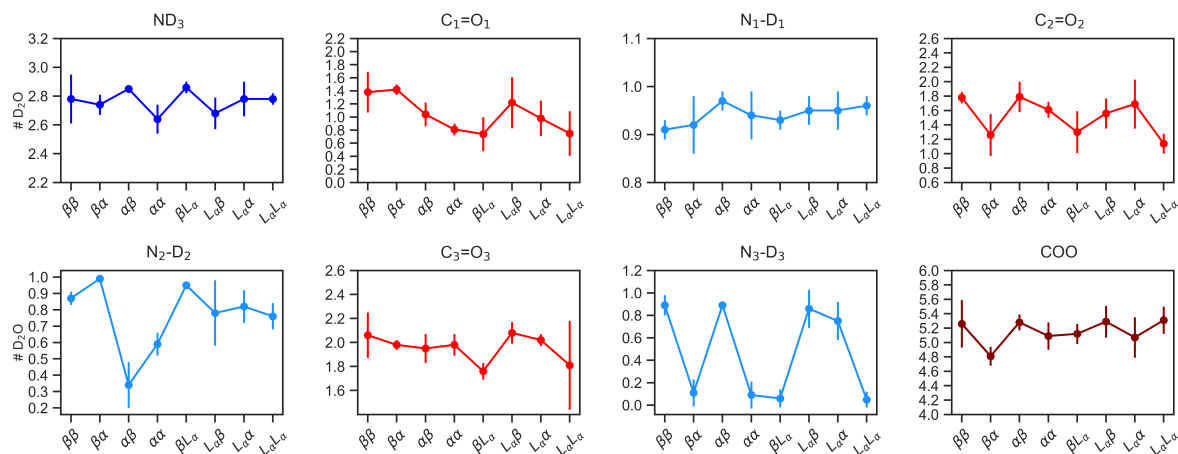


Figure 6.9: Average number of hbonds per polar group from the first-principle MD simulations of metastable-conformations of ALAL.

Table 6.5: Probability and type of intramolecular hydrogen bonds present in first principle simulations of metastable-conformations of ALAL.

Conformation	Probability	Type
β, β	~ 0	$C_2=O_2 \cdots N_3=D_3$
β, α	~ 0	$C_2=O_2 \cdots N_3=D_3$
α, β	~ 0	$C_1=O_1 \cdots N_2=D_2$
α, α	0.4	$C_1=O_1 \cdots N_3=D_3$
	0.1	$C_1=O_1 \cdots N_2=D_2$
	~ 0	$C_2=O_2 \cdots N_3=D_3$
β, L_α	0.5	$C_1=O_1 \cdots N_3=D_3$
	~ 0	$C_2=O_2 \cdots N_3=D_3$
L_α, β	~ 0	$C_1=O_1 \cdots N_2=D_2$
L_α, α	~ 0	$C_1=O_1 \cdots N_2=D_2$
L_α, L_α	0.6	$C_1=O_1 \cdots N_3=D_3$
	~ 0	$C_1=O_1 \cdots N_2=D_2$

6.3.3.4 Water Bridges

We have determined the water bridges between each combination of the polar groups up to the fifth order, that is with up to five hydrogen-bonded water molecules. A first order water bridge is defined as a water molecule that forms two hydrogen bonds simultaneously with any pair of polar groups. Similarly, a higher order water bridge is defined as more than one hydrogen-bonded water molecules connecting two polar groups via hydrogen bonds.

The probabilities of water bridges orders w.r.t to the total number of bridges is shown in Figure 6.10. The probability increases with an increase in the bridge order for all conformations of ALAL, irrespective of the type of the bridge. The probability of first order water bridges is lowest for the pure β -sheet like conformation and other conformations with any combinations of β, α . On the other hand, pure α -helix, L_α -helix like conformations, and any combinations of L_α with either β or α have a higher probability to form first order water bridges. The probabilities of second order water bridges is comparable for conformations other than the (β, α) , (L_α, α) -conformations, those have relatively low probabilities. The third order water bridges are formed more frequently for the (β, α) , (β, β) , (β, L_α) and (L_α, α) conformations compared to others and least for the (L_α, L_α) -conformation. The fourth order water bridges are less probable and fifth order water bridges are more probable for the (β, β) , (β, α) , (α, β) -conformations compared to the probabilities of water bridges of that order in the other conformations. Each conformation tends to form unique lower order water bridges, such as, a first order water bridge between $C_2 = O_2$ and $N_1 = D_1$ is only formed in the case of (β, β) -conformation (see Figure 6.11). Similarly, the first order water bridges between the charged C-terminal group and $C_2 = O_2$ and $N_3 = D_3$ can be seen for (β, α) -conformation. For the (α, β) -conformation, the first order water bridges between second $N_2 = D_2$ and $N_1 = D_1$, $N_2 = D_2$ and $C_3 = O_3$ are formed frequently. The most frequent such bridges in the case of the (α, α) , (β, L_α) -conformations are between the charged C-terminal group and the second peptide bond of ALAL (see Figure 6.12). The water molecules around the (L_α, β) -conformation lean to form first order bridges between the $N_2 = D_2$ and the charged N-terminal group and couple of other unique, one water bridges. The (L_α, α) -conformation is special when it comes to such bridges, because a first order bridge between $N_3 = D_3$ and $N_2 = D_2$ is almost always present. The conformation that shows the highest population of water bridges is the (L_α, L_α) -conformation with bridges between $C_1 = O_1$, $N_2 = D_2$, $N_3 = D_3$ and termini, in addition to first water bridges between the N-terminal and C-terminal group (see Figure 6.12). Further, note the differences in probabilities of higher order water bridges among the different conformations of ALAL

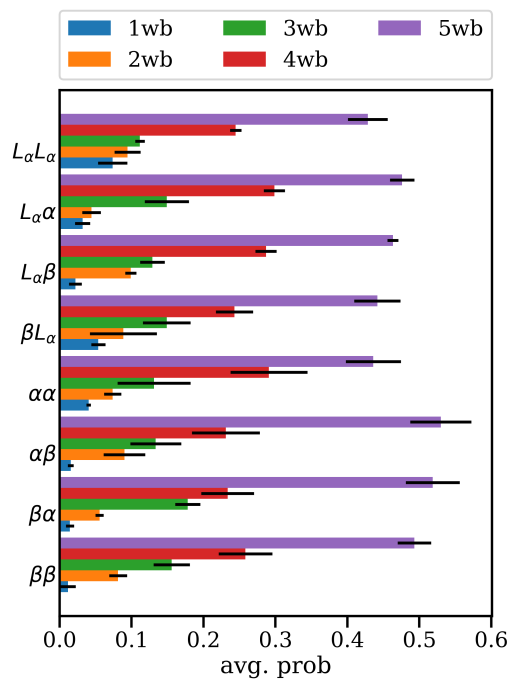


Figure 6.10: Probabilities of water bridges from the first-principle MD simulations of metastable-conformations of ALAL.



Figure 6.11: Water bridges extracted from the first-principle simulations of ALAL conformations of Set-V.

6.3 Results

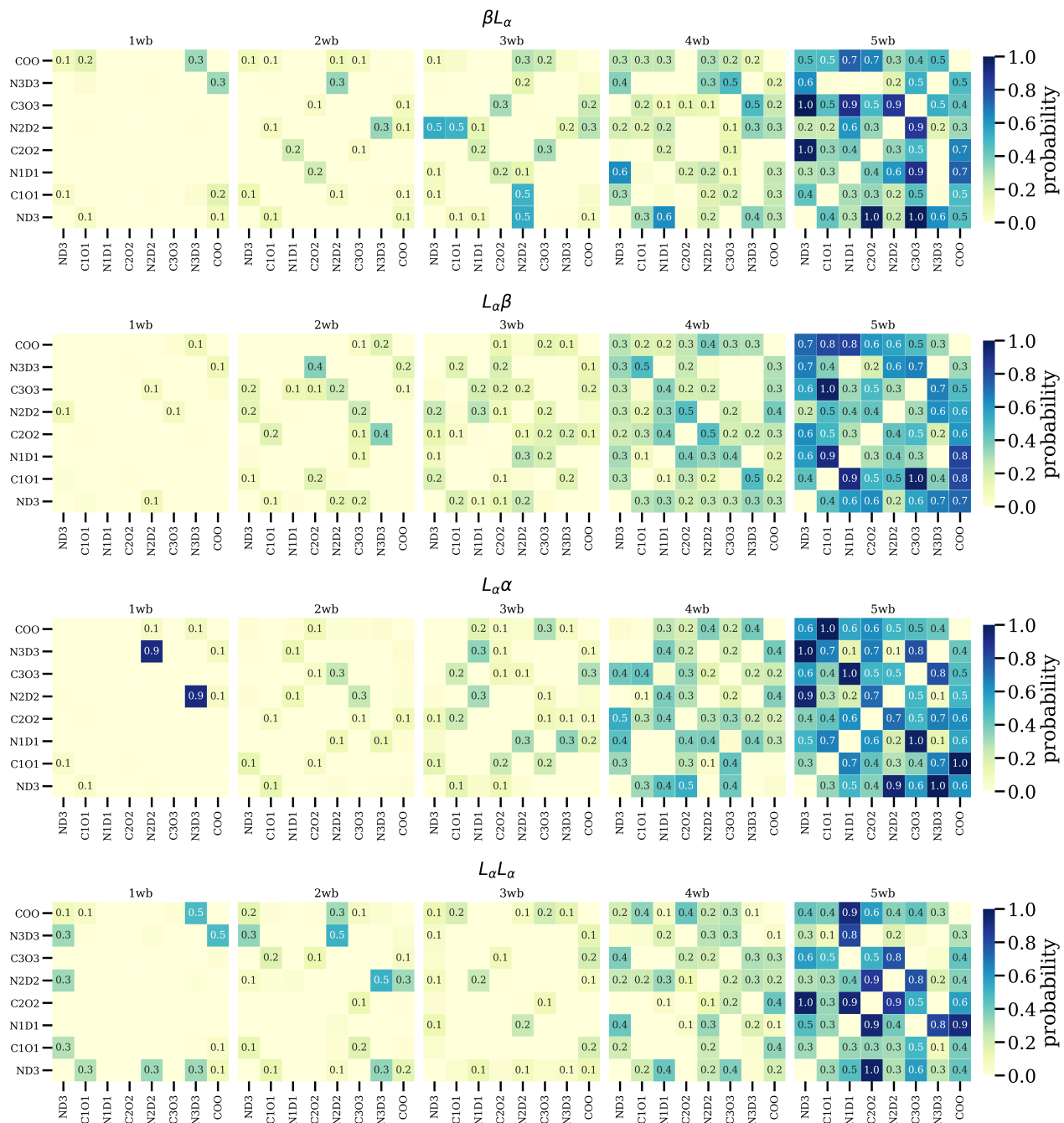


Figure 6.12: Water bridges extracted from the first-principle simulations of ALAL conformations of Set II-IV.

6.3.3.5 Hydrogen Bond Lifetimes

The hydrogen bond lifetimes are reported in Table 6.6. The lifetime of the hydrogen bond between the charged N-terminal and water molecules is half a picosecond for the open, (β, β) -conformation, possibly due to the frequent hydrogen bond exchanges and unstable water bridges around the terminal. The other extreme is the exceptionally long hydrogen bond lifetime of ND_3^+ of (L_α, α) -conformation, which is probably linked to the highly probable (present nearly 90 percent of the total simulation time) first order bridge between $\text{N}_3 = \text{D}_3$ and $\text{N}_2 = \text{D}_2$. For other conformations, the hydrogen bond lifetime of ND_3^+ fluctuates around one picosecond. The lifetime of hydrogen bond of $\text{C}_1 = \text{O}_1$ is on average around half a picosecond for all other conformations except (β, α) , for which is about a picosecond, once again, possibly owing to the impact of the presence of first water bridge between the adjacent $\text{N}_1 = \text{D}_1$ and $\text{N}_2 = \text{D}_2$ for this conformation. This argument also applies for the $\sim 3 \text{ ps}$ hydrogen bond lifetime of $\text{N}_1 - \text{D}_1$ for (α, β) -conformation. The hydrogen bond lifetime of $\text{C}_2 = \text{O}_2$ seems to be impacted by the second order water bridges and first order water bridges. The second order water bridges between the central carbonyl group of (β, β) -conformation and $\text{N}_1 = \text{D}_1$ and $\text{N}_3 = \text{D}_3$ groups are observed frequently, and such bridges with $\text{C}_1 = \text{O}_1$, however, less frequent. This characteristic water network surrounding $\text{C}_2 = \text{O}_2$ of (β, β) -conformation could be the reason for the calculated higher hydrogen bond lifetime. Likewise, the higher lifetime of L_α -helix like conformation is likely due to the stable surroundings i.e., the probability of second order water bridge between $\text{N}_2 = \text{D}_2$ and $\text{N}_3 = \text{D}_3$, $\text{N}_2 = \text{D}_2$ and COO^- is, 0.5 and 0.3, respectively. Identically, we account, the frequently present second order water bridges, between, $\text{N}_2 = \text{D}_2$ and ND_3^+ , $\text{N}_2 = \text{D}_2$ and $\text{N}_3 = \text{D}_3$, for higher hydrogen bond lifetime of $\text{N}_2 = \text{D}_2$ polar group of (β, α) -conformation. The low hydrogen bond lifetime of $\text{C}_3 = \text{O}_3$ of L_α -helix like conformation could be explained by the closeness of flexible Ala and Leu-sidechains, that disturb the water network more often, plus the influence of the COO^- group. Notably, the solvation of $\text{N}_3 = \text{D}_3$ is highly affected by the change in the conformation of ALAL, as can be seen by the nearly zero hydrogen bond lifetime for this group in the case of pure α -helix, L_α -helix like conformations. Lastly, the hydrogen bond lifetime of the C-terminal group also varies between the different conformations of ALAL.

Table 6.6: Hydrogen bond life times (in picoseconds) of polar groups of ALAL.

	ND ₃ ⁺	C ₁ = O ₁	N ₁ - D ₁	C ₂ = O ₂	N ₂ - D ₂	C ₃ = O ₃	N ₃ - D ₃	COO ⁻
β, β	0.50	0.54	0.56	1.43	0.88	0.85	1.05	0.96
β, α	0.89	0.51	1.02	0.28	2.98	0.75	0.09	1.77
α, β	1.09	1.02	3.02	0.55	0.18	1.26	1.69	1.00
α, α	0.93	0.72	1.60	0.76	0.10	0.96	0.01	1.09
β, L_α	1.03	0.53	0.87	1.06	1.67	0.55	0.23	1.80
L_α, β	0.75	0.68	0.89	0.62	1.21	0.93	1.58	1.43
L_α, α	1.24	0.34	0.71	0.52	0.19	1.23	0.86	1.08
L_α, L_α	0.76	0.77	1.90	1.92	0.23	0.30	0.01	1.45

6.3.3.6 Power Spectra

Appendix C, Figure C.29 shows the power spectra of ND₃⁺, C=O, N-D, and COO⁻ groups computed from first-principle MD simulations of the ALAL conformations in deuterated water. The peak values of average power spectra of each group are reported in Table 6.7. The peak of the ND₃⁺ frequency band of the (β, β)-conformation is at $\sim 1160\text{cm}^{-1}$ with a shoulder around $\sim 1190\text{cm}^{-1}$. This band is blue-shifted compared to the (β, β)-conformation for all other conformations of ALAL, more so for the (α, α)-conformation and the (L_α, α)-conformation with the peak values at $\sim 1181\text{cm}^{-1}$ and $\sim 1185\text{cm}^{-1}$, respectively. This might be because the ND₃⁺ polar group, in the case of the (β, β)-conformation, is well exposed to water compared to the other conformations. The (α, α)-conformation and other conformations with α -helix like content, i.e., (β, α), (α, β) and the (L_α, α), give rise to a third shoulder in the range $\sim 1100\text{-}1150\text{cm}^{-1}$. Furthermore, the ND₃⁺ band for the (L_α, L_α)-conformation and the (L_α, β)-conformation does not contain any shoulders, rather a first wide and narrow frequency band, respectively.

The Amide-I frequency range of power spectra of ALAL conformations is relatively diverse, showing that computed spectra of carbonyl groups in explicit solvent vary with the conformation. Mostly, the high intensity peaks are centred around $\sim 1600\text{cm}^{-1}$. Each carbonyl frequency band results from the superposition of the frequency peaks corresponding to three individual carbonyl groups. Their interactions with the surrounding water can be traced to the differences in their frequency. Previously, we reported for the (β, β)-conformation that the probability of the middle carbonyl group forming hydrogen bonds with water is consistent with the observed changes in its calculated stretching frequencies [HFI21]. Therefore, we further need to look at the frequencies of individual carbonyl groups of each conformation because of the varying effect of hydration shell dynamics and charged termini on each. The power spectra of individual carbonyl group of ALAL is shown in Figure 6.13.

The peak value of first carbonyl group for the (β, α) conformation is around $\sim 1598\text{cm}^{-1}$,

which is considerably redshifted compared to the peak values of the $C_1 = O_1$ of all other conformations. The peak values of the central carbonyl group vary significantly with the conformation. In the case of the (β, β) , (β, L_α) , and (L_α, L_α) conformations it vibrates with relatively higher frequency compared to the peak frequency values for other conformations. The peak values for the third carbonyl group are generally lower for all conformations suggesting a reduced bond strength compared to first and second carbonyl bonds.

The other interesting thing to look at is the fluctuations in the frequencies of individual carbonyl groups, i.e., error bands as shaded regions. For each conformation, at least one carbonyl group with higher fluctuations in its frequency indicates frequent changes in the hydrogen-bonded situations. And another carbonyl group with a sharp power spectra peak with minor fluctuations indicates a stable surrounding water network or the presence of intramolecular hydrogen bond with another polar group. The sharper peaks can be seen, for example, for $C_1 = O_1$, $C_2 = O_2$ and $C_3 = O_3$ in case of the (α, α) , (β, β) , and (L_α, L_α) conformations, respectively.

Looking at the Amide-II frequency range, each conformation's N-D frequency bands mainly show two peaks around $\sim 1380\text{cm}^{-1}$ and $\sim 1410\text{cm}^{-1}$ (see Figure S30). However, the band of the (α, α) -conformation is redshifted compared to that of other, as the peak values of all N-D bonds are at lower wavenumber. Lastly, the COO^- frequencies of all conformations are clustered around $\sim 1530\text{cm}^{-1}$ except for the (α, α) -conformation, whose spectrum is blueshifted compared to the spectra of all other conformations. This result is in line with normal modes calculations where the peak frequency of COO^- terminal group of the (α, α) -conformation was off compared to others; however, note that in normal mode calculations, its frequency is lower compared to the COO^- frequency of other conformations (see Table 6.2). Whereas in the power spectra it turns out to be blueshifted w.r.t to the power spectra of C-terminal of remaining conformations.

Table 6.7: Peak values of power spectra of polar groups of each conformation of ALAL.

Conformations	ND_3^+ (cm^{-1})	$C_1 = O_1$ (cm^{-1})	$C_2 = O_2$ (cm^{-1})	$C_3 = O_3$ (cm^{-1})	$\text{N}_1\text{-D}_1$ (cm^{-1})	$\text{N}_2\text{-D}_2$ (cm^{-1})	$\text{N}_3\text{-D}_3$ (cm^{-1})	COO^- (cm^{-1})
β, β	1160	1629	1582	1588	1384	1391	1384	1525
β, α	1171	1598	1610	1572	1415	1344	1399	1531
α, β	1173	1625	1572	1586	1386	1393	1382	1533
α, α	1181	1606	1590	1572	1352	1370	1404	1539
β, L_α	1167	1608	1610	1570	1384	1409	1380	1525
L_α, β	1171	1602	1570	1580	1413	1368	1382	1527
L_α, α	1185	1617	1586	1568	1386	1376	1378	1529
L_α, L_α	1169	1608	1602	1576	1417	1397	1391	1527

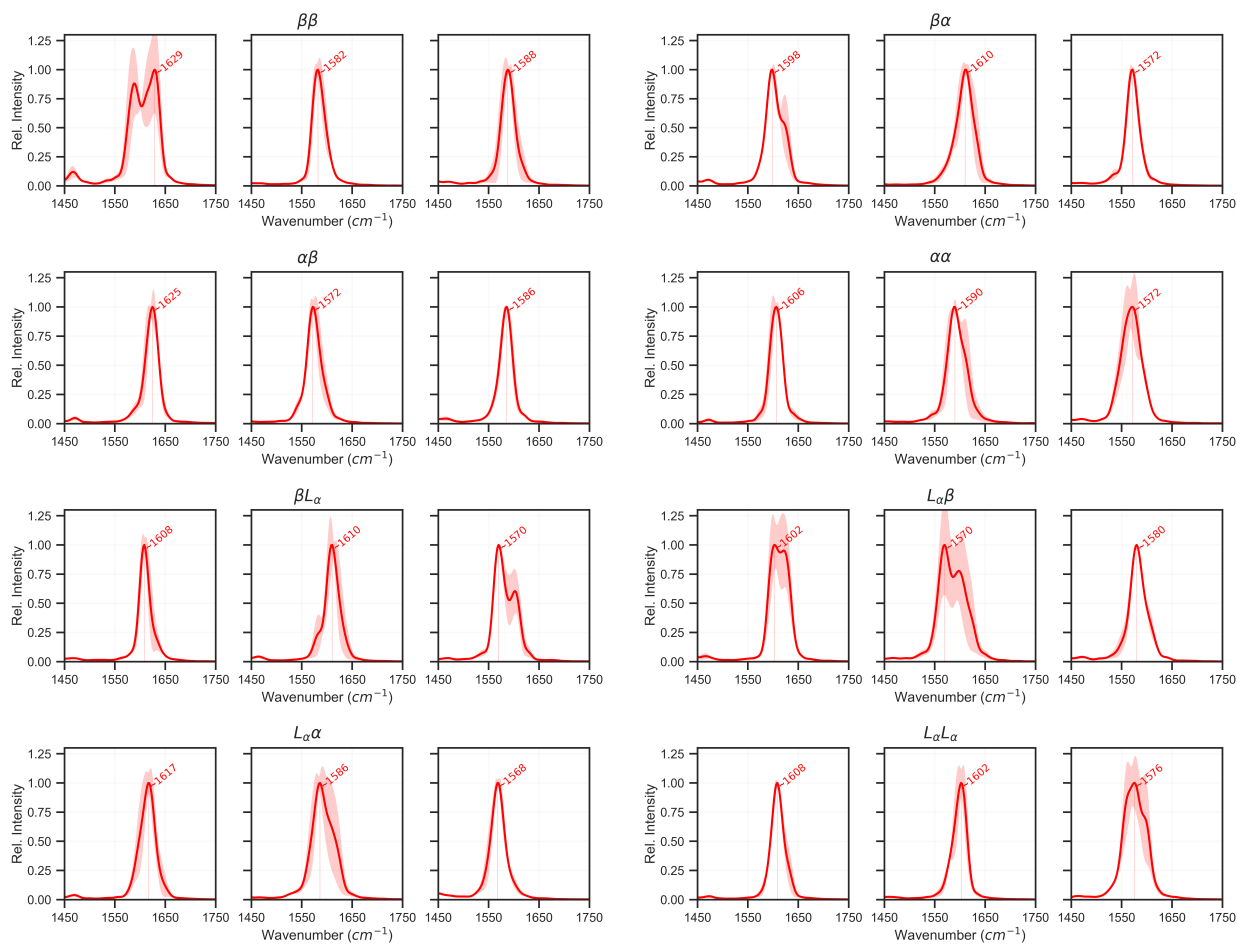


Figure 6.13: Power spectra of carbonyl groups of all conformations of ALAL.

6.3.3.7 Infrared Spectra

The infrared spectra computed from first-principle MD simulations of the ALAL conformations in deuterated water using Voronoi tessellation are shown in Figure 6.14. At first glance, we can see that the Amide-I ($\nu\text{C}=\text{O}$, νCOO^-) spectra of each metastable-conformation differ from one another, reflecting their underlying structural characteristics and specific interactions with water molecules. For the (β, β) -conformation, in the Amide-I region, the most intense peak is located at $\sim 1586\text{cm}^{-1}$ which can be assigned the $\nu\text{C}=\text{O}$ vibrations of the carbonyl groups (also note a bump at $\sim 1632\text{cm}^{-1}$) and the peak at $\sim 1525\text{cm}^{-1}$ originates from the νCOO^- vibrations of charged C-terminal group. The Amide-II region in the range $\sim 1200\text{--}1440\text{cm}^{-1}$ mainly contains the vibrational signals of $\delta\text{N-D}$, $\nu\text{N-C}$, and some contributions from the νCOO^- vibrations. The Amide-III region is the representative of charged $\nu\delta\text{ND}_3^+$ vibrations of N-terminal group. Moreover, the peak at $\sim 1470\text{cm}^{-1}$ is coming from the motion of the sidechains. In the case of the (β, α) -conformation, the carbonyl frequencies band is comparatively wider with

a peak at $\sim 1602\text{cm}^{-1}$ along with an adjacent comparable intensity peak ;also the peak associated with COO^- is very intense. Similarly for the (β, α) -conformation, more intense Amide-II/III regions can be seen compared to (β, β) -conformation. The carbonyl peaks profile of the (α, β) -conformation is almost the mirror image of carbonyl peaks profile of the (β, α) -conformation, however, the two peaks are well-separated. The (α, α) -conformation puts out a first slightly-skewed band in the Amide-I region which seems to be the combination of $\nu\text{C}=\text{O}$ and νCOO^- vibrations. This suggests the possible presence of an intramolecular hydrogen bond between COO^- and $\text{C}=\text{O}$ for the (α, α) -conformation. In the Amide-II region, we can see three distinct peaks of the (α, α) , which is different from the (β, β) , (β, α) , and (α, β) conformations. For the (β, L_α) -conformation, there is a sharp peak corresponding to carbonyl vibrations at $\sim 1606\text{cm}^{-1}$ with a neighbouring bump at $\sim 1560\text{cm}^{-1}$ and an evident peak assigned to COO^- . Next, the situation of Amide-I region of the (L_α, β) -conformation is similar to the (β, α) -conformation to some extent and three pretty distinguished peaks in the Amide-II region. Interestingly, unlike other conformations of ALAL, the IR spectrum of the (L_α, α) -conformation shows two (three including the shoulder) prominent peaks for carbonyl groups. Finally, the (L_α, L_α) -conformation has a clear peak at $\sim 1602\text{cm}^{-1}$ for $\nu\text{C}=\text{O}$ vibrations. However, somewhat similar to the (α, α) -conformation, there is a clear overlap between the $\nu\text{C}=\text{O}$ and νCOO^- frequencies bands. Furthermore, the Amide-II fingerprints of the (L_α, L_α) -conformation are quite distinctive compared to other conformations.

6.3 Results

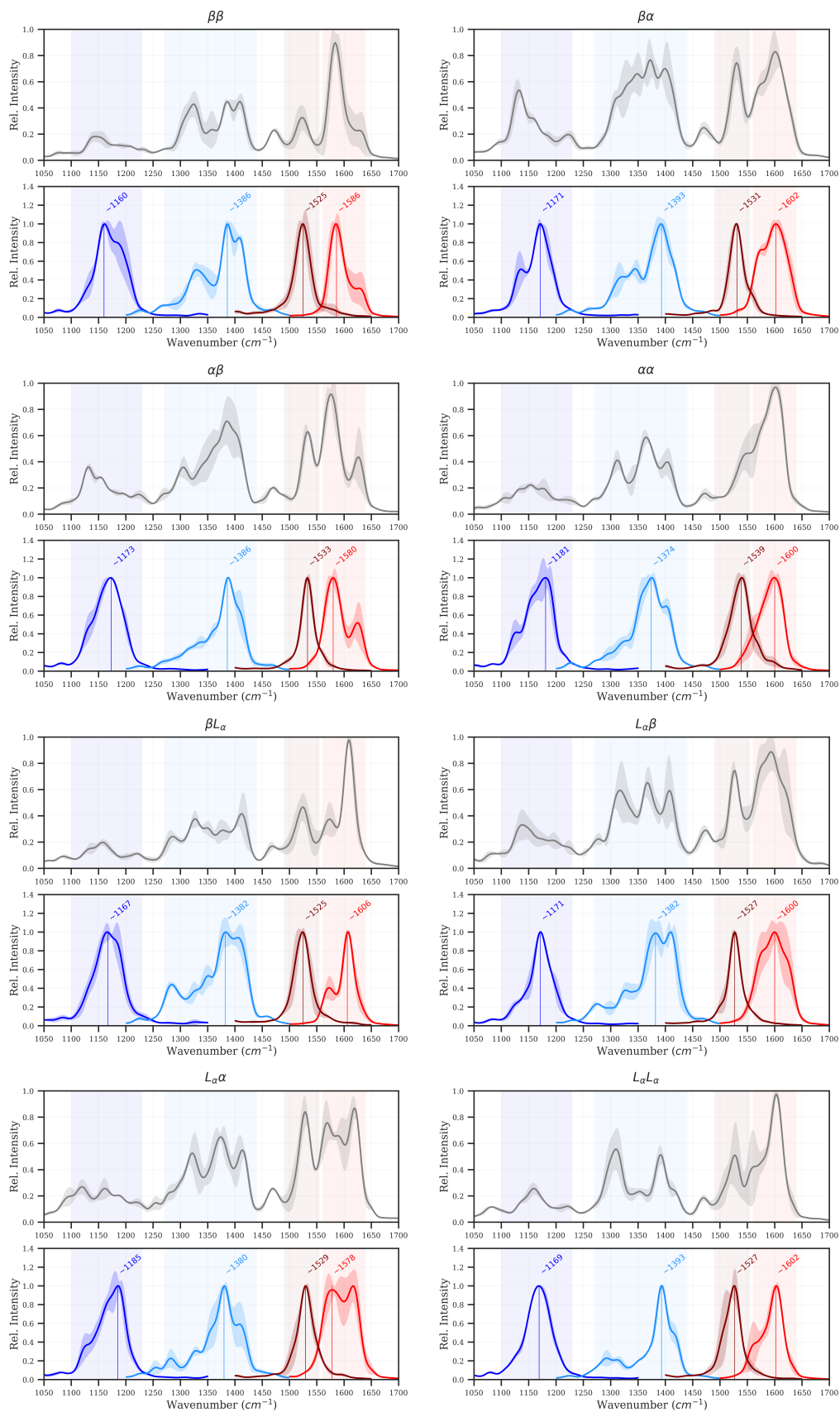


Figure 6.14: Density based infrared spectra of all conformations (each upper panel) with power spectra of polar groups (each below panel).

Further, zooming into the Amide-I region of IR spectra of all conformations of ALAL, and mainly focusing on the $\nu\text{C}=\text{O}$ vibrational bands (see Figure 6.15), we can see that for the (β, β) -conformation, the small bump at $\sim 1632\text{cm}^{-1}$ is originating from the fluctuations in the dipole moments of $\text{C}_1 = \text{O}_1$. The dual peaks in the power spectra of $\text{C}_1 = \text{O}_1$ suggest sampling two completely different hydrogen bonding states, or maybe the complete absence of hydrogen bond for some time-interval, during the simulation runs. For the (α, α) -conformation, the peaks of the power spectra of individual carbonyl groups are well separated i.e., $\text{C}_1 = \text{O}_1$ and $\text{C}_2 = \text{O}_2$ by $\sim 16\text{cm}^{-1}$, and, $\text{C}_2 = \text{O}_2$ and $\text{C}_3 = \text{O}_3$ by $\sim 18\text{cm}^{-1}$, resulting in a broadband in the composed spectrum. Finally, for the (L_α, L_α) -conformation, the peak vibrational frequencies of $\text{C}_1 = \text{O}_1$ and $\text{C}_2 = \text{O}_2$ are close to each other, at $\sim 1608\text{cm}^{-1}$ and $\sim 1602\text{cm}^{-1}$, respectively, whereas $\text{C}_3 = \text{O}_3$ yields a wider band around $\sim 1576\text{cm}^{-1}$. Note that for (α, α) and (L_α, L_α) , the νCOO^- peaks are not easily distinguishable.

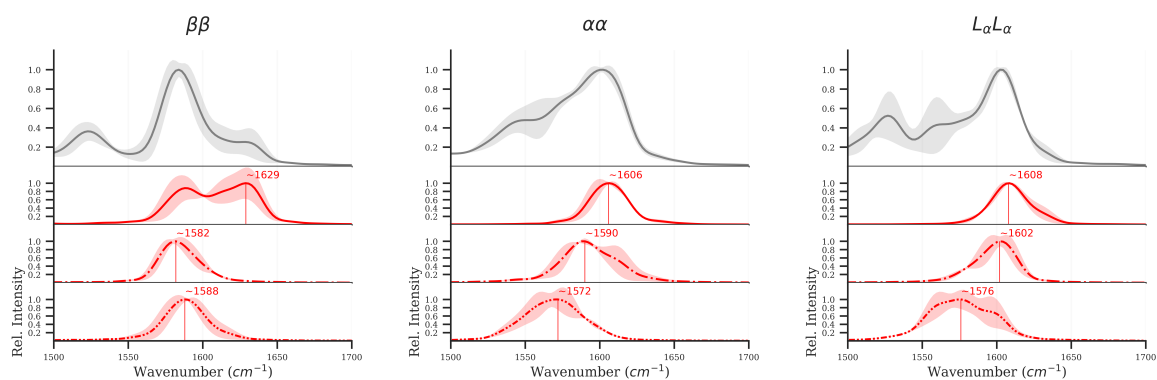


Figure 6.15: Density based infrared spectra of ALAL in Amide-I range (each upper panel) with average power spectra of each carbonyl group (each below panels).

In order to validate the IR spectra calculated using Voronoi tessellation, and for the comparison between the two methods for estimation of molecular dipole moments, we also calculated IR spectra using the commonly used, however computationally more expensive, Wannier localization scheme. In Appendix C, Figure C.35 IR spectra calculated using both methods are presented. The density-based IR spectra for all conformations of ALAL match remarkably with the Wannier centers based IR spectra. The Amide-III region, whose peaks are assigned to the charged $\nu\delta\text{ND}_3^+$ vibrations of the N-terminal group are marginally better resolved in the case of density-based IR spectra. Additionally, the implementation of on-the-fly calculation of dipole moments via Voronoi integration of the total electron density in the CP2K8.1 release makes it more suitable for the calculation of moderately sized model peptides in explicit solvent. This removes the need to save the massive amount of electron density data for later processing. Wannier localization requires an iterative procedure, which is very time-consuming,

particularly for large systems, and is not guaranteed to converge. On the other hand, Voronoi integration is quick and does not require iterations, also no problems with convergence. Therefore, a significant amount of computing time can be saved by switching to Voronoi integration.

6.4 Discussion

Our analyses show the influence of each thermally accessible, structurally heterogeneous, metastable-conformation of the ALAL peptide on the intramolecular hydrogen-bonding interactions and the interaction of the polar groups (ND_3^+ , $\text{C}=\text{O}$, $\text{N}-\text{D}$ and COO^-) with the surrounding water molecules is different. Besides the structural specific, intrinsic difference in the vibrational frequencies of the polar groups of independent conformations, the dynamic changes in their vibrational state are directly related to the topology and dynamics of the hydrating water [HDS⁺18]. Previously, we showed that the interactions between water molecules and the $\text{C}_2=\text{O}_2$ group clearly have an effect on the stretching frequency of this carbonyl group. Thereby, due to the varying peptide-water interactions between the metastable-conformations, the effect of these interactions on the calculated IR spectrum is different. Particularly, the Amide-I region of the IR spectra is sensitive to the underlying structure responsible (and its exposure to water) for the observed spectroscopic signals and makes it possible to distinguish between the spectra of metastable-conformations (see Figure 6.14).

From the distance analysis, clearly the compactness of the conformations varies i.e., the optimised representative structures of metastable conformations differ from each other in terms of the position of the polar groups relative to each other, separation between the Ala-sidechains and Leu-sidechains, radius of gyration, end-to-end distance, etc (see Table 6.1). For example, the end-to-end distance of the β -sheet like conformation (12.7Å, “open conformation”) is largest and shortest for the L_α -helix like conformation (3.7Å, “closed” conformation). These structural characteristics are typically maintained during the first-principles MD simulations. The different structures of the simulated conformations have a different effect on the surrounding water molecules, resulting in a different amount of water being exposed to each metastable-conformation. For example, the integrated number of water molecules for the central carbonyl group ($\text{C}_2 = \text{O}_2$) of β -sheet like conformation is 1.80 ± 0.07 , whereas, for the L_α -helix like conformation is 1.15 ± 0.12 (see Table 6.4).

Previously, we reported the solvation probability of carbonyl groups of β -sheet like conformation of ALAL decreases from $\text{C}_1 = \text{O}_1$, over $\text{C}_2 = \text{O}_2$, to $\text{C}_3 = \text{O}_3$ [HFI21]. This is inline with the RMSF observation of backbone of β -sheet like conformation ALAL that flexibility (positional fluctuations) of its three peptide bonds ($-\text{CO}_{1,2,3}-\text{ND}_{1,2,3}-$), also decreases from $\text{C}_1 = \text{O}_1$, over $\text{C}_2 = \text{O}_2$, to $\text{C}_2 =$

O₃ (see Appendix C, Figure C.15). However, the RMSF is different for different conformations, and the order of flexibility, as manifested by the RMSF of the C=O groups, varies with conformations. It has been reported that the equilibrium fluctuations obtained for the helix conformation α_R and the two extended conformations β and P_{II} of trialanine differ mainly in the vicinity of the central peptide group [MS02]. On the other hand the amount of solvation of the three C=O groups remains in the same order for the different conformations.

From the RDF analysis, it is evident that the coordination number of the carbonyl groups (shown as the horizontal black line in Figure 6.7 and given in Table 6.4), varies from conformation to conformation. Note, for example, the coordination number of C₁ = O₁ for α -helix, L_α -helix like conformations and (β, L_α)-conformation is 0.85 ± 0.07 , 0.83 ± 0.17 , and 0.79 ± 0.27 , respectively, is lower compared to all other conformations with values above one. For C₂ = O₂, it is fairly lower for (L_α, L_α), (β, α) and (β, L_α)-conformations and for C₃ = O₃, the level of solvation is almost the same for each conformation. Moreover, the significantly reduced exposure to water of C₂ = O₂ in case of pure α -helix, L_α -helix like conformations is evident from the CDF analysis. We can relate the low number of water molecules in the vicinity of the central carbonyl group of these conformations to their compact structure, resulting in a frequently present intramolecular hydrogen bond/s that perhaps cause energetically favorable shielding from the attack of water molecules.

In a manner similar to the distribution function analysis, the hydrogen bonding analysis of polar groups with water reveals similar results for all carbonyl groups. The solvation probabilities of the second carbonyl group of (L_α, L_α), (β, α) and (β, L_α)-conformations are low compared to other conformations i.e., 1.14 ± 0.14 , 1.26 ± 0.29 , 1.3 ± 0.29 , respectively. The decreased solvation in these conformations is linked to the presence of highly probable intramolecular hydrogen bonds, C₁=O₁ ··· N₃=D₃, C₁=O₁ ··· N₂=D₂ (see Table 6.5) arising due to the compact nature of these conformations. Another reason for this is the presence of highly stable, unique to the conformation, first order water bridges between the polar groups, such as, the first order water bridges between the charged COO⁻ and C₂ = O₂ and N₃= D₃ can be seen for (β, α)-conformation. This also aids in explaining the lower coordination number of C₂ = O₂ of (β, α). The central carbonyl (C₂ = O₂), is the best representative of a longer peptide or protein because it shows least impact of the termini. For trialanine, it is reported that the probability and lifetime of hydrogen bonds involving the site O8 i.e., the first carbonyl group, are significantly reduced because of the proximity of the charged NH₃⁺ terminal group [MS02].

By looking at the water network surrounding each of the hydrated conformations, relatively stable, unique second and third order water bridges are observed for each conformation. The idea of a conformation dependent topology of the surrounding water is conceivable from the water bridges analysis. Using data science algorithms, it has also been shown for trialanine that different conform-

ations of the solute leaves a strong fingerprint on the surrounding water network [JH18]. Moreover, the hydrogen bond life of individual polar groups varies among the conformation and is linked to the factors described earlier.

From the analysis of the vibrational spectra, the peak values of power spectra values of $C_2 = O_2$ vary significantly with the conformation. The peak frequency values of $C_2 = O_2$ for (L_α, L_α) , (β, α) and (β, L_α) -conformations are $\sim 1602\text{cm}^{-1}$, $\sim 1610\text{cm}^{-1}$ and $\sim 1610\text{cm}^{-1}$, respectively. This is in agreement with solvation probabilities of $C_2 = O_2$ for these conformations. Linear correlation coefficients between peak frequencies and mean average hydrogen bond number for $C_1 = O_1$, $C_2 = O_2$ and $C_3 = O_3$ are 0.10, -0.81, 0.36, respectively. It is easily conceivable how the underlying conformation and conformation-dependent diverse hydration are manifested in the amide-I region of IR spectra.

Therefore it can be argued that the water topology/dynamics affect the vibrational strength of the individual polar bond, impacting the calculated composed IR spectra (see Appendix C for the individual power spectra of carbonyl groups of other conformations).

Everything is connected. From the intrinsic structural differences between the metastable conformations (“open”, “close” or “intermediate”) of ALAL, which result in the presence or absence of intramolecular hydrogen bonds, to their effect on water exposures. This leads to different solvation levels of individual polar groups of each conformation, the formation of unique water bridges, and eventually, the different overall topology of the hydrating water molecules. Consequently, the interaction of the polar groups with water molecules vary, resulting in the differences of the vibrational signatures of different metastable-conformations.

6.5 Conclusions

We showed from the analysis of the interaction energies of individual water molecules with the central carbonyl group of ALAL that shifts in its frequency are directly related to the interactions with the water molecules in the first hydration shell [HFI21]. Thereby, with the help of computed vibrational spectra, we can relate the conformational based diverse topology/dynamics of the hydrated water molecules around the polar groups of metastable-conformations to the frequencies of the three individual carbonyl groups for each conformation. There is a strong correlation between the solvation probabilities and peak frequencies of the central carbonyl group $C_2 = O_2$, another confirmation of its being the best representative of a carbonyl group in a longer peptide or protein, least influenced by the termini and can be used as a marker to differentiate between the IR spectra of metastable-conformations of ALAL. It can be established that the water topology/dynamics affect the vibrational strength of the individual polar bond,

impacting the calculated composed IR spectra. Finally, using distinguishable calculated spectra of estimated metastable-conformations, it is possible to decipher the measured IR spectrum into spectra of its constituent metastable-conformers using this combined approach.

Chapter 7

Effect of Peptide Length

7.1 Introduction

Proteins are relatively large, compact, structurally complex molecules. Their structure-function relationships account for conformational variation. High structural and temporal resolution are required due to the ensemble nature and coupled dynamics nature of protein-water interactions and conformational fluctuations. Furthermore, it is unclear whether protein dynamics timescales are proportional to protein size.

On the other hand, computationally, peptides are a more tractable systems than proteins and exhibit a number of the characteristics and complexities associated with them. It is possible to simulate peptides in explicit solvent for the range of time scales at atomic resolution. Moreover, as discussed in Chapter , a combined approach enables the estimation of metastable conformations and their associated timescales, and the DFT-MD based computational IR spectra contain information about the structure of peptides due to the high sensitivity of Amide vibrational frequencies, particularly the Amide-I mode (C=O stretching), to local atomic organisation (e.g. hydrogen bonds, solvation effects, etc).

In this Chapter, we investigated the effect of peptide length on slow timescales and Amide I spectra using Alanine-Leucine peptides of different lengths, i.e., Ala-Leu (AL), Ala-Leu-Ala (ALA) , Ala-Leu-Ala-Leu (ALAL), Ala-Leu-Ala-Leu-Ala (ALALA), Ala-Leu-Ala-Leu-Ala-Leu (ALALAL), in an explicit solvent using a combination of classical and first-principles MD simulations, as well as MSM's.

7.2 Methods

7.2.1 Classical Molecular Dynamics simulations

We performed long classical MD simulations of the AL, ALA, ALAL, ALALA, ALALAL peptide in a cubic simulation box of explicit water (modeled as TIP3P [JCM⁺83] water) employing the AMBER 99SB-ILDN [HAO⁺06, LLPP⁺10] force field and the gromacs programme [PPS⁺13]. The positions of the solute atoms were saved to file every 0.25 ps. Each system is simulated using a similar setup. For complete details on the simulation setup, see Chapter 5.

Table 7.1: Classical MD simulations details of each system.

System	Cell vectors A,B,C (Å)	# of water molecules ¹	Total # of atoms	Total simulation time
AL	2.95, 2.95, 2.95	844	2564	$3 \times 0.4 \mu s = 1.2 \mu s$
ALA	3.08, 3.08, 3.08	954	2904	$3 \times 3.0 \mu s = 9 \mu s$
ALAL	3.60, 3.60, 3.60	1477	4492	$6 \times 2.5 \mu s = 15 \mu s$
ALALA	4.15, 3.67, 3.13	1059	3248	$10 \times 6.0 \mu s = 60 \mu s$
ALALAL	4.67, 3.19, 3.13	1298	3984	$10 \times 5.0 \mu s = 50 \mu s$

For each system, we used a minimum distance of 1 nm between the solute and the box's periodic boundaries. Water hydrogen atoms and polar hydrogen atoms of the peptide (ND_3 , $N - D$) were modeled with deuterium mass. For Lennard-Jones interactions and electrostatic interactions (Particle-Mesh Ewald [DYP93b, EPB⁺95b] with a grid spacing of 0.16 an interpolation order of 4), we used a cutoff value of 1 nm. A V-rescale [BDP07b] thermostat was applied to control the temperature at 300 K (NVT ensemble). The positions of the solute atoms were saved to file every 0.25 ps. No constraints were applied, and the leap-frog integrator with a time step of 1 fs was employed. Each system was minimized and equilibrated for 500 ps followed by multiple long MD simulations (see Table 7.1 for details of each system).

¹ALALA and ALALAL simulation setup were prepared using the Amber tools, keeping in mind that if required to perform QM/MM MD simulations with CP2K, we can use the same inputs, i.e., forcefield parameters, as they are compatible with CP2K, for consistency. The simulation box is not cubic in order to reduce the number of water molecules. However, a minimum distance of 1 nm between the solute and the periodic boundaries of the box is the same as for the other three peptides.

Moreover, for the representative conformation of the most probable conformation of each system estimated by a MSM, we performed three 1 ns constrained MD simulations for each system using a similar setup. The positions of the solute plus solvent atoms were saved to file every 0.5 fs. The root mean square deviation (RMSD), end-to-end distances, solvent accessible surface area (SASA), radial distribution functions (RDF) are calculated using gromacs utilities.

7.2.2 Markov State Modeling

By following the steps discussed in the previous chapters, namely the selection of essential internal coordinates (i.e., torsion angles), dimensionality reduction (TICA), clustering (k-means), MSM construction, interpretation (PCCA+), and validation (chapman-kolmogorov test), we constructed the MSM of ALA, ALALA and ALALAL. The details of the construction of the MSM and the transition network plots of AL and ALAL are given in the Chapters 4 and 6, respectively. For all the steps involved in the construction and validation of the MSM, we used PyEMMA [STSP⁺15b]. PyEMMA 2.4 was used for AL, ALA and ALAL and PyEMMA 2.5.7 was used for ALALA and ALALAL.

7.2.3 First-Principles Molecular Dynamics Simulations

The details of the first-principles MD simulations setup used for the simulations of β -sheet like conformation of ALAL are given in [HFI21]. The same setup is being used for all other peptides. The default Gaussian and plane waves (GPW [LHP99]) electronic structure method, as implemented in the Quickstep module [VKM⁺05] of the CP2K package [HISV14, KIDB⁺20b], was used. The Geodecker–Teter–Hutter (GTH) norm-conserving pseudopotentials, double zeta valence plus (DZVP) basis set and BLYP with Grimme’s D3 dispersion correction exchange–correlation functional were employed. To keep the computational cost of the first-principles simulations moderate, the box size and the number of water molecules are smaller than in the classical simulations but still large enough to avoid interactions of the periodic images. The cubic simulation box had a minimum distance of 0.4 nm between the solute and the box’s boundaries, and periodic boundary conditions were applied in all three dimensions. See Table 7.2 for details.

First, the system was energy minimized using a conjugate-gradient algorithm where the positions of the solute atoms were fixed. This allows the solvent molecules to relax around the peptide and find energy favorable positions. We performed a 5 ps NVT equilibration run from the minimized system, during which the solute was kept fixed to avoid the transition to an undesired conformation (100 ps without any constraints for ALALA and ALALAL), followed by the production run

of 50 ps in an NVE ensemble. The time-step for the numerical integration was 0.5 fs, and atom positions were saved every step.

Power spectra are calculated using the TRAVIS program [BK11, BTGK20], and in-house, Python scripts.

Table 7.2: First-principles MD simulations details of each system.

System	Cell vectors A,B,C (Å)	# of water molecules	Total # of atoms
AL	17.90, 17.90, 17.90	160	512
ALA	19.60, 19.60, 19.60	199	639
ALAL	22.20, 22.20, 22.20	325	1045
ALALA	27.18, 27.53, 27.81	357	1142
ALALAL	32.26, 26.30, 27.80	406	1308

7.3 Results

7.3.1 Markov State Modeling

7.3.1.1 Ala-Leu-Ala

A Markov state model has been constructed using the trajectory (9 μ s combined) of partially deuterated ALA in deuterated water obtained from classical MD simulations on the conformational space spanned by the torsion angles χ_{Leu} , $[\phi_{Leu}, \psi_{Leu}]$ and $[\phi_{Ala2}, \psi_{Ala2}]$ (highlighted in Figure 7.1). The implied-timescales (ITS) indicate that the slowest process converges at approximately 75 ps for ALA (Figure 7.3), so the MSM was built with a 75ps lag time. It can be seen that the timescale of the slowest process is very well separated from the timescales of remaining slow processes. Moreover, the gap between the timescales of remaining slow processes is also very small. It is significant to note that, the two implied-timescales curves around ~ 1 ns disappear if we exclude the χ_{Leu} from our dihedral angles subspace. The very low population of χ_{Leu} torsion angle around 1 rad is the source of these timescales as can be seen from the probability distribution of χ_{Leu} in Fig 7.2 (left). This is also reflected in the eigenvalue spectrum (Figure 7.3) by the appearance of two noticeable spectral gaps i.e., between 2nd and 3rd

eigenvalue and 6th and 7th eigenvalue. To investigate all of the dynamic processes of ALA, we used the robust Perron Cluster Analysis (PCCA+) to coarse-grained microstates into six metastable sets. [DW05, RW13].

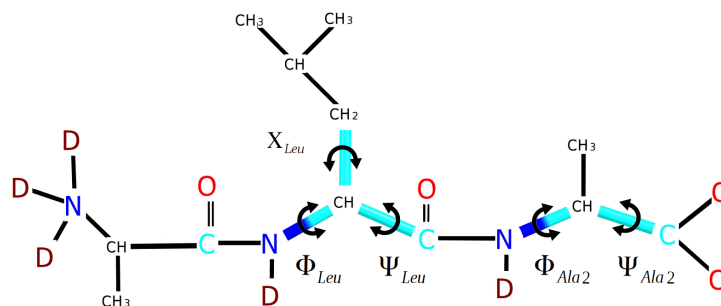


Figure 7.1: (a) Scheme of ALA and torsion angles χ_{Leu} , $[\phi_{Leu}, \psi_{Leu}]$ and $[\phi_{Ala2}, \psi_{Ala2}]$. Note that ψ_{Ala2} of ALA is a pseudo torsion angle and rotation around it result in a chemically equivalent state. Carbon atoms are shown in cyan, nitrogen in blue and oxygen in red. "D" denotes a deuterium atom. The bonds between backbone, sidechain carbon atoms and hydrogen atoms and are not shown.

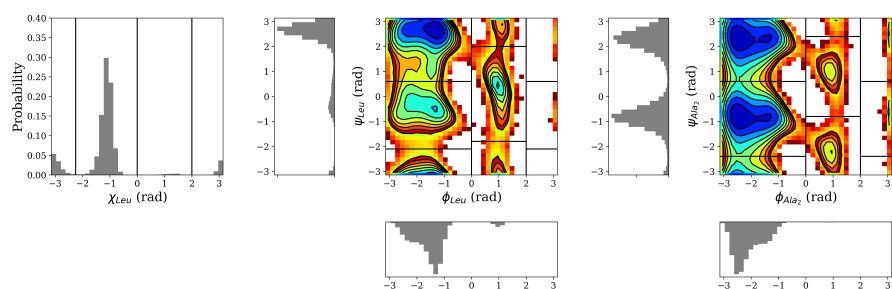


Figure 7.2: One dimensional distribution of χ_{Leu} and 2d-distribution of $[\phi_{Leu}, \psi_{Leu}]$ and $[\phi_{Ala2}, \psi_{Ala2}]$ obtained from the classical MD simulation of partially deuterated ALA in deuterated water.

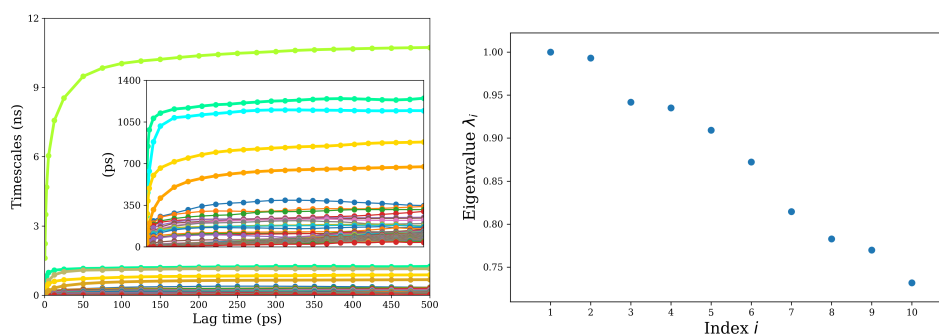


Figure 7.3: (left) Implied time scales, (right) Eigenvalues of the transition matrix sampled with lag time $\tau = 75ps$.

The metastable-sets of ALA identified from the MSM and PCCA+ can be subdivided based on the populated region on Ramachandran plot (i.e., α – *helix*, β – *sheet* and left handed helix (L_α) region) of $[\phi_{Leu}, \psi_{Leu}]$, $[\phi_{Ala2}, \psi_{Ala2}]$ torsion angles of sampled metastable conformations. Table. 7.3 provides the information about conformations based on Ramachandran space of $[\phi_{Leu}-\psi_{Leu}]$ and $[\phi_{Ala2}-\psi_{Ala2}]$ for each of the six metastable-sets. The validation of the MSM is performed using a Chapman-Kolmogrov test (see Figure 7.4).

Table 7.3: Metastable-sets and their corresponding conformations based on Ramachandran space of $[\phi_{Leu}-\psi_{Leu}]$ and $[\phi_{Ala2}-\psi_{Ala2}]$. Note the χ'_{Leu} is used to represent the very low population around 1 *rad*.

Set	Conformation		
	χ_{Leu}/χ'_{Leu}	$[\phi_{Leu}-\psi_{Leu}]$	$[\phi_{Ala2}-\psi_{Ala2}]$
I	χ'_{Leu}	β	Left-handed helix
II	χ'_{Leu}	α	α
	χ'_{Leu}	α	β
III	χ'_{Leu}	β	α
	χ_{Leu}	β	β
IV	χ_{Leu}	L_α	α
	χ_{Leu}	L_α	β
V	χ_{Leu}	α	α
	χ_{Leu}	α	β
VI	χ_{Leu}	β	α
	χ_{Leu}	β	β

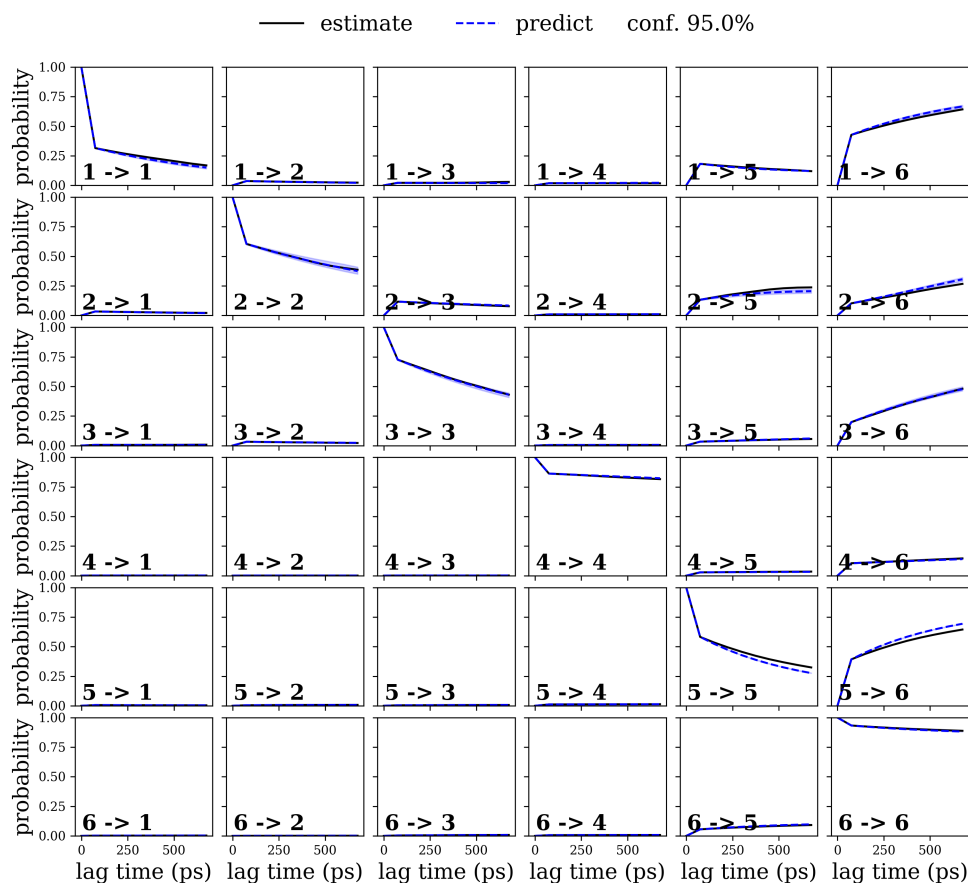


Figure 7.4: Chapman-Kolmogorov test of the Markov state model with six states 1,2,3,4,5,6 corresponding to metastable sets I,II,III,IV,V,VI respectively.

Fig. 7.5 shows the coarse-grained model of ALA as a transition graph between the metastable-sets.

In set-I, the (α, LHH) conformation can be defined however it was least sampled in the classical MD simulation trajectories of ALA. On the other hand, it is clear that the most probable conformations are (β, β) , (β, α) which belong to set-V. The second most probable conformations, (α, β) , (α, α) belong to set-VI. The least probable conformations while at the same time slightly more probable than the conformations, of set I, II and III, are the (LHH, α) and (LHH, β) conformations. The set-I correspond to transition along the pseudo-dihedral angle into Left-handed helix region which results in not only rotation of COO but also the inversion of it. The set II and III comprise the conformations where the value of χ_{Leu} dihedral angle is around 1 rad (third state of χ_{Leu} and denoted by χ'_{Leu}). This conformation is sampled with significantly low probability in the long MD simulations. Note that, this state of χ_{Leu} gives rise to two timescales which other-

wise are missing in the calculation of ITS. This reflects the artifact of classical MD simulations as well as of the MSM's.

As each set consists of more than one type of conformations which can be distinguished only by the rotation of COO group, for each set two representative conformations of the most probable microstates in the set are presented. The transition from any set to set I, II, and III involves transitions of either χ_{Leu} to χ'_{Leu} or a COO transition to Left-handed helix. Both these transitions are extremely unlikely and least sampled during the course of classical MD simulations. A much better sampled and least probable metastable-set which corresponds to left-handed helix conformation of ALA is the set IV. Like in, AL, a transition to this set involves rotation around ϕ_{Leu} and it is the slowest process. The metastable-sets V and VI are the two high probability sets. Set VI comprises of β -sheet like conformations being slightly more probable than the set V which is dominated by the α -helices. The transition between the conformations in these two sets mainly corresponds to the rotation of ψ_{Leu} . With the appearance of a third state of χ_{Leu} , it is not straight forward to distinguish the slow processes and the assignment of the corresponding timescales. However, if we completely neglect the χ_{Leu} for the construction of MSM then the slow processes follow the same pattern as in the case of AL i.e., from slowest to fastest, ϕ_{Leu} , ψ_{Leu} and χ_{Leu} .

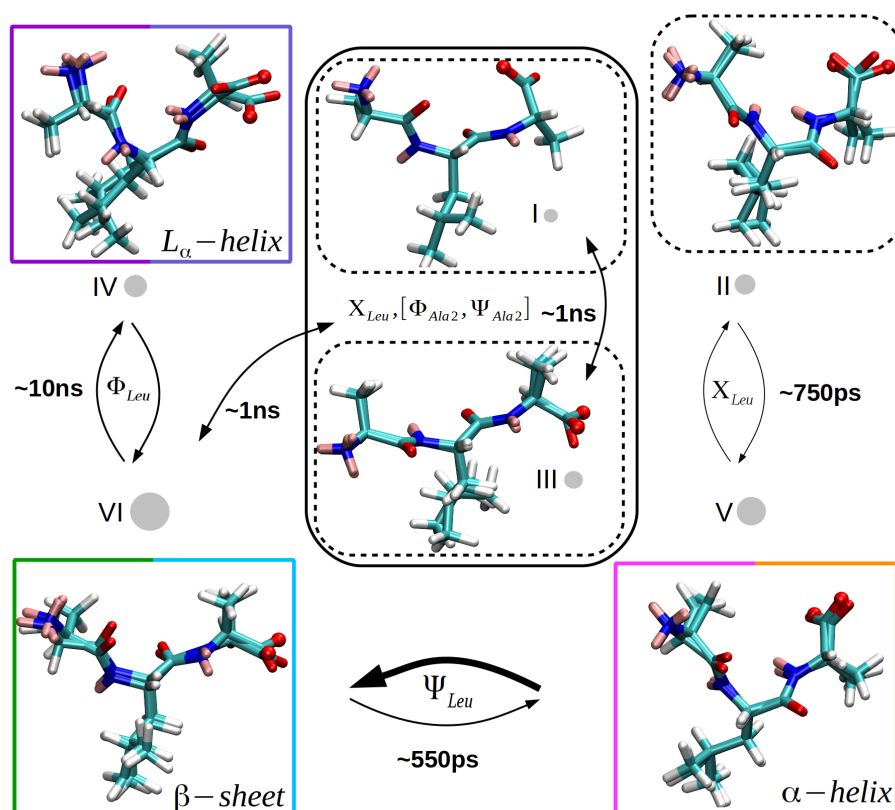


Figure 7.5: Coarse-grained model of the conformational dynamics of ALA in water. Arrows indicate the kinetic processes. The molecular structures in the boxes next to the circles are representative conformations of the most probable conformations in the respective set. The two colors of the boxes correspond to two conformations within the same metastable-set which can be distinguished by the torsion of COO. Set IV in color violet and slateblue corresponds to the representative conformations (LHH, α) and (LHH, β) respectively. Set-V in color magenta and orange corresponds to the representative conformations (α, α) and (α, β) respectively. Set-VI in color green and blue corresponds to the representative conformations (β, α) and (β, β). Carbon atoms are shown in cyan, oxygen atoms in red, nitrogen atoms in blue, hydrogen atoms in white and deuterium atoms in pink. Note that the dotted boxes indicate representative conformations of metastable sets which appear when χ_{Leu} is included as feature along with other torsion angles for the construction of a MSM.

7.3.1.2 Ala-Leu-Ala-Leu-Ala

A Markov state model has been constructed using the trajectory (60 μ s combined) of partially deuterated ALALA (Figure 7.6) in deuterated water obtained from classical MD simulations on the conformational space spanned by the torsion

angles $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$. The Figure 7.7 shows that the conformational space of the backbone torsion angles ψ, ϕ is well sampled by the classical MD simulations.

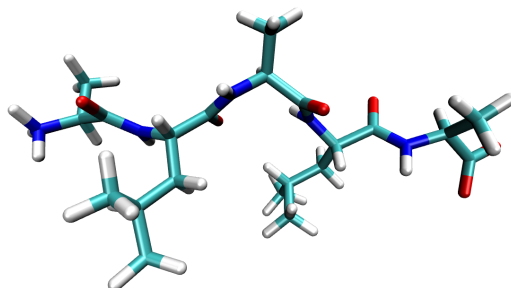


Figure 7.6: The β -sheet like representative conformation of ALALA peptide.

VAMP2 score analysis of backbone torsions was performed by varying the dimension parameter over a range of lag times (5ps to 2ns) (see Appendix.). At a 2ns lag, the score does not improve after the first four dimensions, however, by knowing the timescales of slow process (a few nanoseconds) of similar systems of shorter length (AL, ALA), it is possible to bypass some of the dynamics. As a compromise, we used a 500ps lagtime at which the first seven dimensions contain all relevant information about the slow dynamics as a best heuristic for dimensional reduction using TICA.

Figure 7.8(a) shows the one-dimensional distribution of first four TICA components and a two-dimensional distributions of first two components. Clearly, the first two components project seven high-density regions, which could be estimated metastable sets. The trajectories of the TICA components also nicely resolve the slow transitions as discrete jumps. Using k-means, for the clustering of the TICA coordinates, 150 cluster centers were found to be enough to discretise the dynamics. The ITS are shown in Figure 7.8(b), which are nicely converged, the timescale of the two slow processes i.e., ~ 120 ns, and ~ 47 ns are well separated from the others. Moreover, the convergence of ITS and a function of lagtime over the timescale 0 – 10 ns indicates the presence of numerous other significantly slower processes. A lagtime value of 1 ns is used for the estimation of the transition matrix.

The model is then validated with a Chapman-Kolmogorov test (Figure 7.9) after spectral analysis of the estimated transition matrix. The eigenvalue spectrum

(Figure 7.8(c)), like in the case of ALA, shows more than one spectral gaps between the eigenvalues. By analysing eigenfunctions projected on the first two TICA components, it is possible to define slow processes. Figure 7.8(d) depicts the four slowest processes of the implied timescale plot.

Lastly, the clusters were merged into seven metastable sets using robust Perron Cluster Analysis (PCCA+ [RW13]), based on the 2d-distribution of the first two TICA components, ITS, spectral analysis and the results of a Chapman-Kolmogorov test. Figure 7.8(e) represent the reweighted free energy surface by reweighting the trajectory frames with MSM stationary probabilities and Figure 7.8(f) shows the metastable sets identified by the PCCA+, within the first two TICA components clearly separating the state space. Metastable-sets and their corresponding representative conformations of ALALA based on Ramachandran space of the torsion angles used for the construction of a MSM are given in the Table 7.4. The two most probable metastable set consists of $(\beta, \beta/\alpha, \beta/\alpha, \beta/\alpha)$ and $(L_\alpha, L_\alpha, \beta/\alpha, \beta/\alpha)$ -conformations of ALALA with approximated stationary probability value of 0.55 and 0.26. The transition network plot obtained from a hidden Markov model based coarse-graining of the MSM of ALALA is shown in Appendix.

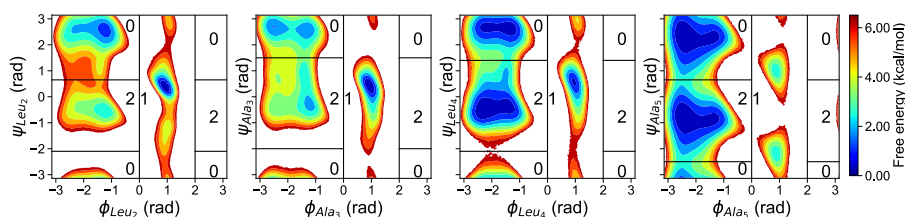


Figure 7.7: Ramachandran plot of $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$ obtained from the classical MD simulation of partially deuterated ALALA in deuterated water.

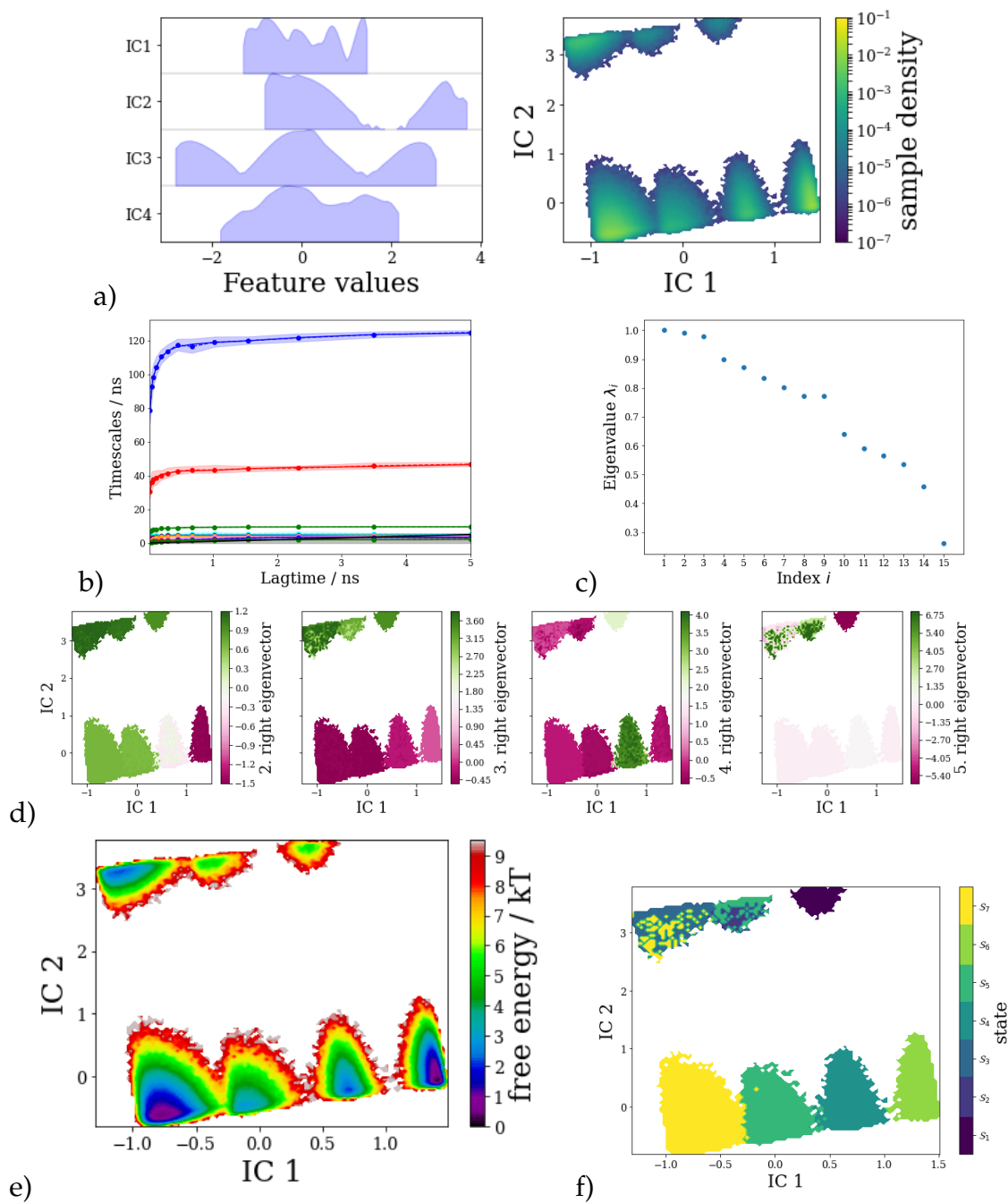


Figure 7.8: a) One-dimensional distribution of first four TICA components of ALALA (left) and the two-dimensional distributions of first two components (right), b) implied time scales, c) eigenvalues of the transition matrix sampled with lag time $\tau = 1\text{ ns}$, d) eigenvectors projections on the first two TICA components. e) re-weighted free energy surface f) metastable sets identified by the PCCA+.

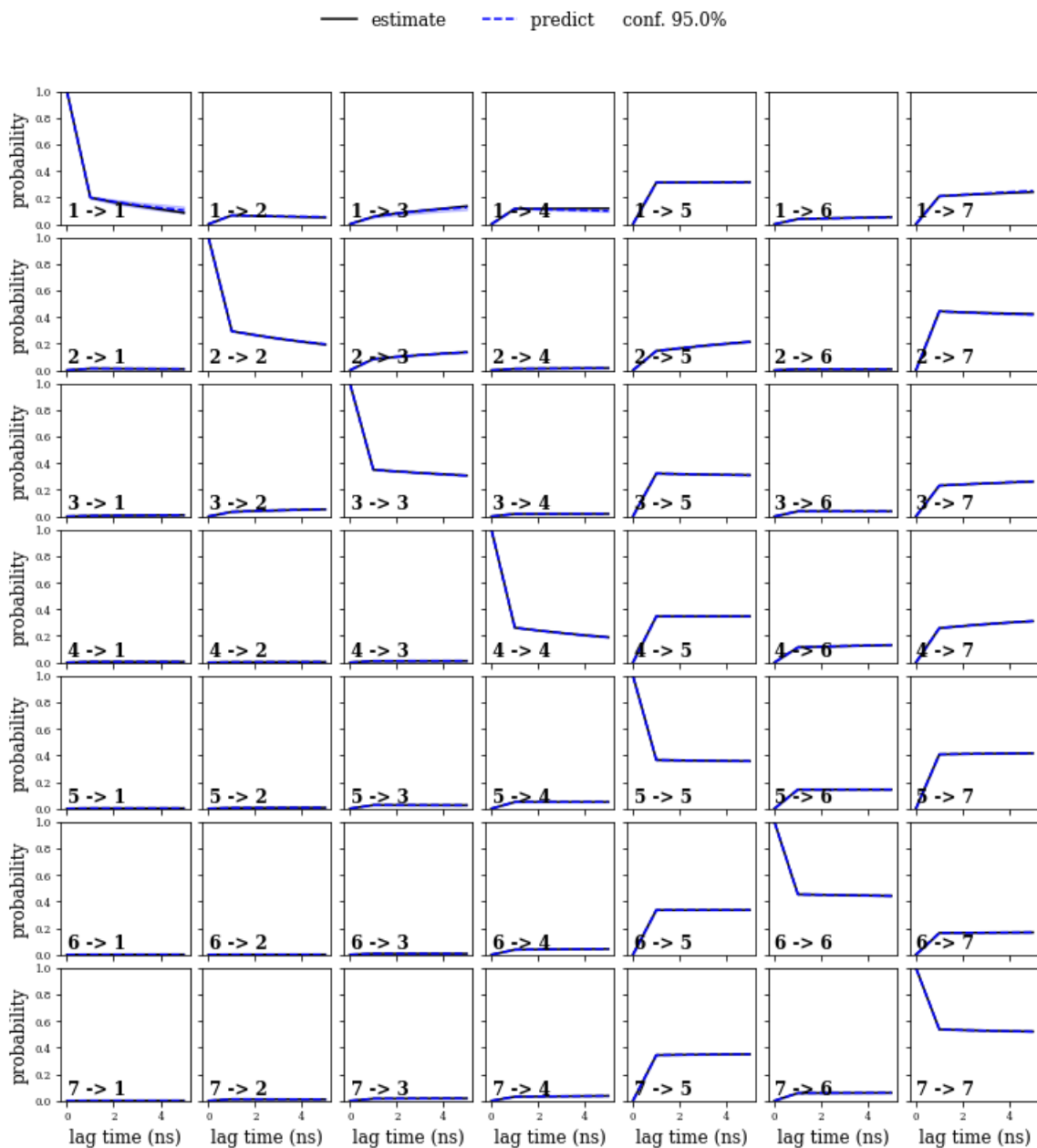


Figure 7.9: A Chapman-Kolmogorov test of MSM of ALALA.

Table 7.4: Metastable-sets and their corresponding conformations of ALALA based on Ramachandran space of $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$.

Set ²	Conformation			
	$[\phi_{Leu2}, \psi_{Leu2}]$	$[\phi_{Ala3}, \psi_{Ala3}]$	$[\phi_{Ala4}, \psi_{Ala4}]$	$[\phi_{Ala5}, \psi_{Ala5}]$
I	β	β/α	β/α	β/α
II	L_α	L_α	β/α	β/α
III	β/α	β	L_α	β/α
IV	β	L_α	β/α	β
V	L_α	L_α	α	α
VI	α	β	β/α	β/α
VII	α	β	L_α	β/α

7.3.1.3 Ala-Leu-Ala-Leu-Ala-Leu

Following the same procedure as for the construction of a MSM of ALALA, a MSM of ALALAL (Figure 7.10) has been constructed using the trajectory (50 μ s combined) of partially deuterated ALALA in deuterated water obtained from classical MD simulations on the conformational space spanned by the torsion angles $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$. The Figure 7.11 shows that the conformational space of the backbone torsion angles ψ, ϕ is well sampled by the classical MD simulations.

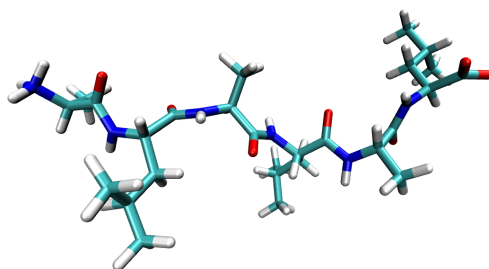


Figure 7.10: The β -sheet like representative conformation of ALALAL peptide.

²Note that a set can be further subdivided based on the representative sample conformations, e.g., set-1 contains four different conformations.

Figure 7.12(a) shows the one-dimensional distribution of first four TICA components, as well as the two-dimensional distributions of first two components. The first two components indicate seven densely populated regions with a few regions that are poorly separated and may contain a sub-region. As with ALALA, the trajectories of the TICA components also nicely resolve the slow transitions as discrete jumps. Using k-means, for the clustering of the TICA coordinates, 350 cluster centers were found to be sufficient to discretise the dynamics. The ITS are shown in Figure 7.12(b), which are nicely converged, the timescale of the three slow processes is ~ 76 ns, ~ 35 ns and ~ 26 ns, respectively. The lagtime value of 1 ns is also used for the estimation of the transition matrix in this case.

After performing spectral analysis of the estimated transition matrix, the model is then validated with a Chapman-Kolmogorov test (Figure 7.9). The eigenvalue spectrum (Figure 7.12(c)) show a clear spectral gap between 9th and 10th eigenvalue and Figure 7.12(d) shows the four slowest processes of the implied timescale plot as eigenvectors projections on the first two TICA components.

For this longer peptide, the clusters were merged into nine metastable sets using robust Perron Cluster Analysis (PCCA+ [RW13]), which was based on ITS, spectral analysis and the results of a Chapman-Kolmogorov test. Figure 7.12(e) represent the re-weighted free energy surface indicating the presence of few rarely explored energy minima and Figure 7.12(f) shows the metastable sets identified by the PCCA+.

Metastable-sets and their corresponding representative conformations of ALALAL based on Ramachandran space of the torsion angles used for the construction of a MSM are given in the Table 7.5. The most probable metastable set consists of $(\beta, \beta/\alpha, \beta/\alpha, \beta)$ -conformations of ALALAL with approximated stationary probability value of 0.82. The transition network plot obtained from a hidden Markov model based coarse-graining of the MSM of ALALAL is shown in Appendix.

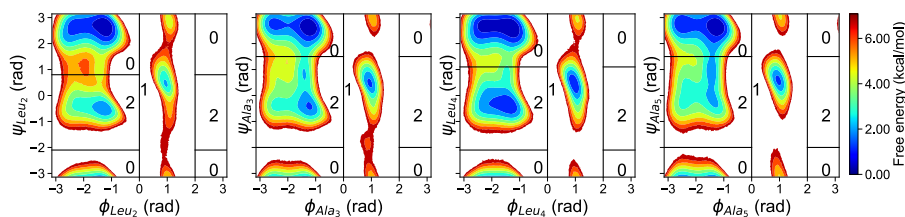


Figure 7.11: Ramachandran plot of $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$ obtained from the classical MD simulation of partially deuterated ALALAL in deuterated water.

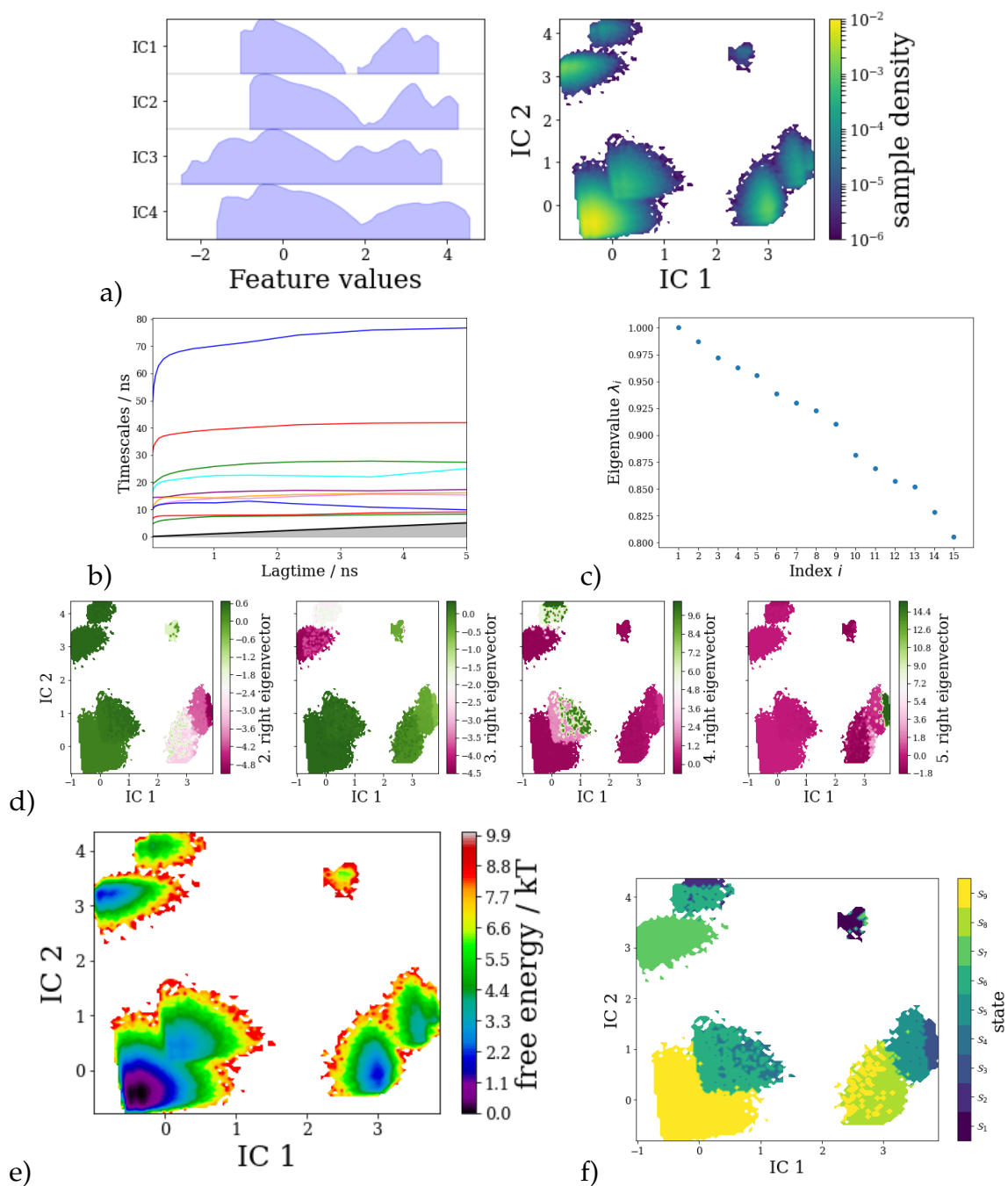


Figure 7.12: a) One-dimensional distribution of first four TICA components of ALALAL (left) and the two-dimensional distributions of first two components (right), b) implied time scales, c) eigenvalues of the transition matrix sampled with lag time $\tau = 1$ ns, d) eigenvectors projections on the first two TICA components. e) re-weighted free energy surface f) metastable sets identified by the PCCA+.

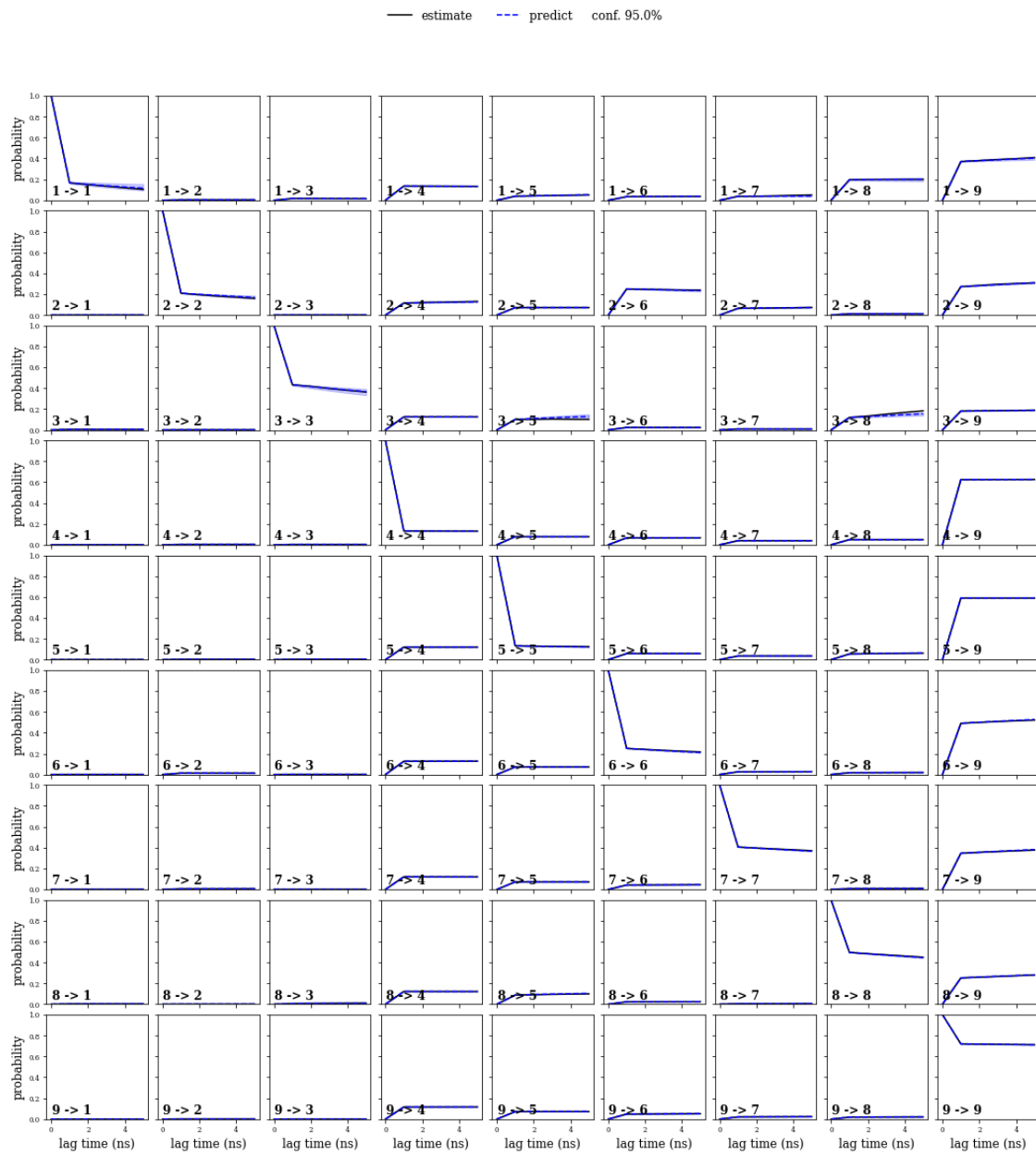


Figure 7.13: A Chapman-Kolmogorov test of MSM of ALALAL.

Table 7.5: Metastable-sets and their corresponding conformations based on Ramachandran space of $[\phi_{Leu2}, \psi_{Leu2}]$, $[\phi_{Ala3}, \psi_{Ala3}]$, $[\phi_{Ala4}, \psi_{Ala4}]$ and $[\phi_{Ala5}, \psi_{Ala5}]$.

Set2	Conformation			
	$[\phi_{Leu2}, \psi_{Leu2}]$	$[\phi_{Ala3}, \psi_{Ala3}]$	$[\phi_{Ala4}, \psi_{Ala4}]$	$[\phi_{Ala5}, \psi_{Ala5}]$
I	β	β/α	β/α	β
II	$\beta/\alpha/L_\alpha$	β	L_α	β/α
III	β/α	β	β/α	L_α
IV	β/α	β/L_α	β/α	β
V	β	L_α	L_α	β/α
VI	L_α	β/L_α	β/α	β
VII	L_α	L_α	L_α	β/α
VIII	β/L_α	L_α	β	L_α
IX	β	α, L_α	L_α	L_α

The time scales of the slowest process of AL, ALA, ALAL, ALALA and ALALAL peptide are given in the Table 7.6.

Table 7.6: Estimated timescale of the slowest process of each peptide.

System	Timescale of the slowest process (ns)
AL	~ 0.75
ALA	~ 10
ALAL	~ 24
ALALA	~ 120
ALALAL	~ 76

7.3.2 Constrained Classical Simulations

7.3.2.1 Structural Analysis

The distribution of the RMSD of the backbone, the end-to-end distance between the COM of the N-terminal and the COM of the C-terminal, and the solvent accessible surface area (SASA) of AL, ALA, ALAL, ALALA, ALALA, and water calculated from the short trajectories of beta-sheet-like conformations are shown in Figure 7.14 (a), (b), (c), respectively. The RMSD distribution of the backbone of AL is bimodal, whereas the width of the RMSD distribution generally increases from ALA to ALALAL. Notably, despite the same ensemble size, RMSD generally increases as the size of the system increases. AL, ALAL, and ALALAL all

have increasing end-to-end distances, whereas ALA and ALALA have decreasing end-to-end distances. Similarly, as the size of the peptide increases, the solvent accessible surface area (SASA) increases, as do the SASA distributions from AL to ALALAL.

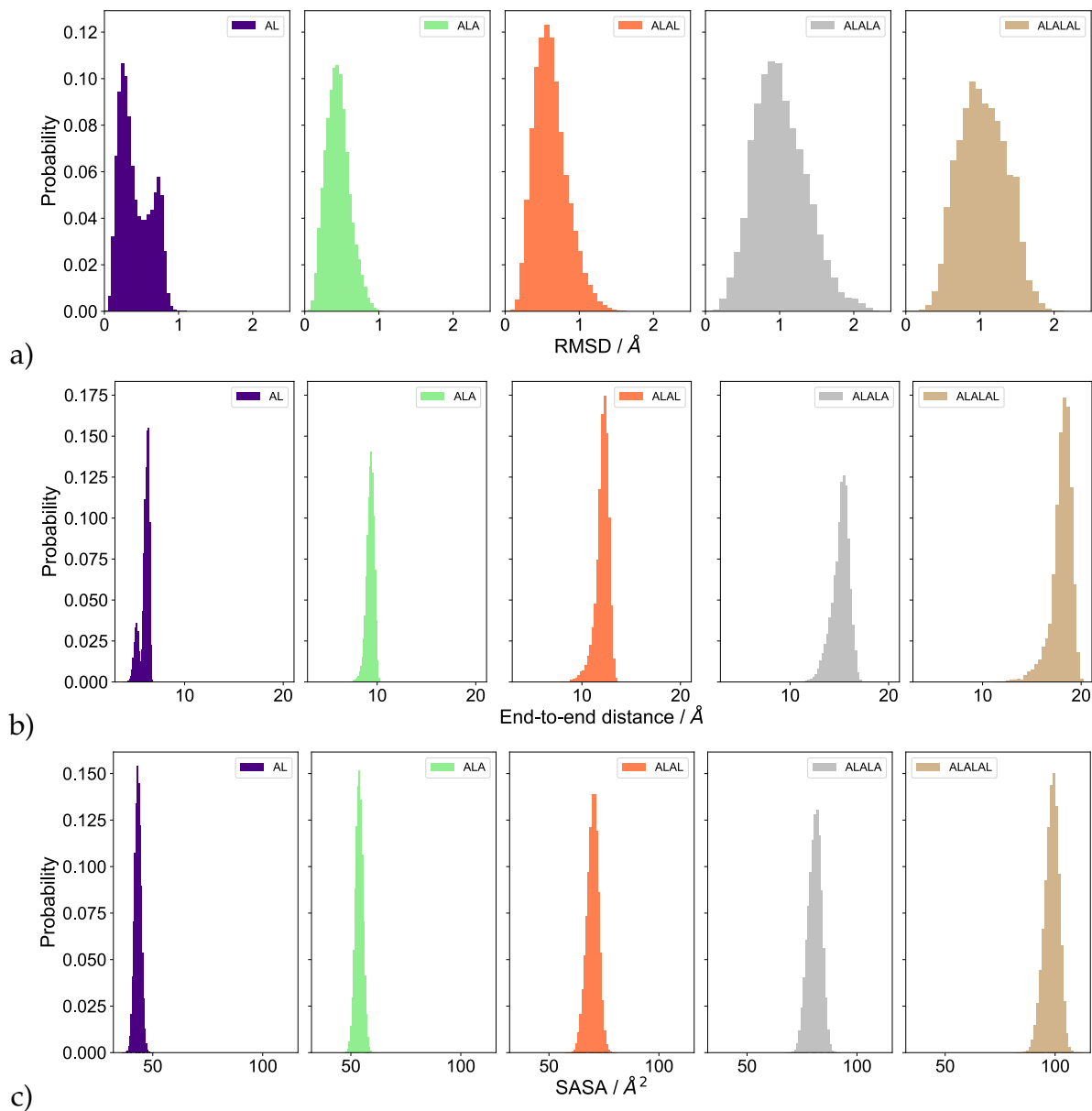


Figure 7.14: a) Distribution of RMSD of backbone, b) end-to-end distance i.e., between COM of N-terminal and COM of C-terminal, c) distribution of SASA of AL, ALA, ALAL, ALALA, ALALA and water.

The mean number of hydrogen bonds is shown in Table C.11. The number of polar groups in a peptide increases as its length increases. Notably, the number of

hydrogen bonds per carbonyl group increases with peptide length, so the second and third solvation shells are well defined for larger peptides.

Table 7.7: Average number of hydrogen bonds for each polar group of the systems. Note that number of carbonyl groups, amine groups increases from 1 to 5 for AL to ALALAL, respectively.

System	Peptide-Water	ND ₃	C _n =O _n <i>n</i> = 1, 2, 3, 4, 5	N _n -D _n <i>n</i> = 1, 2, 3, 4, 5	COO
AL	11.80±1.29	3.20±0.74	1.28±0.52	0.81±0.47	6.50±0.75
ALA	14.39±1.43	3.20±0.75	1.50±0.33	0.80±0.31	6.59±0.78
ALAL	16.59±1.58	3.20±0.74	1.52±0.28	0.75±0.26	6.60±0.77
ALALA	19.03±1.73	3.21±0.74	1.55±0.24	0.76±0.22	6.58±0.78
ALALAL	21.03±1.83	3.23±0.75	1.61±0.21	0.78±0.20	6.62±0.78

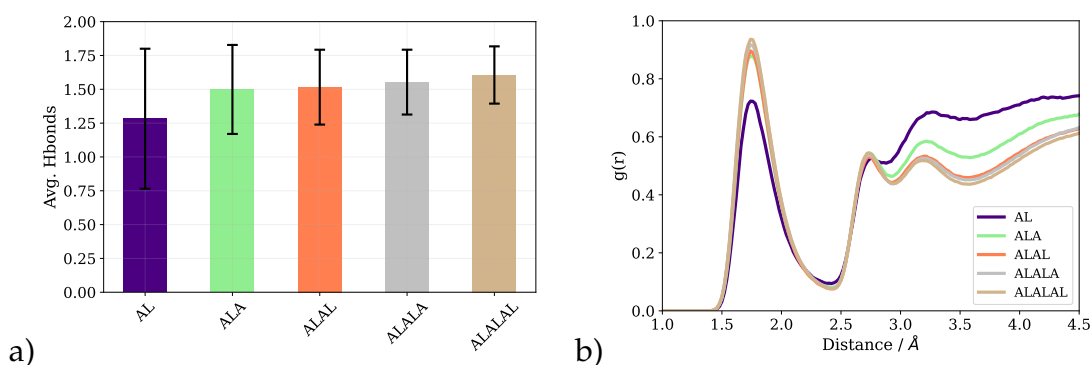


Figure 7.15: a) Average number of hydrogen bonds per carbonyl group for each system i.e., AL, ALA, ALAL, ALALA, ALALA. Note that number of carbonyl groups, amine groups increases from 1 to 5 for AL to ALALAL, respectively, b) radial distribution function between the oxygen atom of the carbonyl groups of AL, ALA, ALAL, ALALA, ALALA and water.

7.3.3 First-principles MD Simulations

7.3.3.1 Power Spectra

The effect of hydration on the power spectra of the carbonyl group is discussed in the previous chapter. Here we focus on how the increase in the length of the peptide effect the power spectra of carbonyl groups. Note that with an increase in length of the Ala-Leu peptide from AL to ALALAL, the number of carbonyl groups increases from one to five.

The computed power spectra (so called Amide-I region (C=O stretching)) of β – *sheet*-like, most probable conformation of AL, ALA, ALAL, ALALA and ALALAL peptides in water using the first-principle MD simulations is shown in Fig.7.16. These spectra are plotted without any corrections applied or normalisation for the sake of comparison in terms of carbonyl band positions, profile and to some extent intensity. Overall, the Amide-I region is well produced and the bands which correspond to the stretch vibrations of carbonyl ($\nu C_n = O_n$) are of high intensity. The bands of stretch vibrations of carbonyl groups are located between $\sim 1540\text{cm}^{-1}$ and $\sim 1640\text{cm}^{-1}$. The wide, high intensity band which correspond to the stretch vibration of C=O's group of ALALAL is located around $\sim 1600\text{cm}^{-1}$, so are the carbonyl bands of other peptides. However, there are differences in terms of band widths and their intensities.

With the increase in peptide length, i.e., number of carbonyl groups, the width of the Amide-I band increases due to the convolution of the spectra signal of more than one polar group (congestion). Similarly, due to increasing SASA with the peptide length, the interactions between the carbonyl groups and the water molecules give higher intensity peaks for longer peptides. The width and intensity of carbonyl bands of AL, ALAL and ALAL peptides gradually increases from AL to ALALAL. Whereas, there is a blue and red shift in the carbonyl bands of ALA and ALALA respectively, compared to the carbonyl band of AL.

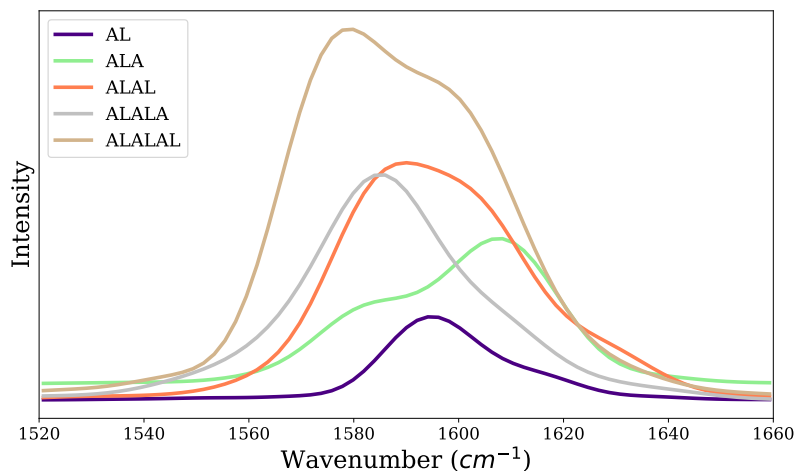


Figure 7.16: Computed Infrared spectrum in the Amide I region of AL, ALA, ALAL, ALALA, ALALAL peptide in water.

7.4 Discussion and Conclusions

The complexity in the conformational dynamics of peptides increases with the increase in the size of the peptide. The longer classical MD simulations are required to explore the full configurational space, along with high saving rates of atomic coordinates for the construction of a reasonable MSM (as it demands ergodicity and detailed-balance conditions) for the estimation of metastable conformations. MSM demands involves many steps and the optimality of the steps is dependent on the modeler. For example, the selection of internal coordinates to capture the dynamics, the selection of number of cluster centers which directly impact the quality of discretisation, etc. In the case of MSM of ALA, the addition of χ_{Leu} torsion angle gives rise to the two implied-timescales curves around ~ 1 ns which otherwise are missing, possibly the artifact of the MSM. Such issues are discussed in [SS18].

Nevertheless, the timescales associated with the slowest process of the studied peptide generally increase with increase in the length of the peptide except for the ALALA-peptide. ALALA tends to be in the close shape as evident by the constituent conformations i.e., $(L_\alpha, L_\alpha, \beta/\alpha, \beta/\alpha)$ -conformations, of the second most probable metastable set for this peptide. It might be due to the presence of the fast moving, two bulky leucine sidechains in such a short peptide. On the other hand pure L_α -like conformation is not identified in any of the nine metastable sets of ALALAL, it may not have been sampled and requires even longer classical MD simulations. Thus, no clear argument can be made on the relationship between the length of the time regime of peptide dynamics and the size of a peptide. Still, it highlights the increase in complexity with the increase in the size of peptide.

From the constrained classical MD simulations, it becomes evident that with the increase in the length of the peptide the properties of the hydration shell improve i.e., average number of hydrogen bonds per carboxyl group and the well-defined second and third hydration shells (better representative of the solvation of a bulk system). This dictates that the choice of the model system to study the dynamical properties of the proteins is also critical.

The power spectra calculated from the first-principle simulations of the peptides in explicit solvent shows the typical trend, i.e., the more the number of carbonyl groups gives rise to a more broaden Amide-I band and increased interaction of water molecules enhance the intensity features of the power spectra.

As except for ALALA, the timescales of the slowest process are increasing, so are the end-to-end distances (naturally), SASA, and the number of hydrogen bonds with polar groups. This may lead to more friction/slower conformational changes, and the more red-shifted and more intense bands in the power spectra are may be due to the same effect. Further investigation is needed to get a clear picture.

Chapter 8

Discussion

Proteins inherit conformational diversity, which is critical for their function in the aqueous environment. Due to the structural heterogeneity of proteins, their thermally accessible conformations (so-called metastable conformations) undergo interconversion on a complex free energy landscape [BNW09]. While the observed timescales of complex conformational transitions are longer (up to tens of seconds), their origins are in dynamics on the femto to picosecond time scale. The experimental IR spectroscopy techniques such as the Fourier transform IR spectroscopy [YYK⁺15], IR difference spectroscopy [LF20] and recently developed multidimensional ultrafast spectroscopy [Wan17] provide a wealth of information about the protein structures and dynamics [GOZ17, KLFH17, Ark06, TT97, Bar07, BZ02] on a variety of timescales.

However, the analysis of measured IR spectra is not straightforward as it contains averaged spectral properties of all the intermediates, weighted by their population evolution over time, plus its sensitivity to virtually all of the elements found in samples (a large number of vibrational modes including non IR active modes) makes it difficult to interpret and demand aid from static and dynamic calculations. For example, the measured Amide I band is usually a single wide band and it is extremely difficult to resolve bands of interesting vibrational modes of carbonyl groups. The spectral characteristics that contain the conformational information are suppressed due to congestion, which makes the assignment of the spectrum difficult.

Furthermore, other than the interesting slow metastable conformations and associated timescales, the faster hydration shell dynamics (i.e., changes in hydrogen bonding situations) also directly influences the vibrational properties/strength of the individual polar bonds (e.g., C=O), which impacts the measured IR spectra. It becomes even more difficult to extract information about hydration shells such as water topology and hydrogen bonding at the atomic level from measured spectra.

In order to understand the vibrational and scaling cascades of peptides, in this work we studied the Ala_{*n*}-Leu_{*n*} model peptides by classical MD simulations, MSM's and vibrational spectroscopy (experimental and calculated using first-principles MD simulations).

Classical MD simulations followed by the construction of a MSM, when used in conjunction with first-principles MD simulations i.e., *the combined approach* enabled us to understand the conformational dynamics of model floppy peptide AL and to interpret its measured vibrational fingerprints in terms of the estimated metastable conformations. The calculated IR spectrum of AL in water combined from the Boltzmann-weighted average of spectra computed for each metastable conformation and experimental spectrum was in good agreement. The first step of this approach comes with the advantage of an extensive sampling of the solvated model peptide's conformational space to explore the possible conceivable conformations. The second enables the estimation of the metastable conformations, and the third step provides the reliable IR spectrum of a metastable conformation in one single calculation.

Despite that, the purpose of the second step of the combined approach is to estimate the metastable conformations and associated timescales and not to extrapolate the long-time dynamics of the peptides, it is necessary to highlight some of the issues associated with the construction of a MSM. For example, for ALA, the addition of χ_{Leu} torsion angle give rise to the two implied-timescales curves around $\sim 1 ns$ which otherwise are missing (Section 7.3.1.1). The MSM construction involves the number of steps (from the selection of the internal coordinates to the coarse-graining into intuitively understandable models). It requires making many decisions, such as choosing a particular method (e.g., for clustering), method parameters, etc., from the modeler, which induces systemic errors and sometimes introduction of memory (i.e., violation of the Markovian assumption) [PWS⁺11, NWPP13, WPN15, GVE16]. These issues are addressed [PWS⁺11, NWPP13, WPN15, GVE16], however a modern framework that combines the whole data processing pipeline in a single end-to-end framework and provides easily interpretable few-state kinetic model could be employed [MPWN18]. Moreover, the larger the peptide, the longer the classical MD simulations required to explore the entire configurational space for the construction of a MSM to accurately estimate all possible metastable conformations.

The effect of the hydration shell cannot be ignored. For ALAL (which contains three distinct carbonyl groups), the frequency differences between its carbonyl groups are due to their interactions with the surrounding water. The probabilities of groups forming hydrogen bonds with water are consistent with the observed shifts in computed stretching frequencies. Zooming into the individual carbonyl group (the central carbonyl group ($C_2 = O_2$), as it is the least affected carbonyl group by the charged termini), the one, two or mixed hydrogen bonded states of the $C_2 = O_2$ group exhibited a clear trend of the red-shift of the $C_2 = O_2$ vibrational frequencies with the averaged number of hydrogen-bonded water molecules. Furthermore, the interaction energies analysis showed that the interaction of the second water molecule determines the amount of the (additional) red-shift, as the first water molecule almost always interact strongly. Thus, it is critical to consider not just the peptide itself but also its interaction with water,

as this has an effect on the experimental observables, namely the positions (and intensities) of the IR bands.

Many previous studies also highlighted the importance of hydration shell. For example, A combined experimental and computational approach on dialanine in water found that the spectral diffusion of the Amide-I vibration is dictated by water solvation dynamics [FT17a]. The combination of classical MD simulations with several data science algorithms showed the significance of water bridges around the peptide, trialanine [JH18]. [KNG04] studied the role of the water network during the formation of β -turns. [KNS10] and [JGH17] showed the importance of the formation of critical water bridges that define the peptide's structure.

On the other hand, the hydration shell is also affected by the underlying conformation. Utilizing the combined approach for ALAL and a closer look at the hydration shells revealed that the interaction of the polar groups with water molecules vary, resulting in the differences of the vibrational signatures of different metastable conformations ("open", "close" or "intermediate"). First, the metastable conformations have intrinsic structural differences that result in the presence or absence of intramolecular hydrogen bonds. Second, these differences cause individual polar groups of each conformation to have different solvation levels. Third, the formation of unique water bridges is caused by the intrinsic structure of metastable conformations combined with varying solvation levels. Finally, differences in the vibrational signatures of metastable conformations result from the hydrating water molecules' different overall topology/dynamics. It is easily conceivable that everything is connected. The two contrary observations (hydration shell affect the peptide and vice versa) for the same peptide highlights the entangled nature and importance of peptide-water interactions.

The MSM's of peptide of different lengths, i.e., AL, ALA, ALAL, ALALA, ALALAL, show that the most probable conformation of all peptides is the β -sheet like and the timescale of the slow transition increases with the peptide length (from AL to ALA). This might also be true for ALALA and ALALAL, we find it otherwise. Maybe 1) due to longer classical trajectories of ALALA (60 μ s long combined) than ALALAL (60 μ s long combined) used for the construction of MSM's, 2) It is intrinsic to ALALA, the MSM showed that second most probable metastable set consist of ($L_\alpha, L_\alpha, \beta/\alpha, \beta/\alpha$) -conformations which is in contrast to all other peptides. Moreover, with the increase in the length of the peptide (β -sheet like representative conformations) the properties of the hydration shell improve i.e., average number of hydrogen bonds per carboxyl group and the well-defined second and third hydration shells. Therefore, it is arguable that the choice of the model system to study the dynamical properties of the proteins is also critical. As an example, 1) the α -helix and β -sheet like conformation of AL are chemically equivalent, 2) only the central carbonyl group of ALAL behaves as in the larger system, and others are affected by the termini, 3) ALA and ALALA show some unusual dynamics which is reflected in their MSMs, possibly due to the fast-moving, one and two bulky leucine sidechains in such a short peptides.

Lastly, the power spectra of the carbonyl groups calculated from the first-principle simulations of the peptides in explicit solvent show the typical trend, i.e., the more the number of carbonyl groups gives rise to a more broader Amide-I band (congestion) and increased interaction of water molecules enhance the intensity features of the power spectra. Once again this highlights the importance of calculations for the interpretation of experimental results.

Chapter 9

Conclusions and Outlook

The experimentally observed vibrational signatures of peptides/proteins can be assigned/dissected into spectra of their constituent metastable-conformers using the combined approach. It makes possible extensive sampling of the solvated model peptide's conformational space, estimation of metastable conformations, and calculation of a reliable vibrational spectrum. The first-principles DFT-MD based vibrational spectra are reliable because the calculations rely on the time evolution of the electric dipole moment of the molecular system in the explicit solvent at finite temperature. Unlike quantum chemical calculations of normal modes spectra which rely on the curvature of the potential energy surface at the minima.

The estimation of all possible metastable conformations with the help of a MSM require the exhaustive sampling of molecular configuration space. Due to the established processing workflow, the estimation/validation of MSM relies on the technical expertise of the modeler, and an incorrect choice/decision at any phase may result in significant modeling errors.

The dynamics and topology of hydrating waters cannot be neglected. The frequency differences between the carbonyl groups of ALAL are due to their interactions with the surrounding water. The hydrogen bond probabilities of carbonyl groups match the calculated stretching frequency shifts. The central carbonyl group ($C_2 = O_2$) of ALAL showed a clear trend of the red-shift in its vibrational frequencies with the averaged number of hydrogen-bonded water molecules. The amount of the (additional) red-shift is determined by the interaction of the second water molecule.

Conversely, the interaction of the polar groups with water molecules varies for different metastable conformations of the same peptide (ALAL) resulting in the differences of the observed vibrational signatures. The different water topology/dynamics of the different metastable conformations affect the vibrational strength of the individual polar bond differently, impacting the calculated composed IR spectrum of the peptide.

Effect of hydration shell dynamics on the vibrational signatures of carbonyl groups and changes in polar groups-water interaction due to change in the underlying conformation indicate the entangled nature of peptide-water dynamics.

The most probable conformation of all studied Ala_n - Leu_n peptides is the β -sheet like, and the timescale of the slow transition increases with the peptide length (from AL to ALA). The choice of the model system to study the dynamical properties of the proteins is also critical. The increase in the length of the peptide improves the properties of the hydration shell at the cost of increased complexity. A trade-off should be made during the selection of relevant model peptides to study the protein dynamics using MD simulations. The larger the peptide length, the more broadened the Amide-I band and enhanced intensity features of the power spectra.

As an outlook, to further understand the vibrational and scaling cascades of suitable, possibly longer peptides in water, a modern framework that combines the whole data processing pipeline for the construction of a MSM in a single end-to-end framework to estimate metastable conformations could be employed [MPWN18]. Hybrid quantum mechanical/molecular mechanical MD simulations could be employed to calculate reliable IR spectra of longer peptides/proteins [MAdAF⁺18]. In order to reveal coupling between the vibrational modes and to further understand the interplay between peptide dynamics and water dynamics, ultrafast multidimensional spectroscopy (such as 2D time-resolved IR spectra) techniques [Hoc07] could be used. Like the combined approach, if combined with the calculated 2D-IR spectra (first-principles based on the wavelet method or using empirical maps), such vibrational spectroscopy techniques can provide a wealth of information. Furthermore, the so-called compressed modes [OLCO13] (sparse and spatially localized with compact support) can be used in conjunction with static/dynamic IR spectra calculations to better understand the measured spectra features. Moreover, the dynamical properties of water with respect to the change in the metastable conformation can be further analysed by performing short (avoiding unwanted transitions to a different metastable conformation) classical MD simulations with high saving rates of both solute and solvent coordinates. Followed by the construction of a MSM's including solvent dynamics [GCM⁺13]. For better accuracy, as the timescales of water dynamics are accessible in first-principles MD simulations, the desired conformation can be sampled in water long enough to build such a MSM.

Appendix A

Supplementary material for: A Combined Approach

A.1 Conformational analyses

A.1.1 Time series of torsion angles and micro-states in the first-principles MD simulations

Table A.1: Parts of first-principles trajectories of conformational clusters 0, 2, 4, and 6, respectively, used for the computation of spectra. The labels of the runs refer to the conformation in which the simulation was initiated.

0	2	4	6
0-run2-(0-20ps)	2-run1-(3.5-19ps)	0-run1-(34-56ps)	2-run3-(2-36ps)
0-run4-(0-20ps)	2-run1-(21-30ps)	4-run1-(30-50ps)	4-run1-(60-82ps)
2-run1-(10-50ps)	2-run2-(4-30ps)	6-run4-(0-24ps)	4-run3-(26-52ps)
2-run3-(0-20ps)			4-run4-(26-53ps)
2-run4-(0-20ps)			
6-run1-(10-50ps)			

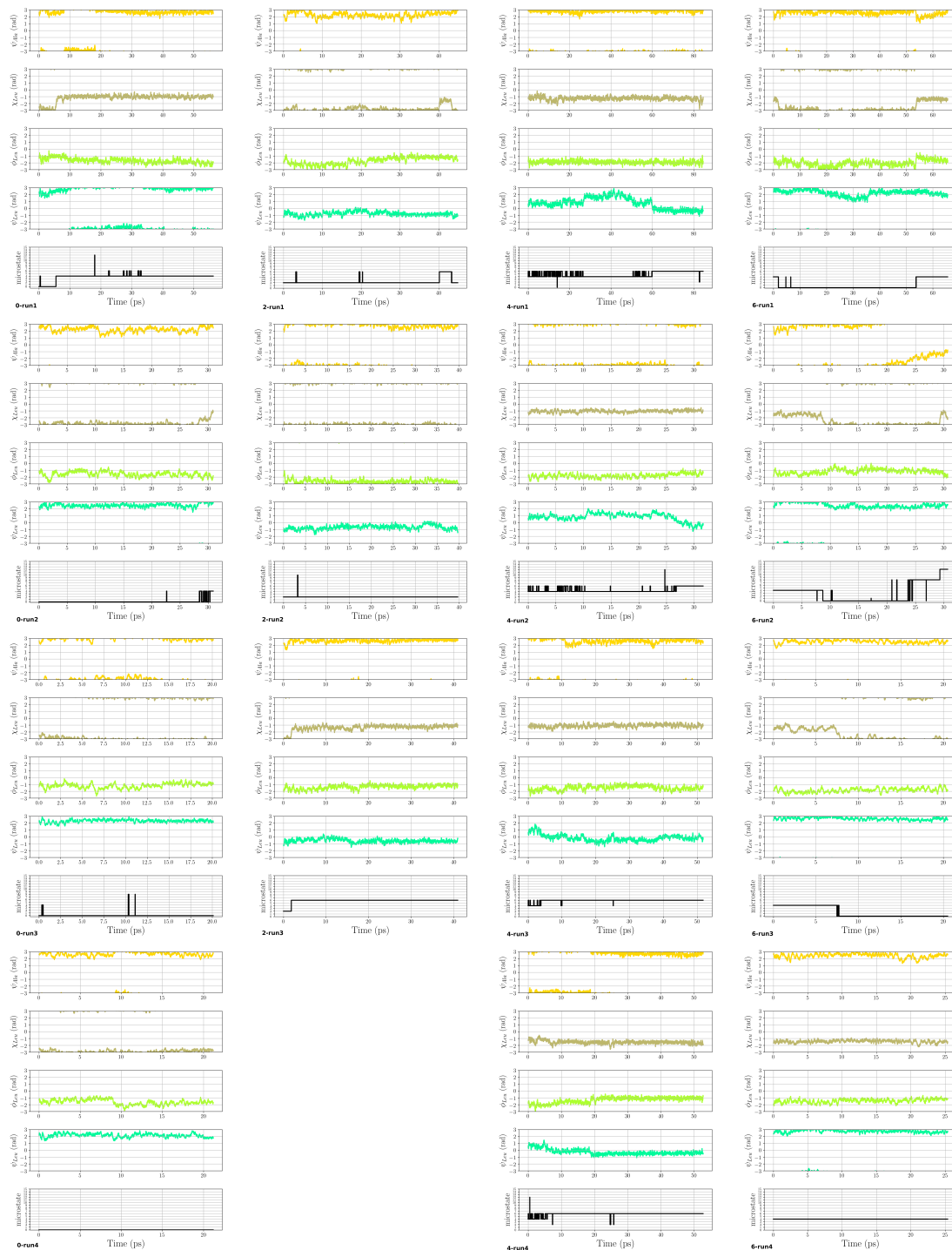


Figure A.1: Time series of torsion angles ψ_{Ala} , χ_{Leu} , ϕ_{Leu} and ψ_{Leu} , and conformational states defined by them along the first-principles simulations. Note that a jump between conformational states is defined by the micro-states they belong to based on the initial discretisation. Such a jump may in fact be only a small change in one torsion angle.

A.1.2 Hydrogen bond analysis

The radial distribution functions of water molecules around the polar groups (Figures A.2 and A.3) show a significant probability for a first solvation shell and thus a water molecule to be within hydrogen bond distance for the terminal groups, ND3, and COO-, in all micro-states, albeit with significant variance between individual simulations of micro-states 0 and 6. The central peptide groups, CO and ND, in contrast, exhibit a lower density of water molecules within hydrogen-bond distance for micro-states 0, 2 and 4, indicating a less ordered water structure around these groups and a decreased likelihood of hydrogen-bond formation between these groups and water molecules. Micro-state 2 shows a more pronounced peak of water density around the CO group, yet smaller than those observed for the terminal carboxyl group. The reduced probability to find water in a first solvation shell around the central groups CO and ND, may be explained by a competition for binding water molecules with the charged termini which are in close vicinity in this short peptide.

This observation also explains the relatively low number of hydrogen bonds (1 to 1.5) between water and the CO group, compared to the two acceptor possibilities by the two lone-pairs at the oxygen atom. Other first-principles MD studies on small capped peptides [Gai09, KC15], report an average number of hydrogen bonds of 2 to 2.5 per CO group. In those systems, however, there is no nearby charged terminal group, to which the surrounding water molecules are attracted instead. Indeed, a combined spectroscopic and computational study on uncapped di-alanine, find 1.5 to 2 hydrogen bonds to the amide carbonyl group, depending on the peptide conformation [FT17b].

For each of the polar groups in the Ala-Leu peptide, i.e. the terminal ND3 group, CO, ND, and the two oxygen atoms of the terminal carboxyl group separately, we have analysed the hydrogen bonds with water. A hydrogen bond is defined based on geometrical criteria: the donor-acceptor distance is below 3.5Å and the donor-hydrogen...acceptor angle is larger than 135°.

In order to test whether the hydrogen bond probabilities are affected by the number of water molecules in the simulation setup, we have computed the average number of hydrogen bonds between the polar groups and water molecules from the three classical MD simulations (v.s.) with 844 water molecules and from another set of classical simulations (3 runs for each of the 4 microstates) with the system setup used in the first-principles simulations, i.e. 123–160 water molecules. As can be seen from Figure A.4, the average number of hydrogen bonds is not affected by the number of water molecules and the box size.

Figure A.5 lists the distribution of number of water molecules that are hydrogen-bonded to the polar groups as obtained from the individual first-principles simulations of the different micro-states.

Figure A.6 shows the distribution of distances between the donor/acceptor atom

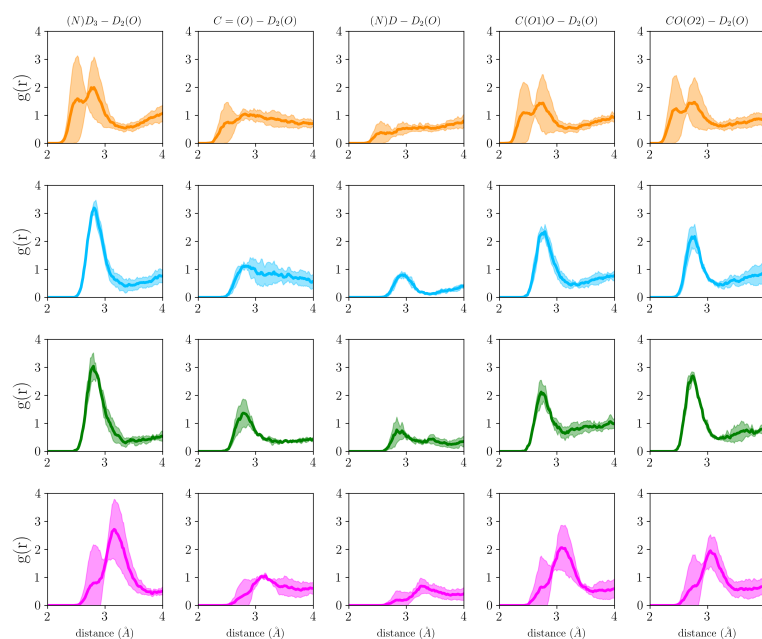


Figure A.2: Radial distribution function, $g(r)$, of water oxygen atoms around the polar groups, analysed separately for ND3, CO, ND, and the two carboxyl oxygen atoms (COO1 and COO2, respectively). The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively, averaged over the individual simulation runs.

of the polar group and the oxygen atom of the water molecules that are hydrogen-bonded to that group.

Figure A.7 shows the angular distribution of donor-hydrogen...acceptor angles for all water molecules that are within hydrogen-bond distance. Note that due to the angle criterion these water molecules are not necessarily forming a hydrogen bond to the polar groups of Ala-Leu.

The distribution of donor-acceptor distances within a hydrogen bond, ($\leq 3.5 \text{ \AA}$) shows small differences between the individual simulations of some micro-states for hydrogen bonds with the CO and the ND group. Whereas in most simulations the majority of hydrogen bonds is towards shorter distances to the CO group, simulations 0-run2, 0-run3 2-run3, 4-run2, 6-run1 6-run3 also show a significant probability for longer hydrogen-bond distances with CO. Hydrogen bonds with ND show distributions shifted towards larger distances mainly for simulations 0-run1, 2-run3, and 6-run4.

Analysis of the donor-hydrogen-acceptor angle distribution for water molecules that are within a hydrogen-bond distance, reveals the majority of donor-acceptor pairs to fulfill both hydrogen bond criteria, i.e. distance and angle. Still, a significant number of donor-acceptor pairs within hydrogen-bond distances exhibits

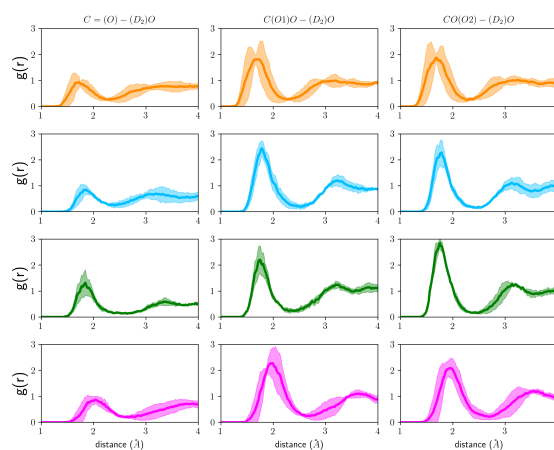


Figure A.3: Radial distribution function, $g(r)$, of (a) water deuterium atoms around the oxygen atoms of the C=O group and the two carboxyl group (oxygen atoms COO1 and COO2, respectively). The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively, averaged over the individual simulation runs.

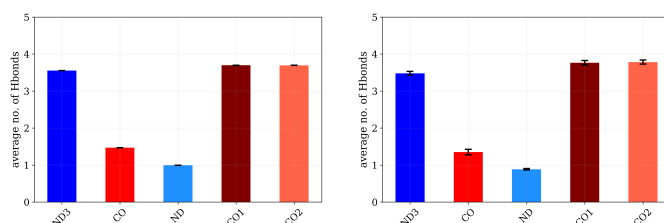


Figure A.4: Number of hydrogen bonds between the polar groups of Ala-Leu and water from classical dynamics simulations (a) in a box with 844 water molecules and (b) in a box with 123–162 water molecules.

donor-hydrogen-acceptor angles that are far from linear (up to 90 degree). This effect is most pronounced for hydrogen bonds with CO (in particular in simulations 4-run2, 2-run2, and all runs of micro-state 0). For hydrogen bonds with the ND group, simulations 0-run1 0-run5, 2-run1, 2-run2, 6-run3, 6-run4, exhibit smaller donor-hydrogen-acceptor angles, but no such effect is observed for any simulation of micro-state 4. These donor-acceptor pairs are not counted as hydrogen bonds, but likely still have a significant interaction, somewhat contributing to a weakening of the CO or ND bond, respectively.

The joint distributions of donor-hydrogen-acceptor angles and hydrogen-peptide oxygen distances (Figure A.8) further support this idea. All micro-states exhibit a significant, rather localised, population of angles and distances that correspond to hydrogen bonds between water and the peptide oxygen atoms. In addition, a broader distribution at angles lower than the threshold we used for defining hydrogen bonds and below 4 Å with probabilities that vary between individual

simulations can be observed in all micro-states. These interactions likely also have an effect on the vibration frequencies of $\nu\text{C}=\text{O}$ and νCOO , broadening the corresponding bands in the infrared spectrum.

The hydrogen-bonds with longer distances are supposedly weaker than those with shorter distances and thus contribute less to a shift in the CO and ND frequencies, respectively. But no such effect can be observed for micro-states 0, whereas for micro-state 2, number of hydrogen bonds and distribution of their distances agrees with the observation of a blue shifted $\nu\text{C}=\text{O}$ band. For micro-state 4 in run2 the $\nu\text{C}=\text{O}$ band is blue-shifted, seemingly contrasting the generally longer hydrogen-bonds. However, for the simulation 4-run2 the donor-hydrogen-acceptor angles significantly deviate from linearity, explaining the comparably low number and thus weak impact of hydrogen-bonded water molecules on the CO group. As for micro-state 6, only simulation 6-run4 shows a relation between the donor-acceptor orientation and CO frequency. In this simulation hydrogen-bond angles are far from linear, and thus hydrogen-bond interactions are weaker than in the other simulations of micro-state 6. Consequently, 6-run4, exhibits the $\nu\text{C}=\text{O}$ band with the highest frequency among the simulations of micro-state 6. The precise hydrogen-bond distance, in contrast, appears not to have a large effect on the $\nu\text{C}=\text{O}$ vibration frequencies.

Hydrogen bond distributions



Figure A.5: Probability distribution of number of water molecules hydrogen-bonded to the peptide, analysed separately for ND3, CO, ND, and the two carboxyl oxygen atoms (COO1 and COO2, respectively). The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively. Each row of a sub-figure represents an individual simulation of that micro-state.



Figure A.6: Probability distribution of hydrogen-bond distances between the peptide and water molecules, analysed separately for ND3, CO, ND, and the two carboxyl oxygen atoms (COO1 and COO2, respectively). The dashed line shows the distance threshold used as hydrogen bond criterion. The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively. Each row of a sub-figure represents an individual simulation of that micro-state.

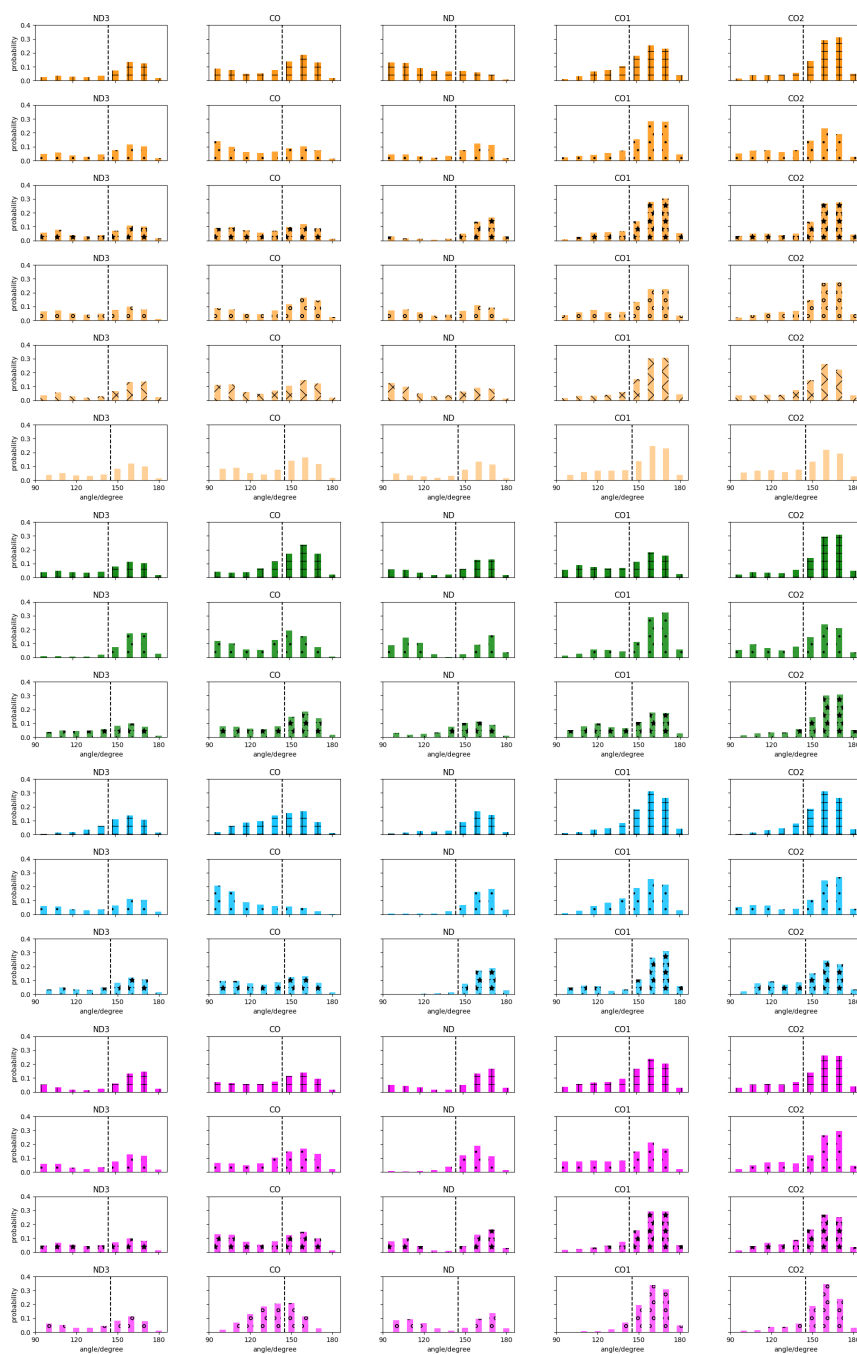


Figure A.7: Probability distribution of donor-hydrogen-acceptor angles between the peptide and water molecules within hydrogen-bond distance, analysed separately for ND3, CO, ND, and the two carboxyl oxygen atoms (COO1 and COO2, respectively). The dashed line shows the angle threshold used as hydrogen bond criterion. The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively. Each row of a sub-figure represents an individual simulation of that micro-state.

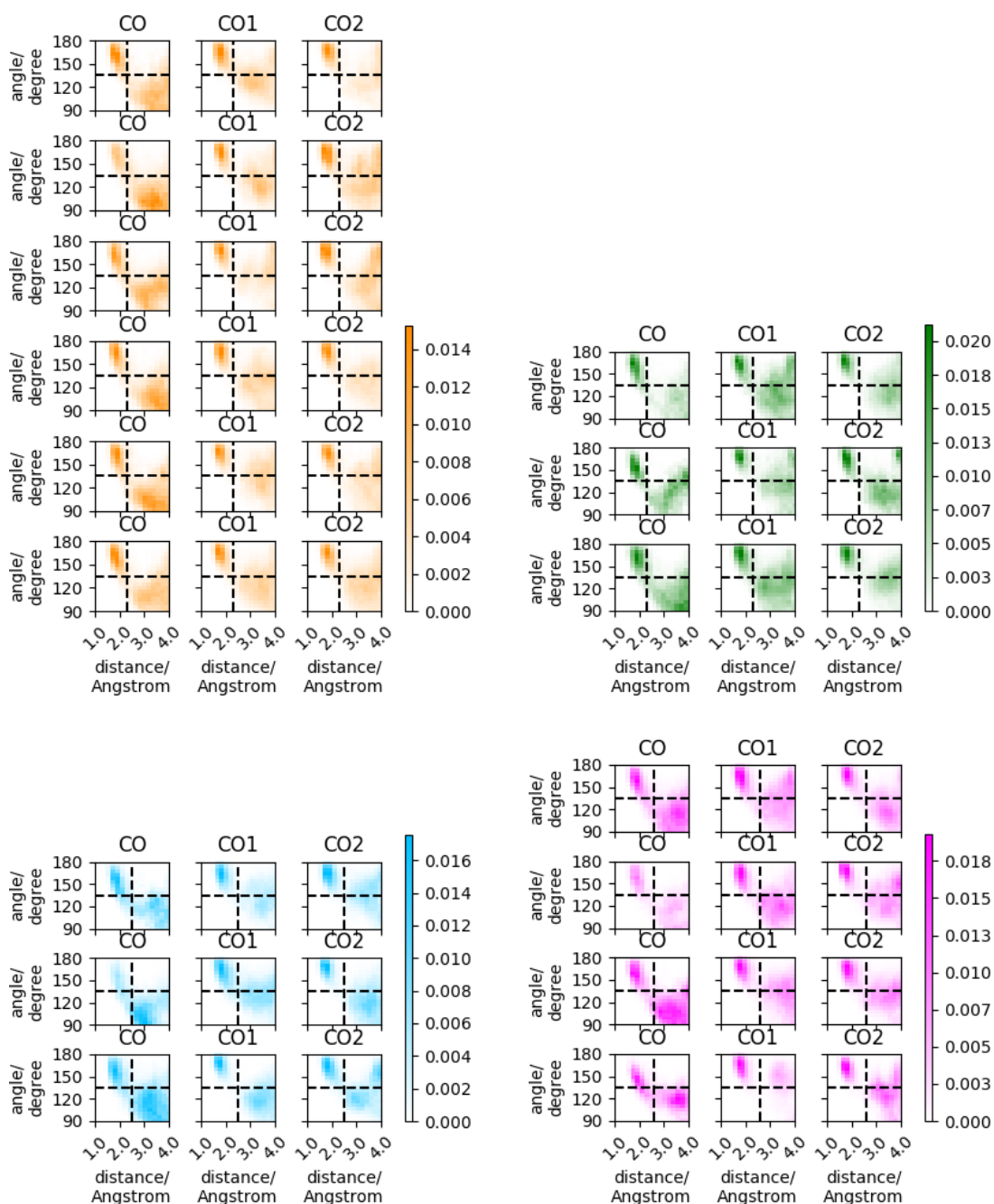
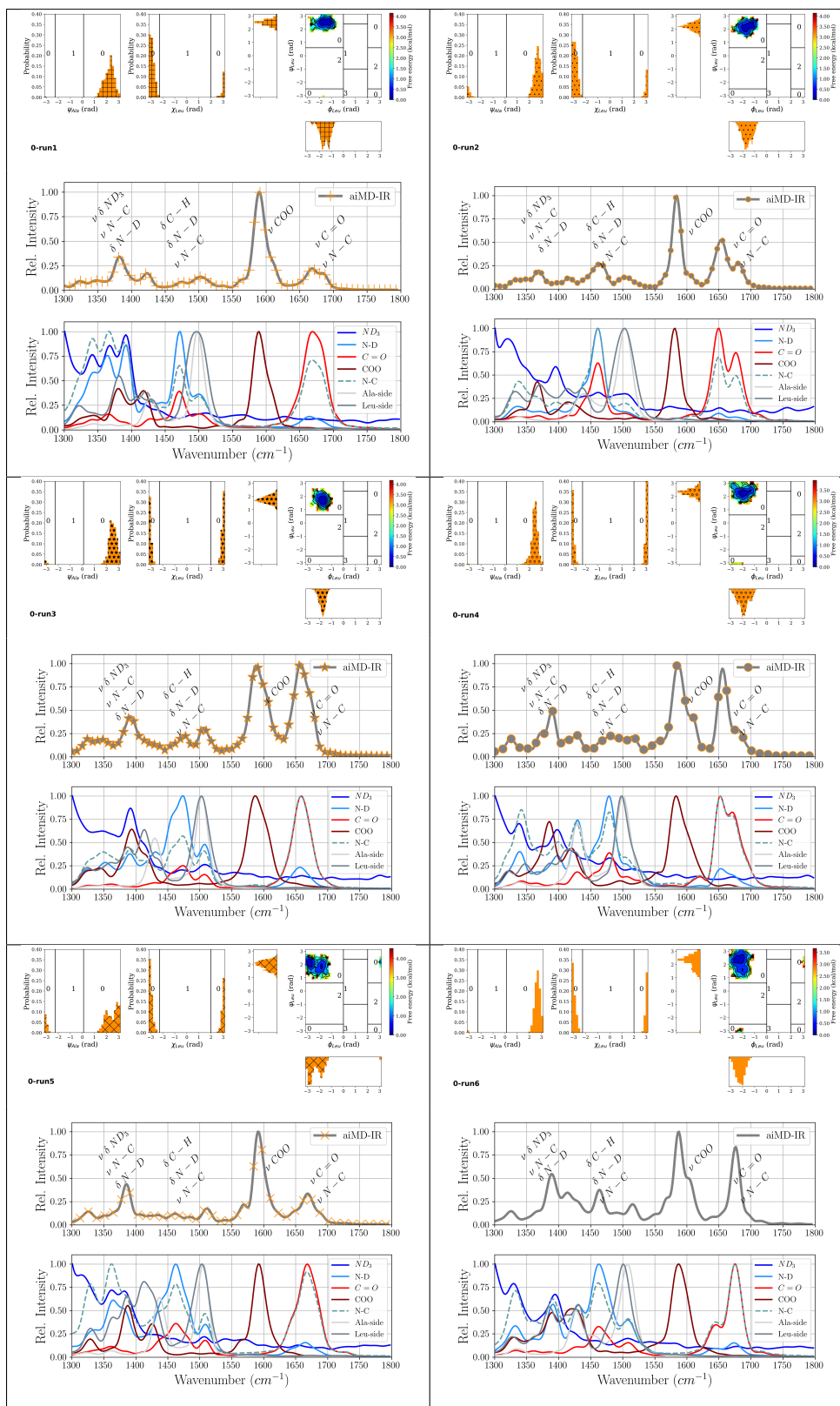
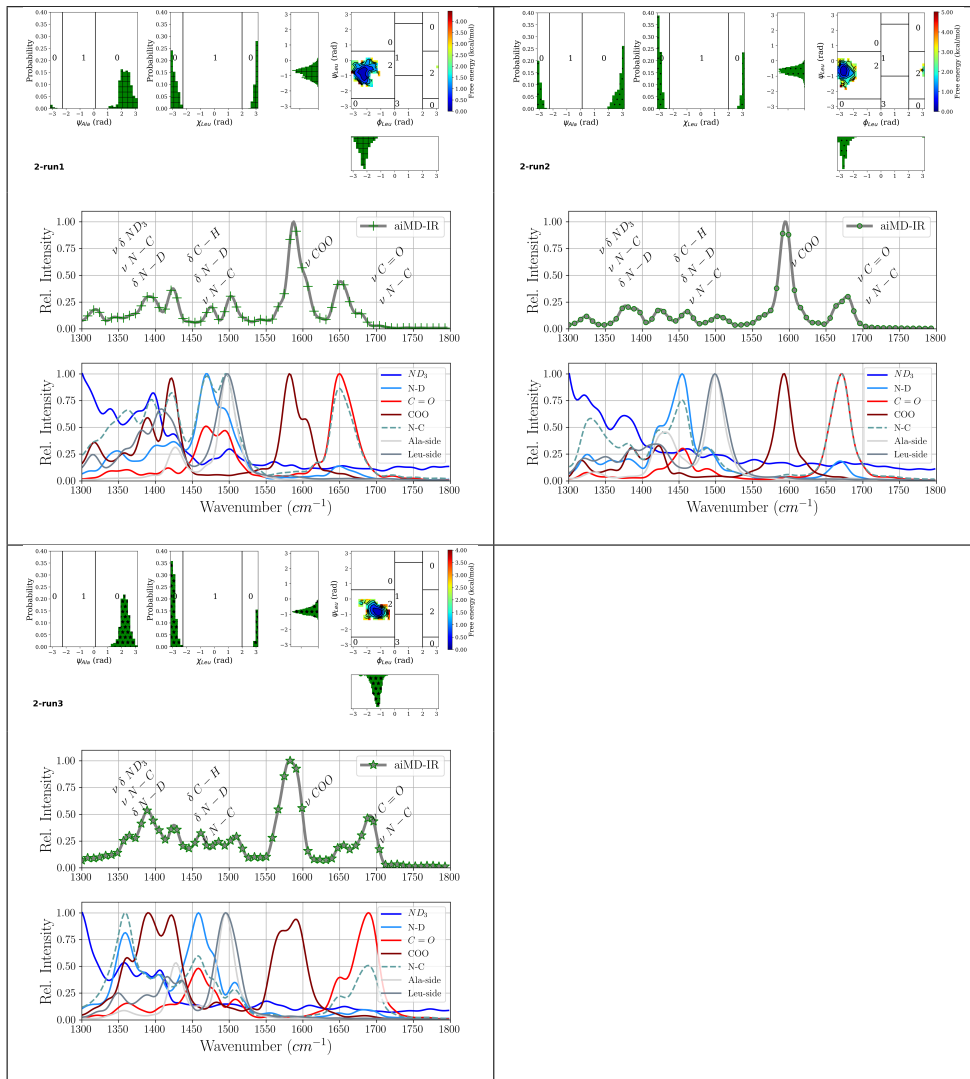
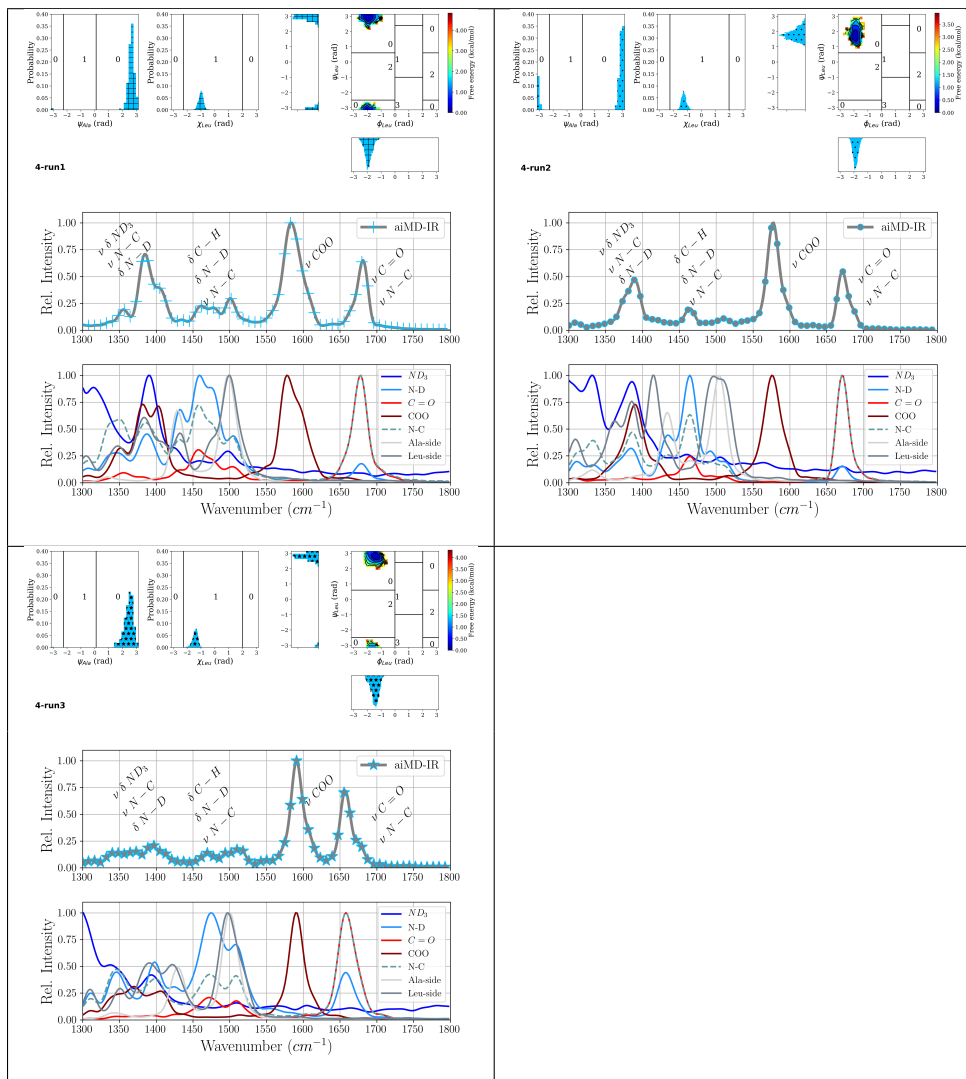


Figure A.8: Joint distribution of angles between peptide oxygen atoms (C=O, COO1, and COO2, respectively), water deuterium, and water oxygen atoms with distances between peptide oxygen and water deuterium atoms. The four micro-states 0, 2, 4, and 6 are shown in orange, green, blue, and magenta, respectively. Each row of a sub-figure represents an individual simulation of that micro-state. The horizontal dashed line shows the angle threshold used as hydrogen bond criterion whereas the vertical dashed line indicates the first minimum of the radial distribution function shown in Figure A.3 that can be understood as the limiting distance to find a water deuterium atom hydrogen-bonded to a peptide oxygen atom.

A.1.3 Torsion angle distributions, IR and power spectra from first principles MD simulations







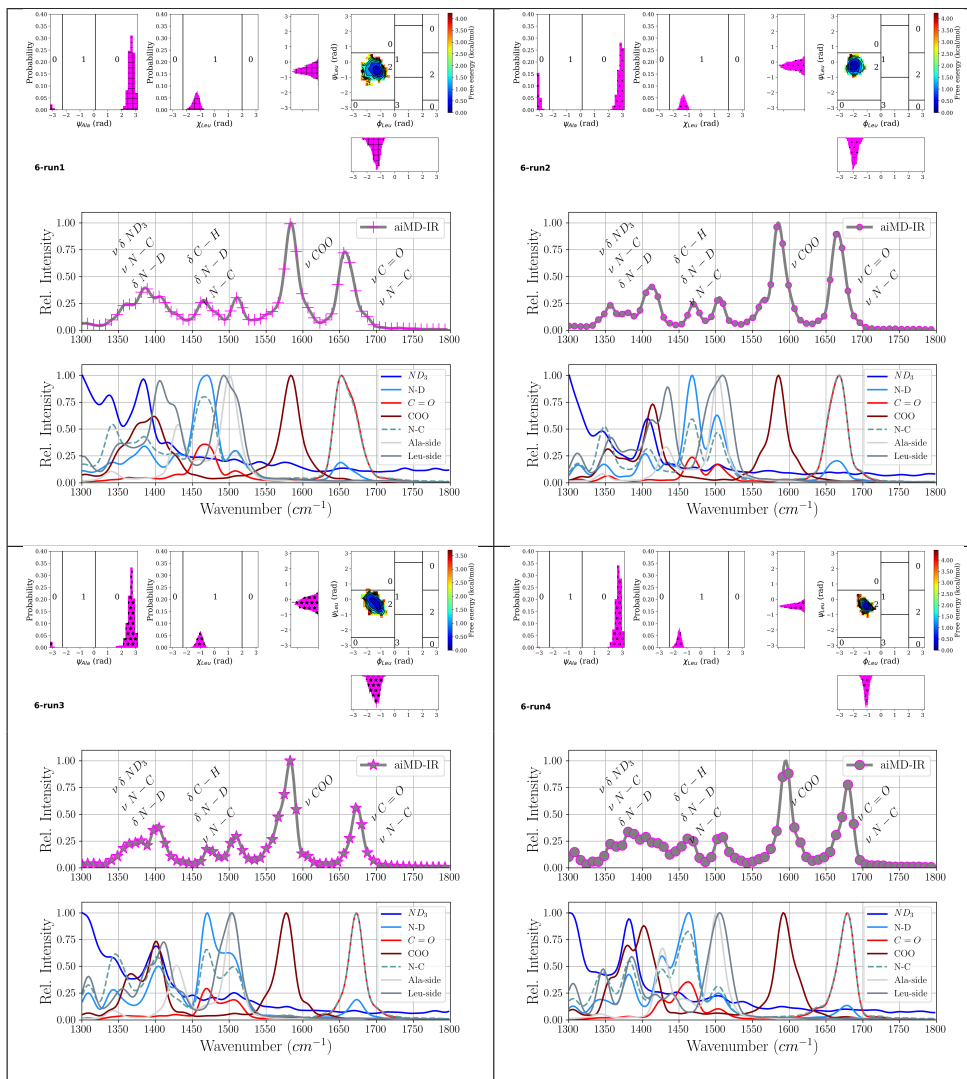


Figure A.9: Torsion angle distributions (top), IR spectra (middle) and power spectra (bottom), obtained from the individual first-principles MD simulations of representative Ala-Leu conformations in water. The colours in distributions and the IR spectra correspond to the different clusters as 0 (orange), 2 (green), 4 (purple), and 6 (magenta), respectively, with all runs of one cluster shown next to each other (run1 . . . run3, run4, or run6, respectively, from left to right). The colours in the power spectra correspond to the different groups of atoms. The leucine backbone torsion angles, as shown in the ramachandran plot, would correspond to β -sheet for cluster 0 and 4, and to α -helix for clusters 2 and 6, respectively, but are chemically equivalent conformations due to the second oxygen atom in the carboxyl terminus.

Appendix B

Supplementary material for: Hydration Shell Effect

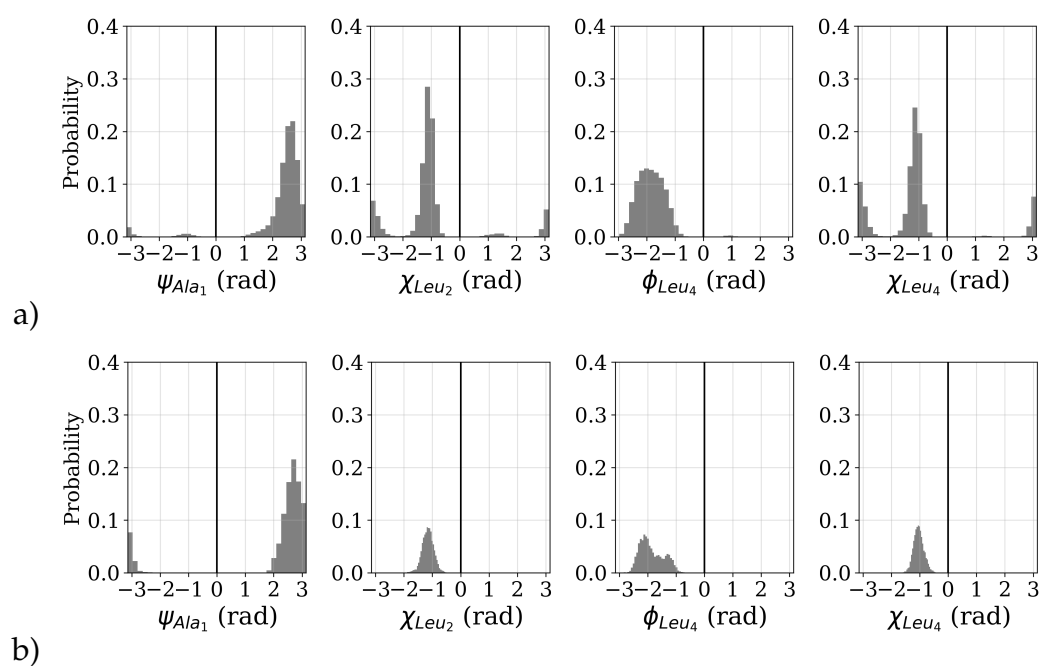


Figure B.1: Probability distribution of the χ_1 side chain torsion angles of the Leu residues, χ_{Leu2} and χ_{Leu4} , and the first and last backbone torsion angles, ψ_{Ala1} and ϕ_{Leu4} , computed from a) the classical and b) the first-principles MD simulation of the ALAL peptide in water.

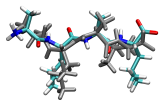
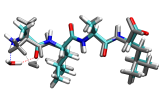
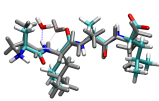
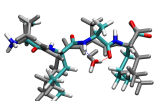
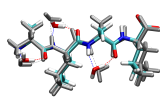
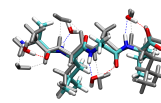
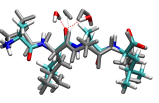
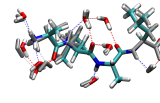
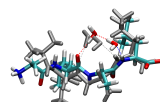
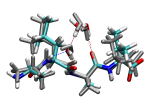
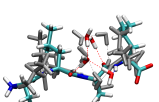

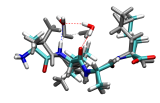
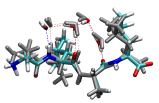
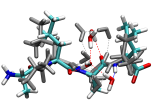
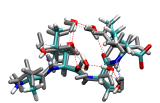
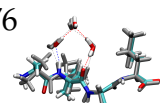
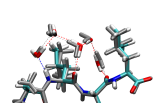
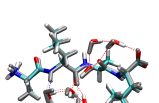
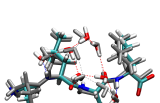
unbound	one hydrogen bond $C_1 = O_1$	one hydrogen bond $C_2 = O_2$	one hydrogen bond $C_3 = O_3$
			
$C_1 = O_1 : 1645$ $C_2 = O_2 : 1631$ $C_3 = O_3 : 1619$ RMSD: 2.15	$C_1 = O_1 : 1617$ $C_2 = O_2 : 1639$ $C_3 = O_3 : 1623$ RMSD: 0.83	$C_1 = O_1 : 1636$ $C_2 = O_2 : 1601$ $C_3 = O_3 : 1627$ RMSD: 1.34	$C_1 = O_1 : 1645$ $C_2 = O_2 : 1628$ $C_3 = O_3 : 1591$ RMSD: .57
C=O groups solvated	C=O groups and COO ⁻ group	two hydrogen bonds $C_2 = O_2$	all polar groups solvated
			
$C_1 = O_1 : 1620$ $C_2 = O_2 : 1604$ $C_3 = O_3 : 1588$ RMSD: 0.60	solvated $C_1 = O_1 : 1622$ $C_2 = O_2 : 1604$ $C_3 = O_3 : 1588$ RMSD: 1.16	$C_2 = O_2 : 1582$ RMSD: 1.05	$C_1 = O_1 : 1595$ $C_2 = O_2 : 1577$ $C_3 = O_3 : 1620$ RMSD: 0.24
0-0-2	0-3-0	0-2-2	0-3-3
			
$C_2 = O_2 : 1604$ RMSD: 1.48	$C_2 = O_2 : 1632$ RMSD: 0.58	$C_2 = O_2 : 1591$ RMSD: 1.83	$C_2 = O_2 : 1581$ RMSD: 0.39
2-0-0	2-3-0	0-3-2	2-3-3
			
$C_2 = O_2 : 1613$ RMSD: 1.63	$C_2 = O_2 : 1611$ RMSD: 0.94	$C_2 = O_2 : 1592$ RMSD: 1.83	$C_2 = O_2 : 1596$ RMSD: 0.86
3-0-0	3-3-0	4-3-0	3-3-3
			
176 $C_2 = O_2 : 1615$ RMSD: 0.60	$C_2 = O_2 : 1609$ RMSD: 0.30	$C_2 = O_2 : 1624$ RMSD: 0.29	$C_2 = O_2 : 1612$ RMSD: 0.57

Figure B.2: Frequencies (cm^{-1}) of the carbonyl stretching vibrations, computed by normal mode analysis of different ALAL-water clusters. The hydrogen bonding topology is described or indicated by labels x - y - z for connections between $C_2 = O_2 \cdots N_1 - D_1$ with x water molecules, $C_2 = O_2 \cdots C_3 = O_3$ with y water molecules and $C_2 = O_2 \cdots N_3 - D_3$ with z water molecules, respectively. RMSD is the root mean square deviation (\AA) between the initial snapshot (grey) and the optimised (coloured) structure.

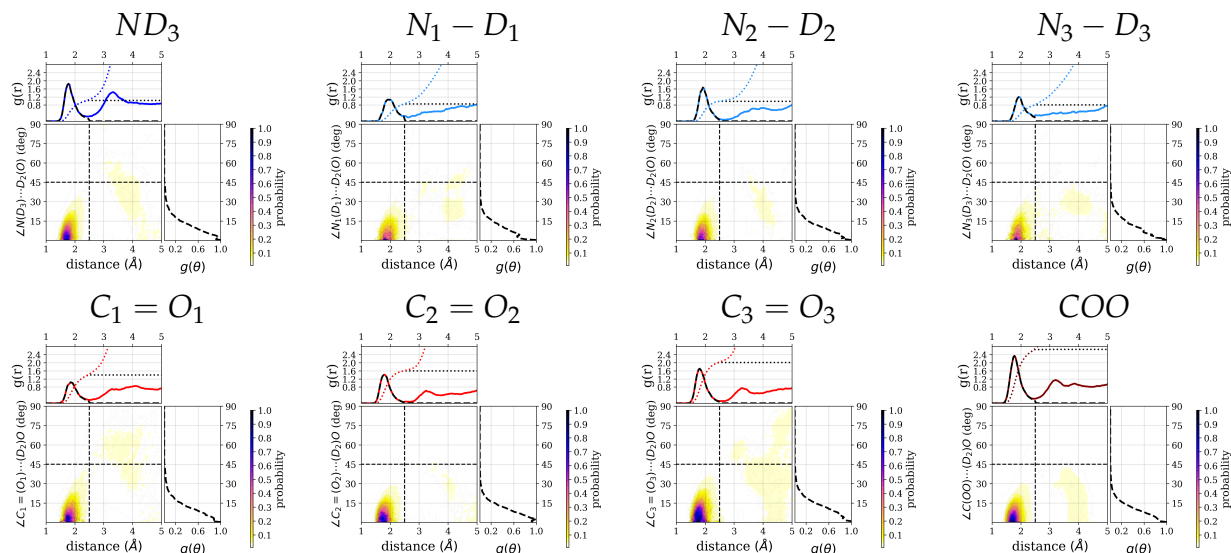


Figure B.3: Combined radial distribution functions, $g(r)$, and angular distribution functions, $g(\theta)$, of hydrogen-bonded water (D-atoms) around the polar groups of the ALAL peptide. Each top marginal plot shows $g(r)$ and right marginal plot shows $g(\theta)$ for the respective distribution function. Black dashed line-style is used for to show the restriction to hydrogen bond criteria. In each $g(r)$ plot, the black and red dotted curves represent the running integration of hydrogen-bonded water molecules and of all water molecules, respectively.

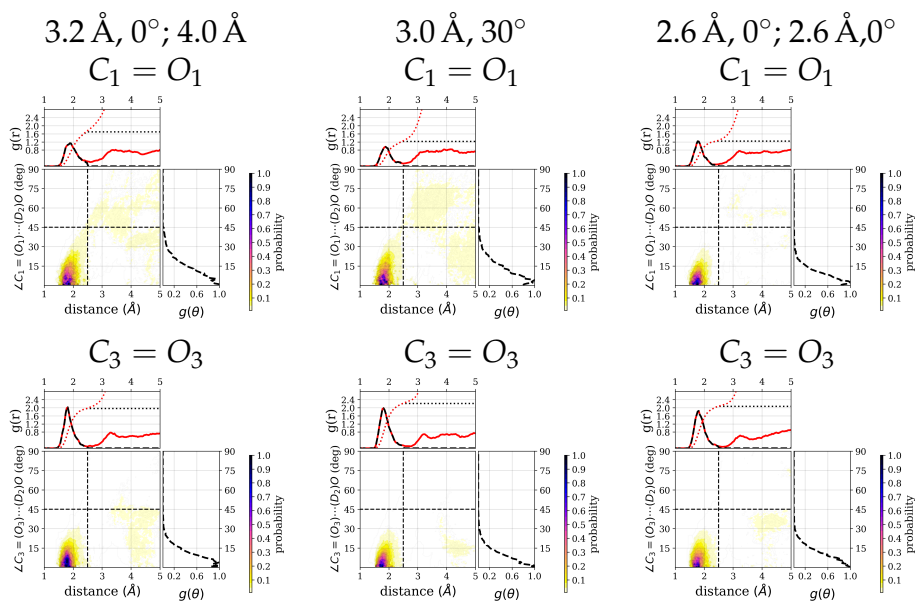


Figure B.4: Combined radial distribution functions, $g(r)$, and angular distribution functions, $g(\theta)$, of hydrogen-bonded water (D-atoms) around the $C_1 = O_1$ (top) and $C_3 = O_3$ (bottom) group of the ALAL peptide, computed for the constrained simulations $3.2 \text{ \AA}, 0^\circ$; 4.0 \AA , $3.0 \text{ \AA}, 30^\circ$, and $2.6 \text{ \AA}, 0^\circ$; $2.6 \text{ \AA}, 0^\circ$. Each top marginal plot shows $g(r)$ and right marginal plot shows $g(\theta)$ for the respective distribution function. Black dashed line-style is used for to show the restriction to hydrogen bond criteria. In each $g(r)$ plot, the black and red dotted curves represent the running integration of hydrogen-bonded water molecules and of all water molecules, respectively.

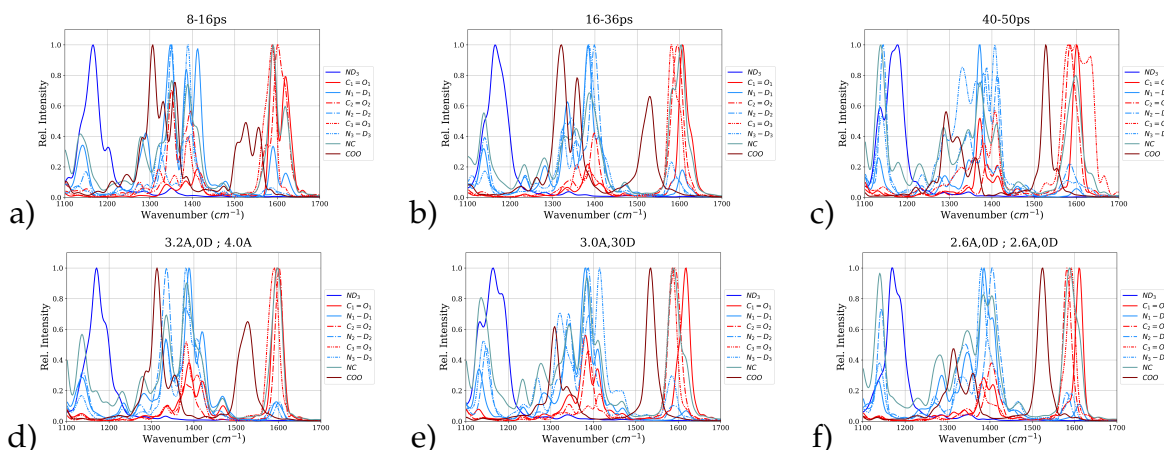


Figure B.5: Power spectra computed from first principles calculations of the ALAL peptide in deuterated water from windows a) 8–16 ps, b) 16–36 ps, and c) 40–50 ps of unrestrained simulation, as well as from simulations with restraints (see methods for details) d) $3.2 \text{ \AA}, 0^\circ$; 4.0 \AA , e) $3.0 \text{ \AA}, 30^\circ$, and f) $2.6 \text{ \AA}, 0^\circ$; $2.6 \text{ \AA}, 0^\circ$.

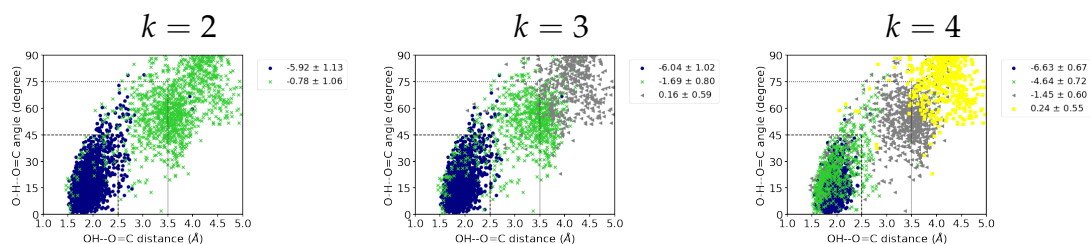


Figure B.6: Results of k -means clustering the interaction energy values, computed from the snapshots of an unconstrained simulation of the ALAL peptide in water.

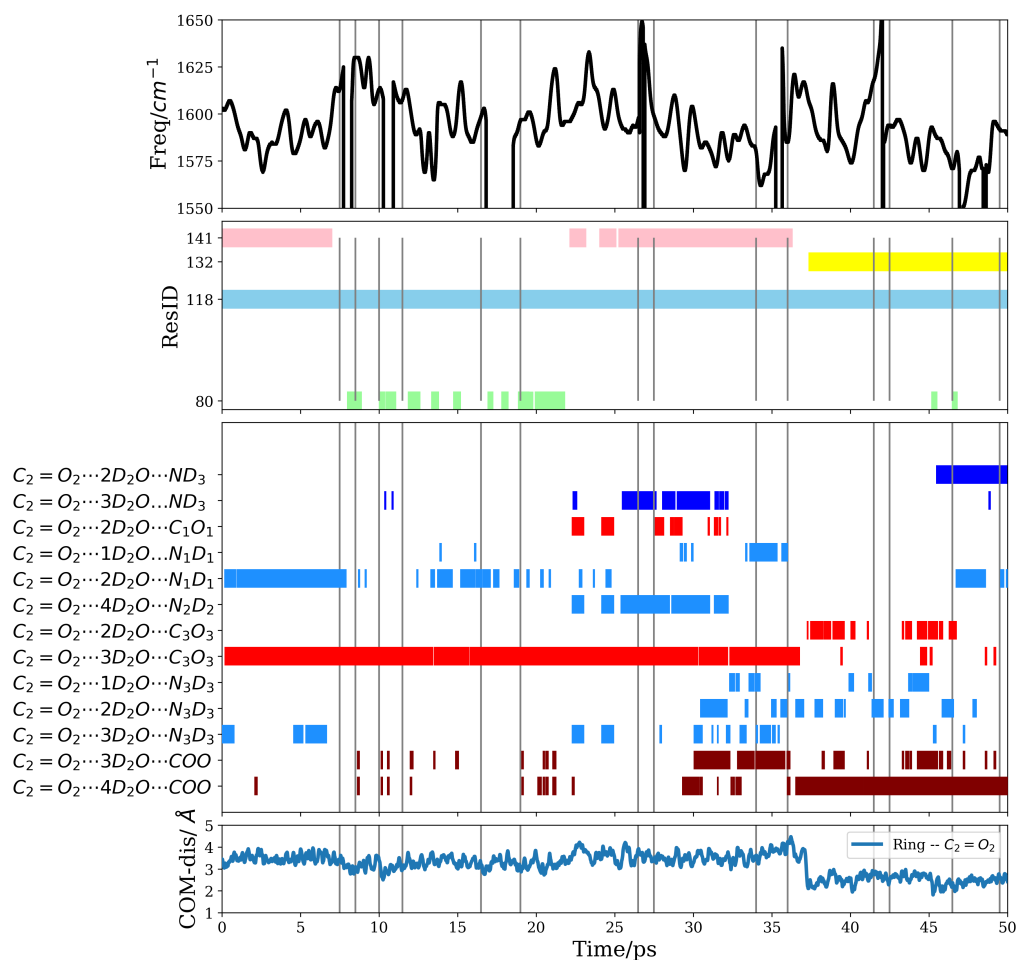


Figure B.7: Time series of the instantaneous frequencies from a wavelet analysis, individual water molecules hydrogen-bonded to the C₂ = O₂ group (see also Figure 6 in the main text), water bridges between the C₂ = O₂ group and the other polar groups, and distance of the centre of mass of a three-water ring (see Figure 6a) in the main text), connecting the C₂ = O₂ and C₃ = O₃ group to the centre of mass of the C₂ = O₂ group.

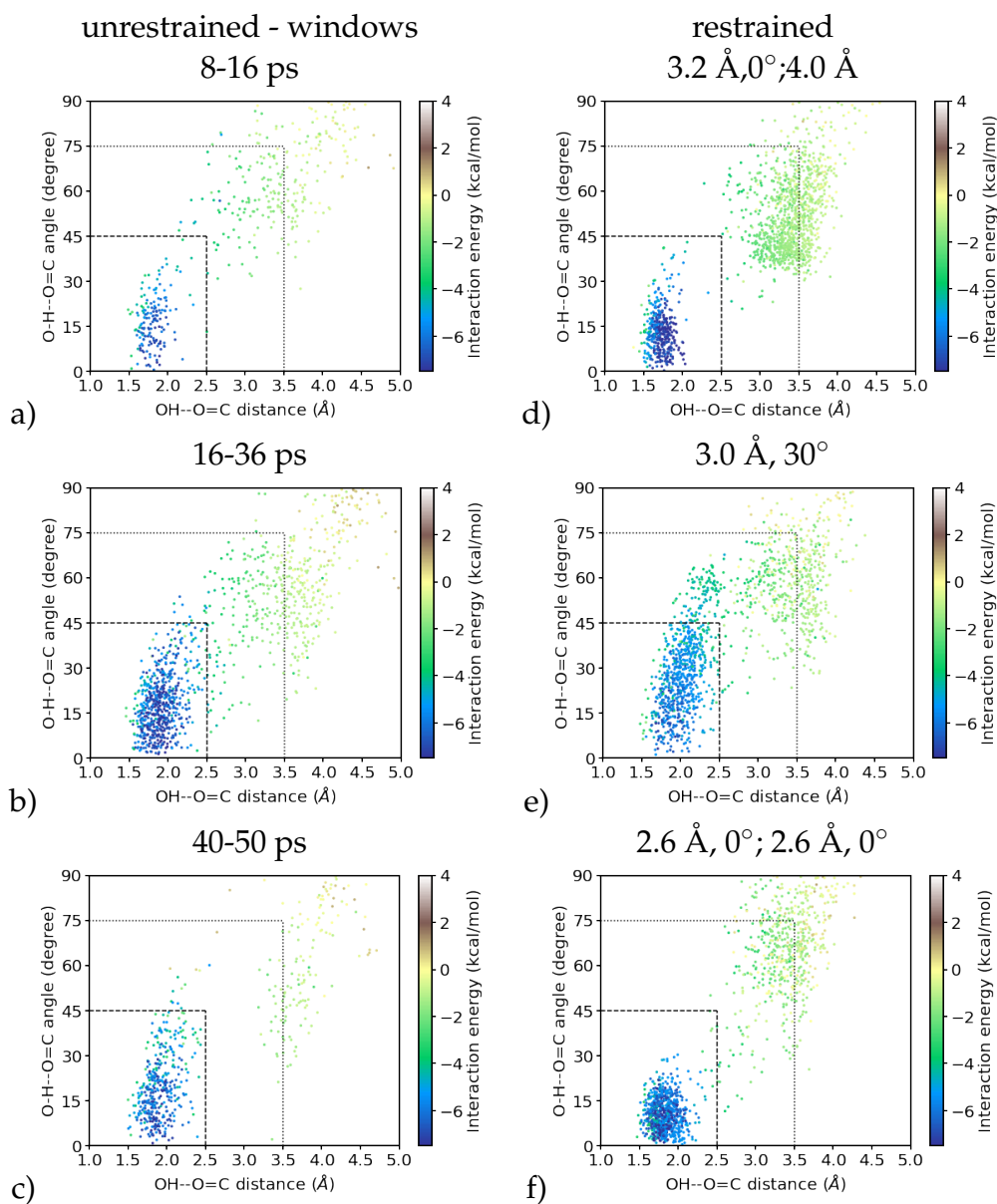


Figure B.8: Interaction energy distributions for the four water molecules closest to the $C_2 = O_2$ group from windows if an unrestrained simulation a) 8–16 ps, b) 16–36 ps, and c) 40–50 ps, as well as simulations with restraints (see methods for details) d) $3.2 \text{ \AA}, 0^\circ; 4.0 \text{ \AA}$, e) $3.0 \text{ \AA}, 30^\circ$, and f) $2.6 \text{ \AA}, 0^\circ; 2.6 \text{ \AA}, 0^\circ$. Note that the number of points are different due to the different simulation or window lengths.

Table B.1: Correlation between interaction energies and $C_2 = O_2 \cdots H - Ow$ distances or $C_2 = O_2 \cdots H - Ow$ angles.

Restraint	Energy-Distance Correlation				Energy-Angle Correlation			
	W1	W2	W3	W4	W1	W2	W3	W4
None (full)	-0.09	0.89	0.80	0.40	0.43	0.83	-0.78	0.47
None (8–16 ps)	-0.22	0.86	0.80	0.30	0.49	0.67	-0.64	0.32
None (16–36 ps)	-0.10	0.91	0.78	0.24	0.38	0.77	-0.69	0.31
None (40–50 ps)	0.03	0.65	0.70	0.74	0.51	0.75	-0.72	0.76
3.0 Å, 30°	0.14	0.76	0.78	0.67	0.49	0.79	0.84	0.70
3.2 Å, 0°; 4.0 Å	-0.56	0.73	0.79	0.75	0.38	0.46	0.66	0.75
2.6 Å, 0°; 2.6 Å, 0°	-0.21	0.27	0.69	0.72	0.01	-0.01	0.65	0.67

Appendix C

Supplementary material for: Metastable-Conformations vs. Hydration Shell

C.1 Markov state modelling

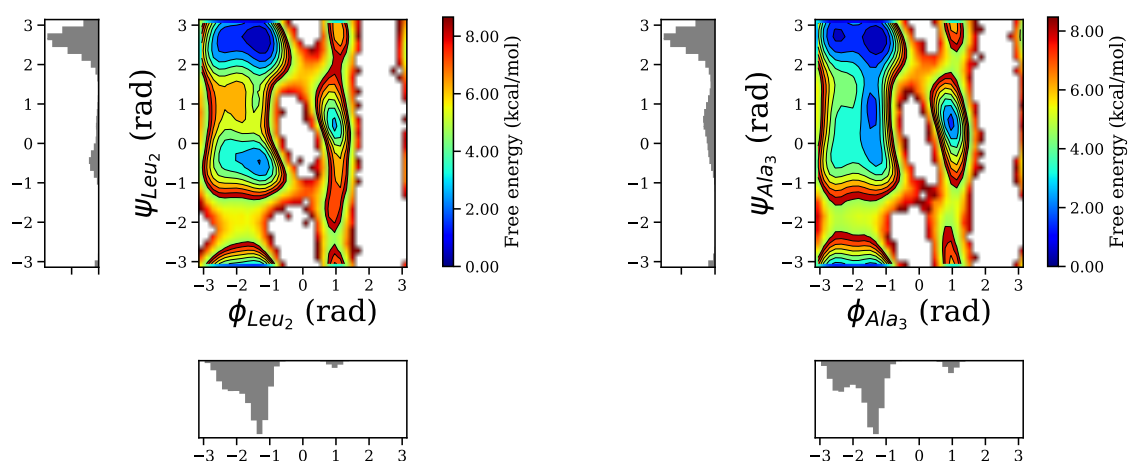


Figure C.1: Probability distribution of different backbone conformations as observed in the classical MD simulations, the first label refers to the first peptide bond, i.e., the ψ_{Leu2}, ϕ_{Leu2} -pair, and the second one to the second peptide bond, i.e., the ψ_{Ala3}, ϕ_{Ala3} -pair.

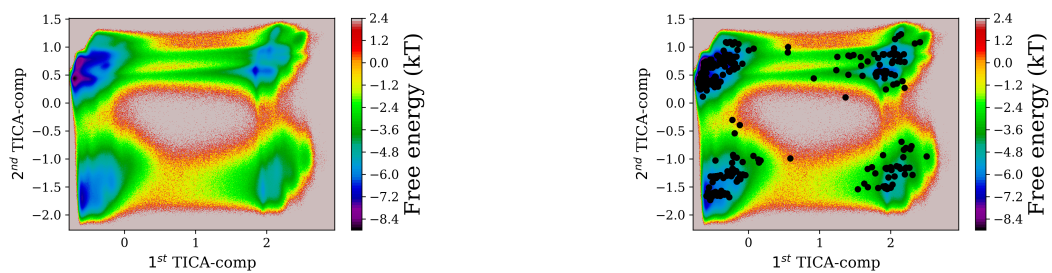


Figure C.2: (left) Free energy surface between first two TICA dimissions. (right) Distribution of k-means cluster centeriods on FES between the first two TICA dimissions.

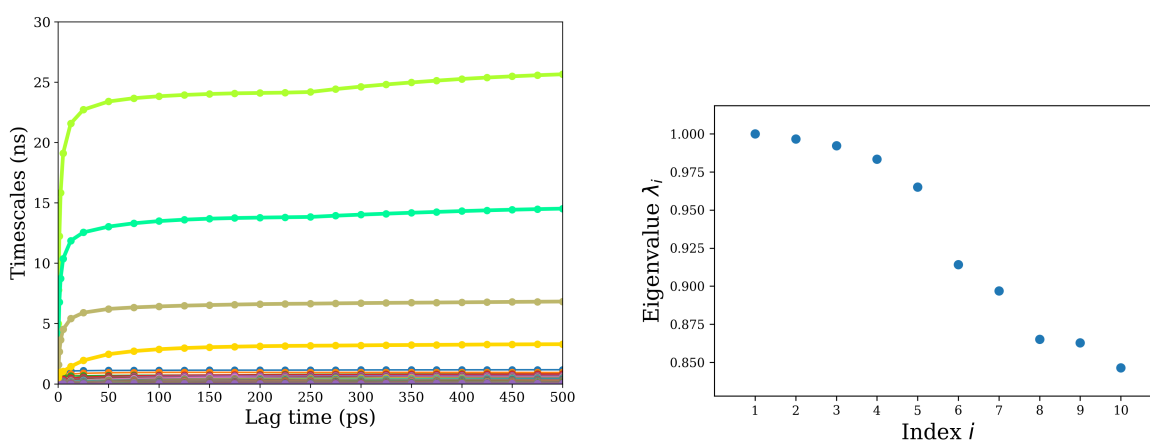


Figure C.3: (left) Implied time scales, (right) Eigenvalues of the transition matrix sampled with lag time $\tau = 100ps$.

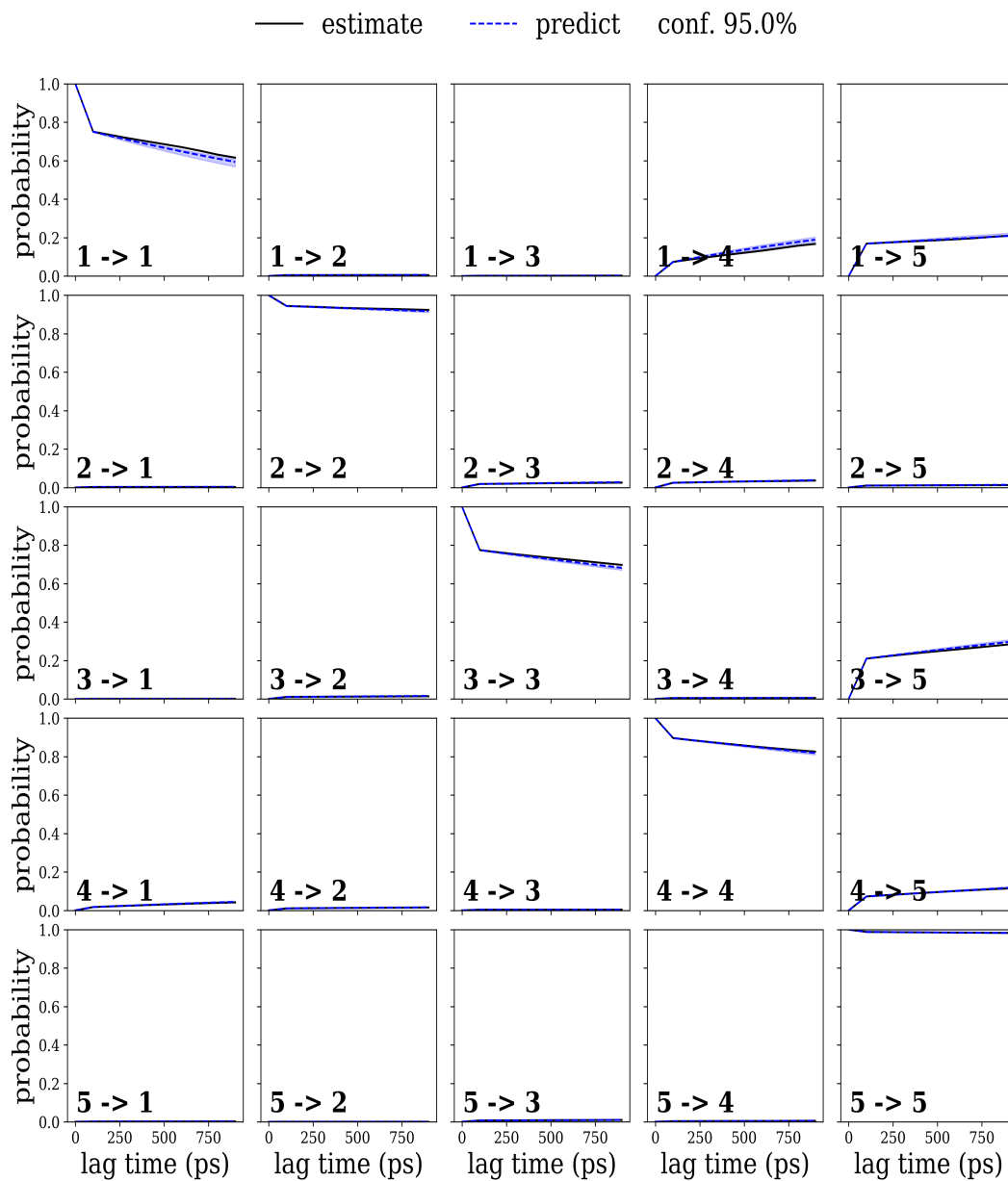


Figure C.4: Chapman-Kolmogorov test of the Markov state model (MSM) with four states 1,2,3,4,5 corresponding to meta-stable sets I,II,III,IV,V respectively.

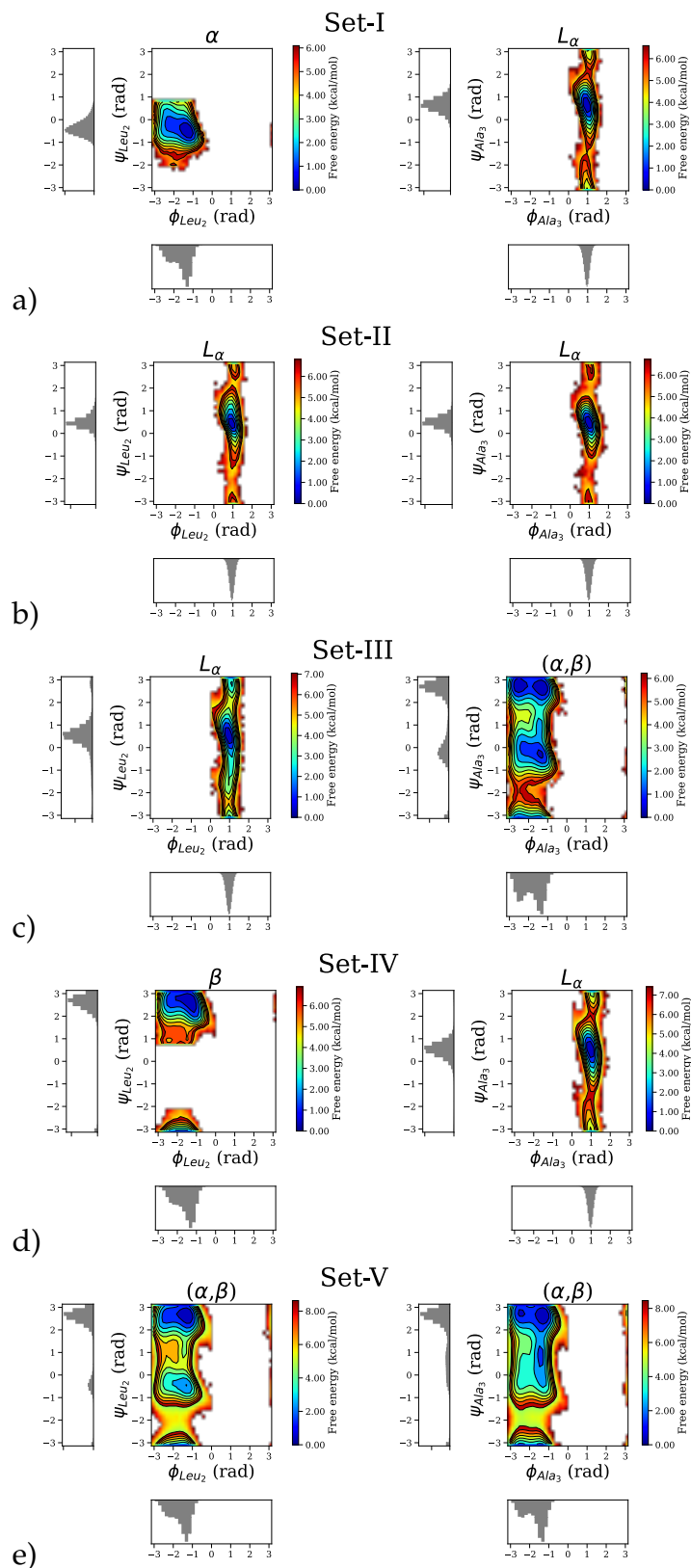


Figure C.5: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the representative structures of I, II, III, IV and V, metastable-sets.

Table C.2: Torsion angles of the representative conformation of each metastable-set used for the first-principle calculations.

rep. conformation	ψ_{Ala1}	$[\phi_{Leu2}-\psi_{Leu2}]$	χ_{Leu2}	$[\phi_{Ala3}-\psi_{Ala3}]$	ϕ_{Leu4}	χ_{Leu4}
Set-I (α, LHH)	2.45	-1.36 , -0.34	-1.20	1.02 , 0.35	-1.36	-1.00
Set-II (LHH, LHH)	2.43	1.06 , 0.29	-1.08	1.01 , 0.42	-1.60	-1.46
Set-III (LHH, α)	2.31	1.03 , 0.42	-1.06	-1.21 , -0.24	-2.09	-1.10
Set-III (LHH, β)	2.35	1.08 , 0.27	-0.86	-1.48 , 2.62	-2.12	-1.06
Set-IV (β, LHH)	2.41	-1.39 , 2.57	-1.25	1.04 , 0.46	-1.83	-1.46
Set-V (α, α)	2.38	-2.12 , -0.40	-1.01	-1.61 , -0.05	-1.44	-1.09
Set-V (α, β)	2.46	-1.36 , -0.40	-1.05	-1.61 , 2.62	-1.95	-1.24
Set-V (β, α)	2.49	-1.37 , 2.74	-1.22	-1.29 , -0.63	-2.26	-1.28
Set-V (β, β)	2.46	-1.76 , 2.55	-1.31	-1.42 , 2.62	-1.72	-1.13

C.2 Normal modes

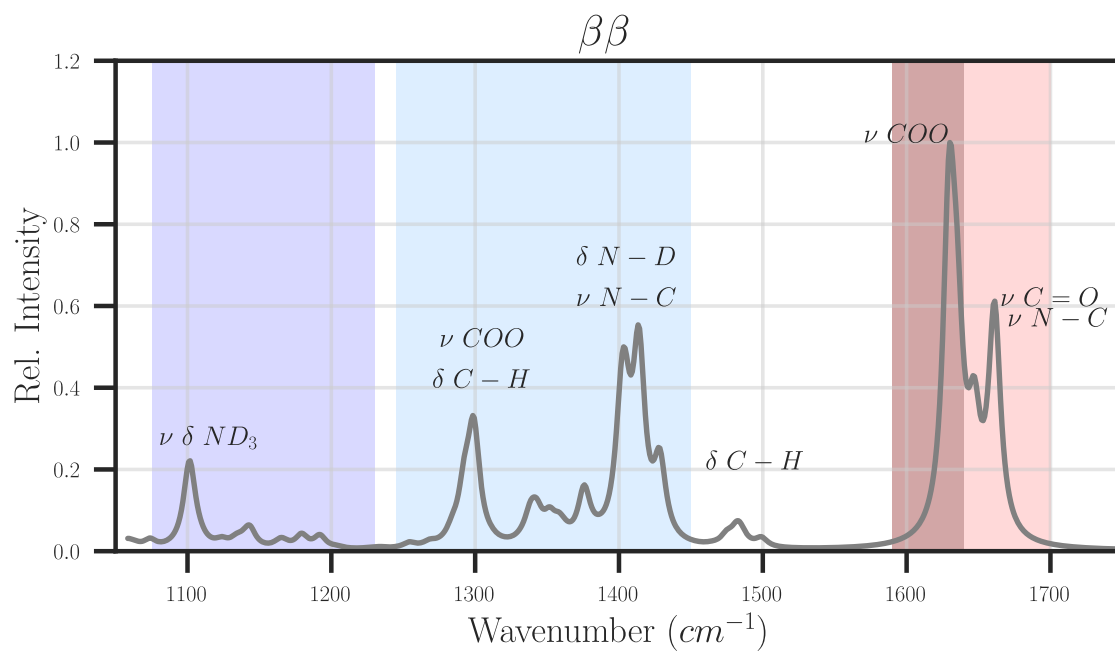


Figure C.6: Normal modes spectra of (β, β)-conformation of ALAL.

C.3 Torsion angle distributions

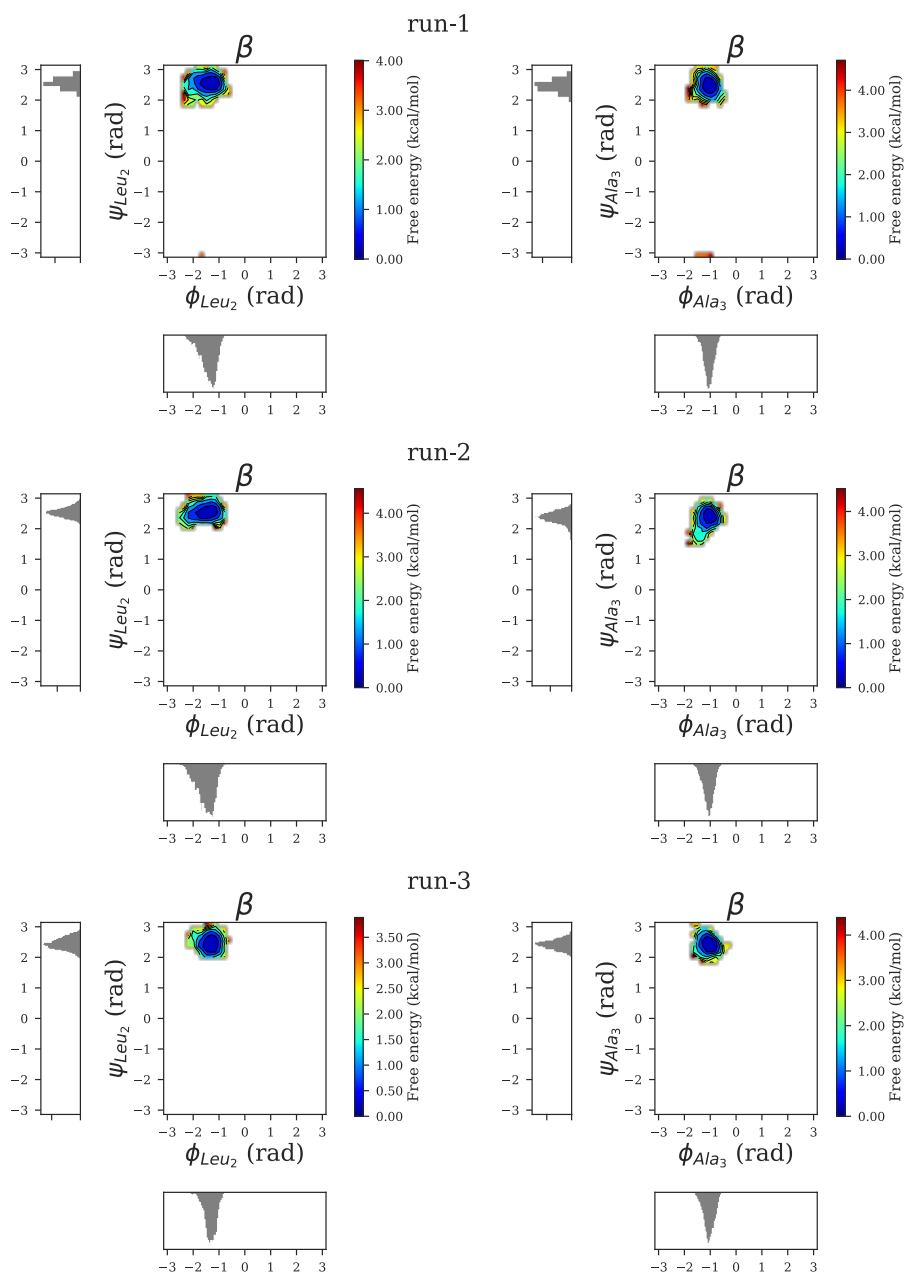


Figure C.7: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (β, β) -conformation.

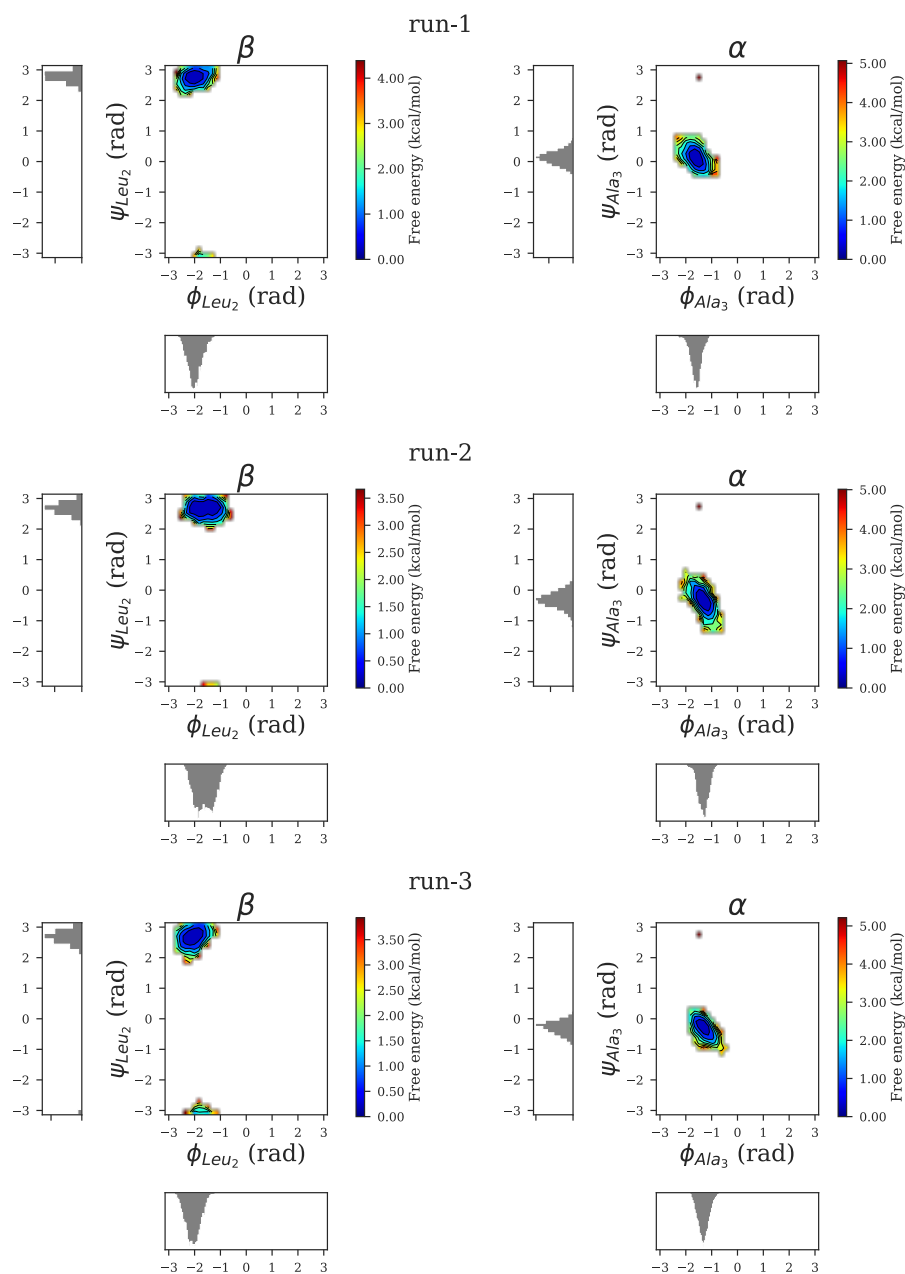


Figure C.8: Ramachandran plot of the torsion angles $[\phi_{Leu_2}-\psi_{Leu_2}]$, $[\phi_{Ala_3}-\psi_{Ala_3}]$ extracted from the first-principle simulations of (β, α) -conformation.

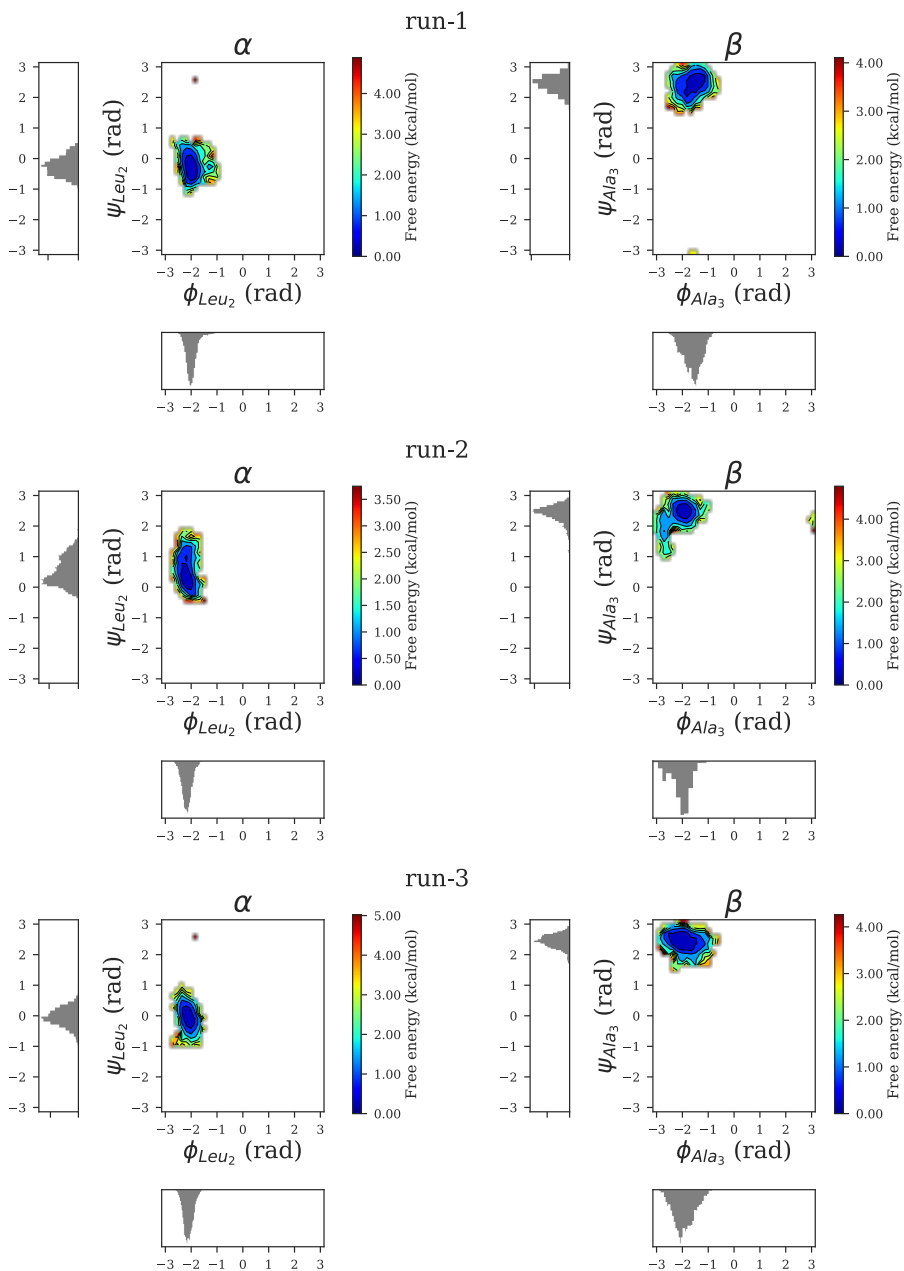


Figure C.9: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (α, β) -conformation.

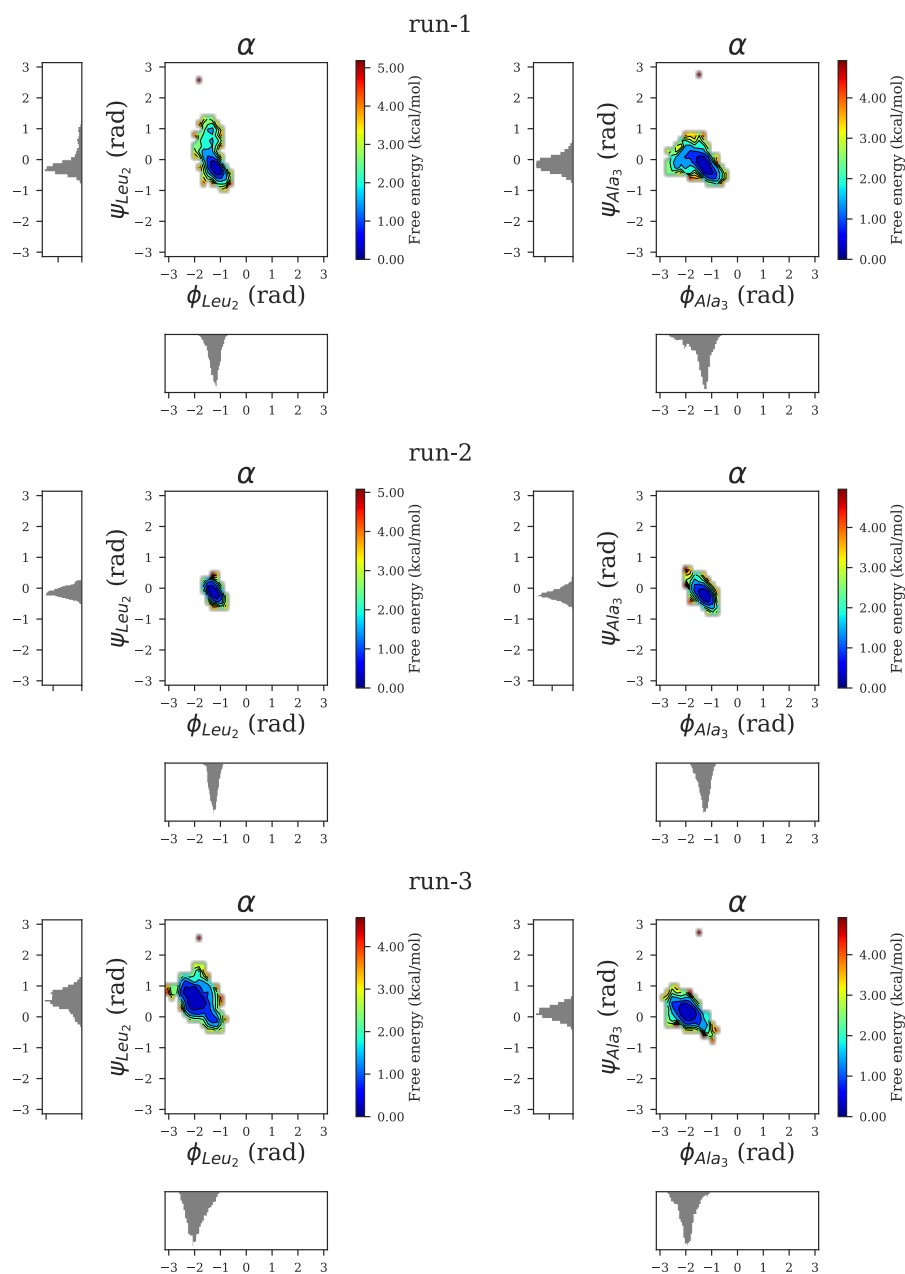


Figure C.10: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (α, α) -conformation.

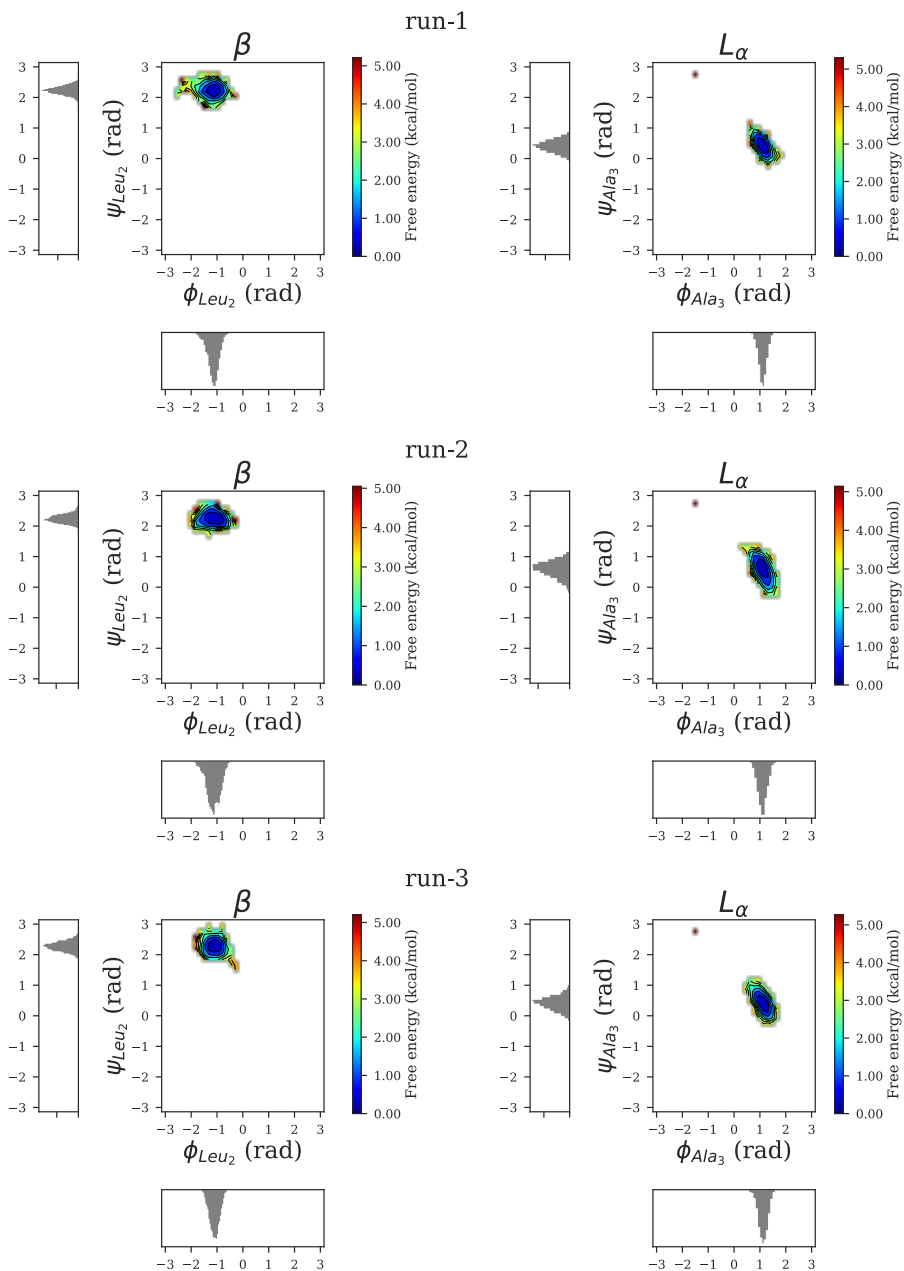


Figure C.11: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (β, L_α) -conformation.

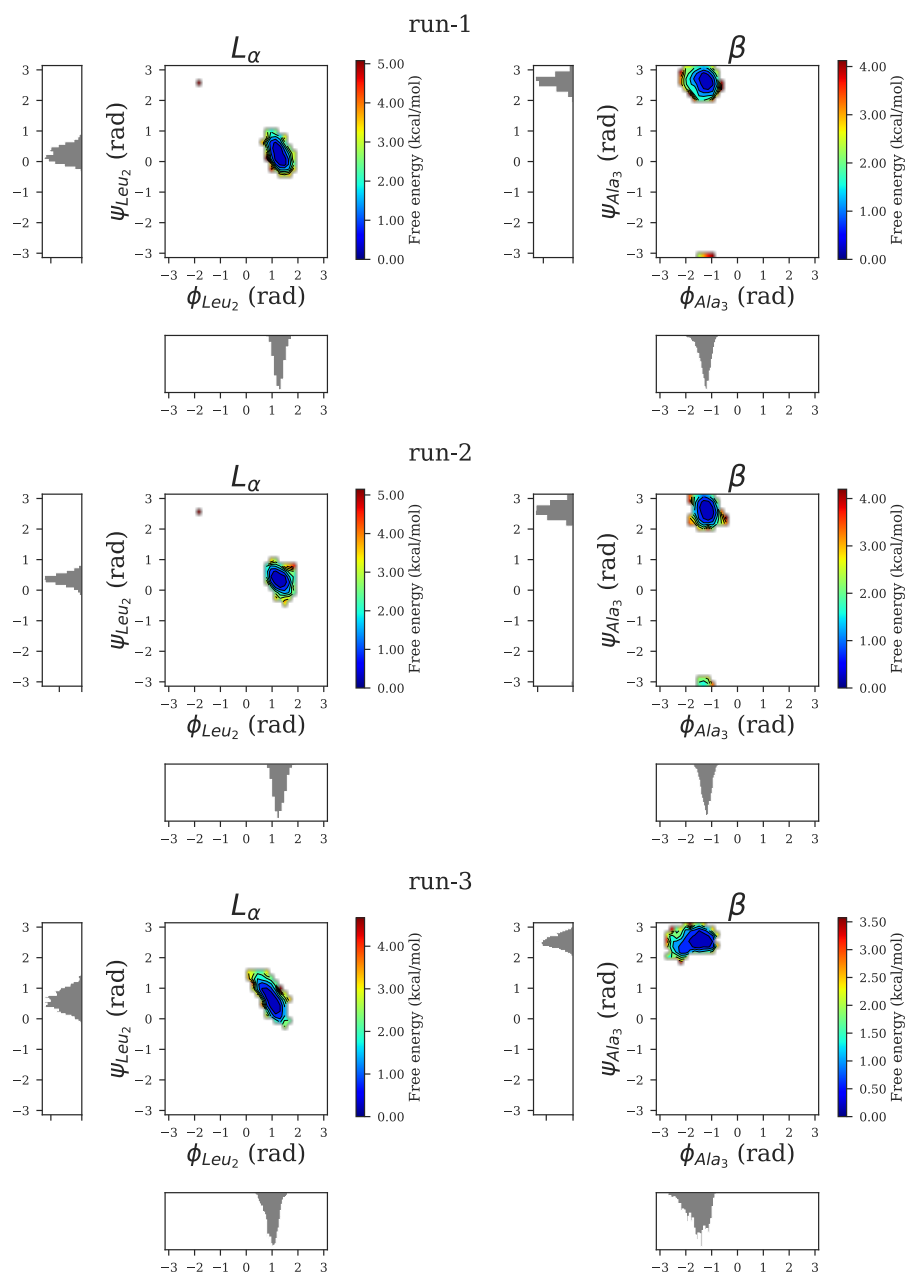


Figure C.12: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (L_α, β) -conformation.

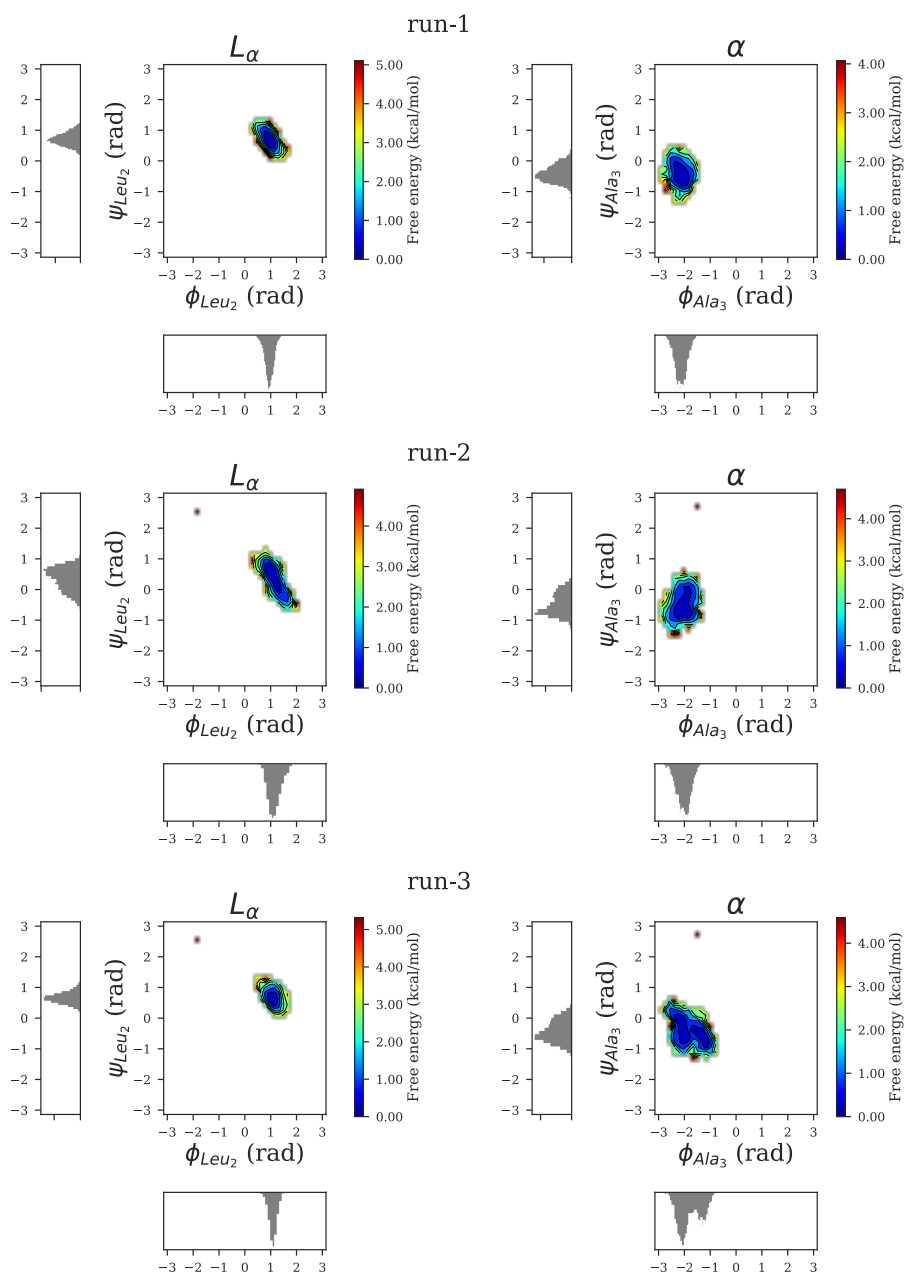


Figure C.13: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (L_α, α) -conformation.

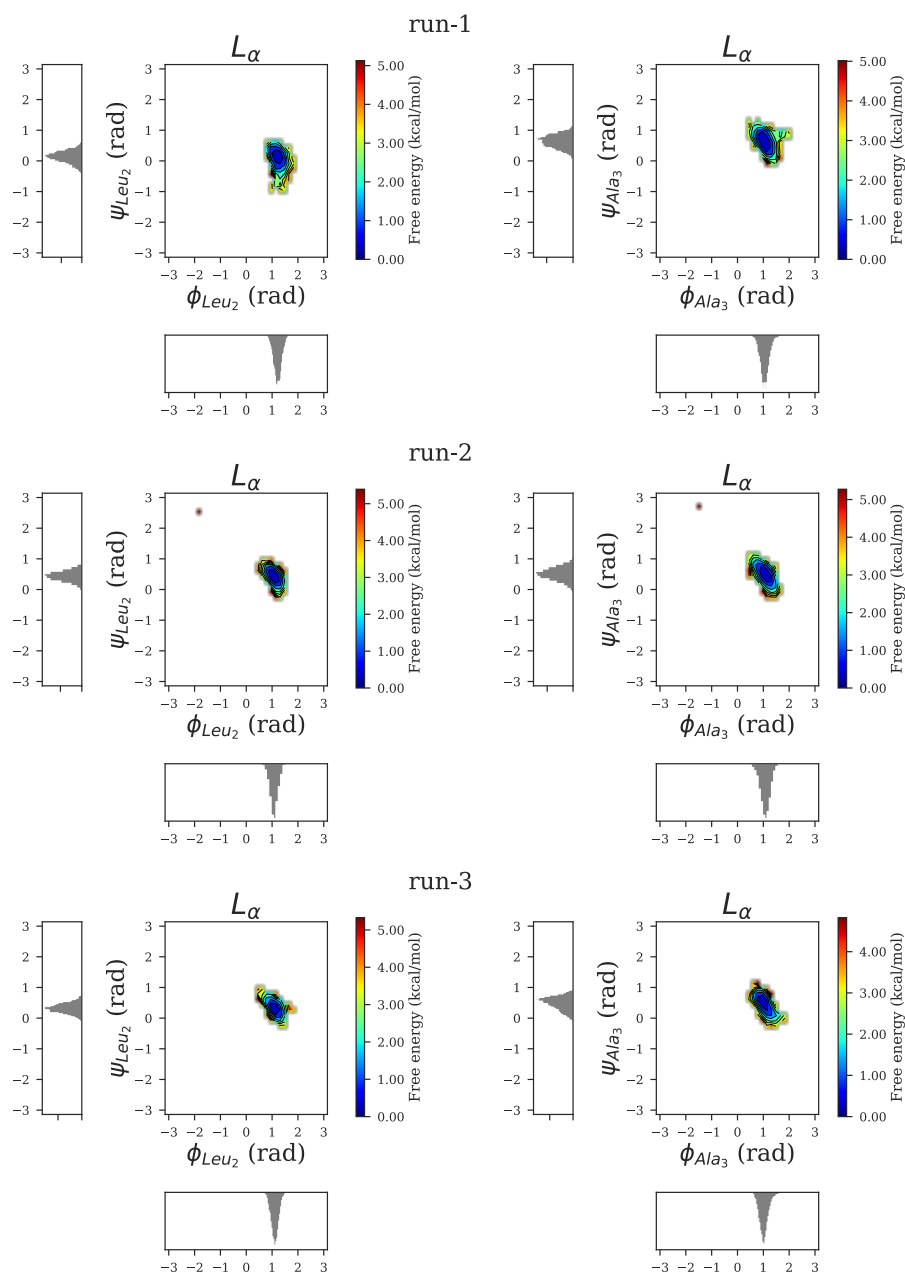


Figure C.14: Ramachandran plot of the torsion angles $[\phi_{Leu2}-\psi_{Leu2}]$, $[\phi_{Ala3}-\psi_{Ala3}]$ extracted from the first-principle simulations of (L_α, L_α) -conformation.

C.4 Root mean square fluctuations

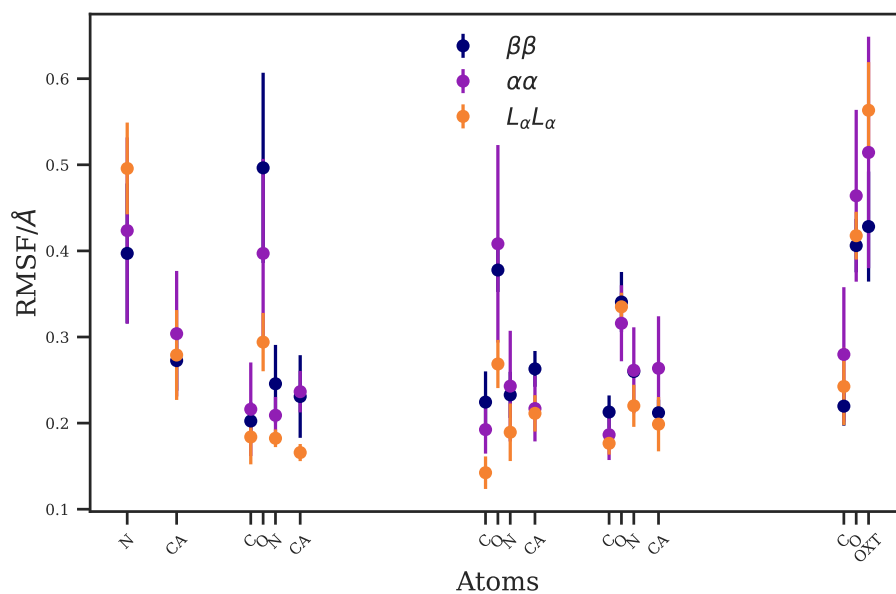


Figure C.15: Root mean square fluctuations of backbone atoms from first-principle calculations.

C.5 Radial distribution functions

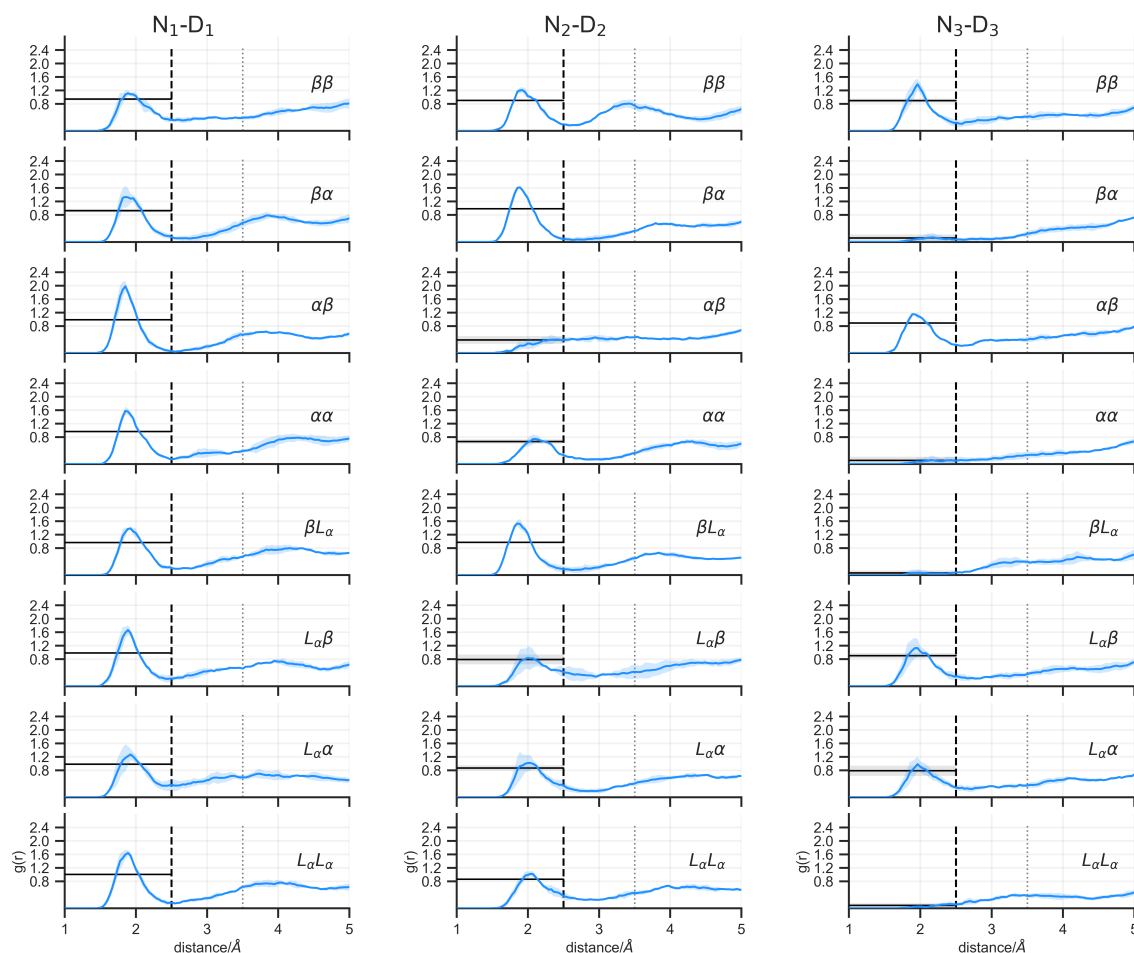


Figure C.16: Mean RDF plot with shaded region as standard deviation of N-D groups (first,second,third from left to right) , vertical dashed line 2.5Å (black) around first minimum represents the 1st hydration shell i.e., strongly interacting zone., (note that it also the defined geometric criteria for hydrogen bond calculations), other vertical dotted line at 3.5Å (gray) represents so called moderately interacting zone. Horizontal black line represents the mean coordination number with gray shaded region as error band.

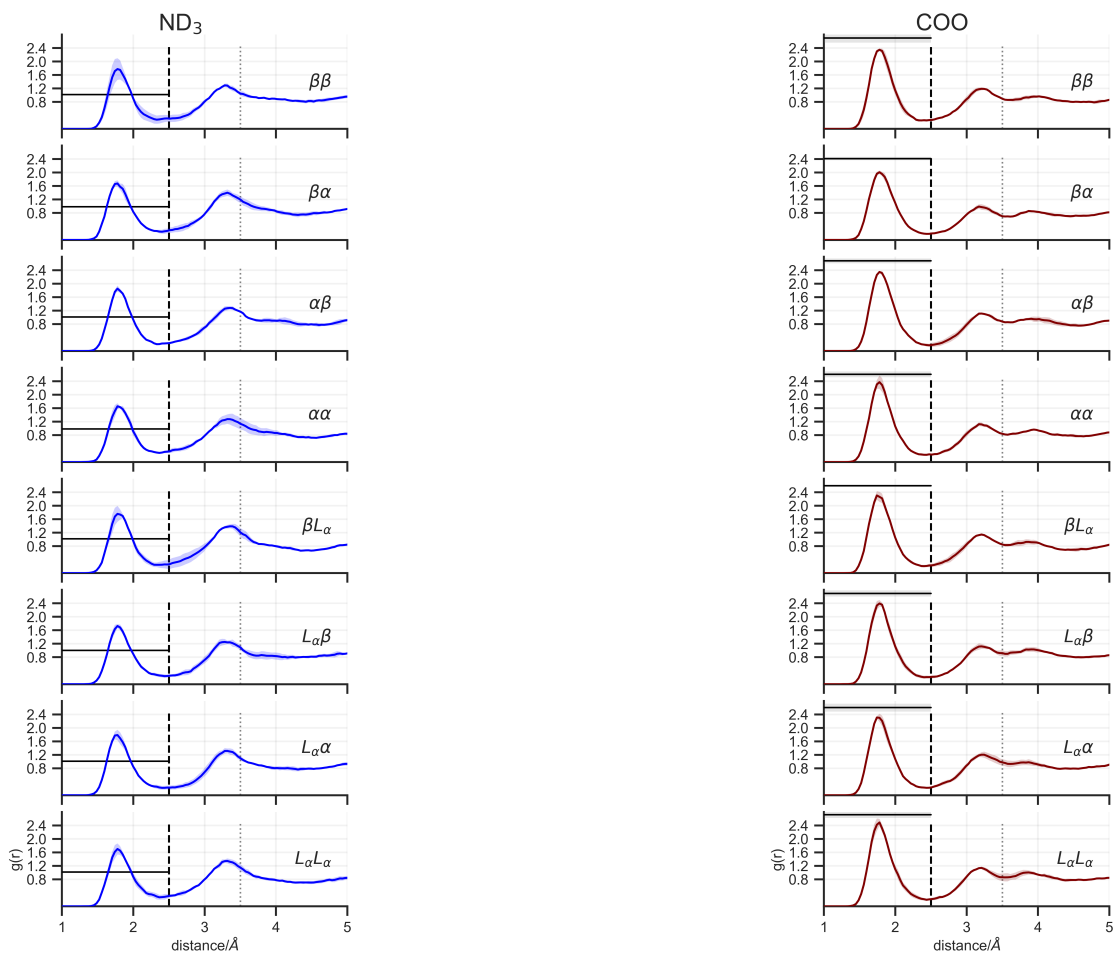


Figure C.17: Mean RDF plot with shaded region as standard deviation of terminal groups ND₃ (left) and COO (right) , vertical dashed line 2.5Å (black) around first minimum represents the 1st hydration shell i.e., strongly interacting zone., (note that it also the defined geometric criteria for hydrogen bond calculations), other vertical dotted line at 3.5Å (gray) represents so called moderately interacting zone. Horizontal black line represents the mean coordination number with gray shaded region as error band.

C.6 Angular distribution functions

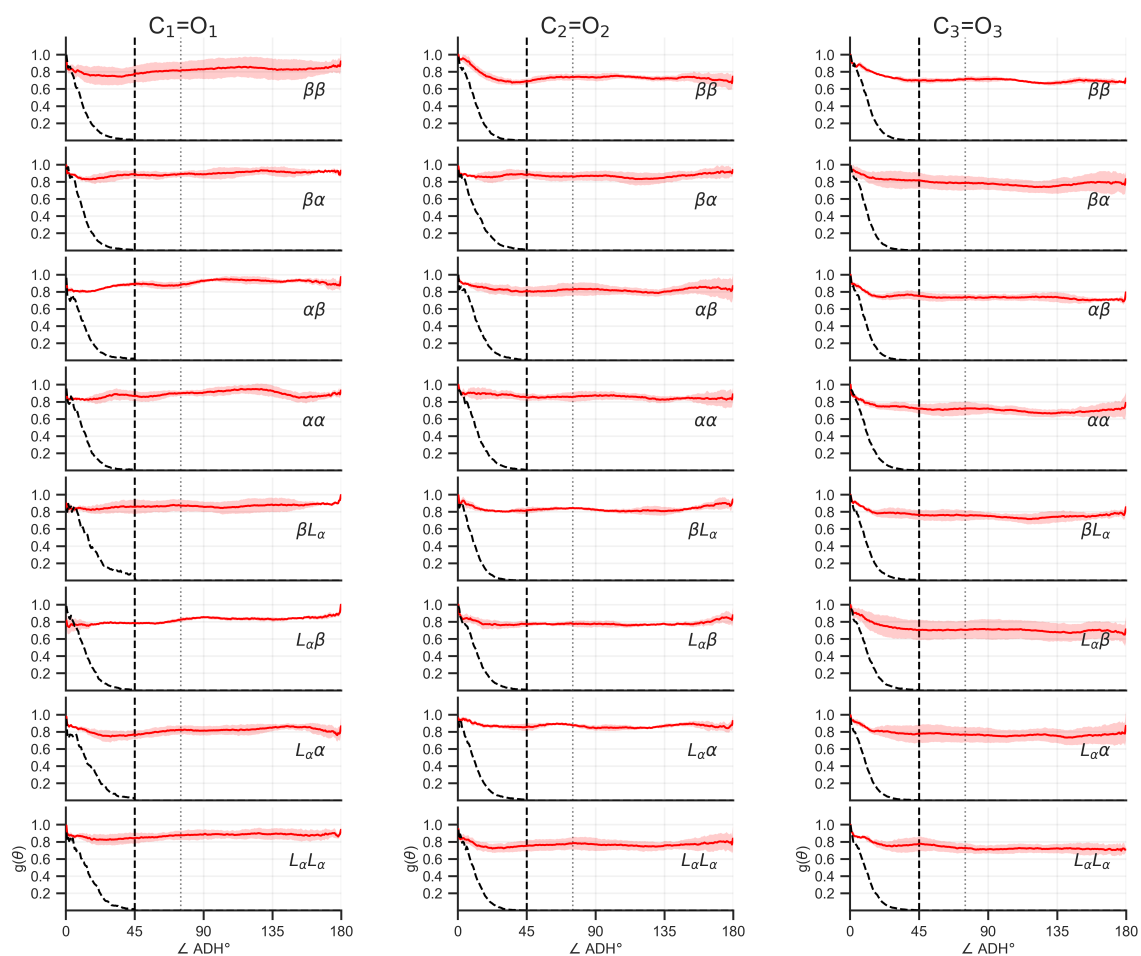


Figure C.18: Mean ADF plot with shaded region as standard deviation of all carbonyl groups (first,second,third from left to right) , vertical dashed line 45° (black) is to represent the more likely hydrogen bond angle values of strongly interacting zone (0-45°) while other vertical dotted line at 75° (gray) shows moderately interacting zone (45° - 75°)

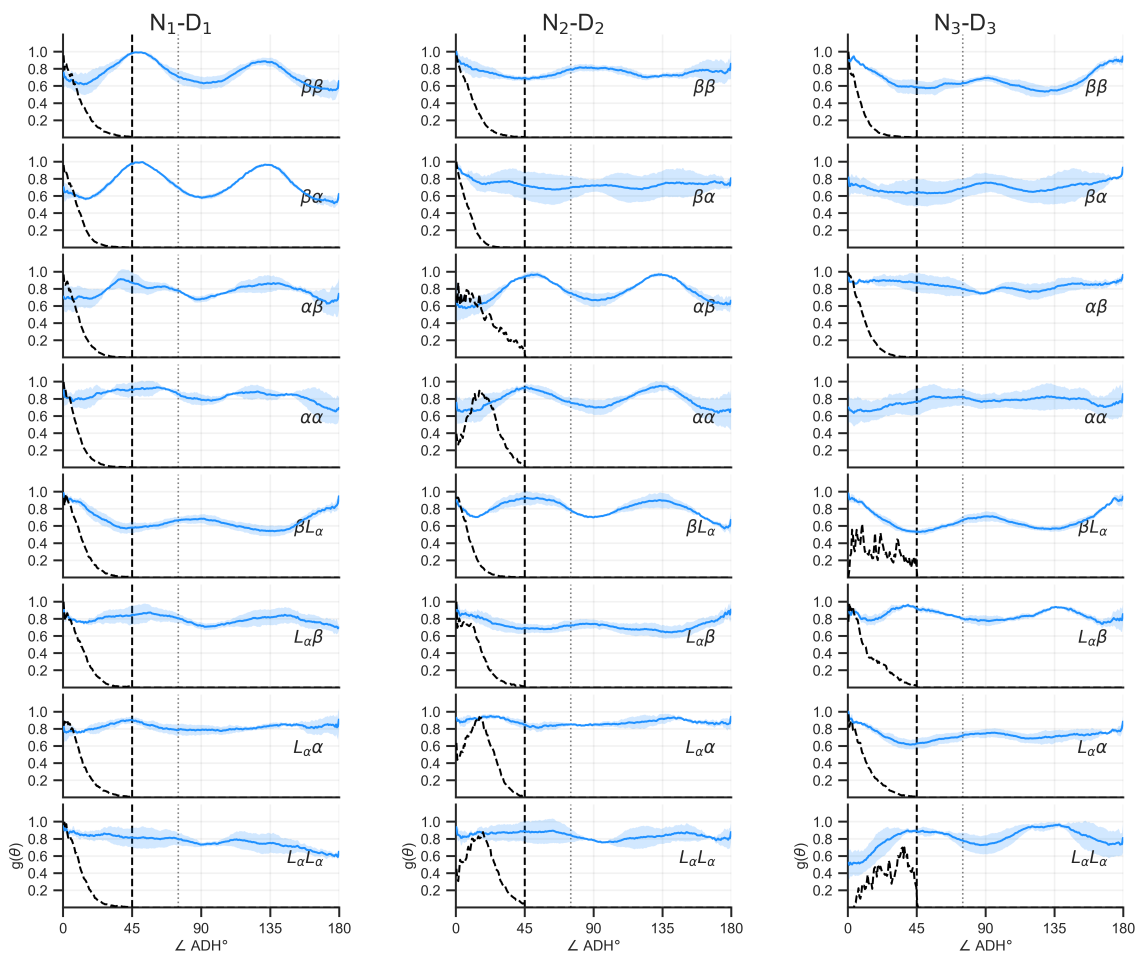


Figure C.19: Mean ADF plot with shaded region as standard deviation of all N-D groups (first,second,third from left to right) , vertical dashed line 45° (black) is to represent the more likely hydrogen bond angle values of strongly interacting zone ($0-45^\circ$) while other vertical dotted line at 75° (gray) shows moderately interacting zone ($45^\circ - 75^\circ$)

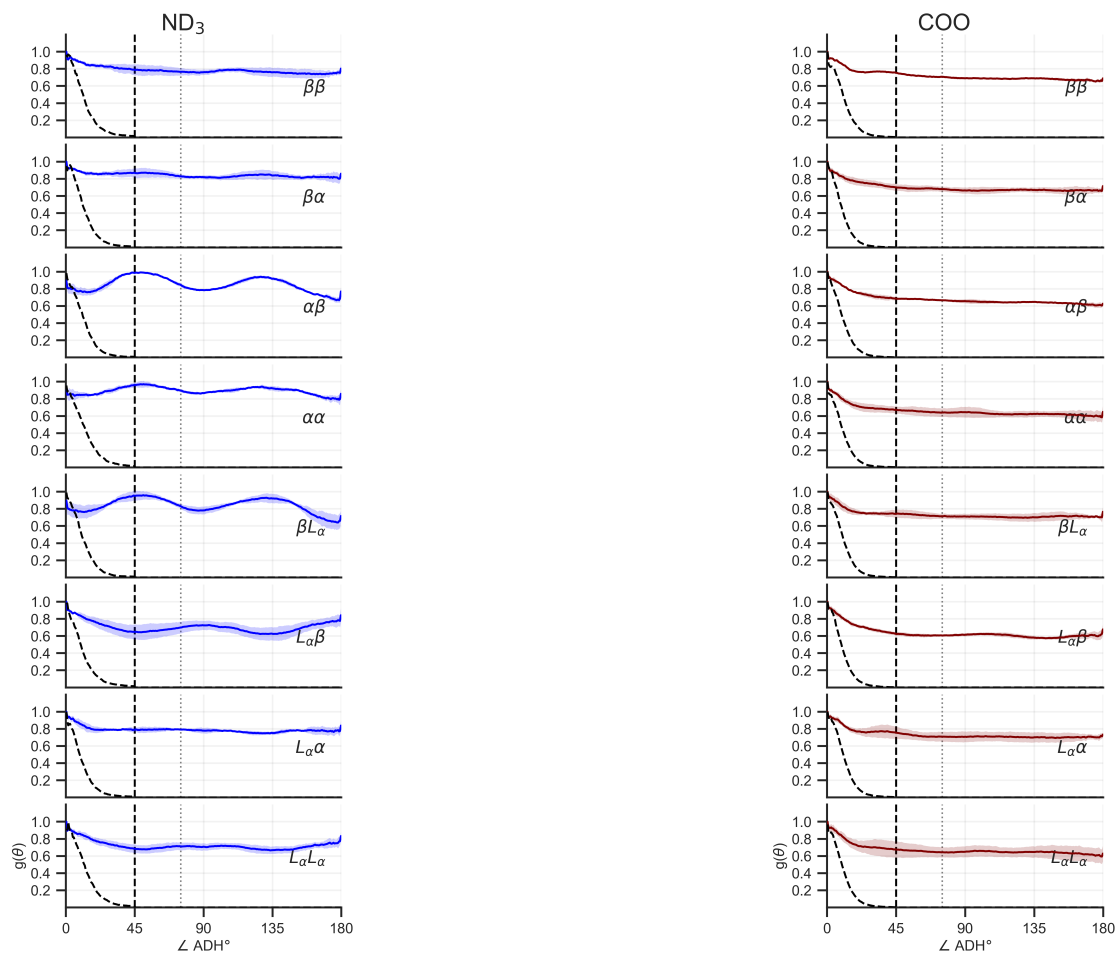


Figure C.20: Mean ADF plot with shaded region as standard deviation of terminal groups, vertical dashed line 45° (black) is to represent the more likely hydrogen bond angle values of strongly interacting zone (0-45 degree) while other vertical dotted line at 75° (gray) shows moderately interacting zone (45-75 degree)

C.7 Combined distribution functions

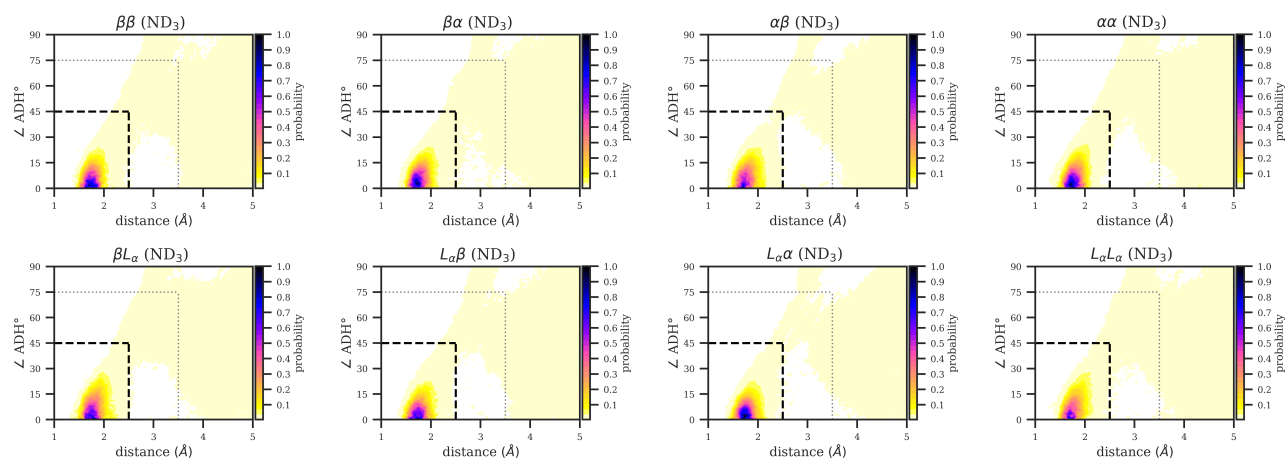


Figure C.21: CDF of N-terminal group using data from all three runs of each conformation of ALAL.

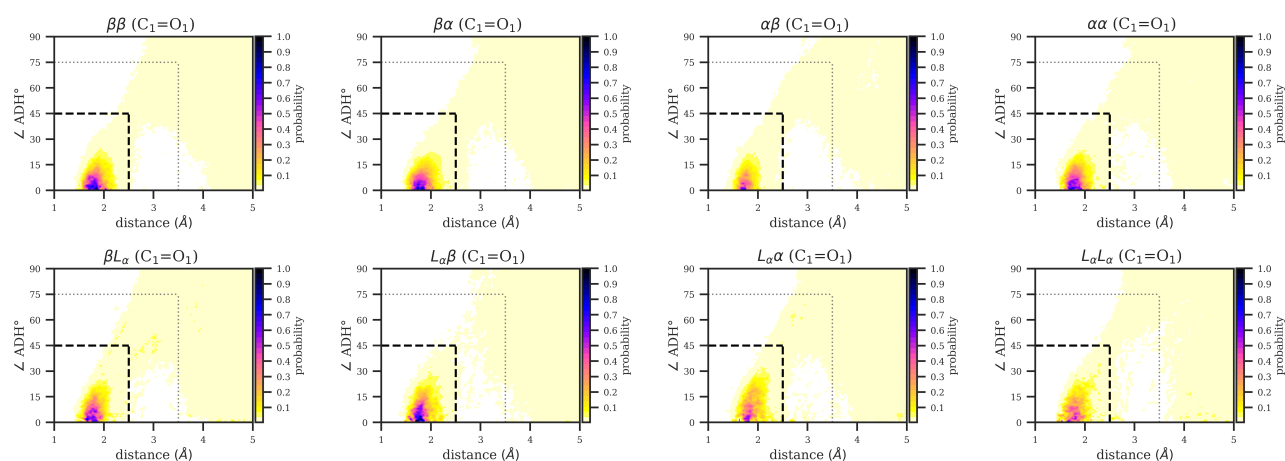


Figure C.22: CDF of first carbonyl group using data from all three runs of each conformation of ALAL.

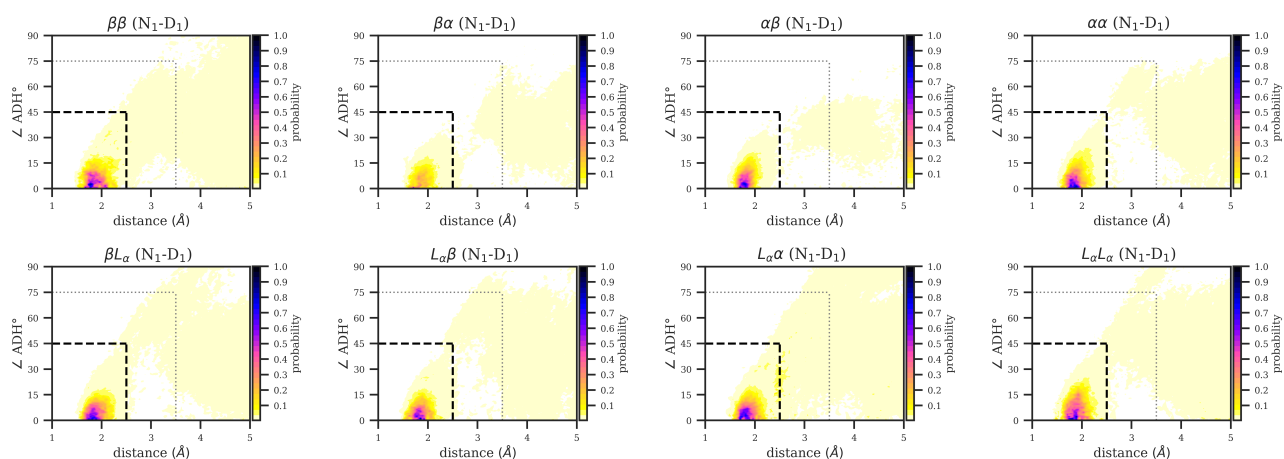


Figure C.23: CDF of first amine group using data from all three runs of each conformation of ALAL.

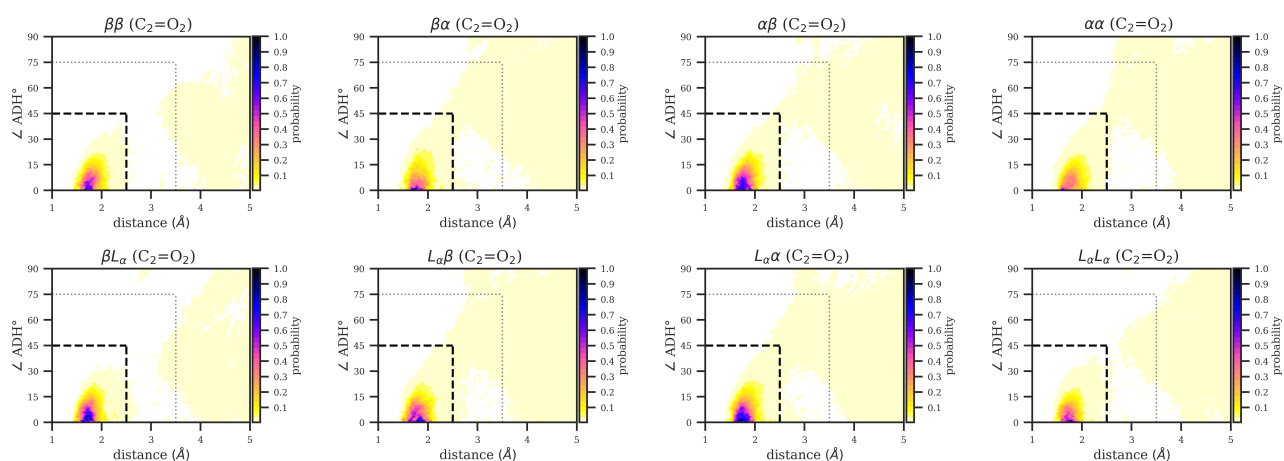


Figure C.24: CDF of second carbonyl group using data from all three runs of each conformation of ALAL.

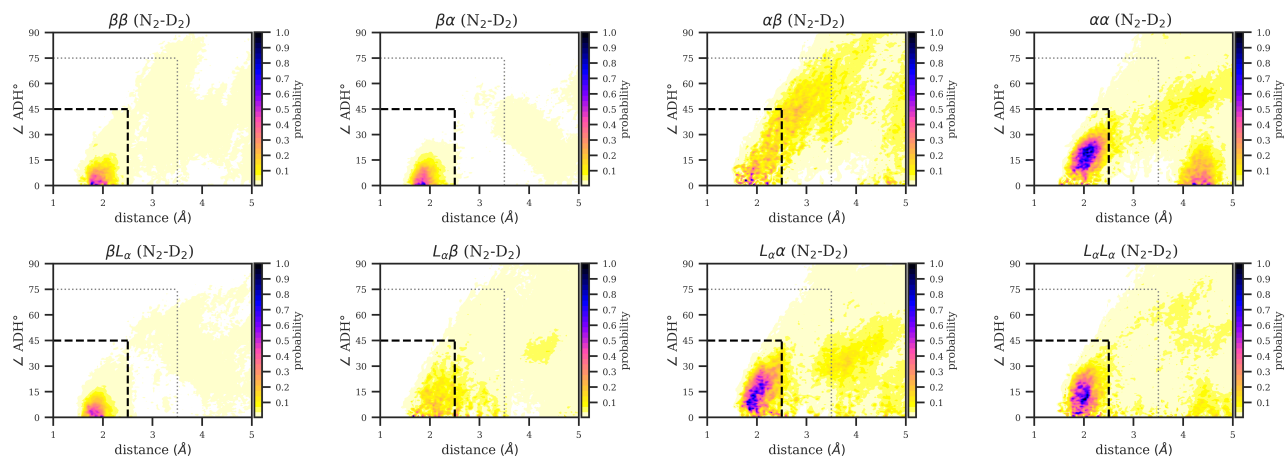


Figure C.25: CDF of second amine group using data from all three runs of each conformation of ALAL.

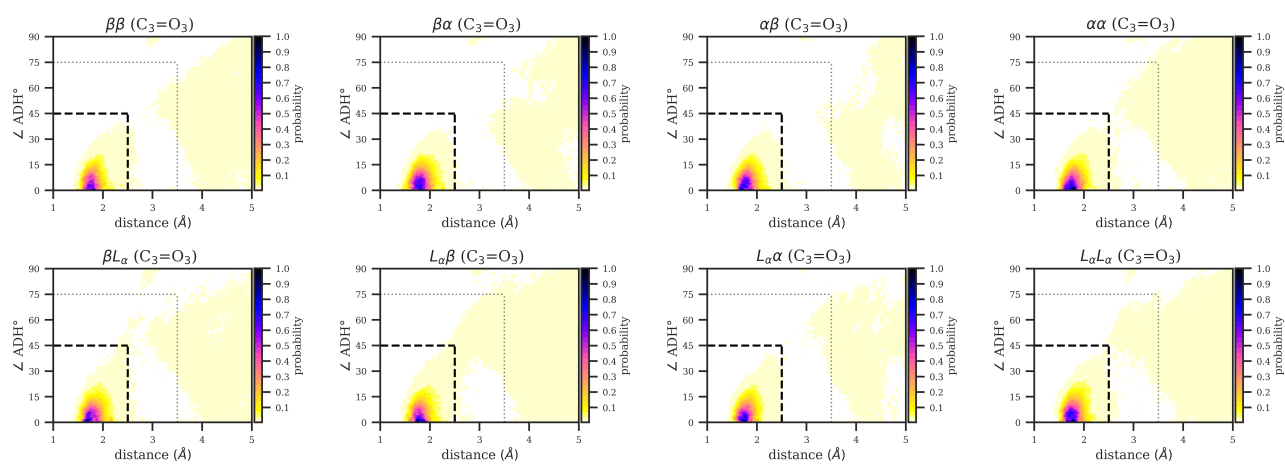


Figure C.26: CDF of third carbonyl group using data from all three runs of each conformation of ALAL.

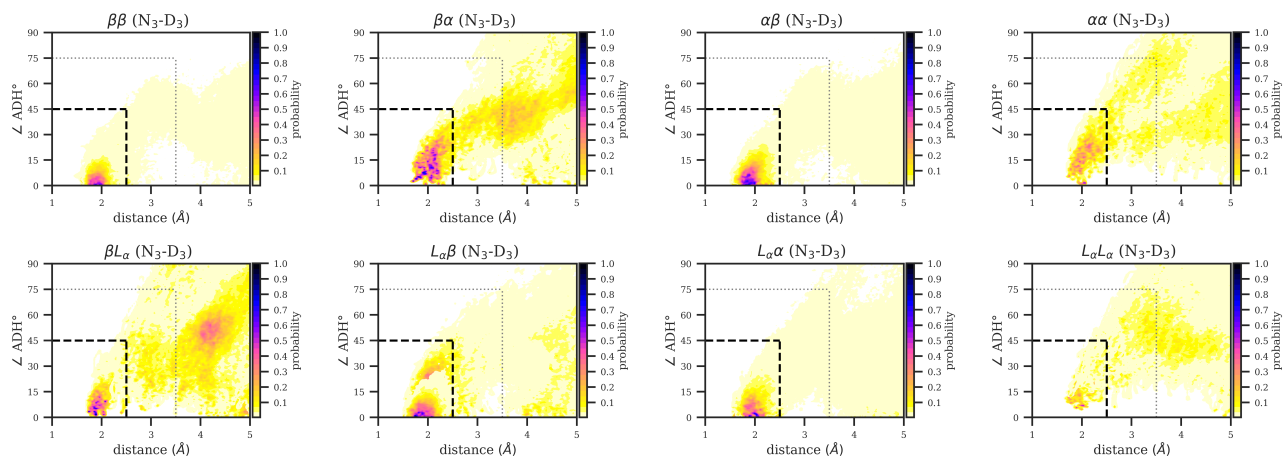


Figure C.27: CDF of third amine group using data from all three runs of each conformation of ALAL.

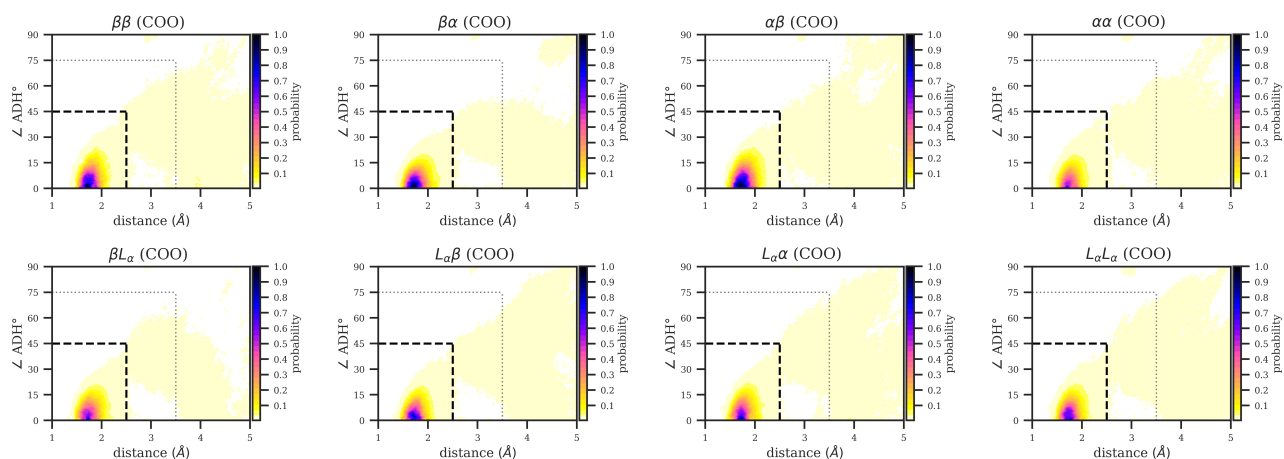


Figure C.28: CDF of C-terminal group using data from all three runs of each conformation of ALAL.

C.8 Hydrogen bonds

Table C.11: Average number of hbonds per polar group.

	ND ₃ ⁺	C ₁ =O ₁	N ₁ -D ₁	C ₂ =O ₂	N ₂ -D ₂	C ₃ =O ₃	N ₃ -D ₃	COO ⁻
β,β	2.78±0.17	1.38±0.31	0.91±0.02	1.78±0.08	0.87±0.04	2.06±0.19	0.89±0.09	5.26±0.33
β,α	2.74±0.07	1.42±0.08	0.92±0.06	1.26±0.29	0.99±0.01	1.98±0.04	0.11±0.12	4.81±0.13
α,β	2.85±0.01	1.04±0.18	0.97±0.02	1.79±0.21	0.34±0.14	1.95±0.12	0.89±0.02	5.28±0.11
α,α	2.64±0.1	0.81±0.09	0.94±0.05	1.61±0.11	0.59±0.07	1.98±0.09	0.09±0.12	5.09±0.19
β,L_α	2.86±0.04	0.74±0.26	0.93±0.02	1.3±0.29	0.95±0.02	1.76±0.07	0.06±0.08	5.12±0.14
L_α,β	2.68±0.11	1.22±0.39	0.95±0.03	1.56±0.21	0.78±0.2	2.08±0.09	0.86±0.17	5.29±0.22
L_α,α	2.78±0.12	0.98±0.27	0.95±0.04	1.69±0.34	0.82±0.1	2.02±0.05	0.75±0.17	5.07±0.28
L_α,L_α	2.78±0.04	0.75±0.34	0.96±0.02	1.14±0.14	0.76±0.08	1.81±0.37	0.05±0.07	5.31±0.19

Table C.12: Probability and type of intramolecular hydrogen bonds present in the individual first principle simulations of metastable-conformations of ALAL.

Conformation	Simulation run		
	1	2	3
β,β	-	-	0.00242 C ₂ =O ₂ ··· N ₃ =D ₃
β,α	0.0121 C ₂ =O ₂ ··· N ₃ =D ₃	0.00092 C ₂ =O ₂ ··· N ₃ =D ₃	-
α,β	-	0.00258 C ₁ =O ₁ ··· N ₂ =D ₂	-
α,α	0.4702 C ₁ =O ₁ ··· N ₃ =D ₃	0.7336 C ₁ =O ₁ ··· N ₃ =D ₃	0.0519 C ₁ =O ₁ ··· N ₃ =D ₃
	0.0936 C ₁ =O ₁ ··· N ₂ =D ₂	-	0.0642 C ₁ =O ₁ ··· N ₂ =D ₂
	0.0086 C ₂ =O ₂ ··· N ₃ =D ₃	0.0010 C ₂ =O ₂ ··· N ₃ =D ₃	-
	-	-	-
β,L_α	0.6464 C ₁ =O ₁ ··· N ₃ =D ₃	0.2783 C ₁ =O ₁ ··· N ₃ =D ₃	0.5124 C ₁ =O ₁ ··· N ₃ =D ₃
	-	-	0.0007 C ₂ =O ₂ ··· N ₃ =D ₃
L_α,β	-	-	0.0021 C ₁ =O ₁ ··· N ₂ =D ₂
L_α,α	-	0.0053 C ₁ =O ₁ ··· N ₂ =D ₂	-
L_α,L_α	0.1654 C ₁ =O ₁ ··· N ₃ =D ₃	0.8846 C ₁ =O ₁ ··· N ₃ =D ₃	0.8436 C ₁ =O ₁ ··· N ₃ =D ₃
	0.0258 C ₁ =O ₁ ··· N ₂ =D ₂	-	-

C.9 Power spectra

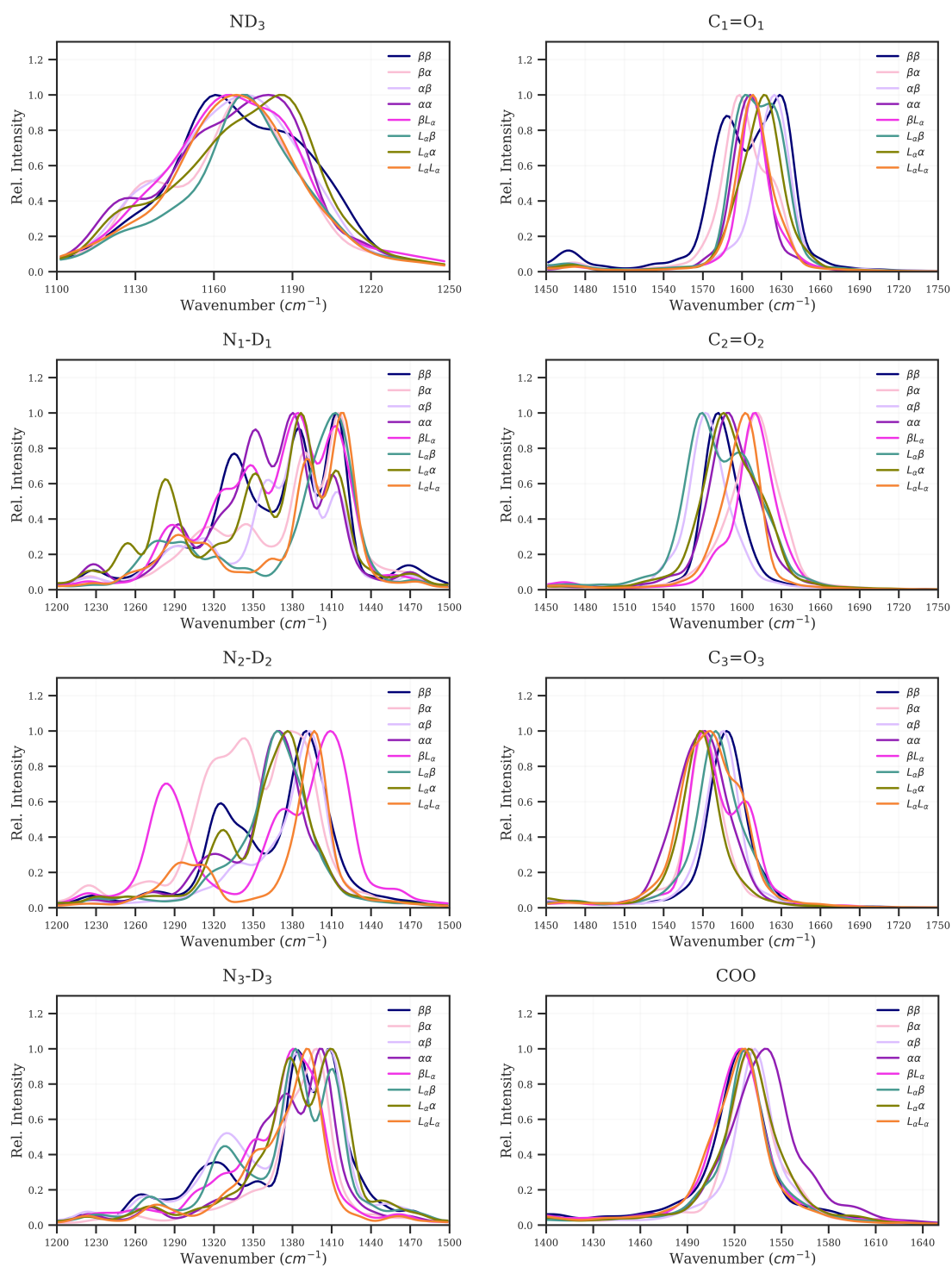


Figure C.29: Group wise power spectra of all conformations of ALAL.

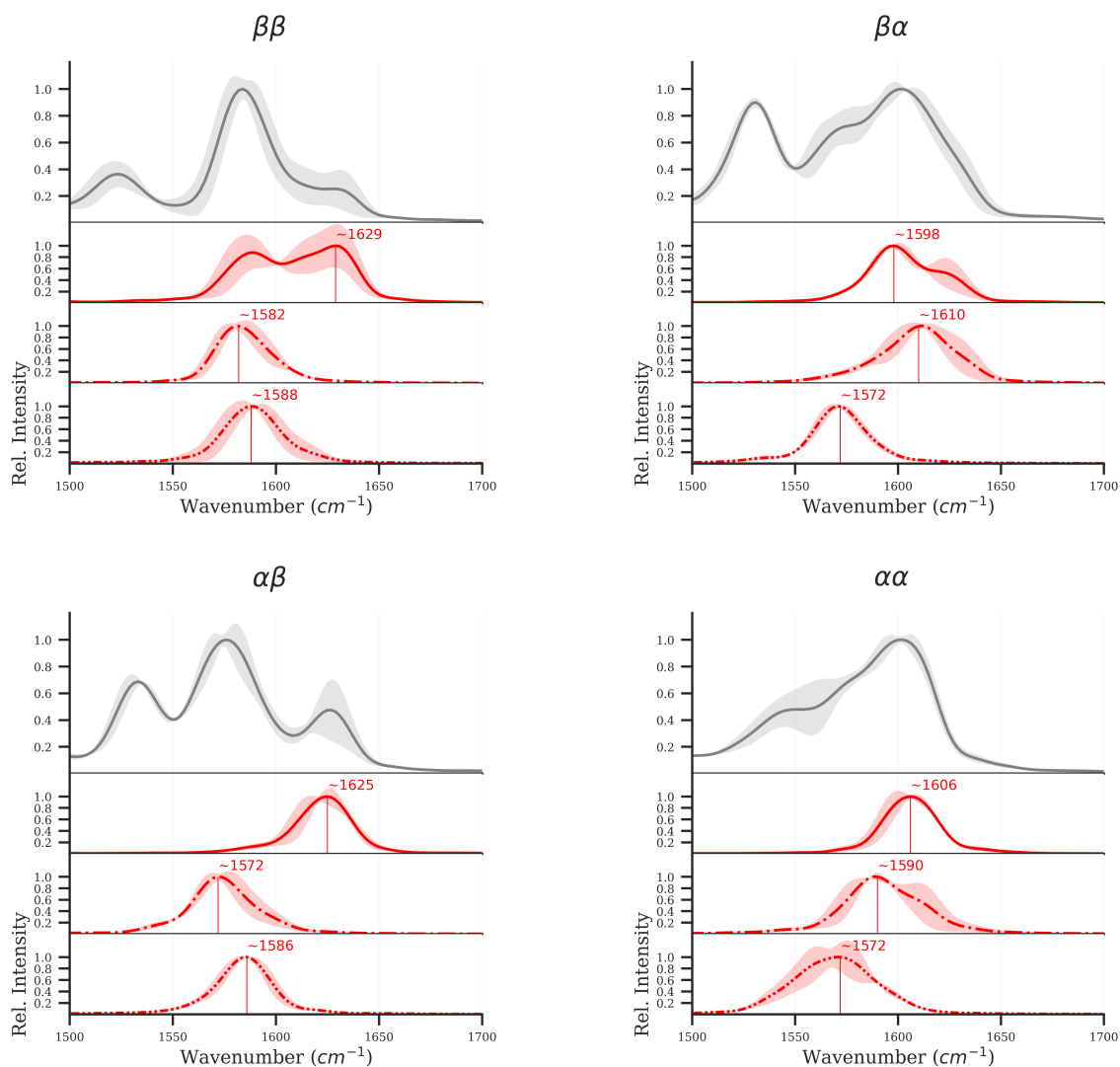


Figure C.30: Density based infrared spectra of representative conformations of Set-V of ALAL, in the amide I range (each upper panel) with power spectra of each carbonyl group (each below panels).

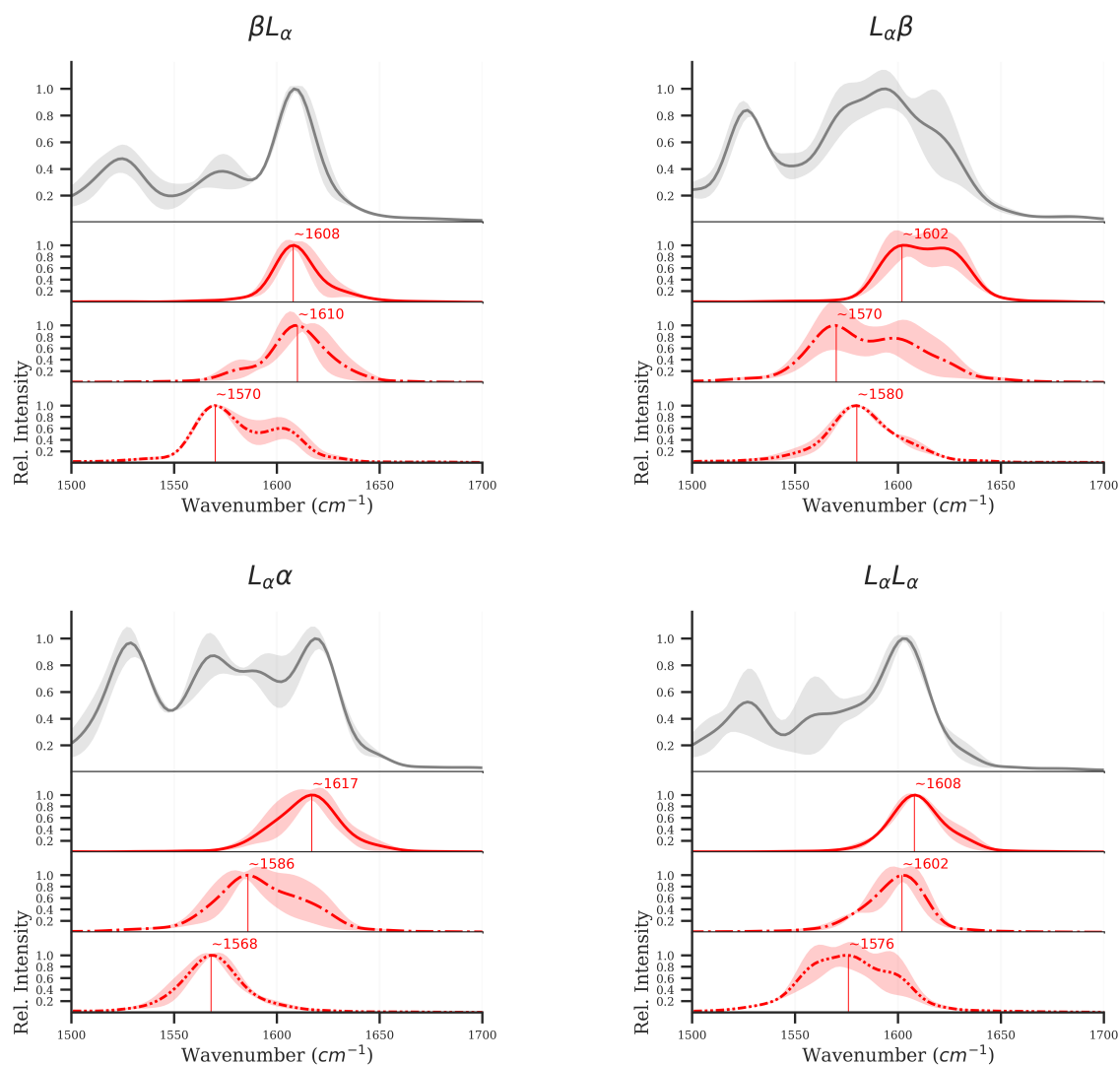


Figure C.31: Density based infrared spectra of representative conformations of Set-II to Set-IV of ALAL, in the amide I range (each upper panel) with power spectra of each carbonyl group (each below panels).

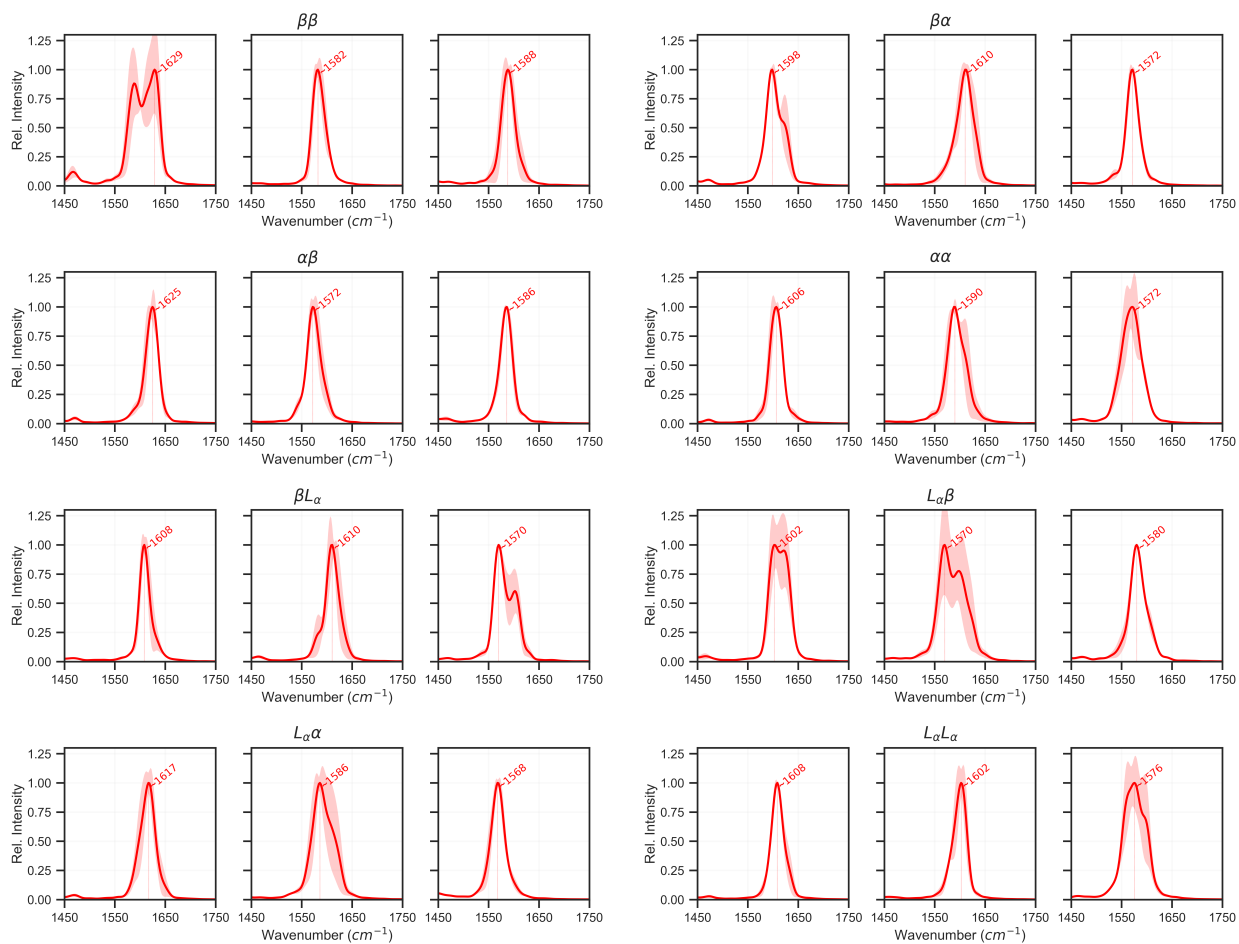


Figure C.32: Power spectra of carbonyl groups with SD of all conformations of ALAL.

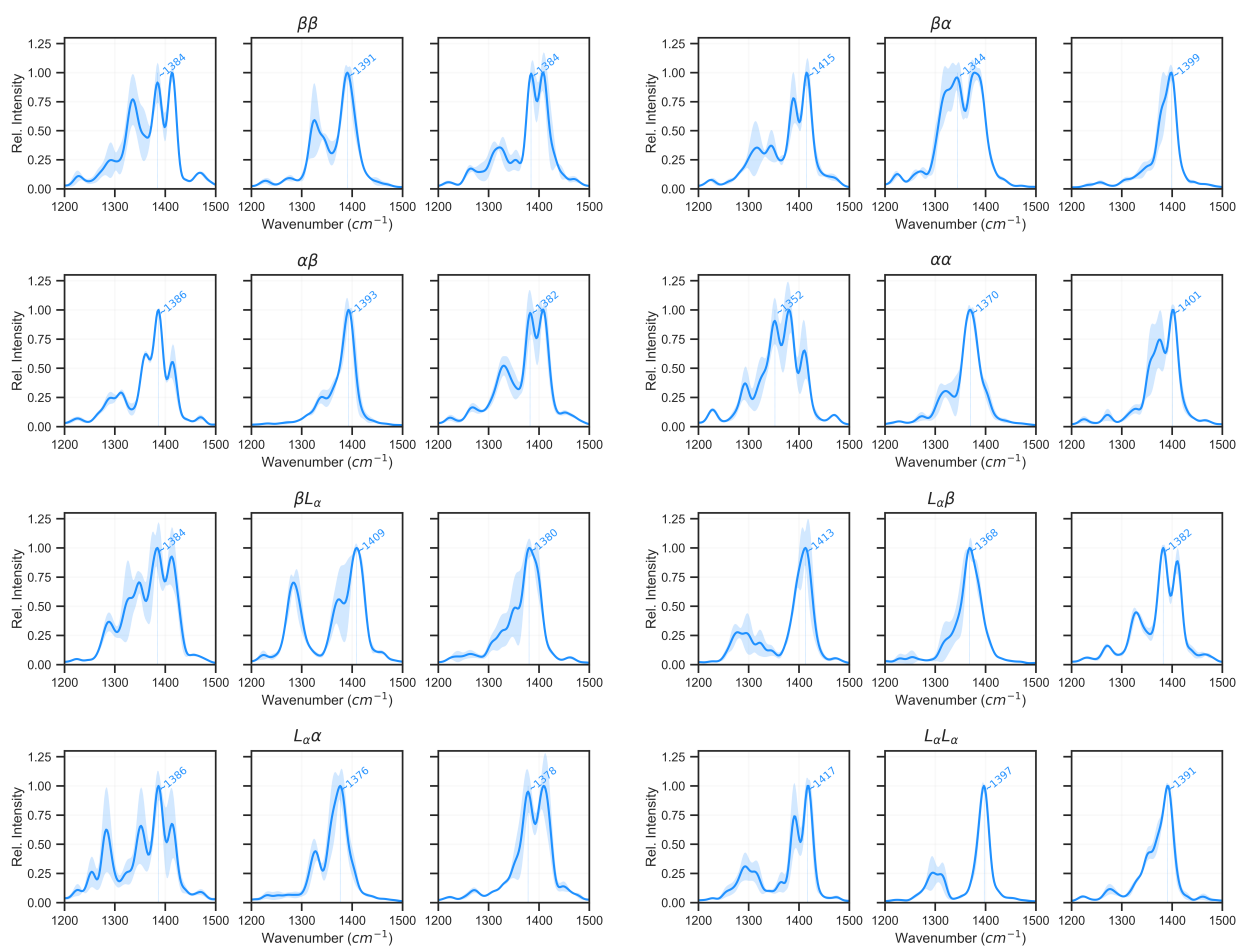


Figure C.33: Power spectra of amine groups, (first,second,third from left to right) of all conformations of ALAL.

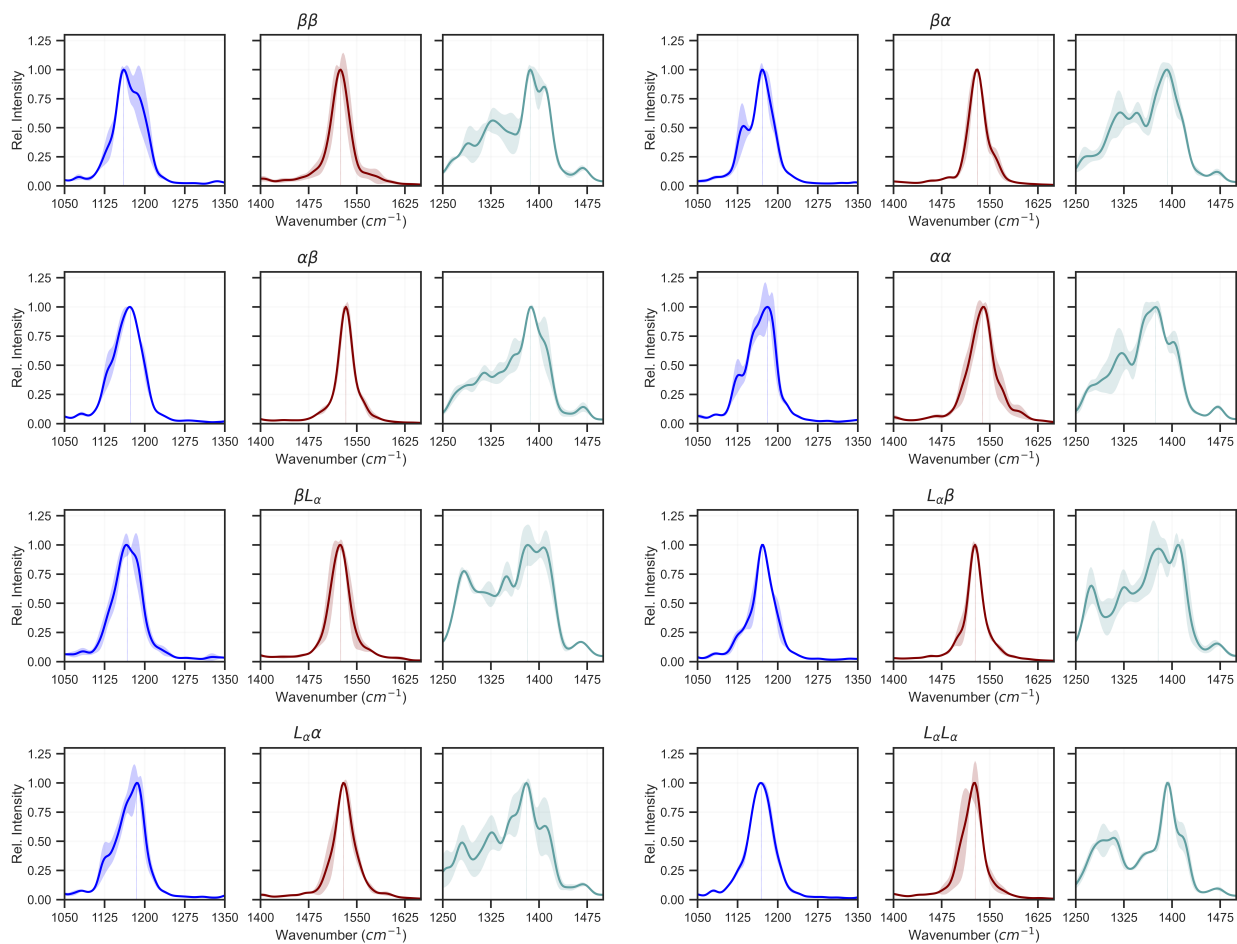


Figure C.34: Power spectra of terminal groups (ND₃,COO) and N-C (from left to right) of all conformations of ALAL.

C.10 Density-based vs Wannier centers-based IR spectra

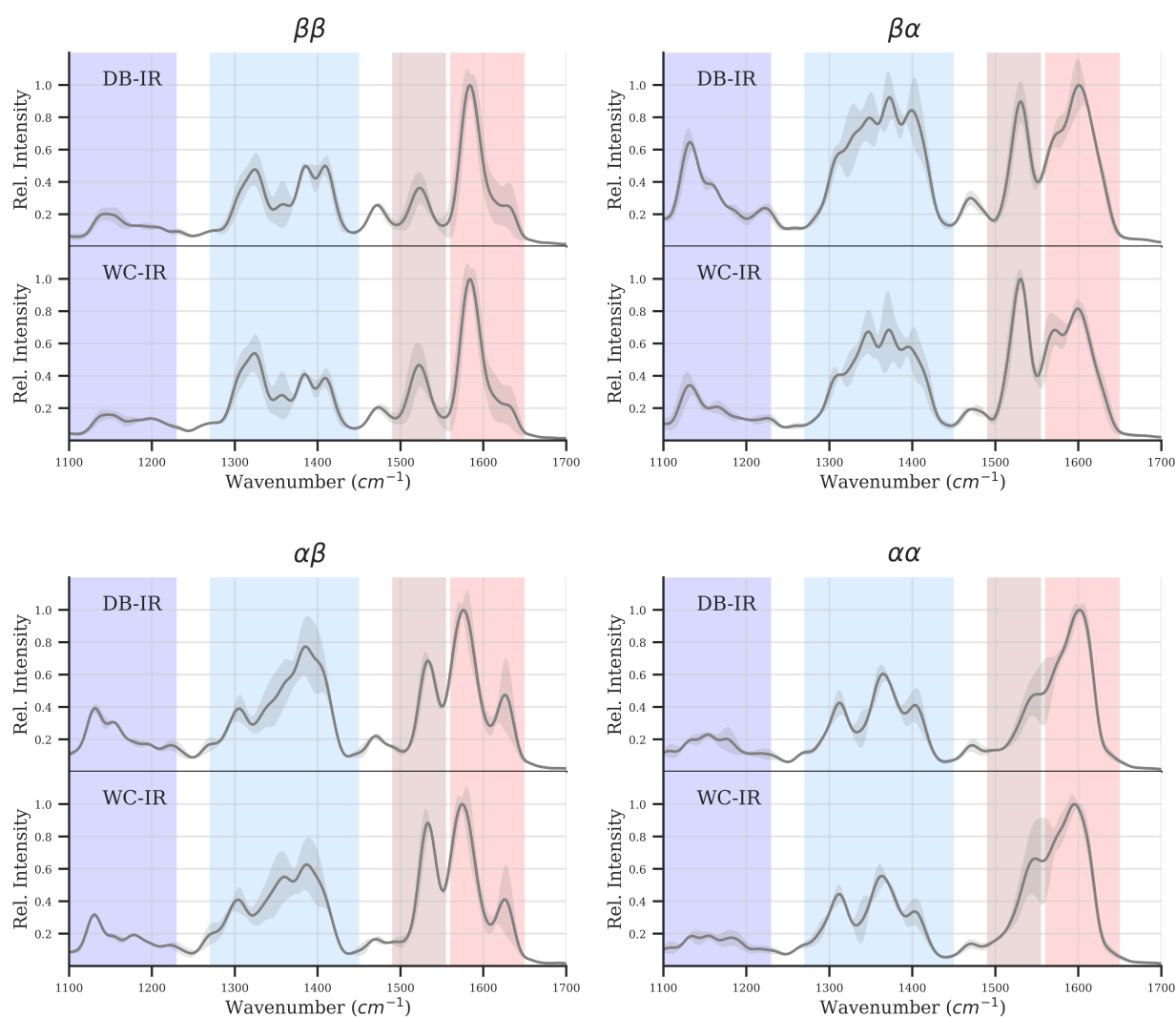


Figure C.35: Density based infrared spectra of representative conformations of Set-V of ALAL (above), and Wannier function localisation based infrared spectra of all conformations (below).

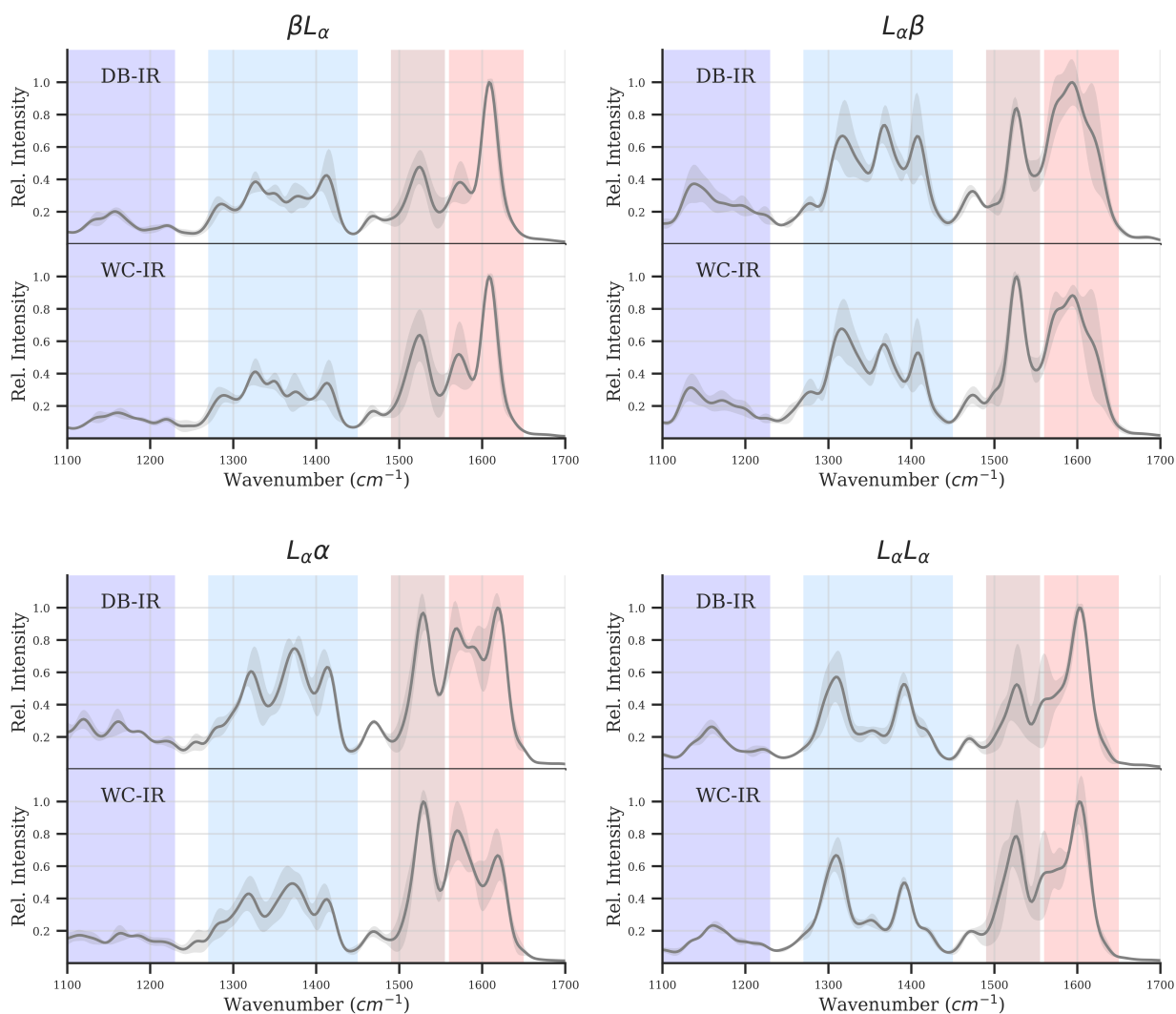


Figure C.36: Density based infrared spectra of representative conformations of Set-II to Set-IV of ALAL (above) and Wannier function localisation based infrared spectra of all conformations (below).

Appendix D

Supplementary material for: Effect of Peptide Length

D.1 Markov State Modeling

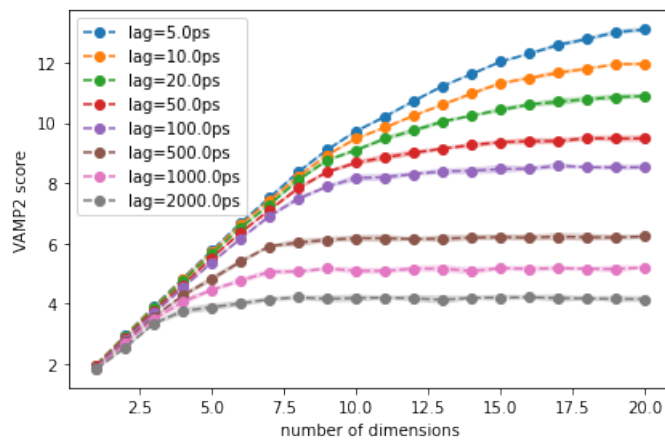


Figure D.1: VAMP2 score for a range of lagtimes.

D.1.1 Alanine-Leucine-Alanine

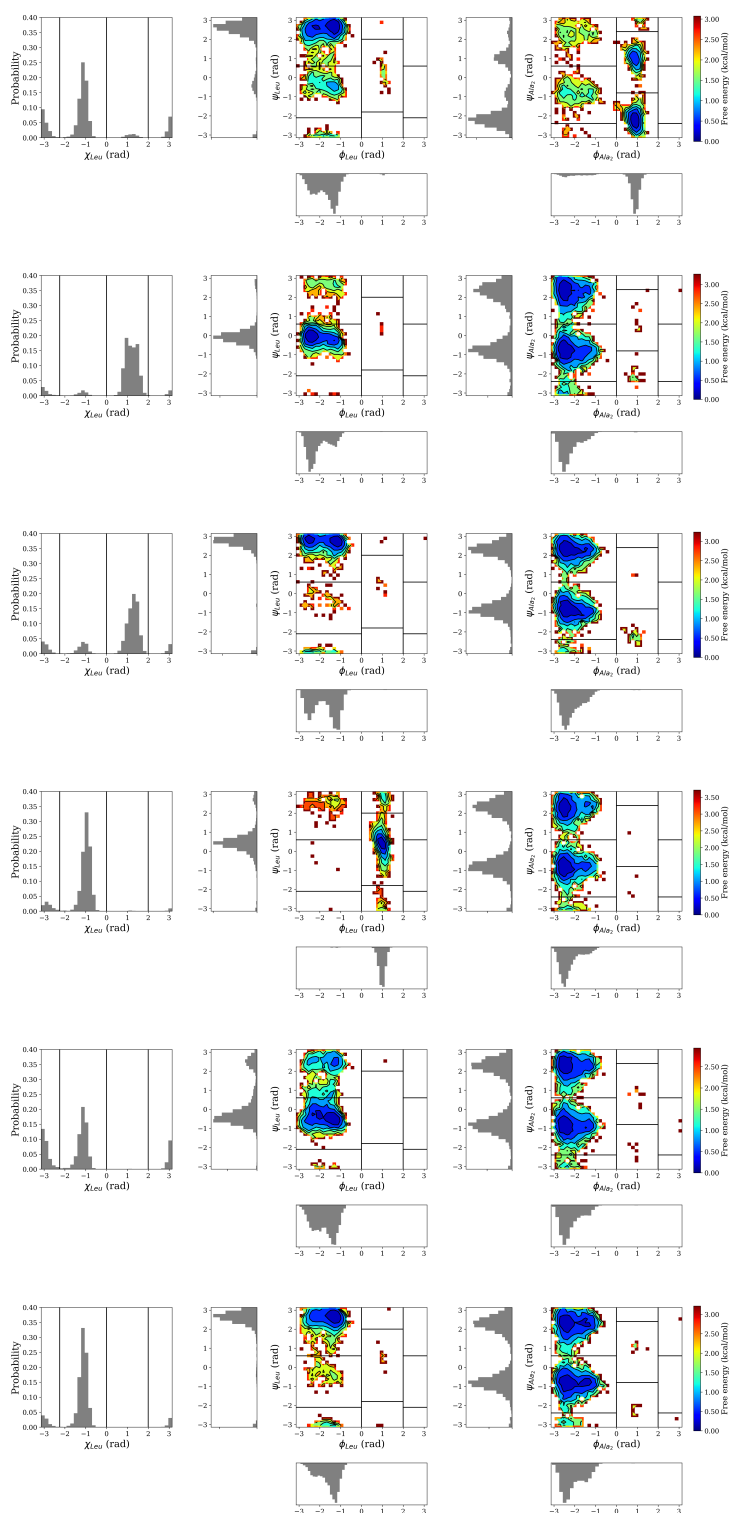


Figure D.2: Probability distributions of dihedral angles used for TICA extracted from the trajectories of metstable-sets I to VI from top to bottom.

D.1.2 Alanine-Leucine-Alanine-Leucine-Alanine

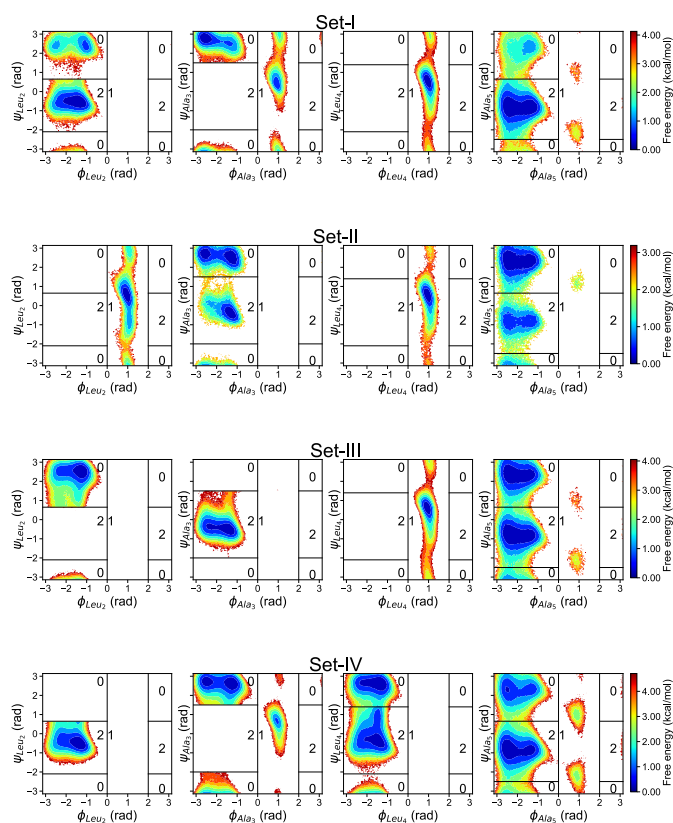


Figure D.3: Probability distributions of dihedral angles used for TICA extracted from the trajectories of metstable-sets I to IV from top to bottom.

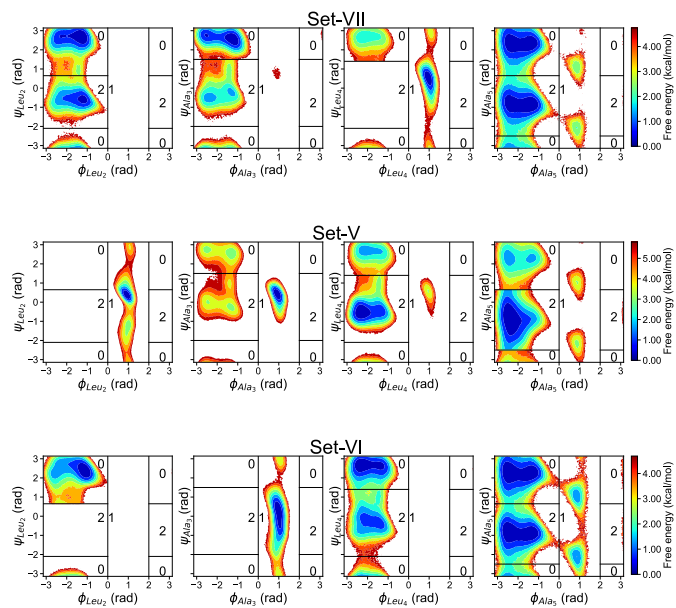


Figure D.4: Probability distributions of dihedral angles used for TICA extracted from the trajectories of metstable-sets V to VII from top to bottom.

D.1.3 Alanine-Leucine-Alanine-Leucine-Alanine-Leucine

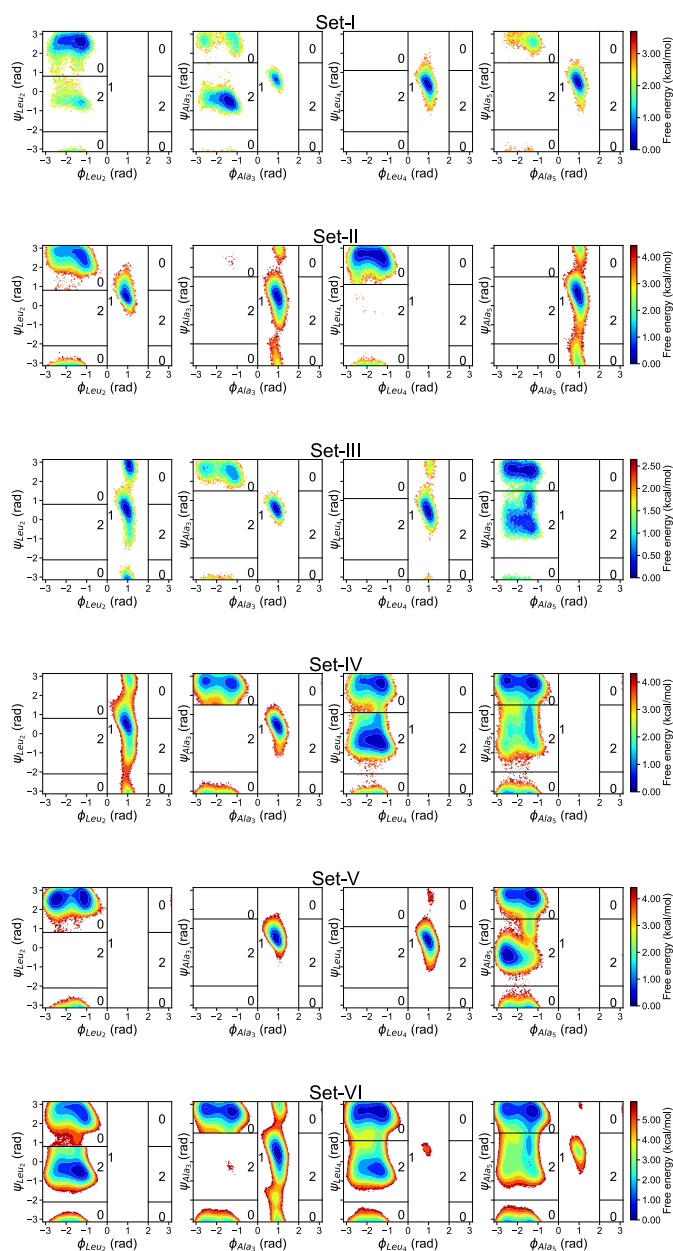


Figure D.5: Probability distributions of dihedral angles used for TICA extracted from the trajectories of metstable-sets I to VI from top to bottom.

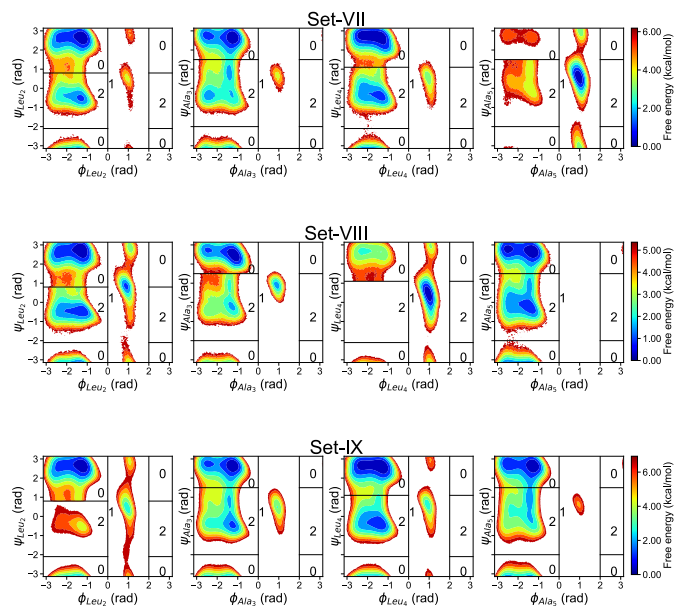


Figure D.6: Probability distributions of dihedral angles used for TICA extracted from the trajectories of metstable-sets VII to IX from top to bottom.

The transition network plots of ALALA and ALALAL are shown in Figure. D.7.

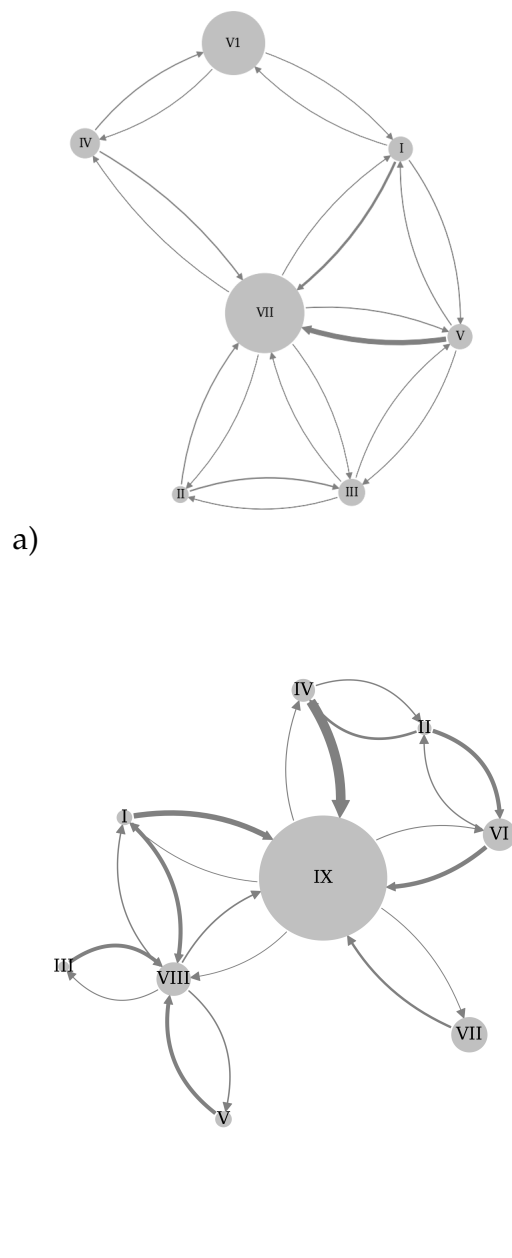


Figure D.7: The transition network plot obtained from a hidden Markov model based coarse-graining of the MSM of a) ALALA, b) ALALAL.

D.2 Constrained Classical Simulations

D.2.1 Dihedral Distributions

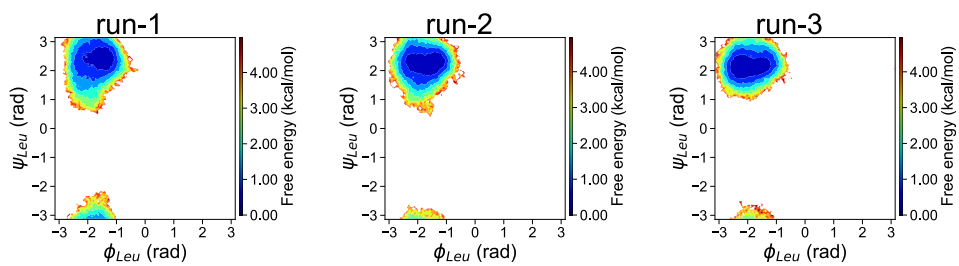


Figure D.8: Dihedral distribution of torsion angles of AL.

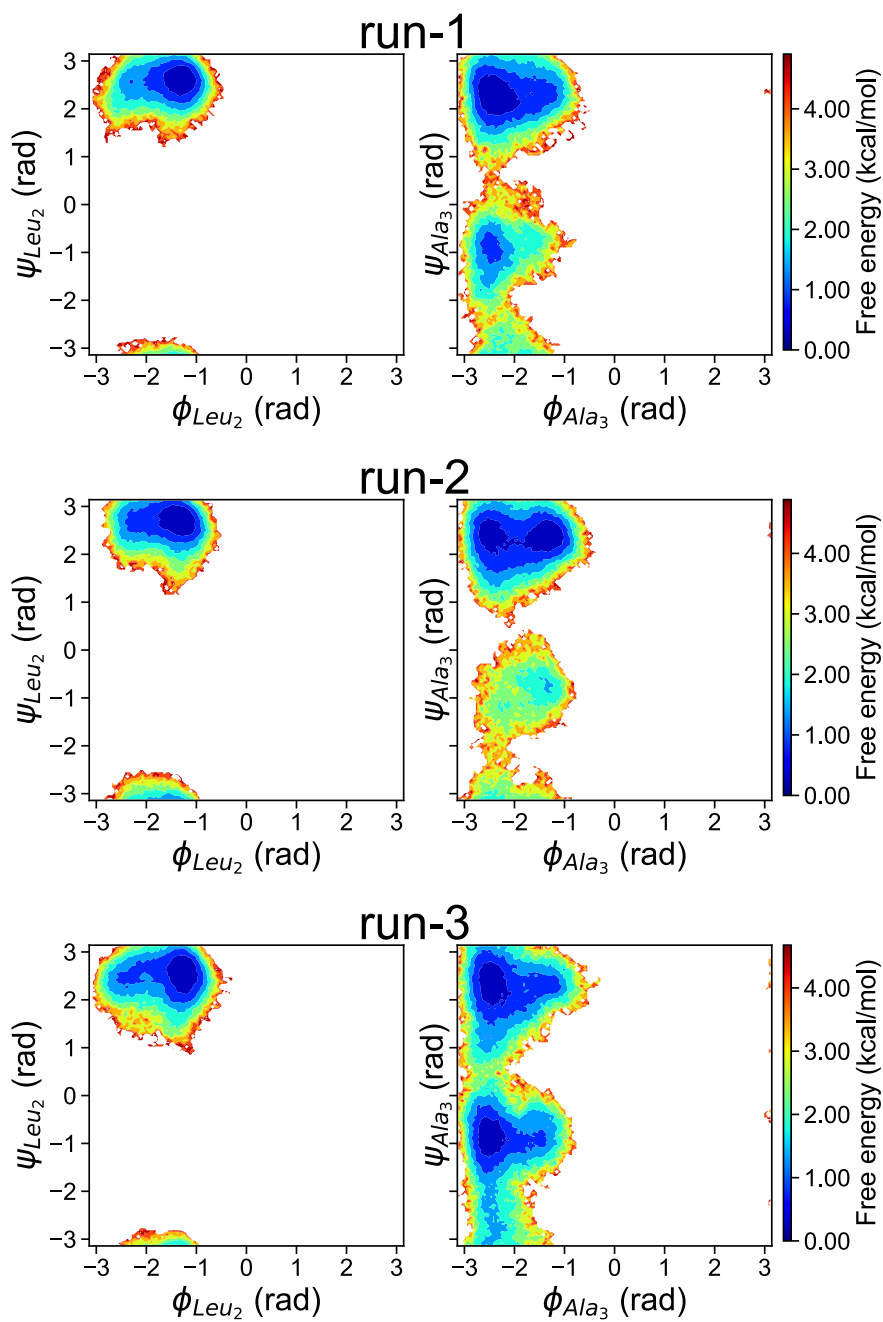


Figure D.9: Dihedral distribution of torsion angles of ALA.

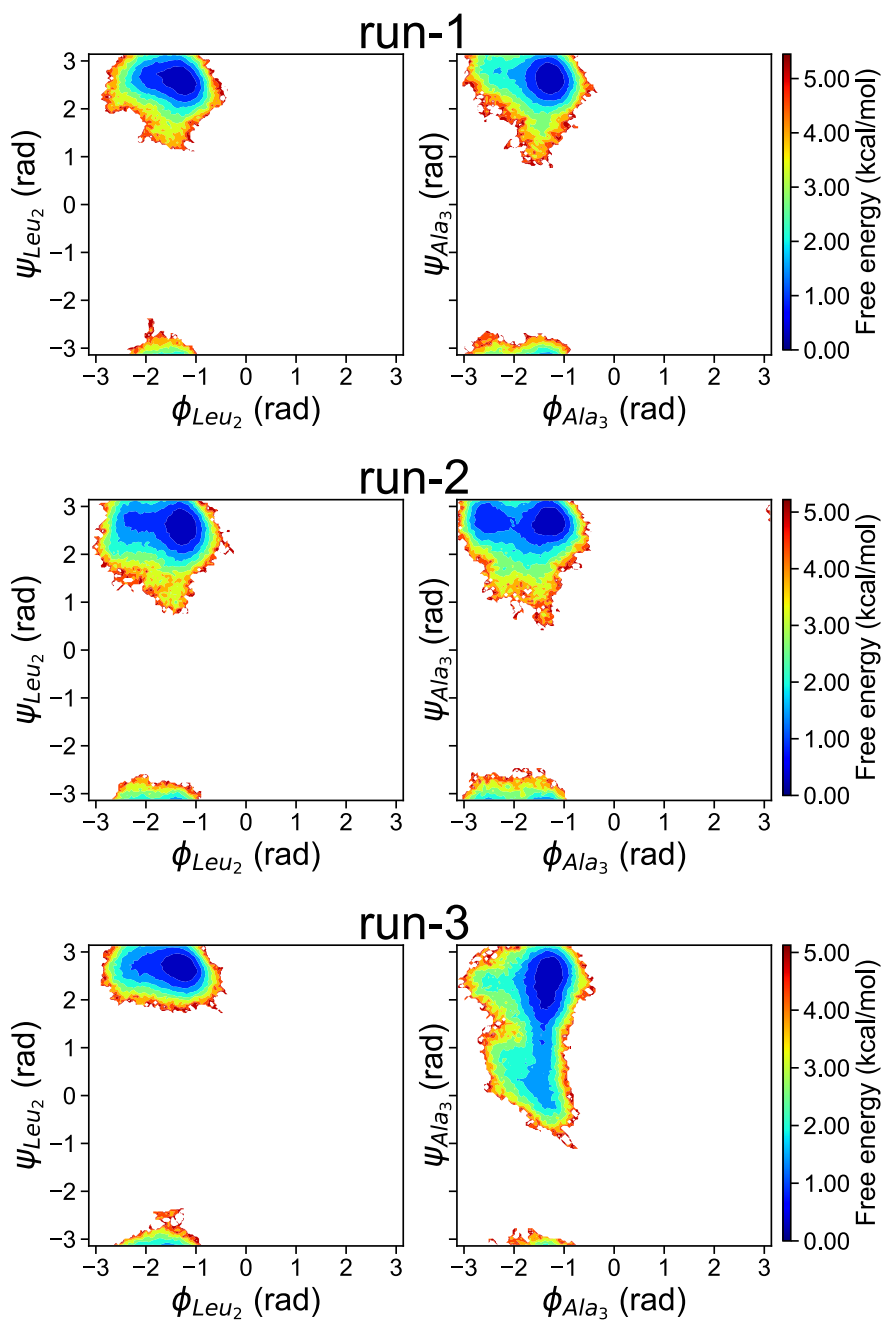


Figure D.10: Dihedral distribution of torsion angles of ALAL.

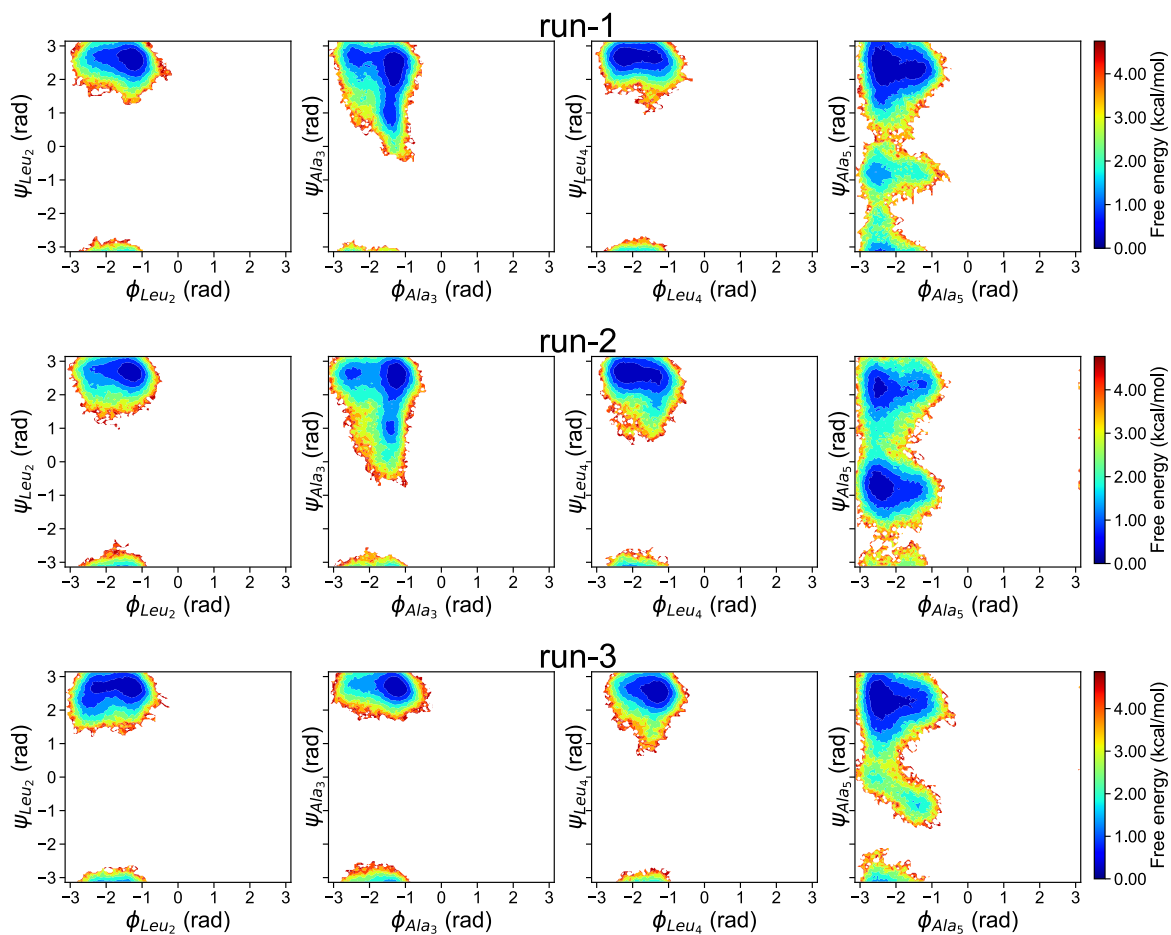


Figure D.11: Dihedral distribution of torsion angles of ALALA.

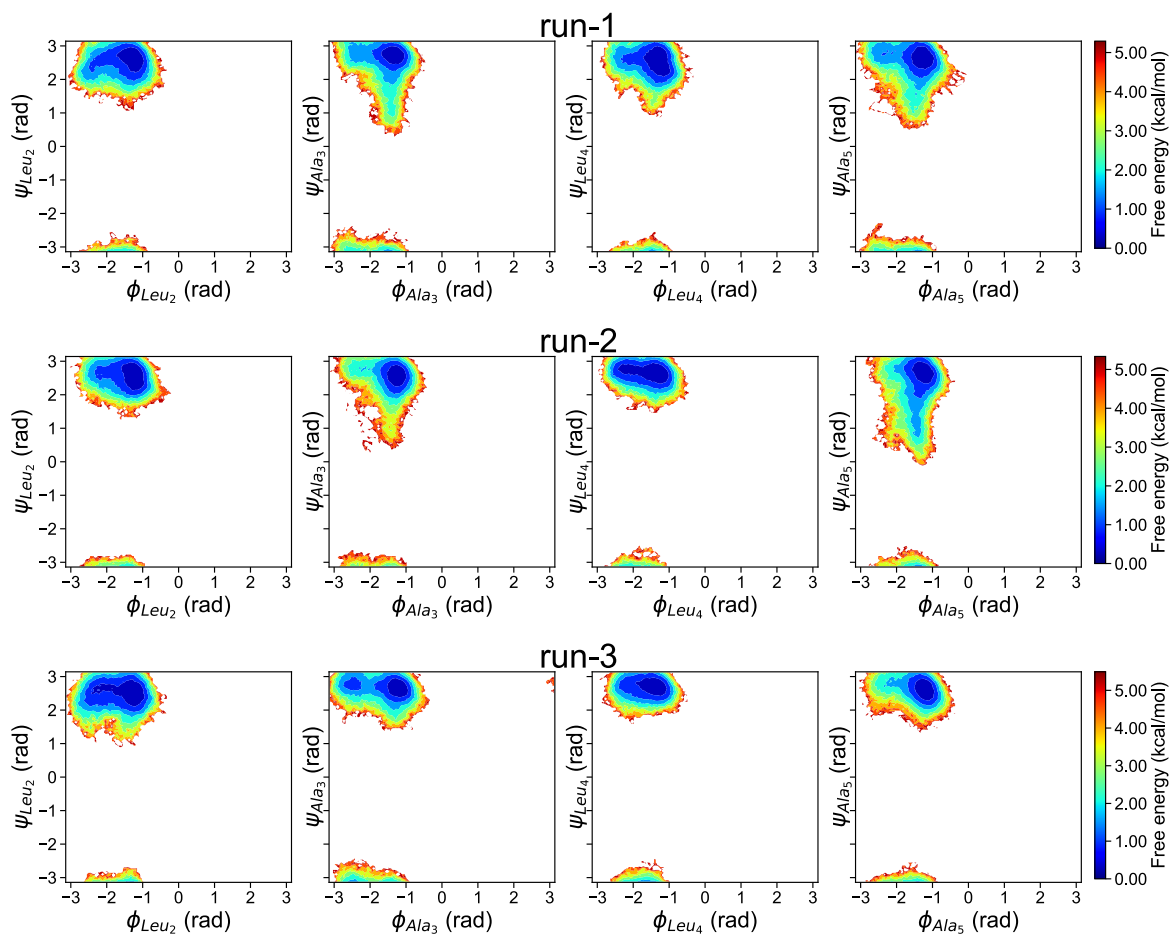


Figure D.12: Dihedral distribution of torsion angles of ALALAL.

Acknowledgments

I would like to express my most profound appreciation to my supervisor Prof. Dr. Petra Imhof, for providing me an opportunity to do my doctoral studies in her research group. Without her constant support, it would not be possible for me to complete the thesis. I am really grateful for her assistance, professional guidance, encouragement, and scientific discussions during the whole period of my doctorate. Unfortunately, she moved from FU Berlin a few years ago. Still, her virtual guidance and frequent visits to Berlin kept me motivated, and she was always willing to discuss new results and ideas.

Second, I am thankful to Prof. Dr. Bettina G. Keller for her willingness to be my co-supervisor. She introduced me to Markov State Models and always provided related support and guidance. The collaborative work done with her doctorate student Luca Donati was invaluable, and I am also thankful to him. I would also like to mention Francesca Vitalini of her group for the introductory, beneficial discussions on Markov State Models.

I would like to thank my experimental collaborator, Prof. Dr. Karsten Heyne, and his group member Till Stensitzki, for measuring the vibrational spectra of the Ala-Leu peptide. I thank my colleague Federica Ferraro for the combined work that has produced significant results. I appreciate my colleagues for their helpful comments and discussions. Everybody was accommodating and always eager to help/advice, science related or otherwise. I would also like to thank Prof. Dr. Roland Netz for providing the workspace in the absence of my supervisor.

I acknowledge the constant support of the IT service of our department (particularly Mr. Jens Dreger), and Ms. Annette Schumann-Welde for her administrative assistance. I also thank the CRC 1114 and Freie Universität Berlin for funding my doctorate. I am very thankful to all my Berlin friends, especially my university fellows Tauqir Shinwari and Hamid Khurshid, for a memorable time during my stay in Berlin.

In the end, I acknowledge and thank all members of my family, particularly my parents (Liaqat Ali Khan and Ghulam Zahra), my siblings (Basit Ali Khan and Saima Rubab) who have provided emotional, moral, and financial support throughout my life. I would also like to extend my deepest gratitude to my wife Memoona Riaz for her unfailing, unconditional support and taking care of our lovely little sons, Ahmad Hassan, and Ali Hassan. I want to express my gratitude for my uncle Ch. Abdul Rehman who always kept check on the progress of my academic career.

Bibliography

- [ADK⁺11] Elangannan Arunan, Gautam R. Desiraju, Roger A. Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C. Clary, Robert H. Crabtree, Joseph J. Dannenberg, Pavel Hobza, Henrik G. Kjaergaard, Anthony C. Legon, Benedetta Mennucci, and David J. Nesbitt. Defining the hydrogen bond: An account (iupac technical report). *Pure and Applied Chemistry*, 83(8): 1619–1636, 2011, <https://doi.org/10.1351/PAC-REP-10-01-01>.
- [AGGH11] Mazen Ahmad, Wei Gu, Tihamér Geyer, and Volkhard Helms. Adhesive water networks facilitate binding of protein interfaces. *Nature communications*, 2(1): 1–7, 2011.
- [AK21] T Reid Alderson and Lewis E Kay. Nmr spectroscopy captures the essential role of dynamics in regulating biomolecular function. *Cell*, 184(3): 577–595, 2021.
- [AON⁺08] Alexandros Altis, Moritz Otten, Phuong H. Nguyen, Rainer Hegger, and Gerhard Stock. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.*, 128(24): 245102, 2008.
- [Ark06] Isaiah T Arkin. Isotope-edited ir spectroscopy for the study of membrane proteins. *Current opinion in chemical biology*, 10(5): 394–401, 2006.
- [Bag05] Biman Bagchi. Water dynamics in the hydration layer around proteins and micelles. *Chemical Reviews*, 105(9): 3197–3219, 2005.
- [Bar07] Andreas Barth. Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1767(9): 1073–1101, 2007.
- [BBB⁺20] Carlos R Baiz, Bartosz Błasiak, Jens Bredenbeck, Minhaeng Cho, Jun-Ho Choi, Steven A Corcelli, Arend G Dijkstra, Chi-Jui Feng, Sean Garrett-Roe, Nien-Hui Ge, et al. Vibrational spectroscopic map, vibrational spectroscopy, and intermolecular interaction. *Chemical reviews*, 120(15): 7152–7218, 2020.
- [BDP07a] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1): 014101+, January 2007, <http://dx.doi.org/10.1063/1.2408420>.
- [BDP07b] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1): 014101, 2007.
- [Bec88] Axel D Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6): 3098, 1988.

- [BER78] ROBERT W BERRY. Function and metabolism of neuronal proteins: comparisons among the identified neurons of aplysia. In *Biochemistry of Characterised Neurons*, pages 283–308. Elsevier, 1978.
- [BH08] Nicolae-Viorel Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B*, 112(19): 6057–6069, 2008.
- [BK11] Martin Brehm and Barbara Kirchner. Travis—a free analyzer and visualizer for monte carlo and molecular dynamics trajectories. *Journal of Chemical Information and Modeling*, 51(8): 2007–2023, 2011.
- [Blo29] Felix Bloch. Bemerkung zur elektronentheorie des ferromagnetismus und der elektrischen leitfähigkeit. *Zeitschrift für Physik*, 57(7): 545–555, 1929.
- [BLS⁺13] Benjamin Breiten, Matthew R Lockett, Woody Sherman, Shuji Fujita, Mohammad Al-Sayah, Heiko Lange, Carleen M Bowers, Annie Heroux, Goran Krilov, and George M Whitesides. Water networks contribute to enthalpy/entropy compensation in protein–ligand binding. *Journal of the American Chemical Society*, 135(41): 15579–15584, 2013.
- [BM17] Sohag Biswas and Bhabani S. Mallik. Time-dependent vibrational spectral analysis of first principles trajectory of methylamine with wavelet transform. *Phys. Chem. Chem. Phys.*, 19: 9912–9922, 2017, <http://dx.doi.org/10.1039/C7CP00412E>.
- [BNW09] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11): 789–796, 2009.
- [BO27] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der physik*, 389(20): 457–484, 1927.
- [BSS15] Sebastian Buchenberg, Norbert Schaudinnus, and Gerhard Stock. Hierarchical biomolecular dynamics: Picosecond hydrogen bonding regulates microsecond conformational transitions. *Journal of chemical theory and computation*, 11(3): 1330–1336, 2015.
- [BT12] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Science, 2012.
- [BT21] Martin Brehm and Martin Thomas. Optimized atomic partial charges and radii defined by radical voronoi tessellation of bulk phase simulations. *Molecules*, 26(7): 1875, 2021.
- [BTGK20] M. Brehm, M. Thomas, S. Gehrke, and B. Kirchner. Travis—a free analyzer for trajectories from molecular simulation. *The Journal of Chemical Physics*, 152(16): 164105, 2020.
- [Buc58] Amyand David Buckingham. Solvent effects in infra-red spectroscopy. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 248(1253): 169–182, 1958.
- [BZ02] Andreas Barth and Christian Zscherp. What vibrations tell about proteins. *Quarterly reviews of biophysics*, 35(4): 369–430, 2002.

- [CHT98] René Carmona, Wen-Liang Hwang, and Bruno Torresani. *Practical Time-Frequency Analysis: Gabor and wavelet transforms, with an implementation in S*, volume 9. Academic Press, 1998.
- [CLS⁺11] Cong Chen, Wei Zhong Li, Yong Chen Song, Lin Dong Weng, and Ning Zhang. The effect of geometrical criteria on hydrogen bonds analysis in aqueous glycerol solutions. *J. Mol. Imag. Dynamic.*, 1: 1, 2011.
- [CMNF21] Carmelo Corsaro, Domenico Mallamace, Giulia Neri, and Enza Fazio. Hydrophilicity and hydrophobicity: Key aspects for biomedical and technological purposes. *Physica A: Statistical Mechanics and its Applications*, page 126189, 2021.
- [CN14] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25: 135–144, 2014.
- [Col14] Michael A Collins. Molecular forces, geometries, and frequencies by systematic molecular fragmentation including embedded charges. *The Journal of chemical physics*, 141(9): 094108, 2014.
- [Cra13] Christopher J Cramer. *Essentials of computational chemistry: theories and models*. John Wiley & Sons, 2013.
- [CSP⁺07] John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *The Journal of chemical physics*, 126(15): 155101, 2007.
- [CYK⁺12] Jyoti Roy Choudhuri, Vivek K Yadav, Anwesa Karmakar, Bhabani S Mallik, and Amalendu Chandra. A first-principles theoretical study of hydrogen-bond dynamics and vibrational spectral diffusion in aqueous ionic solution: Water in the hydration shell of a fluoride ion. *Pure and Applied Chemistry*, 85(1): 27–40, 2012.
- [Dau90] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5): 961–1005, 1990.
- [DHS⁺19] John F Darby, Adam P Hopkins, Seishi Shimizu, Shirley M Roberts, James A Brannigan, Johan P Turkenburg, Gavin H Thomas, Roderick E Hubbard, and Marcus Fischer. Water networks can determine the affinity of ligand binding to proteins. *Journal of the American Chemical Society*, 141(40): 15818–15826, 2019.
- [Dir30] Paul AM Dirac. Note on exchange phenomena in the thomas atom. In *Mathematical proceedings of the Cambridge philosophical society*, volume 26, pages 376–385. Cambridge University Press, 1930.
- [DW05] Peter Deuffhard and Marcus Weber. Robust perron cluster analysis in conformation dynamics. *Linear algebra and its applications*, 398: 161–184, 2005.
- [DYP93a] T. Darden, Darrin York, and Lee G. Pedersen. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98: 10089–10092, 1993.

- [DYP93b] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \log n$ method for ewald sums in large systems. *The Journal of chemical physics*, 98(12): 10089–10092, 1993.
- [ECNSS02] Fatma Eker, Xiaolin Cao, Laurence Nafie, and Reinhard Schweitzer-Stenner. Tripeptides adopt stable structures in water. a combined polarized visible raman, ftir, and vcd spectroscopy study. *J. Am. Chem. Soc.*, 124(48): 14330–14341, 2002.
- [EPB⁺95a] Ulrich Essman, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103: 8577–8503, 1995.
- [EPB⁺95b] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19): 8577–8593, 1995.
- [Ewa21] Paul P Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der physik*, 369(3): 253–287, 1921.
- [FBBB15] Teresa Fornaro, Diletta Burini, Malgorzata Biczysko, and Vincenzo Barone. Hydrogen-bonding effects on infrared spectra from anharmonic computations: uracil–water complexes and uracil dimers. *The Journal of physical chemistry A*, 119(18): 4224–4236, 2015.
- [FDT18] Chi-Jui Feng, Balamurugan Dhayalan, and Andrei Tokmakoff. Refinement of peptide conformational ensembles by 2d ir spectroscopy: Application to ala–ala–ala. *Biophysical journal*, 114(12): 2820–2832, 2018, <https://www.sciencedirect.com/science/article/pii/S000634951830571X>.
- [FHK⁺16] Yuan Feng, Jing Huang, Seongheun Kim, Ji Hyun Shim, Alexander D. MacKerell, and Nien-Hui Ge. Structure of penta-alanine investigated by two-dimensional infrared spectroscopy and molecular dynamics simulation. *The Journal of Physical Chemistry B*, 120(24): 5325–5339, 2016.
- [FL14] Aoife C Fogarty and Damien Laage. Water dynamics in protein hydration shells: The molecular origins of the dynamical perturbation. *The Journal of Physical Chemistry B*, 118(28): 7715–7729, 2014.
- [FRLB⁺15] Marwa H Farag, Manuel F Ruiz-Lopez, Adolfo Bastida, Gerald Monard, and Francesca Ingrosso. Hydration effect on amide i infrared bands in water: An interpretation based on an interaction energy decomposition scheme. *The Journal of Physical Chemistry B*, 119(29): 9056–9067, 2015.
- [FSPT21] Chi-Jui Feng, Anton Sinitskiy, Vijay Pande, and Andrei Tokmakoff. Computational ir spectroscopy of insulin dimer structure and conformational heterogeneity. *The Journal of Physical Chemistry B*, 125(18): 4620–4633, 2021.
- [FT17a] Chi-Jui Feng and Andrei Tokmakoff. The dynamics of peptide-water interactions in dialanine: an ultrafast amide i 2d ir and computational spectroscopy study. *The Journal of chemical physics*, 147(8): 085101, 2017.
- [FT17b] Chi-Jui Feng and Andrei Tokmakoff. The dynamics of peptide-water interactions in dialanine: An ultrafast amide I 2D IR and computational spectroscopy study. *J. Chem. Phys.*, 147: 085101, 2017.

- [FTS⁺] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, A. J. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision E.01. Gaussian Inc. Wallingford CT 2009.
- [FTS⁺¹⁶] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.
- [GAEK10] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15): 154104, 2010.
- [Gai08] M.-P. Gaiote. Alanine polypeptide structural fingerprints at room temperature: What can be gained from non-harmonic car-parrinello molecular dynamics simulations. *The Journal of Physical Chemistry A*, 112(51): 13507–13517, 2008.
- [Gai09] Marie-Pierre Gaiote. Unravelling the conformational dynamics of the aqueous alanine dipeptide with first-principles molecular dynamics. *J. Phys. Chem. B.*, 113: 10059–10062, 2009.
- [Gai10a] Marie-Pierre Gaiote. Infrared spectroscopy of the alanine dipeptide analog in liquid water with dft-md. direct evidence for β conformations. *Phys. Chem. Chem. Phys.*, 12: 10198–10209, 2010, <http://dx.doi.org/10.1039/C003485A>.

- [Gai10b] Marie-Pierre Gaigeot. Infrared spectroscopy of the alanine dipeptide analog in liquid water with dft-md. direct evidence for π/β conformations. *Physical Chemistry Chemical Physics*, 12(35): 10198–10209, 2010.
- [Gai10c] Marie-Pierre Gaigeot. Infrared spectroscopy of the alanine dipeptide analog in liquid water with dft-md. direct evidence for π/β conformations. *Phys. Chem. Chem. Phys.*, 12: 10198–10209, 2010, <http://dx.doi.org/10.1039/C003485A>.
- [Gai10d] Marie-Pierre Gaigeot. Theoretical spectroscopy of floppy peptides at room temperature. a dftmd perspective: gas and aqueous phase. *Physical Chemistry Chemical Physics*, 12(14): 3336–3359, 2010.
- [Gai21] Marie-Pierre Gaigeot. Some opinions on md-based vibrational spectroscopy of gas phase molecules and their assembly: an overview of what has been achieved and where to go. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, page 119864, 2021.
- [GCC⁺21] Molly SC Gravett, Ryan C Cocking, Alistair P Curd, Oliver Harlen, Joanna Leng, Stephen P Muench, Michelle Peckham, Daniel J Read, Jarvellis F Rogers, Robert C Welch, et al. Moving in the mesoscale: Understanding the mechanics of cytoskeletal molecular motors by combining mesoscale simulations with imaging. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, page e1570, 2021.
- [GCM⁺13] Chen Gu, Huang-Wei Chang, Lutz Maibaum, Vijay S Pande, Gunnar E Carlsson, and Leonidas J Guibas. Building markov state models with solvent dynamics. In *BMC bioinformatics*, volume 14, pages 1–9. Springer, 2013.
- [GF82] BJ Gellatly and John L Finney. Calculation of protein volumes: an alternative to the voronoi procedure. *Journal of molecular biology*, 161(2): 305–322, 1982.
- [GMV07] Marie-Pierre Gaigeot, Michaël Martinez, and Rodolphe Vuilleumier. Infrared spectroscopy in the gas and liquid phase from first principle molecular dynamics simulations: application to small peptides. *Molecular Physics*, 105(19-22): 2857–2878, 2007.
- [GOZ17] Ayanjeet Ghosh, Joshua S Ostrander, and Martin T Zanni. Watching proteins wiggle: Mapping structures with two-dimensional infrared spectroscopy. *Chemical reviews*, 117(16): 10726–10759, 2017.
- [GPS02] Herbert Goldstein, Charles Poole, and John Safko. *Classical mechanics*. American Association of Physics Teachers, 2002.
- [GS03] Marie-Pierre Gaigeot and Michiel Sprik. Ab initio molecular dynamics computation of the infrared spectrum of aqueous uracil, 2003.
- [GS18] David J Griffiths and Darrell F Schroeter. *Introduction to quantum mechanics*. Cambridge university press, 2018.
- [GTH96] Stefan Goedecker, Michael Teter, and Jürg Hutter. Separable dual-space gaussian pseudopotentials. *Physical Review B*, 54(3): 1703, 1996.

- [GVE16] Enrico Guarnera and Eric Vanden-Eijnden. Optimized markov state models for metastable systems. *The Journal of Chemical Physics*, 145(2): 024102, 2016.
- [GVSB05] Marie-Pierre Gaigeot, Rodolphe Vuilleumier, Michiel Sprik, and Daniel Borgis. Infrared spectroscopy of n-methylacetamide revisited by ab initio molecular dynamics simulations. *Journal of chemical theory and computation*, 1(5): 772–789, 2005.
- [HAO⁺06] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3): 712–725, 2006.
- [HDS⁺18] Irtaza Hassan, Luca Donati, Till Stensitzki, Bettina G. Keller, Karsten Heyne, and Petra Imhof. The vibrational spectrum of the hydrated alanine-leucine peptide in the amide region from ir experiments and first principles calculations. *Chemical Physics Letters*, 698: 227–233, 2018.
- [HF19] Andreas Heßelmann and Federica Ferraro. Study of the wilcox torsion balance in solution for a tröger’s base derivative with hexyl-and heptyl substituents using a combined molecular mechanics and quantum chemistry approach. *Journal of molecular modeling*, 25(3): 69, 2019.
- [HFI21] Irtaza Hassan, Federica Ferraro, and Petra Imhof. Effect of the hydration shell on the carbonyl vibration in the ala-leu-ala-leu peptide. *Molecules*, 26(8): 2148, 2021.
- [HGH98] Christian Hartwigsen, Sephen Gødecker, and Jürg Hutter. Relativistic separable dual-space gaussian pseudopotentials from h to rn. *Physical Review B*, 58(7): 3641, 1998.
- [HISV14] Jürg Hutter, Marcella Iannuzzi, Florian Schiffmann, and Joost VandeVondele. cp2k: atomistic simulations of condensed matter systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1): 15–25, 2014.
- [HK64] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B): B864, 1964.
- [Hoc07] Robin M Hochstrasser. *Multidimensional ultrafast spectroscopy*, 2007.
- [Hoo85] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3): 1695, 1985.
- [HR06] Carmen Herrmann and Markus Reiher. First-principles approach to vibrational spectroscopy of biomolecules. In *Atomistic approaches in modern biology*, pages 85–132. Springer, 2006.
- [HRGM⁺16] Zahra Heidari, Daniel R Roe, Rodrigo Galindo-Murillo, Jahan B Ghasemi, and Thomas E Cheatham III. Using wavelet analysis to assist in identification of significant events in molecular dynamics simulations. *Journal of chemical information and modeling*, 56(7): 1282–1291, 2016.
- [Hun07] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007.

- [Hun09] NT Hunt. Ultrafast 2d-ir spectroscopy–applications to biomolecules. *Chem. Soc. Rev*, 38: 1837–1848, 2009.
- [HVS01] Mathew D. Halls, Julia Velkovski, and H. Bernhard Schlegel. Harmonic frequency scaling factors for Hartree-Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 with the Sadlej pVTZ electric property basis set. *Theoretical Chemistry Accounts*, 105(6): 413–421, May 2001.
- [HW79] John A. Hartigan and Manchek A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108, 1979.
- [HWK07] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172): 964–972, 2007.
- [JCM⁺83] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2): 926–935, 1983.
- [JGH17] KwangHyok Jong, Luca Grisanti, and Ali Hassanali. Hydrogen bond networks and hydrophobic effects in the amyloid β 30–35 chain in water: A molecular dynamics study. *Journal of chemical information and modeling*, 57(7): 1548–1562, 2017.
- [JH18] KwangHyok Jong and Ali A Hassanali. A data science approach to understanding water networks around biomolecules: the case of tri-alanine in liquid water. *The Journal of Physical Chemistry B*, 122(32): 7895–7906, 2018.
- [KB86] Samuel Krimm and Jagdeesh Bandekar. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Prot. Chem.*, 38: 181–364, 1986.
- [KBP96] Walter Kohn, Axel D Becke, and Robert G Parr. Density functional theory of electronic structure. *The Journal of Physical Chemistry*, 100(31): 12974–12980, 1996.
- [KC03] Joo-Hee Kim and Min-Haeng Cho. Interplay of the intramolecular water vibrations and hydrogen bond in n-methylacetamide-water complexes: Ab initio calculation studies. *Bulletin of the Korean Chemical Society*, 24(8): 1061–1068, 2003.
- [KC13] Heejae Kim and Minhaeng Cho. Infrared probes for studying the structure and dynamics of biomolecules. *Chemical reviews*, 113(8): 5817–5847, 2013.
- [KC15] Vivek Kumar and Amalendu Chandra. First-principles simulation study of vibrational spectral diffusion and hydrogen bond fluctuations in aqueous solution of N-methylacetamide. *J. Phys. Chem. B*, 119: 9858–9967, 2015.
- [KH04] Barbara Kirchner and Jürg Hutter. Solvent effects on electronic properties from wannier functions in a dimethyl sulfoxide/water mixture. *The Journal of chemical physics*, 121(11): 5133–5142, 2004.
- [KIDB⁺20a] Thomas D Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z Khaliullin,

- Ole Schütt, Florian Schiffmann, et al. Cp2k: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19): 194103, 2020.
- [KIDB⁺20b] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, Tiziano Müller, Robert Schade, Manuel Guidon, Samuel Andermatt, Nico Holmberg, Gregory K. Schenter, Anna Hehn, Augustin Bussy, Fabian Belleflamme, Gloria Tabacchi, Andreas Glensk, Michael Lass, Iain Bethune, Christopher J. Mundy, Christian Pleschl, Matt Watkins, Joost Vandevondele, Matthias Krack, and Jörg Hutter. Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19): 194103, 2020, <https://doi.org/10.1063/5.0007045>.
- [KLAH09] Yung Sam Kim, Liu Liu, Paul H Axelsen, and Robin M Hochstrasser. 2d ir provides evidence for mobile water molecules in β -amyloid fibrils. *Proceedings of the National Academy of Sciences*, 106(42): 17751–17756, 2009.
- [KLFH17] Tilman Kottke, Victor A Lorenz-Fonfria, and Joachim Heberle. The grateful infrared: Sequential protein structural changes resolved by infrared difference spectroscopy. *The Journal of Physical Chemistry B*, 121(2): 335–350, 2017.
- [KNG04] George Karvounis, Dmitry Nerukh, and Robert C Glen. Water network dynamics at the critical moment of a peptide's β -turn formation: A molecular dynamics study. *The Journal of chemical physics*, 121(10): 4925–4935, 2004.
- [KNS10] Maja Kobus, Phuong H Nguyen, and Gerhard Stock. Infrared signatures of the peptide dynamical transition: A molecular dynamics simulation study. *The Journal of chemical physics*, 133(3): 034512, 2010.
- [Kra05] Matthias Krack. Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts*, 114(1): 145–152, 2005.
- [KS65] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A): A1133, 1965.
- [KSS07] R. Kumar, J. R. Schmidt, and J. L. Skinner. Hydrogen bonding definitions and dynamics in liquid water. *The Journal of Chemical Physics*, 126(20): 204107, 2007.
- [KSV93] RD King-Smith and David Vanderbilt. Theory of polarization of crystalline solids. *Physical Review B*, 47(3): 1651, 1993.

- [KTH12] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*, volume 31. Springer Science & Business Media, 2012.
- [KWH05] Yung Sam Kim, Jianping Wang, and Robin M. Hochstrasser. Two-dimensional infrared spectroscopy of the alanine dipeptide in aqueous solution. *The Journal of Physical Chemistry B*, 109(15): 7511–7521, 2005.
- [LCW12] Marie L. Laury, Matthew J. Carlson, and Angela K. Wilson. Vibrational frequency scale factors for density functional theory and the polarization consistent basis sets. *Journal of Computational Chemistry*, 33(30): 2380–2387, 2012, <http://dx.doi.org/10.1002/jcc.23073>.
- [LDAI⁺03] Simon C Lovell, Ian W Davis, W Bryan Arendall III, Paul IW De Bakker, J Michael Word, Michael G Prisant, Jane S Richardson, and David C Richardson. Structure validation by $c\alpha$ geometry: ϕ , ψ and $c\beta$ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3): 437–450, 2003.
- [LF20] Victor A Lorenz-Fonfria. Infrared difference spectroscopy of proteins: from bands to bonds. *Chemical Reviews*, 120(7): 3466–3576, 2020.
- [LHP99] Gerald Lippert, Jürg Hutter, and Michele Parrinello. The gaussian and augmented-plane-wave density functional method for ab initio molecular dynamics simulations. *Theoretical Chemistry Accounts*, 103(2): 124–140, 1999.
- [LJ31] John E Lennard-Jones. Cohesion. *Proceedings of the Physical Society (1926-1948)*, 43(5): 461, 1931.
- [LLPP⁺10] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8): 1950–1958, 2010.
- [LPM97] By Gerald Lippert, JURG HUTTER PARRINELLO, and MICHELE. A hybrid gaussian and plane wave density functional scheme. *Molecular Physics*, 92(3): 477–488, 1997.
- [LYP88] Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2): 785, 1988.
- [MAA08] Nataliya S Myshakina, Zeeshan Ahmed, and Sanford A Asher. Dependence of amide vibrations on hydrogen bonding. *The Journal of Physical Chemistry B*, 112(38): 11873–11877, 2008.
- [MAdAF⁺18] Uriel N Morzan, Diego J Alonso de Armino, Nicolas O Foglia, Francisco Ramirez, Mariano C Gonzalez Lebrero, Damian A Scherlis, and Dario A Estrin. Spectroscopy in complex environments from qm–mm simulations. *Chemical reviews*, 118(7): 4071–4113, 2018.
- [MADWB11] Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, 32(10): 2319–2327, 2011.

- [Mar20] Richard M Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.
- [Mat07] Masakazu Matsumoto. Relevance of hydrogen bond definitions in liquid water. *The Journal of Chemical Physics*, 126(5): 054503, 2007.
- [MBH⁺15] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical J.*, 109(8): 1528 – 1532, 2015.
- [MCG07] Smita Mukherjee, Pramit Chowdhury, and Feng Gai. Infrared study of the effect of hydration on the amide I band and aggregation properties of helical peptides. *The Journal of Physical Chemistry B*, 111(17): 4596–4602, 2007.
- [McQ00] D.A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.
- [MGD⁺06] DC Marinica, G Gregoire, C Desfrancois, JP Schermann, D Borgis, and MP Gaigeot. Ab initio molecular dynamics of protonated dialanine and comparison to infrared multiphoton dissociation experiments. *The Journal of Physical Chemistry A*, 110(28): 8802–8810, 2006.
- [MH09] Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press, 2009.
- [MH16] Oinam Romesh Meitei and Andreas Heßelmann. Molecular energies from an incremental fragmentation method. *The Journal of Chemical Physics*, 144(8): 084109, 2016.
- [MJRG15] Jerome Mahe, Sander Jaecx, Anouk M. Rijs, and Marie-Pierre Gaigeot. Can far-ir action spectroscopy combined with bond simulations be conformation selective? *Phys. Chem. Chem. Phys.*, 17: 25905–25914, 2015, <http://dx.doi.org/10.1039/C5CP01518A>.
- [MMPCS11] Francesco Muniz-Miranda, Marco Pagliai, Gianni Cardini, and Vincenzo Schettino. Wavelet transform for spectroscopic analysis: Application to diols in water. *Journal of chemical theory and computation*, 7(4): 1109–1118, 2011.
- [MMR07] Jeffrey P. Merrick, Damian Moran, and Leo Radom. An evaluation of harmonic vibrational frequency scale factors. *The Journal of Physical Chemistry A*, 111(45): 11683–11700, 2007.
- [MPWN18] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1): 1–11, 2018.
- [MS94] Lutz Molgedey and Heinz Georg Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23): 3634, 1994.

- [MS02] Yuguang Mu and Gerhard Stock. Conformational dynamics of trialanine in water: a molecular dynamics study. *The Journal of Physical Chemistry B*, 106(20): 5294–5301, 2002.
- [MSC08a] Bhabani S Mallik, A Semparithi, and Amalendu Chandra. A first principles theoretical study of vibrational spectral diffusion and hydrogen bond dynamics in aqueous ionic solutions: D₂O in hydration shells of Cl⁻ ions. *The Journal of chemical physics*, 129(19): 194512, 2008.
- [MSC08b] Bhabani S Mallik, A Semparithi, and Amalendu Chandra. Vibrational spectral diffusion and hydrogen bond dynamics in heavy water from first principles. *The Journal of Physical Chemistry A*, 112(23): 5104–5112, 2008.
- [MSM58] Tatsuo Myazawa, Takehiko Shimanouchi, and San-Ichiro Mizushima. Normal vibrations of N-Methylacetamide. *J. Chem. Phys.*, 29: 611–616, 1958.
- [MT13] Luca Monticelli and D Peter Tieleman. Force fields for classical molecular dynamics. *Biomolecular simulations*, pages 197–213, 2013.
- [MV97] Nicola Marzari and David Vanderbilt. Maximally localized generalized wannier functions for composite energy bands. *Physical review B*, 56(20): 12847, 1997.
- [MZS⁺17] Helen Miller, Zhaokun Zhou, Jack Shepherd, Adam JM Wollman, and Mark C Leake. Single-molecule techniques in biophysics: a review of the progress in methods and applications. *Reports on Progress in Physics*, 81(2): 024601, 2017.
- [NDK⁺07] Erik TJ Nibbering, Jens Dreyer, Oliver Kühn, Jens Bredenbeck, Peter Hamm, and Thomas Elsaesser. Vibrational dynamics of hydrogen bonds. In *Analysis and control of ultrafast photoinduced reactions*, pages 619–687. Springer, 2007.
- [NHSS07] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126(15): 155102, 2007.
- [NK13] Dmitry Nerukh and Sergey Karabasov. Water-peptide dynamics during conformational transitions. *The journal of physical chemistry letters*, 4(5): 815–819, 2013.
- [Nos84a] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2): 255–268, 1984.
- [Nos84b] Shuichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52(2): 255–268, 1984.
- [Nos84c] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1): 511–519, 1984.
- [NRKH02] Johannes Neugebauer, Markus Reiher, Carsten Kind, and Bernd A Hess. Quantum chemical calculation of vibrational spectra of large molecules—raman and ir spectra for buckminsterfullerene. *Journal of computational chemistry*, 23(9): 895–910, 2002.

- [NWPP13] Frank Noé, Hao Wu, Jan-Hendrik Prinz, and Nuria Plattner. Projected and hidden markov models for calculating kinetics and metastable states of complex molecules. *The Journal of chemical physics*, 139(18): 11B609_1, 2013.
- [OKK18] D. Ojha, K. Karhan, and T. D. Kühne. On the hydrogen bond strength and vibrational spectroscopy of liquid water. *Sci Rep.*, 8: 16888, 2018, <https://doi.org/10.1038/s41598-018-35357-9>.
- [OLCO13] Vidvuds Ozoliņš, Rongjie Lai, Russel Caflisch, and Stanley Osher. Compressed modes for variational problems in mathematics and physics. *Proceedings of the National Academy of Sciences*, 110(46): 18368–18373, 2013.
- [oPC09] International Union of Pure and Applied Chemistry. Iupac compendium of chemical terminology – the gold book, 2009.
- [OVMVG04] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, 25(13): 1656–1676, 2004.
- [OZC14] Abdullah Ozkanlar, Tiecheng Zhou, and Aurora E. Clark. Towards a unified description of the hydrogen bond network of liquid water: A dynamics based approach. *The Journal of Chemical Physics*, 141(21): 214107, 2014.
- [PGSR13] Diego Prada-Gracia, Roman Shevchuk, and Francesco Rao. The quest for self-consistency in hydrogen bond definitions. *The Journal of Chemical Physics*, 139(8): 084501, 2013.
- [PHPG⁺13] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1): 015102, 2013.
- [PMMC⁺10] Marco Pagliai, Francesco Muniz-Miranda, Gianni Cardini, Roberto Righini, and Vincenzo Schettino. Hydrogen bond dynamics of methyl acetate in methanol. *The Journal of Physical Chemistry Letters*, 1(19): 2951–2955, 2010.
- [PPS⁺13] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David Van Der Spoel, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7): 845–854, 2013.
- [PTA⁺92] Mike C. Payne, Michael P. Teter, Douglas C. Allan, T.A. Arias, and J.D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Modern Physics*, 64(4): 1045, 1992.
- [PWS⁺11] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17): 174105, 2011.

- [Res94] Raffaele Resta. Macroscopic polarization in crystalline dielectrics: the geometric phase approach. *Reviews of modern physics*, 66(3): 899, 1994.
- [RMH02] Rossend Rey, Klaus B Møller, and James T Hynes. Hydrogen bond dynamics in water and ultrafast infrared spectroscopy. *The Journal of Physical Chemistry A*, 106(50): 11993–11996, 2002.
- [RRS63] GN Ramachandran, C Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1): 95–99, 1963.
- [RW13] Susanna Röblitz and Marcus Weber. Fuzzy spectral clustering by pcca+: application to markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2): 147–179, 2013.
- [SAB⁺17] Francesca Spyrakis, Mostafa H Ahmed, Alexander S Bayden, Pietro Cozzini, Andrea Mozzarelli, and Glen E Kellogg. The roles of water in the protein matrix: a largely untapped resource for drug discovery. *Journal of medicinal chemistry*, 60(16): 6781–6827, 2017.
- [SABW82] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of chemical physics*, 76(1): 637–649, 1982.
- [SBG⁺04] Pankaj Sinha, Scott E. Boesch, Changming Gu, Ralph A. Wheeler, and Angela K. Wilson. Harmonic vibrational frequencies: Scaling factors for HF, B3LYP, and MP2 methods in combination with correlation consistent basis sets. *J. Phys. Chem. A*, 108(42): 9213–9217, 2004.
- [Sch26] Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6): 1049, 1926.
- [SFHD99] Ch Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuffhard. A direct approach to conformational dynamics based on hybrid monte carlo. *Journal of Computational Physics*, 151(1): 146–168, 1999.
- [SH18] H Lee Sweeney and Erika LF Holzbaur. Motor proteins. *Cold Spring Harbor Perspectives in Biology*, 10(5): a021931, 2018.
- [SHD01] Ch Schütte, Wilhelm Huisinga, and Peter Deuffhard. Transfer operator approach to conformational dynamics in biomolecular systems. In *Ergodic theory, analysis, and efficient simulation of dynamical systems*, pages 191–223. Springer, 2001.
- [SHZ09] Gideon Schreiber, Gilad Haran, and H-X Zhou. Fundamental aspects of protein- protein association kinetics. *Chemical reviews*, 109(3): 839–860, 2009.
- [SKM⁺21] Amberley D Stephens, Johanna Kolbel, Rani Moons, Michael T Ruggerio, Najet Mahmoudi, Talia A Shmool, Thomas M McCoy, Daniel Nietlispach, Alexander F Routh, Frank Sobott, et al. The role of water mobility in protein misfolding. *bioRxiv*, 2021.

- [SN11] JJ Sakurai and J Napolitano. Modern quantum mechanics, 2: nd edition. *Person New International edition*, 2011.
- [SP99a] Pier Luigi Silvestrelli and Michele Parrinello. Structural, electronic, and bonding properties of liquid water from first principles. *The Journal of chemical physics*, 111(8): 3572–3580, 1999.
- [SP99b] Pier Luigi Silvestrelli and Michele Parrinello. Water molecule dipole in the gas and in the liquid phase. *Physical Review Letters*, 82(16): 3308, 1999.
- [SP13] Christian R Schwantes and Vijay S Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of nt19. *Journal of chemical theory and computation*, 9(4): 2000–2009, 2013.
- [SPS04a] William C Swope, Jed W Pitera, and Frank Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *The Journal of Physical Chemistry B*, 108(21): 6571–6581, 2004.
- [SPS⁺04b] William C. Swope, Jed W. Pitera, Frank Suits, Mike Pitman, Maria Eleftheriou, Blake G. Fitch, Robert S. Germain, Aleksandr Rayshubski, T. J. C. Ward, Yuriy Zhestkov, and Ruhong Zhou. Describing protein folding kinetics by molecular dynamics simulations. 2. example applications to alanine dipeptide and a β -hairpin peptide. *J. Phys. Chem. B*, 108(21): 6582–6594, 2004.
- [SS18] Florian Sittel and Gerhard Stock. Perspective: Identification of collective variables and metastable states of protein dynamics. *The Journal of chemical physics*, 149(15): 150901, 2018.
- [SS19] Amberley D Stephens and Gabriele S Kaminski Schierle. The role of water in amyloid aggregation kinetics. *Current opinion in structural biology*, 58: 115–123, 2019.
- [STSP⁺15a] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Perez-Hernandez, Moritz Hoffmann, Nuria Plattne, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: a software package for estimation, validation, and analysis of markov models. *J. Chem. Theory Comput.*, 11(11): 5525–5542, 2015.
- [STSP⁺15b] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo PÁlrez-HernÁandez, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank NoÁl. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11: 5525–5542, October 2015, <http://dx.doi.org/10.1021/acs.jctc.5b00743>.
- [SW08] Birgit Strodel and David J. Wales. Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.*, 466(4): 105 – 115, 2008, <http://www.sciencedirect.com/science/article/pii/S0009261408014796>.
- [TBF⁺13] Martin Thomas, Martin Brehm, Reinhold Fligg, Peter Vöhringer, and Barbara Kirchner. Computing vibrational spectra from ab initio molecular dynamics. *Physical Chemistry Chemical Physics*, 15(18): 6608–6622, 2013.

- [TBK15] Martin Thomas, Martin Brehm, and Barbara Kirchner. Voronoi dipole moments for the simulation of bulk phase vibrational spectra. *Physical Chemistry Chemical Physics*, 17(5): 3207–3213, 2015.
- [TC98] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1): 61–78, 1998.
- [TT97] Lukas K Tamm and Suren A Tatulian. Infrared spectroscopy of proteins and peptides in lipid bilayers. *Quarterly reviews of biophysics*, 30(4): 365–429, 1997.
- [TT02] M Tarek and DJ Tobias. Role of protein-water hydrogen bond dynamics in the protein dynamical transition. *Physical Review Letters*, 88(13): 138101, 2002.
- [VDD⁺15] Valerio Vitale, Jacek Dziedzic, Simon M.-M. Dubois, Hans Fangohr, and Chris-Kriton Skylaris. Anharmonic infrared spectroscopy through the fourier transform of time correlation function formalism in onetep. *Journal of Chemical Theory and Computation*, 11(7): 3321–3332, 2015.
- [Ver67] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1): 98, 1967.
- [VHA⁺10] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4): 671–690, 2010.
- [VKM⁺05] Joost VandeVondele, Matthias Krack, Fawzi Mohamed, Michele Parrinello, Thomas Chassaing, and Jürg Hutter. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications*, 167(2): 103–128, 2005.
- [Wal04] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, 2004.
- [Wan37] Gregory H Wannier. The structure of electronic excitation levels in insulating crystals. *Physical Review*, 52(3): 191, 1937.
- [Wan17] Jianping Wang. Ultrafast two-dimensional infrared spectroscopy for molecular structures and dynamics with expanding wavelength range and increasing sensitivities: from experimental and computational perspectives. *International Reviews in Physical Chemistry*, 36(3): 377–431, 2017.
- [WCW⁺03] Scott T.R. Walsh, Richard P. Cheng, Wayne W. Wright, Darwin O.V. Alonso, Valerie Daggett, Jane M. Vanderkooi, and William F. DeGrado. The hydration of amides in helices; a comprehensive picture from molecular dynamics, ir, and nmr. *Protein Science*, 12(3): 520–531, 2003.
- [WDC80] Edgar Bright Wilson, John Courtney Decius, and Paul C Cross. *Molecular vibrations: the theory of infrared and Raman vibrational spectra*. Courier Corporation, 1980.

- [WH00] S. Woutersen and P. Hamm. Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J Phys. Chem. B*, 104(47): 11316–11320, 2000.
- [WPN15] Hao Wu, Jan-Hendrik Prinz, and Frank Noé. Projected metastable markov processes and their estimation with observable operator models. *The Journal of chemical physics*, 143(14): 10B610_1, 2015.
- [WWC⁺04] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9): 1157–1174, 2004.
- [XB01] Huafeng Xu and BJ Berne. Hydrogen-bond kinetics in the solvation shell of a polypeptide. *The Journal of Physical Chemistry B*, 105(48): 11929–11932, 2001.
- [YKCC12] Vivek K Yadav, Anwesa Karmakar, Jyoti Roy Choudhuri, and Amalendu Chandra. A first principles molecular dynamics study of vibrational spectral diffusion and hydrogen bond dynamics in liquid methanol. *Chemical Physics*, 408: 36–42, 2012.
- [YYK⁺15] Huayan Yang, Shouning Yang, Jilie Kong, Aichun Dong, and Shaoning Yu. Obtaining information about protein secondary structures in aqueous solution using fourier transform ir spectroscopy. *Nature protocols*, 10(3): 382–396, 2015.
- [ZAH01] Martin T. Zanni, Matthew C. Asplund, and Robin M. Hochstrasser. Two-dimensional heterodyned and stimulated infrared photon echoes of n-methylacetamide-d. *J. Chem. Phys.*, 114(10): 4579–4590, 2001.
- [ZBC⁺10] Hui Zhu, Martine Blom, Isabel Compagnon, Anouk M Rijs, Santanu Roy, Gert Von Helden, and Burkhard Schmidt. Conformations and vibrational spectra of a model tripeptide: change of secondary structure upon micro-solvation. *Physical Chemistry Chemical Physics*, 12(14): 3415–3425, 2010.

