Utilizing Optimal Transport Theory to Model Peptide Conformational Distributions and Address the Levinthal Problem

By

Vigneshwaran K

PHYS10201604007

The Institute of Mathematical Sciences, Chennai

A thesis submitted to the

Board of Studies in Physical Sciences

In partial fulfillment of requirements

for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

Homi Bhabha National Institute

Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by Vigneshwaran K entitled "Utilizing Optimal Transport Theory to Model Peptide Conformational Distributions and Address the Levinthal Problem" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairman - Satyavani Vemparala — Date: - 26 - August - 2025 -

Guide/Convenor - S R Hassan — Date: - 26 - August - 2025 -

Examiner - N. S. Vidhyadhiraja — Date: - 26 - August - 2025 -

Member 1 - Sanatan Digal — Date: - 26 - August - 2025 -

Member 2 - Sibasish Ghosh — Date: - 26 - August - 2025 -

Member 3 - Venkata Suryanarayana Nemani — Date: - 26 - August - 2025 -

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I hereby certify that I have read this thesis prepared under my direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: - 26 - August - 2025 -

Place: IMSc, Chennai

Guide : S.R.Hassan

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Vigneshwaran K

DECLARATION

I hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.

Vigneshwaran K

LIST OF PUBLICATIONS ARISING FROM THE THESIS

Publications:

- Published

    1. Vigneshwaran Kannan, Ramesh Anishetty, and S. R. Hassan. 'Optimal Trans port Technique to Understand Peptide Conformations.' published in Compu tational Biology and Chemistry **98** (2022): 107684.

- Submitted

    1. Vigneshwaran Kannan and S.R.Hassan, 'Tackling the Levinthal Problem with Recursive Distributional Optimal Transport' submitted to the journal and it is under review.

List of presentations and participations at conferences

1. Presented a poster titled *'Multi-marginal Optimal Transport technique to study tetrapeptide conformations'* at the STATPHYS-Kolkata XI conference, held virtually from March 21 to 25, 2022. The conference was organized by IISER Kolkata.

2. Participated in and presented a poster titled *'Generating Hexapeptide Backbone Distributions using Optimal Transport Theory'* at the 45th Indian Biophysical Society Meeting. The conference was held from March 27 to 29, 2023, at NCBS-TIFR Bangalore

To My Family and Friends

ACKNOWLEDGEMENTS

I am deeply grateful to the many individuals who supported me during challenging times, without whom this thesis would not have been possible. First and foremost, I wish to express my heartfelt gratitude to my research supervisor, Dr. S. R. Hassan, for his invaluable guidance and unwavering support throughout my Ph.D. journey. I am also sincerely thankful to Dr. Ramesh Anishetty for the insightful discussions and valuable inputs that enriched my work.

I extend my gratitude to the doctoral committee chairman, Dr. Satyavani Vemparala, and the committee members, Dr Sibasish Ghosh, Dr Sanatan Digal, and Dr. Venkata Suryanarayana Nemani, for their constructive criticism and constant support throughout my research work.

I would like to acknowledge the administrative staff of IMSc for their invaluable support. In particular, I am deeply thankful to Mrs. Indira and Mrs. Prema for their prompt responses and for efficiently resolving all official matters in a timely manner. I also express my appreciation to IMSc for providing the necessary computational resources, and I am especially grateful to Mr. Srinivasan for his assistance in resolving software-related issues. I am profoundly thankful to my friend, Gopal, for his assistance with TeX-related issues and for engaging in fruitful discussions on the subject. I also express my gratitude to Dr. Nana Siddharth and Sudharshan for generously sharing their knowledge on Python, computational skills, and software-related technical expertise. I would like to thank Dr. Thirukumaran, Dr. Ramarajan, and Mr. Kanmani for their invaluable support during this journey. Additionally, I extend my sincere thanks to Dr. V. Murugan, Retired Professor from Vivekananda College, for his assistance in proofreading my research articles, which significantly enhanced their clarity and quality.

I would also like to thank Dr. Giri for his encouragement and financial support during my journey. Finally, I am deeply grateful to my wife for her unwavering support during challenging times, as well as to my brother and mother for their constant support and encouragement.

Abstract

Peptide conformation studies are essential due to their role in biological functions like cell signaling and drug design, as well as their importance in protein structure prediction. Peptides form secondary structures such as alpha helices and beta hairpins, which can serve as building blocks for predicting three-dimensional protein structures. However, peptides exhibit structural flexibility, adopting a range of conformations, with only specific low-energy conformations being bioactive for particular functions. Constructing conformational distributions for longer peptides is challenging due to limited data from experimental sources like the Protein Data Bank (PDB), which mainly provides information for shorter peptides like dipeptides and tripeptides.

In this thesis, we address this challenge by using optimal transport techniques to construct conformational distributions for longer peptides. Starting with dipeptide distributions, we develop a method to generate tetrapeptide conformational distributions by minimizing the expectation value of interaction energy functions. Applying this approach to tetrapeptides composed of alanine and glycine reveals preferences for right-handed alpha helices in alanine-rich sequences (e.g., AAAA, AAAG) and beta turns in glycine-dominated ones (e.g., GGGG, GAGG). Extending this method recursively, we generate conformational probabilities for longer peptides, enabling efficient prediction of their structural behavior. This approach provides an innovative solution for exploring peptide flexibility and bioactive conformations.

Contents

Synopsis

Abstract:

In this thesis, we present a conformational study of peptides using Optimal Transport Theory. We conceptualize peptides as a collection of tetrapeptides. Our procedure involves determining the conformational probability of tetrapeptides first, using multi-marginal transport and the input distribution of dipeptides from the PDB distribution database. We then build our approach using tetrapeptide distribution to generate the extended peptide conformational probability. This approach is used to fuse shorter peptides to create longer peptides. By doing this recursively and solving distributional optimal transport, we obtain the desired conformational probability of longer peptides.

Introduction

Peptides, which are short chains of amino acids linked by peptide bonds, perform essential functions in various biological processes. They act as hormones, like insulin that manage blood sugar levels, and function as antimicrobial compounds, crucial parts of the immune system, and sometimes even as toxins.[1, 2] The way in which peptides fold, varying from solid three-dimensional forms to highly flexible shapes, is fundamentally connected to their biological functions. Grasping these structures is vital for purposes such as protein structure forecasting, loop region prediction, and the creation of peptide-based medications.[3, 4]

The shape of peptides is described by their backbone torsional angles, which are referred to as Ramachandran angles. These angles dictate the three-dimensional architecture of the peptides, which can differ greatly depending on the protein environment surrounding them.

Conventional approaches to investigating peptide conformations include using empirical data from the Protein Data Bank (PDB) and computational techniques such as molecular dynamics (MD) and Monte Carlo (MC) simulations. However, these approaches face obstacles such as sparse data and substantial computational requirements.[5–13]

The latest advances in artificial intelligence, especially the creation of AlphaFold 2 (AF2), have transformed the prediction of protein structures [14]. Despite AF2's high accuracy in predicting proteins, its use for peptides is still restricted, particularly for predicting areas with bends and flexibility [15].

This thesis introduces a novel methodology based on optimal transport theory (OT) to effectively compute backbone Ramachandran angle distributions for extended peptides [16].The approach utilizes PDB data to create conditional dipeptide distributions, which are subsequently employed in a numerical framework based on Multi-Marginal Transport (MOT) to generate tetrapeptide distributions. By considering tetrapeptides as the primary building blocks, this method is further applied to longer peptides, thereby addressing the computational issues related to dipeptide-centric techniques.

The initial section of the thesis describes the use of MOT to derive tetrapeptide distributions, verified against PDB data for alanine and glycine peptides.[17] The subsequent section presents a dual-phase iterative method to expand these distributions to hexapeptides and larger peptides, using a concatenation technique influenced by the density matrix renormalization group (DMRG) method of physics [18, 19]. This novel methodology addresses the challenges of peptide conformation analysis, offering a scalable and efficient approach to examining longer peptides.

Through the incorporation and modification of methods from computational biology and theoretical physics, this study introduces a solid framework for analyzing peptide conformations. This multidisciplinary strategy not only improves our understanding of peptide dynamics but also paves the way for practical uses in drug development and protein engineering.

Method

Employing the foundational principles of variational quantum mechanics, an intrinsic minimization framework, we harness optimal transport theory, an analogous minimization construct closely aligned with the principles of quantum mechanics. Both paradigms are anchored in probabilistic approaches. Initially, we elucidate the convergences and divergences between these theoretical frameworks. Subsequently, we delve into the iterative deployment of optimal transport theory to tackle the intricate challenge of elucidating the conformational distributions of peptides, especially within extensive molecular systems. This iterative methodology facilitates the effective management of the inherent complexity in modelling peptide configurations, offering a robust strategy for addressing the "Levinthal problem."

Variational techniques within quantum mechanics are essential methods in computational physics, aiding in the estimation of ground-state energies and wave functions. For any chosen wave function Ψ, the expectation value of the Hamiltonian can be calculated as follows:,

$$E[\Psi] = \frac{\langle \Psi | H | \Psi \rangle}{\langle \Psi | \Psi \rangle} = \frac{\int dx \Psi(x) H \Psi(x)}{\int dx \Psi(x) \Psi(x)} \tag{1}$$

, Here, E is a functional of the wave function Ψ, providing a numerical result. Energy eigenvalues are obtained by optimizing $E[\Psi]$ with respect to Ψ. This optimization involves finding a Ψ that minimizes $E[\Psi]$ by proposing a "trial" wavefunction with adjustable parameters. These parameters are iteratively adjusted until the trial wave function's energy is minimized, offering approximations to the exact wave function and energy through the variational method. In variational quantum mechanics, boundary conditions are crucial. The wave function Ψ must adhere to physical requirements like square integrability and normalization, ensuring it stays within the valid solution space.

Optimal transport theory, having originated independently within another discipline and exhibiting extensive applicability across various domains such as Mathematics, Computer Science, Physics, Chemistry, Biology, and Economics, strives to minimize a cost function. Conventionally, optimal transport theory focuses on minimizing a cost function that encapsulates pairwise interactions between two distributions. In contrast, the multimarginal optimal transport theory generalizes this minimization to cost functions that encompass

multipoint interactions among multiple distributions. The enforcement of constraints in optimal transport theory, akin to boundary conditions in quantum mechanics, ensures that solutions remain physically coherent and conform to the pre-specified marginals of the involved distributions.

Within the realm of extended peptides, the use of a similar methodology proves to be pertinent. The configurational landscape of a peptide of length n, characterized by the sequence $A_1A_2A_3...A_n$, is encapsulated in a multidimensional space $X_n = \{\bar{x}_n\}$, where $\bar{x}_n$ denotes a tuple $(x_1, x_2, x_3, ..., x_n)$, with each $x_i = (\phi_i, \psi_i)$ representing Ramachandran angles. Here, A_i designates any residue from the set of 20 amino acids at position i. For a given value of n, the number of potential sequence permutations is equivalent to 20^n. Although individual values of x_i for each residue A_i fluctuate within specific bounds, resulting in a distribution $\rho(x_i)$ over the torsional angles of the backbone, these distributions alone do not adequately specify the conformational distribution $\Pi_n(\bar{x}_n)$ of peptides due to the lack of multipoint correlations between the Ramachandran angles of the backbone, which are essential for delineating the permissible conformation space. Analogously to the principles of variational quantum mechanics, one can infer $\Pi_n(\bar{x}_n)$ by evaluating the expectation value of a potential energy function $K_n(\bar{x}_n)$ for a peptide of length n with a specified sequence:

$$E_n[\Pi_n] = \langle K_n \rangle = \sum_{\bar{x}_n} K_n(\bar{x}_n)\Pi_n(\bar{x}_n), \tag{2}$$

In this context, $\Pi_n(\bar{x}_n)$ signifies the probability distribution of the peptide conformation in X_n, analogous to $|\Psi(\bar{x}_n)|^2$ in quantum mechanics. Consequently, the objective entails identifying a set of conformations that minimize $E_n[\Pi_n]$. In quantum mechanics, the wave function Ψ adheres to the boundary condition of square integrability; here, we enforce the boundary condition on $\Pi(\bar{x}_n)$ through empirical input. The Protein Data Bank (PDB) provides $\rho_i(x_i)$ for each residue A_i as a conditional probability, which can serve as a marginal for $\Pi_n(\bar{x}_n)$, acting as a boundary condition. A formulation paralleling that of variational quantum mechanics can thus be articulated:

$$E_n[\Pi_n] = \min_{\Pi} \sum_{\bar{x}_n} K_n(\bar{x}_n)\Pi_n(\bar{x}_n) \tag{3}$$

Subject to the constraints for each i:

$$\sum_{\bar{x}_n \setminus x_i} \Pi_n(\bar{x}_n) = \rho_i(x_i), \text{ and } i = 1, 2, \cdots, n \tag{4}$$

Multi Marginal Optimal (MOT) for Exploring Tetrapeptide Distribution

To create a tetrapeptide distribution via Optimal Transport, input distributions from the PDB(D) database are used [20]. This database contains distributions of dihedral angles (ϕ, ψ) for dipeptide amino acids. The conditional distributions for a central amino acid C given neighbouring amino acids R (right) and L (left) are indicated by $\hat{f}(\phi, \psi \mid C, R)$ and $\hat{f}(\phi, \psi \mid C, L)$. In addition, it includes neighbor-independent distributions $\hat{f}(\phi, \psi \mid C)$, leading to 800 potential distributions. The conditional triplet probability distribution $P(\phi, \psi \mid C, L, R)$ is calculated as:

$$P(\phi, \psi \mid C, L, R) = \frac{\hat{f}(\phi, \psi \mid C, L)\hat{f}(\phi, \psi \mid C, R)}{\mathscr{N}\hat{f}(\phi, \psi \mid C)}$$

where $\mathscr{N}$ is a normalization constant.

For simplicity, the peptide sequence is denoted as $i, i-1, i+1$. The conformation of this sequence is described by the Ramachandran angles $\{(\phi_i, \psi_i)\}$. The triplet distribution $P(\phi_i, \psi_i \mid i, i-1, i+1)$ depends on neighboring amino acids $i-1$ and $i+1$. Our goal is to find the probability distribution Π for a sequence of N amino acids, minimizing the interaction energy $K(\{(\phi_i, \psi_i)\})$:

$$K(\{(\phi_i, \psi_i)\}) = \sum_{i<j} \frac{\epsilon_{ij}}{\lambda} \left[\left(\frac{r_{0_{ij}}}{|\vec{R}_i - \vec{R}_j|} \right)^6 - 1 \right]^2 + \sum_{i<j} \frac{q_i q_j}{D|\vec{R}_i - \vec{R}_j|}$$

The parameters ϵ_{ij}, $r_{0_{ij}}$, q_i, and D represent the potential depth, the van der Waals radii, the partial charge and the dielectric constant, respectively.

The input conditional triplet distributions $P(\phi_i, \psi_i \mid i, i-1, i+1)$ are converted into $\{\rho_i(\phi_i)\}$ and $\{\tilde{\rho}_i(\psi_i)\}$:

$$\sum_{\psi_i} P(\phi_i, \psi_i \mid i, i-1, i+1) = \rho_i(\phi_i)$$

$$\sum_{\phi_i} P(\phi_i, \psi_i \mid i, i-1, i+1) = \tilde{\rho}_i(\psi_i)$$

We seek $\Pi(\{\phi_i\}, \{\psi_i\})$ that minimizes the expected cost $E[\Pi]$:

$$E[\Pi(\{\phi_i\}, \{\psi_i\})] = \min_{\{\Pi\}} \sum_{\{\Phi_i\}, \{\Psi_i\}} K(\{\phi_i\}, \{\psi_i\}) \Pi(\{\phi_i\}, \{\psi_i\})$$

where $\Pi(\{\phi_i\}, \{\psi_i\})$ satisfies:

$$\sum_{\{\psi_i, \phi_i\}, \phi_i \neq \phi_k} \Pi(\{\phi_i\}, \{\psi_i\}) = \rho_k(\phi_k)$$

$$\sum_{\{\phi_i, \psi_i\}, \psi_i \neq \psi_k} \Pi(\{\phi_i\}, \{\psi_i\}) = \tilde{\rho}_k(\psi_k)$$

This optimization is cast as a linear program (LP). Solving LP yields Π, with the most probable peptide conformations corresponding to the peaks in Π. This approach is used for tetrapeptides, hexapeptides, and octapeptides but becomes infeasible for longer peptides. Thus, we focus on tetrapeptides composed of Ala and Gly due to their simple structure. The obtained distributions are four-variable functions, making direct plotting impractical. To visualize them, we define virtual bond vectors connecting the nearest C_α atoms in the tetrapeptide. The vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$ connect $C_{\alpha_0} - C_{\alpha_1}$, $C_{\alpha_1} - C_{\alpha_2}$, and $C_{\alpha_2} - C_{\alpha_3}$, respectively. The angles between these vectors, θ_1 (between $\vec{d}_1$ and $\vec{d}_2$) and θ_2 (between $\vec{d}_2$ and $\vec{d}_3$), along with the dihedral angle α on the axis of $\vec{d}_2$, define the three-dimensional structure of the tetrapeptide. Consequently, Π_4 is expressed in terms of virtual angles θ_1, θ_2, and α.

The tetrapeptide distributions were compared with the PDB data provided by A. Sharmila (PDB(S)). These distributions are presented in terms of $\cos\theta_1 (= \vec{d}_1 \cdot \vec{d}_2)$, $\cos\theta_2 (= \vec{d}_2 \cdot \vec{d}_3)$, and $V (= (\vec{d}_1 \times \vec{d}_2) \cdot \vec{d}_3)$, where V represents the volume of the tetrapeptide and ranges from -1 to +1. This volume V is expressed as $V = \sin\theta_1 \sin\theta_2 \cos\alpha$, redefining Π_4 as $\Pi_4(\theta_1, \theta_2, V)$.To plot these distributions, Π_4 is summed over θ_1 and θ_2 to obtain $\Gamma^{MOT}(V)$ as a function of V. The corresponding PDB(S) distribution is $\Gamma^{PDB}(V)$. Distributions are obtained for 16 tetrapeptides composed of Ala and Gly.

Observations indicate that alanine-rich tetrapeptides (e.g., AAAA, AAAG, AGAA, GAAA) tend to form right-handed alpha helices. Tetrapeptides with a higher glycine content show significant flexibility, leading to various conformations due to the absence

of a C_β atom in glycine. Some tetrapeptides (e.g., GGGG, GGAG, AAGG, AAGA) form various β turns, with GGGG showing Type I, I', and II' turns, and GGAG, AAGG, and AAGA forming Type II turns.

For tripeptides, the distributions τ_1 and τ_2 are derived from Π_4, as tetrapeptides can be viewed as concatenations of two tripeptides. The torsional angles of the central amino acids in tripeptides define their conformations. Summing $\Pi_4(\phi_1, \psi_1, \phi_2, \psi_2)$ over (ϕ_2, ψ_2) and (ϕ_1, ψ_1) yields $\tau_1(\phi_1, \psi_1)$ and $\tau_2(\phi_2, \psi_2)$. Using these pairs with θ_1 and θ_2, we convert τ_1 and τ_2 to $\gamma_1^{MOT}(\cos\theta_1)$ and $\gamma_2^{MOT}(\cos\theta_2)$. The distributions of the PDB data are converted similarly to $\gamma_1^{PDB(D)}(\cos\theta_1)$ and $\gamma_2^{PDB(D)}(\cos\theta_2)$, with the corresponding PDB (S) distributions as $\gamma_1^{PDB(S)}(\cos\theta_1)$ and $\gamma_2^{PDB(S)}(\cos\theta_2)$.

Recursive Optimal Transport (ROT) for Exploring the Conformational Space of Extended Peptides

Using Multi-Marginal Optimal Transport (MOT) to ascertain the distribution of Ramachandran angles for peptides longer than a tetrapeptide through PDB(D) is computationally infeasible. This is because the cost function grows exponentially as $(m)^{2p}$, where p denotes the number of amino acids in the sequence and m stands for the number of data points in ρ' of PDB(D). With an assumed accuracy of 80%, 40 points must be chosen for each marginal. As a result, the cost function for a decapeptide matrix would be on the order of $(40)^{16}$, which is excessively large. Creating enough decapeptide structures to calculate this cost function is virtually impossible, rendering MOT via PDB(D) an unviable approach for addressing these computational issues.

To address these obstacles, we adopted an alternative approach by utilizing tetrapeptides as the foundational units for synthesizing longer peptides. Recognizing that any peptide can be assembled from tetrapeptide segments, we employ the MOT-derived distributions for tetrapeptides to construct hexapeptide distributions via two-marginal Optimal Transport (OT). In this context, the marginals represent two tetrapeptide distributions. By combining a hexapeptide with an additional tetrapeptide and applying two-marginal OT once more, we produce octapeptide distributions. This recursive method can be extended to peptides of any desired length. In the latter part of the thesis, we introduce this novel

framework to address computational issues, demonstrating that the cost function increases at a computationally feasible rate for peptides of varying lengths.

Recursive Optimal Transport Theory: The Core of Our Methodology

To tackle this issue, we suggest a procedure to determine $\Pi_n(\bar{x}_n)$. Initially, a tetrapeptide is derived from two dipeptide distributions. Following this, the tetrapeptide is treated as a foundational element for forming a hexapeptide. Our method involves converting the MOT of peptides of size n into two marginal problems in a recursive manner. This method, which reformulates MOT as a two-marginal OT at each step, is called 'recursive optimal transport' (ROT). Our method for deriving the probability distribution $\Pi_n(\bar{x}_n)$ for a peptide of size n with the sequence $A_1A_2A_3...A_n$ is as follows:

1. To understand the recursive process of optimizing the potential energy function, we first consider the case of a single tetrapeptide. This serves as a foundation for understanding how the methodology extends to longer peptide chains through fusion.

2. **Single Tetrapeptide Optimization:**

   The procedure for identifying the distribution Π_4 of the tetrapeptide sequence $A_1A_2A_3A_4$ has been covered earlier. We represent the non-zero entries in the resultant distribution as m elements.

3. **Fusion Process:** To improve this method, we combine peptides of varying lengths to create larger ones. For instance, the tetrapeptide $A_1A_2A_3A_4$ is fused with another tetrapeptide $A'_1A'_2A_5A_6$.

   a) **Fusing Condition:** The tetrapeptide $A_1A_2A_3A_4$ is combined with $A'_1A'_2A_5A_6$ in such a way that A'_1 is identical to A_3 and A'_2 is identical to A_4. This results in the creation of a new hexapeptide $A_1A_2A_3A_4A_5A_6$, referred to as the fusion condition. This condition is applicable to peptides of different lengths.

   b) **Gathering the Optimized Angles Set and Constructing Cost Function:** We describe the process here specifically for a hexapeptide, but the same method should be applied to any peptide. m_1 and m_2 denote the non-zero components in $\Pi_4(\bar{x}_4)$ and $\Pi'_4(\bar{x'}_4)$ respectively. The set $\bar{X}_4 = \{\bar{x}_4\}=\{(x_2^{\text{opt}}, x_3^{\text{opt}})\}$,

which includes m_1 optimized angles, corresponds to the two central residues of the initial tetrapeptide. In a similar manner, the set $Y_4 = \{\bar{x'}_4\} = \{(x_{2'}^{\text{opt}}, x_5^{\text{opt}})\}$, which contains m_2 optimized angles, pertains to the central residues of the subsequent tetrapeptide. Afterwards, the cost function $K_6(\bar{x}_4, \bar{x'}_4)$ is calculated, resulting in an $m_1 \times m_2$ matrix, with the Ramachandran angles for the end residues A_1 and A_6 determined by their most probable values.

4. **Establishing Two-Marginal Optimal Transport for Peptides:** Begin with the initial marginal distributions $\Pi_4(\bar{x}_4)$ and $\Pi'_4(\bar{x'}_4)$, along with the cost function $K_6(\bar{x}_4, \bar{x'}_4)$ as described in **step** 2 and 3. The distribution $\Pi_6(\bar{x}_4, \bar{x'}_4)$ is derived by employing a two-marginal optimal transport technique. In conventional optimal transport, an individual point from the source distribution is moved to an individual point in the target distribution to reduce the cost. In this case, however, the transport occurs between the tuples $\bar{x}_4$ and $\bar{x'}_4$, rather than between single points. This means that sets of points from $\bar{x}_4$ in the source marginals to $\bar{x'}_4$ in the target marginals are simultaneously considered to determine an optimal transport plan. The approach involves solving a linear programming problem with the goal of minimizing the overall transport cost given by $K_6(\bar{x}_4, \bar{x'}_4)$. The constraints ensure that the resulting distribution $\Pi_6(\bar{x}_4, \bar{x'}_4)$ aligns with the marginal distributions of the original tetrapeptides. A key part of this process is finding a basic feasible solution, which provides a simplified yet effective representation of the transport plan. In this scenario, the basic feasible solution will contain $m_1 + m_2 - 1$ non-zero entries. This figure represents the maximum number of non-zero entries in the distribution $\Pi_6(\bar{x}_4, \bar{x'}_4)$.

5. **Creating Longer Peptide Chains:** To form an octapeptide, return to **step 2** and combine the hexapeptide with an extra tetrapeptide. The hexapeptide offers $m_1 + m_2$ optimized angles, while the tetrapeptide contributes m_3 optimized angles. In **step 3**, a $K_8(\bar{x}_8)$ matrix with dimensions $(m_1+m_2)\times m_3$ is created. Next, in **step 4**, a $\Pi_8(\bar{x}_8)$ matrix is constructed, containing up to $m_1+m_2+m_3-1$ non-zero values. This method is repeated iteratively to build longer peptides, such as decapeptides, by adding tetrapeptides one after another. The cost function increases as $(\sum_{i=1}^{p-1} m_i) \times$

m_p, where p stands for the number of tetrapeptides in the sequence. Conversely, when combining an even number of peptides, like hexa-hexa or deca-deca fusion, the cost function rises as $(\sum_{i=1}^{p/2} m_i) \times \sum_{i=1}^{p/2} m_i$. In both cases, the cost function grows significantly less compared to the original multimarginal transport problem, which escalates as $(m)^{4p}$—a prohibitive exponential growth. This key realization tackles the exponentially large configuration space issue by greatly minimizing the number of possible configurations. As a result, our approach sidesteps the difficulty of handling a vast number of configurations that other methods encounter.

6. **Importance Sampling:** Although the number of configurations grows efficiently at each fusion stage in the current method, the size of $(\sum_{i=1}^{p-1} m_i)$ can be reduced further through importance sampling, thus retaining the most critical configurations. These selected probabilities $\Pi_n(\bar{x}_n)$ are then normalized to 1 for the subsequent iteration. This crucial step addresses the problem of the exponentially vast configuration space by significantly reducing the number of potential configurations while keeping the most likely ones. This iterative process ensures that only a few configurations with higher probabilities are considered, thereby enhancing the computational efficiency of the optimization.

This concludes the description of the recursive implementation of optimal transport and the fusion process.

Assessing Optimal Fusion Techniques

To emphasize that combining two peptides can be performed in multiple ways and can alter the distribution, the following procedures can be followed:

1. **Specify Peptides of Size n and Potential Fusions q:** Determine the sequences along with the possible fusion points of the two peptides, taking into account both terminal and internal positions.

2. **List Combinations:** Create a comprehensive list of all potential fusions, each representing a unique combination of the two peptides. Designate the poten-

tial q number of fusions as: $F_1, F_2, F_3, \ldots, F_q$ and their respective distributions as $\Pi_n(F_1), \Pi_n(F_2), \Pi_n(F_3), \ldots, \Pi_n(F_q)$.

3. **Compute Energy for Each Fusion:** Using optimal transport, calculate the energy $E_n(F_1)$, $E_n[\Pi_n(F_1)], E_n[\Pi_n(F_2)], E_n[\Pi_n(F_3)], \ldots, E_n[\Pi_n(F_q)]$ for each fusion.

4. **Contrast Energies:** Arrange the calculated energies and discern $\Pi_n(F_{\text{optimal}})$ such that $E_n[\Pi_n(F_{\text{optimal}})]$ is the lowest, indicating the most stable and optimal fusion.

This process is essential for identifying the fusion approach that yields the lowest energy configuration, particularly for larger peptides. By systematically evaluating all possible fusion approaches and calculating their respective energies, we can deem the approach with the lowest energy to be the most effective. This tactic is crucial for enhancing our understanding of peptide fusion and improving peptide synthesis methods. The evaluation of different fusion types has yet to be performed; it could be a subject for future research.

Structural Clustering and Data Visualization

Our approach generates multivariate distributions, but their visualization and plotting are difficult. Using Ramachandran angles for direct plotting is infeasible due to the complex and multivariate nature of the distributions. To address this issue, we apply a clustering method that leverages the three-dimensional structure of peptides and relies on the root mean square deviation (RMSD) values of C_α. RMSD serves as an efficient quantitative measure for evaluating the structural similarity between two peptide conformations in 3D space.

To calculate RMSD, two peptide structures are required: one is the reference structure, while the other is superimposed onto the reference PDB structure. The calculation includes the length of the peptide, the coordinates C_α in the reference peptide (denoted N and $\vec{R}_{\alpha_i}$), and the coordinates C_α in the target peptide (denoted $\vec{\tilde{R}}_{\alpha_i}$). The process iteratively minimizes the RMSD value to ensure accurate structural alignment.

The clustering process begins by selecting the most probable structure from Π_n of a given peptide as initial reference. RMSD values are computed for each structure relative to this reference. Based on these RMSD values, the structures are divided into two sets: one with

RMSD values below 1 Å, designated as cluster 0, and another with RMSD values equal to or greater than 1 Å.

From the remaining structures, the most probable is selected as the new reference, and the RMSD computation process is repeated. This iterative process continues until all peptide structures are grouped into distinct clusters.

After the clustering is complete, the probability associated with each cluster is determined by aggregating the probabilities of its members. These aggregated cluster probabilities are denoted as Γ_n, distinct from Π_n, which is the probability distribution obtained directly from IOT data and assigned to each peptide structure. Γ_n is assigned to each cluster.

The probabilities are visualized by plotting them against their respective clusters in a bar graph, providing a clear representation of the distribution clustering. This visualization reveals insights into the behaviour of peptides in different clusters.

Results and Discussion

We implemented a study of the conformations of a select few hexapeptides, octapeptides, decapeptides and 18 residue peptides. For any given sequence $A_1A_2A_3...A_n$, ROT generates several configurations, each with an associated probability of occurrence. These configurations are visualized in a band-like diagram, which we refer to as the "band diagram of configuration with a probability of occurrence." Here, we examine all possible outcomes and their interpretations, which we observe as follows.

1. **Dominant Configurations**: One, two, or three configurations exhibit significantly higher Π_n values compared to the remaining configurations. These dominant configurations are the most likely conformations to occur.

2. **Disordered Peptides**: A few distinct configurations have nearly similar Π_n values, while the remaining configurations have much lower values. This pattern suggests that the peptide may be disordered, with no single configuration predominating.

3. **No Unique Stabilizing Structure**: Many distinct configurations exhibit Π_n values ranging from the highest to some intermediate value. In this case, there is no

unique structure that can stabilize the peptide, indicating a lack of a predominant conformation.

By analyzing the probability distribution of these configurations, as depicted in the band diagram, we can infer the stability and structural characteristics of the peptides under study. The dominant configurations indicate potential stable conformations, while disordered patterns or a wide range of probable configurations suggest instability or structural variability.

Hexapeptide

We obtain the conformation of 13 hexapeptides: AAAAAA, AAAAAG, AAAGGA, AAAGGG, AAGGAA, AGAAAA, AGAGAG, AGGAAA, GAAAAA, GAGAGA, GGAAGG, GGGAAA and GGGGGG.

Dominant Structure

Among the 13 hexapeptides analyzed, dominant configurations were observed for AAAAAA, AAAAAG, AAAGGG, and AGAAAA. As illustrated in Figure0.1, the first column shows a band plot depicting the probability distributions (Π_6) of these peptides, while the second column shows bar graphs representing the probability distributions of each cluster (Γ_6).

A significant probability gap was observed either after the second most dominant structure, as in AAAAAA and AAAGGG, or after the most dominant structure, as in AAAAAG and AGAAAA. For example, the most dominant structure in AAAAAG corresponds to a right-handed alpha helix with a Π_6 gap of approximately 0.01, similar to AGAAAA.

For AAAGGG and AAAAAA, a notable gap exists between the second and third most probable structures. In AAAAAA, the first two dominant structures, both right-handed alpha helices, have very close Π_6 values.

Hydrogen bonds in the helical structures within cluster 0 are often bicentered. Comparing these peptides to the reference α helix and 3_{10} helix structures using the root mean square deviation (RMSD) measures shows that some peptides in cluster 0 are closer to the 3_{10} helix. For example, in AAAAAA, the total probability of cluster 0 (Γ_6) is nearly

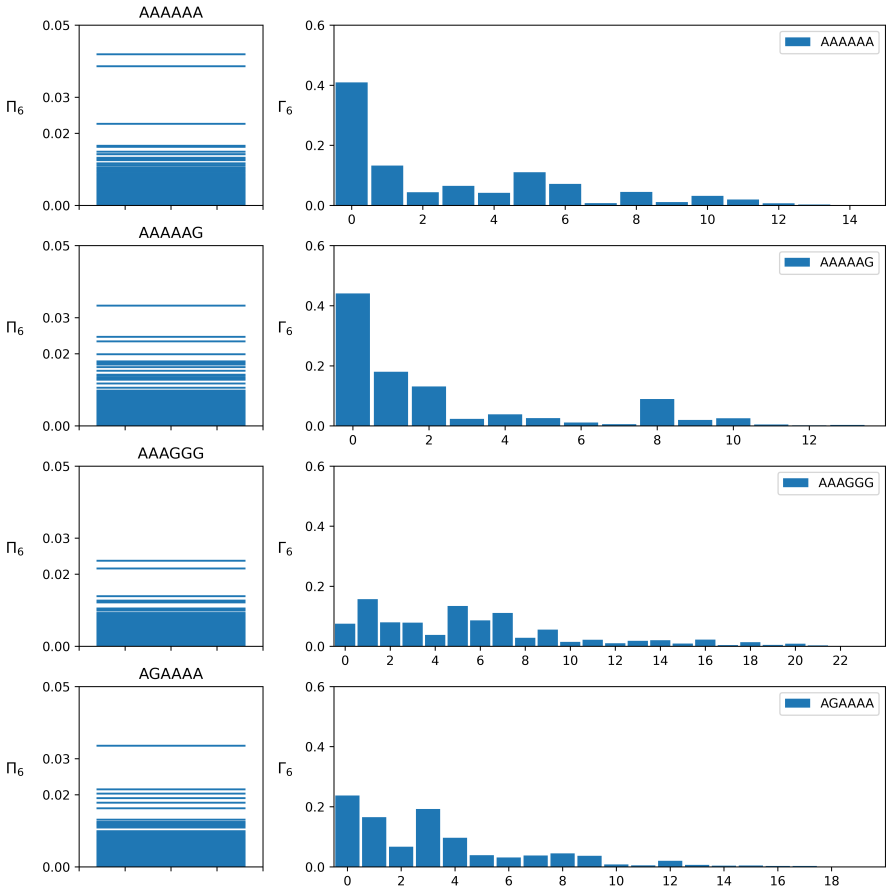


Figure 0.1: The figure shows peptides with a few dominant configurations having large band gaps. Each row corresponds to a particular peptide. The first and second columns display the band plot and bar plot, respectively.

evenly split between structures that resemble the 3_{10} helix (0.209) and the α-helix (0.20). Similarly, in AGAAAA, the total Γ_6 value of 0.23 consists primarily of structures closer to the 3_{10} helix (0.19), with the remainder closer to the α-helix.

In AAAGGG, the most probable conformation is a 3_{10} helix, followed by a type II' β turn in the middle residues. The probability gap between the second and third most probable structures in AAAAAA and AAAGGG is approximately 0.01.

The dominant structures of AAAAAA and AAAGGG fall into clusters 0 and 1, respectively. For AGAAAA and AAAAAG, the dominant structures are located within cluster 0, with Γ_6 values corresponding to right-handed alpha helices.

Additional structures such as PPII helices and type I and II' turns are observed in these hexapeptides. PPII helices appear in cluster 1 of AAAAAA and AAAAAG. Type I turns are seen in the last four residues of AAAAAA and AAAAAG.

In some cases, the peptides in cluster 0 of AAAAAG are closer to the 3_{10} helix. Type II β turns are observed in clusters of AAAGGA, AAGGAA, AGAGAG, and GAGAGA. Type II' β turns appear in the last four residues of AAAGGA and AGGAAA.

Eight of the 13 hexapeptides (e.g., AAAGGA, AAGGAA, AGAGAG) fall into the category of disordered peptides, with minimal probability gaps. In AAAGGA, the most probable structure is a random configuration followed by a 3_{10} helix. AAGGAA has a 3_{10} helix and a type I turn.

Other structures such as type I and type II turns are observed in various hexapeptides. For example, type I turns are found in the last four residues of AAAAAA, AAAAAG, and GAAAAA. Type II β turns are seen in clusters of AAAGGA and AAGGAA.

Only GGGGGG among the analyzed hexapeptides shows negligible gaps between its most probable structures. Type II' and I' turns are observed in GGGGGG, with the type II' turn appearing in different residue positions.

This comprehensive analysis shows that peptides can exhibit diverse secondary structures with varying probabilities, influenced by their specific sequences and resulting in different structural clusters.

Analysis of Octapeptides

Dominant Structure

The octapeptide distributions for AAAAAAAA, AGAGAGAG, GAGAGAGA, and AGAAGAGG are obtained by concatenating the hexapeptide distributions of the first 6 and last 4 residues.

AAAAAAA: The most probable structure is a right-handed alpha helix ($\Pi_8 = 0.0419$) with subsequent structures also being right-handed alpha helices. The gaps between the most probable structures are 0.007 and 0.012. The first three structures fall into cluster 0 ($\Gamma_8 = 0.26$).

AGAGAGAG: The most probable structure is a repeated type II β turn ($\Pi_8 = 0.023$), with a gap of 0.013 between the first and second most probable structures. This structure falls into cluster 0 ($\Gamma_8 = 0.099$).

AGAAGAGG: The most probable structure involves two repeated type II β turns and a random structure ($\Pi_8 = 0.021$), with a gap of 0.006 between the first and second structures. This structure is in cluster 0 ($\Gamma_8 = 0.092$).

For AAAAAAAA, structural comparisons show that 60% of the peptides are closer to the alpha helix and 40% are closer to the 3_{10} helix. Various other structures, such as the 3_{10} and PPII helices, are also observed in AAAAAAAA and AGAGAGAG. Beta turns, like Types I and II, are present, with some parts adopting random structures.

Disordered Structure

GAGAGAGA fits into the disordered category. The most probable structure is a 3_{10} helix ($\Pi_8 = 0.012$), with a small gap of 0.001 to the second structure. The most probable structure falls into cluster 0 ($\Gamma_8 = 0.143$). Distorted 3_{10} helices with weak hydrogen bonds are observed. Type II turns are found in various clusters, and repeated Type II turns (beta bends) span the entire backbone in some clusters.

Decapeptides

Dominant Configuration: Only AAAAAAAAAA falls into this category, forming a right-handed alpha helix with significant gaps between its most probable structures. It also exhibits 3_{10} and PPII helices.

Disordered Peptides: The remaining nine decapeptides (AAGGAAGGAG, AAGGAGAAGG, AGGAGAAGGA, GAGAAGGAAG, AGAAGGAAGG, GGAGAAGGAA, AGAGAGAGAG, GAGAGAGAGA, GGGGGGGGGG) display disordered configurations, often forming various types of β turns, 3_{10} helices, and random coils, with small gaps between their most probable structures. Multiple clusters demonstrate these structural motifs, highlighting their diversity and lack of dominant structure.

18-Residue Peptide (A_8GA_9)

Short Helical Structures: Clusters display short 3_{10} helices transitioning into longer helices or α helices, linked by random coils.

Helix-Coil Combinations: Some clusters exhibit combinations of helices and coils, including PPII helices and Type II turns.

Helices Connected by Turns: Clusters feature helices connected by various turns (Type I, II, and II').

Transitions Between Helices and Random Structures: Clusters show transitions involving 3_{10}, α, and PPII helices connected by turns or random structures.

Beta Turns and Random Structures: Various clusters are characterized by combinations of beta turns and random structures, with single turns linking random segments.

In general, the analysis reveals a range of structural motifs and configurations in both decapeptides and the 18-residue peptide, highlighting the complexity and variability of peptide structures.

Conclusion

In summary, we demonstrated the construction of peptide conformations using optimal transport theory, utilizing tetrapeptide distributions derived from multi-marginal transport based on PDB dipeptide distributions. This new method termed the "Recursive Optimal Transport" (ROT) technique, is specifically designed to address the complex issue of managing the vast configurational spaces of elongated peptides, which is a well-known obstacle in computational simulations. Unlike AlphaFold, which employs deep learning models to predict protein structures from sequence data, our approach utilizes optimal transport theory to systematically and efficiently navigate these large spaces. The core of the ROT approach lies in the deconstruction of the intricate structure of long peptides into smaller, more manageable units, such as dipeptides and tetrapeptides. These smaller units are incrementally assembled to form longer peptide chains. This method focuses on identifying the optimal configurations from a limited set of relevant options, effectively breaking down a high-dimensional challenge into several lower-dimensional problems, thereby significantly boosting computational efficiency. Sequentially, these smaller configurations are integrated through a series of optimal transport calculations to forecast the structure of larger peptides, with the cost function determined by physical interactions and probabilistic constraints.

The ROT method has been utilized to simulate a few hexapeptides, octapeptides, a decapeptide, and an 18-residue sequence, proving its capability to capture peptide con-

formations. This recursive technique streamlines the modelling workflow and enhances predictive accuracy, offering a comprehensive understanding at each stage of peptide formation. However, because of the scarce and fragmented experimental data available for larger peptides, making meaningful comparisons is not feasible.

List of Figures

Chapter 1

Introduction and Motivation

Peptides and Their Structural Insights

Peptides, illustrated at the top of Fig. 1.1, are chains of amino acids connected by peptide bonds, performing crucial biological functions. In this diagram, circles represent amino acids, while lines denote the bonds linking them. The internal atomic structure of an amino acid is shown at the bottom of the figure. Backbone torsional angles, ϕ and ψ, are included to illustrate structural variability, which will be explored in detail in Chapter 2. Peptides, typically comprising 2–50 amino acids, are shorter than proteins but are equally significant in biological processes [23]. They act as hormones, antimicrobial agents, and metabolic regulators [1–3, 24]. Their structural diversity arises from the 20 naturally occurring amino acids, allowing a wide range of functions [3].

In drug development, peptides are advantageous due to their lower synthesis costs and simpler production compared to protein-based therapeutics [3, 25, 26]. Over 7,000 unique peptides have been identified in humans, highlighting their roles in cell signaling, neurotransmission, and the composition of toxins and venoms [1, 2, 24]. Their structural versatility makes them critical in molecular biology and pharmaceutical sciences [27].

This structural versatility stems from the ability of peptides to adopt various conformations, ranging from disordered states to well-defined secondary structures. These conformations, including α helices, β hairpins, and PPII helices, enable peptides to interact with a wide array of biomolecules, thus underpinning their biological significance [28, 29]. Understanding these secondary structures is key to exploring the functional roles of peptides

in biological systems.

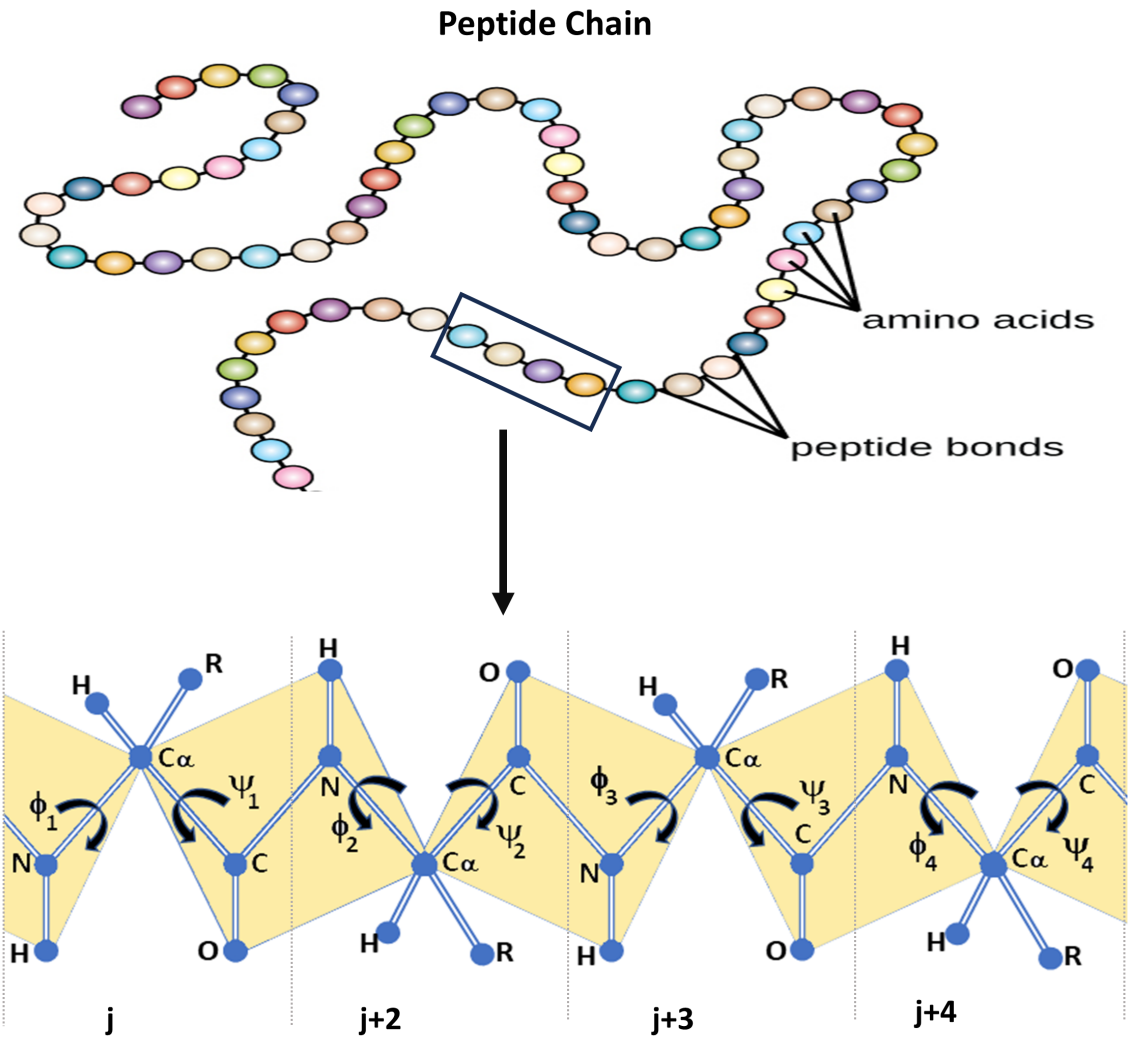


Figure 1.1: Illustration of a peptide chain. Each circle in the peptide chain represents an amino acid, with the internal structure of four residues shown below the chain. 'R' in the backbone indicates the side chain.

Secondary Structures of Peptides

Peptides adopt various secondary structures, such as α helices, β hairpins, PPII helices, and several types of β turns, which enhance their stability and functionality [30–32]. These structures are defined by repeating backbone torsional angles (ϕ, ψ) and specific hydrogen-bonding patterns, making them vital for biological interactions.

The α Helix

The α helix, depicted in Fig. 1.2, is a coiled structure stabilized by hydrogen bonds between the carbonyl group of the i^{th} residue and the amine group of the $(i+4)^{\text{th}}$ residue. A single turn contains approximately 3.6 amino acids, with a vertical rise of 5.4 Å per turn. The backbone torsional angles for a right-handed α helix are typically $(-60°, -45°)$, while a left-handed α helix adopts angles around $(60°, 45°)$ [33, 34].

Other Helices

The 3_{10} helix and PPII helix (Fig. 1.2) differ in geometry and hydrogen-bonding patterns:

- 3_{10} **Helix**: Characterized by hydrogen bonds between the i^{th} residue and the $(i+3)^{\text{rd}}$ residue, with three residues per turn. Backbone angles are approximately $(-49°, -26°)$[35, 36].
- **PPII Helix**: Formed without hydrogen bonds, PPII helices are observed in proline-rich peptides and alanine-rich sequences, favoring backbone angles $(-75°, 150°)$ [37, 38].

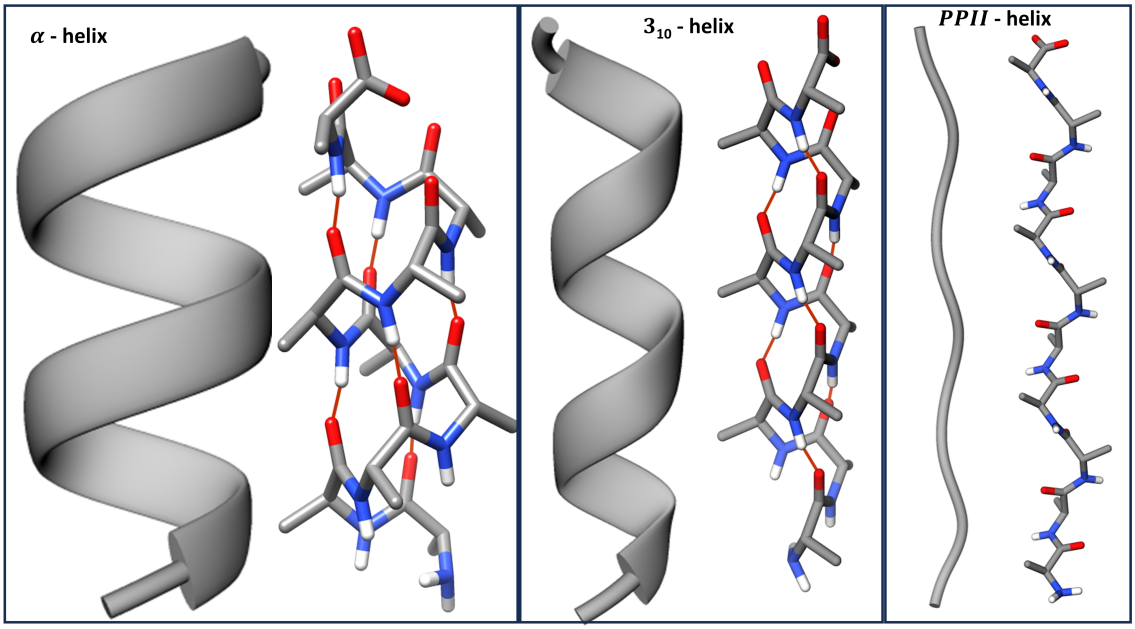


Figure 1.2: Various helical structures: α helix (left), 3_{10} helix (middle), and PPII helix (right). All-atom backbone structures and cartoon depictions are shown.

The β Hairpin

The β hairpin consists of two β-strands linked by hydrogen bonds along their backbone (Fig. 1.3). Its typical backbone angles are $(-135°, 135°)$ [39, 40].

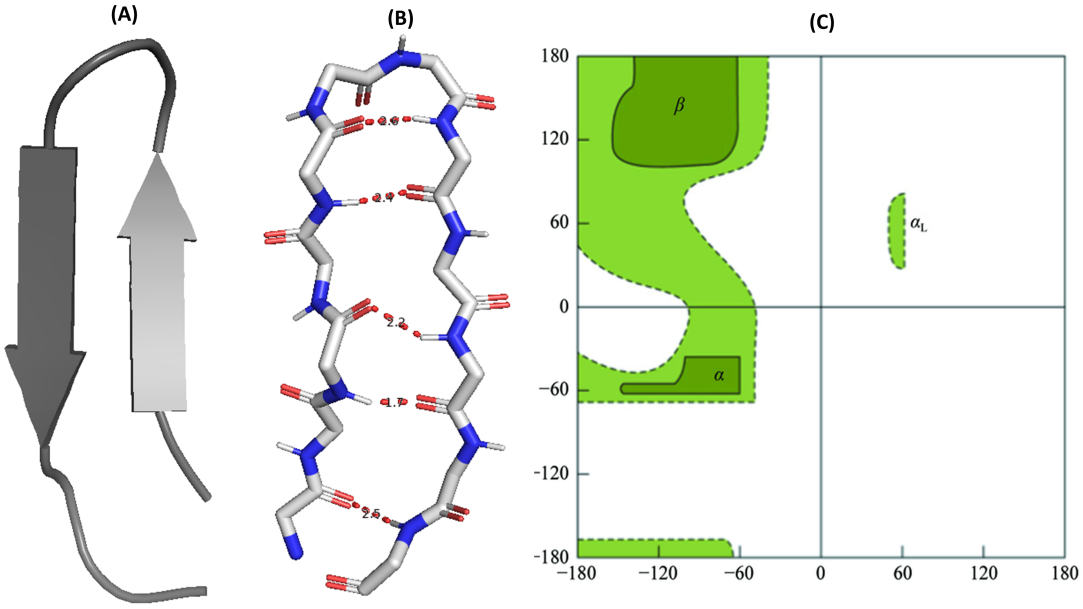


Figure 1.3: Beta-hairpin structure: (A) Backbone cartoon image, (B) all-atom backbone structure, (C) Ramachandran plot showing allowed (green) and disallowed (white) regions. Alpha helix and beta-sheet angles are marked. Ramachandran plot adapted from [21].

β Turns

β turns reverse peptide chain direction and stabilize compact protein structures. A typical β turn involves four residues, with a hydrogen bond between the carbonyl oxygen of the i^{th} residue and the amide hydrogen of the $(i+3)^{\text{rd}}$ residue (Fig.1.4). Common β-turn types include:

- **Type I**: $(\phi_{i+1}, \psi_{i+1}) = (-60°, -30°)$, $(\phi_{i+2}, \psi_{i+2}) = (-90°, 0°)$
- **Type II**: $(\phi_{i+1}, \psi_{i+1}) = (-60°, 120°)$, $(\phi_{i+2}, \psi_{i+2}) = (80°, 0°)$
- **Type I'**: $(\phi_{i+1}, \psi_{i+1}) = (60°, 30°)$, $(\phi_{i+2}, \psi_{i+2}) = (90°, 0°)$
- **Type II'**: $(\phi_{i+1}, \psi_{i+1}) = (60°, -120°)$, $(\phi_{i+2}, \psi_{i+2}) = (-80°, 0°)$

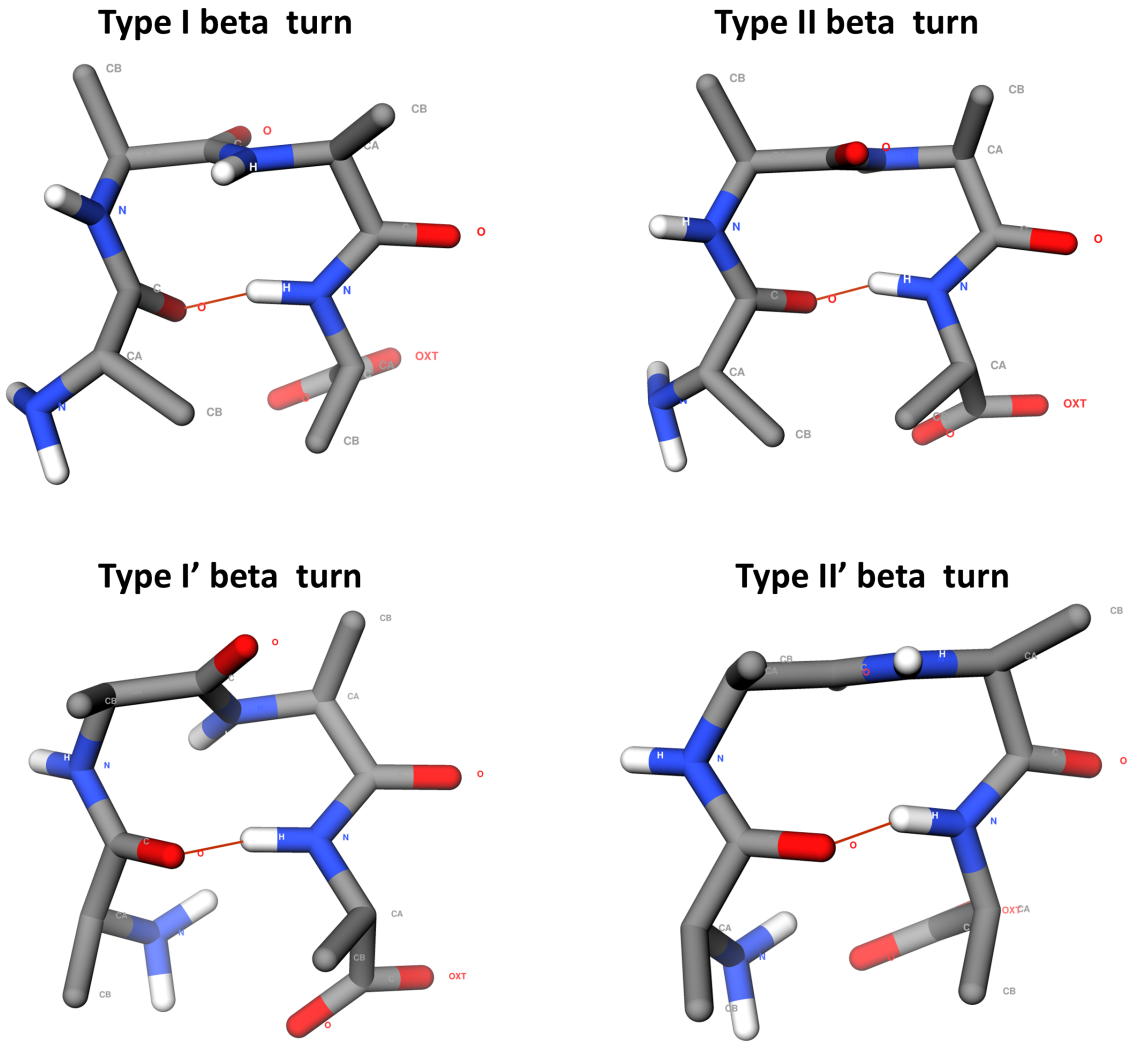


Figure 1.4: Four types of β turns, with hydrogen bonding between the first and fourth residues.

Conformational Flexibility of Peptides

Peptides exhibit remarkable conformational flexibility, transitioning between disordered states and well-defined 3D shapes [28]. This adaptability allows them to interact with diverse biomolecules, such as proteins and lipids, and is influenced by their environment [41].

The energy landscape framework is a key tool for studying peptide conformations, providing insights into preferred states and the mechanisms driving transitions [42]. By mapping the potential energy surface (PES) of peptides, researchers can identify bioactive forms and understand the structural dynamics underlying biological functions [43, 44].

Ramachandran Plot and Structural Insights

The Ramachandran plot (Fig. 1.3) visualizes backbone torsional angles (ϕ, ψ) for each residue, highlighting allowed and disallowed regions based on steric constraints [45]. Specific regions correspond to secondary structures:

- α helices: $(-60°, -45°)$
- β-sheets: $(-135°, 135°)$
- PPII helices: $(-75°, 150°)$

Understanding the structural flexibility and secondary structures of peptides is essential for advancing fields like molecular biology, protein engineering, and drug design. Their ability to adopt specific conformations underpins their interactions and functionalities, making them vital components in biological research and therapeutic development.

1.1 Motivation

The study of peptide conformations is highly valuable, especially for understanding their role in modeling protein folding and molecular interactions. Peptides frequently adopt structural motifs similar to those found in proteins, offering crucial insights into folding pathways and mechanisms essential for protein engineering. Additionally, NMR studies have demonstrated that peptide fragments in solution often replicate the conformations observed in their native protein structures, revealing a natural conformational bias that deepens our understanding of protein architecture [46]. These findings not only improve computational models of protein folding but also contribute to the design of functional biomolecules and novel therapeutic agents.

Furthermore, peptides are effective building blocks for protein structure prediction[47]. Their intrinsic conformational tendencies provide a basis for developing more precise models of intricate protein structures. This modelling capability is vital for progress in the understanding of protein function and supports the design of novel proteins with desired traits. Computational methods that use data from sources like the Protein Data Bank (PDB) to extract and analyze these conformational tendencies are critical, allowing

researchers to refine predictive algorithms and enhance structural prediction accuracy, thereby aiding new drug and therapy development.[48–50] In summary, understanding peptide conformations underpins advances in both structural biology and drug discovery. By using computational tools to predict and map these conformational preferences, one can improve protein structure predictions and utilize peptide attributes to create targeted and effective therapies.

1.2 Energy Landscape and Levinthal Problem

Understanding peptide conformations and the concept of the energy landscape is crucial for improved analysis[51–54]. This framework provides an in-depth perspective on the potential energy surface associated with diverse peptide structures. Each point on this multidimensional surface denotes a specific conformation and its related energy, shedding light on the stability of various states and transitions. Originally created for spin glass systems research, this concept has been adapted to molecular biology to reveal the intricate interactions among peptide conformations, stability, and folding pathways[43, 55].

While quantum mechanical calculations can provide very precise energy landscapes, their usage for large peptide structures is constrained by high computational demands. As a feasible alternative, empirical interaction potentials—typically based on experimental data—are widely utilized. These potentials facilitate efficient modeling that maintains a compromise between computational practicality and the precision essential for accurate representations of peptide behavior. By analyzing the energy landscape, we can pinpoint low-energy conformations that are biologically significant and examine how peptides transition between structural states, thereby complementing the study of conformational preferences obtained from resources like the Protein Data Bank (PDB). This thorough approach advances the development of predictive models for structural biology.

Expanding on the discussion of peptide conformations and their corresponding energy landscapes, Figure 1.5 presents a simplified depiction of a one-dimensional energy landscape characterized by multiple local minima separated by energy barriers. While this visualization aids in understanding the concept, the true energy landscape of peptides is far more complex due to the significant degrees of freedom in their structures. Each

conformation corresponds to a unique position on this multidimensional surface, leading to a multitude of local energy minima that represent stable or semi-stable peptide states. The challenge of pinpointing the global energy minimum, which often correlates with the bioactive form of a peptide. This form is typically associated with the most stable conformation under physiological conditions and is crucial for effective binding and function. However, the "multiple minima problem" [56–58] arises due to the landscape's complexity, as the numerous local minima can mislead search algorithms and make locating the global minimum highly non-trivial.

This complexity becomes exponentially greater as the peptide chain length increases, vastly expanding the number of possible conformations. This exponential growth is epitomized by the "Levinthal Paradox," [59–61] .In 1969, Cyrus Levinthal noted that if a 100-residue protein had just three possible backbone states per residue, it would face

$$3^{100} \approx 5 \times 10^{47}$$

total conformations. Even at an impossibly fast 10^{13} tries per second, a blind search would take over 10^{27} years—yet proteins fold in seconds. This is the *Levinthal paradox.*[59]

Zwanzig, Szabo, and Bagchi (1992) showed with a simple kinetic model that adding a small energy penalty (a few kT) against locally unfavorable shapes reduces folding times to the millisecond–second range. Bryngelson & Wolynes (1987) introduced the idea of a *folding funnel*, in which many downhill pathways guide the chain toward its native state. Onuchic & Wolynes (1997) refined this with *kinetic partitioning*, explaining why some molecules fold faster than others. Later, Markov-state models (Pande & Rokhsar 2002; Noé & Fischer 2008) built networks of key metastable states and transitions, reducing the search to a manageable graph [62–64] .

Our ROT method applies the same principle in probability space. Here the conformational bias is created locally through dipeptide distributions along with interaction bias. Such probabilistic bias created at each step of fusion aids to sample the conformational space of peptides

All these suggests that an exhaustive search for the global minimum would be computationally unfeasible due to the immense conformational space. To navigate this challenge, researchers simplify peptide analysis by focusing on backbone torsional (Ramachandran)

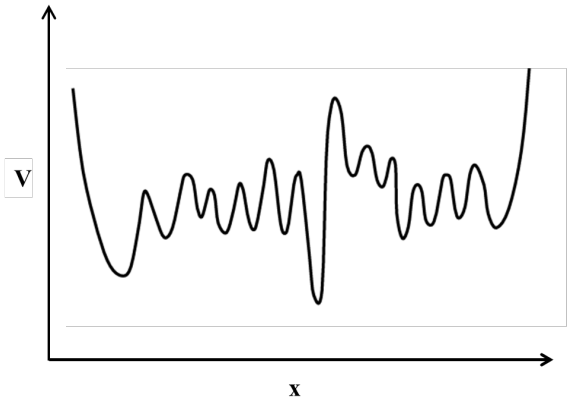


Figure 1.5: The figure shows a dimensional energy landscape with lots of local minima. However, this image is under-representing the number of local minima in the energy landscape.

angles, as bond lengths and bond angles tend to remain relatively constant. This approach narrows the scope of the analysis while still providing a meaningful representation of conformational variability.

Understanding and visualizing the energy landscape, even in simplified forms, is crucial for developing strategies to efficiently explore conformational space. This allows for the identification of energetically favorable pathways that peptides may follow during folding and binding interactions.

1.3 Techniques for Exploring Peptide Conformational Space

Understanding the conformational landscape of peptides requires the integration of experimental and computational methodologies. While experimental techniques such as Nuclear Magnetic Resonance (NMR) provide critical structural insights, they are inherently limited in capturing the full breadth of conformational diversity due to constraints like time resolution and sample preparation. Computational approaches, including molecular dynamics (MD) and Monte Carlo (MC) simulations, are indispensable for complementing experimental data by generating a wide range of conformational states and modelling transitions between them[11, 65–68].

Molecular Dynamics (MD) Simulations: Capturing Time-Dependent Dynamics

Molecular dynamics (MD) simulations provide a time-resolved view of peptide behaviour by modelling the physical motions of atoms under the influence of interatomic forces. By solving Newton's equations of motion, MD simulations predict how atomic positions and velocities evolve over time:

$$\vec{F}_i = -\frac{\partial V(\vec{r}_i, \ldots, \vec{r}_j)}{\partial \vec{r}_i}, \tag{1.1}$$

$$\vec{F}_i = m_i \frac{d^2 \vec{r}_i}{dt^2}, \tag{1.2}$$

where $\vec{F}_i$ is the force on atom i, V is the potential energy function, $\vec{r}_i$ represents atomic coordinates, and m_i is the atomic mass. These equations are solved iteratively using numerical integration methods such as the Verlet or leapfrog algorithms, with small time steps (Δt) ensuring accurate resolution of atomic motions[69].

MD simulations are particularly valuable for exploring conformational transitions and identifying stable and meta-stable states in the energy landscape[9, 10, 66, 68]. They are widely used to study the folding pathways of peptides, their interactions with other molecules, and the effects of environmental conditions. However, in systems with rugged energy landscapes and multiple local minima, MD simulations can become trapped in these minima, limiting their ability to sample the global energy landscape effectively. Enhanced sampling methods, such as replica exchange molecular dynamics (REMD), address this limitation by simulating multiple replicas at different temperatures and allowing configuration exchanges to overcome energy barriers.[70–72]

Monte Carlo (MC) Simulations: Stochastic Sampling of Conformations

Monte Carlo (MC) simulations offer an alternative, stochastic approach to explore peptide conformations[73, 74]. Unlike MD, which tracks the time evolution of atomic positions, MC simulations generate new conformations by randomly perturbing atomic coordinates and evaluate their acceptance based on the Metropolis criterion:

$$P = \min\left(1, \exp\left(-\frac{\Delta V}{k_B T}\right)\right),$$

where ΔV is the energy difference between the current and proposed conformations, k_B is the Boltzmann constant, and T is the temperature. This probabilistic framework enables MC simulations to efficiently sample diverse regions of the conformational space, even in systems with rugged energy landscapes.

MC simulations are particularly well-suited for overcoming local energy minima, as the stochastic nature of the method allows for larger conformational jumps that can cross energy barriers. Techniques such as replica exchange Monte Carlo (REMC) further enhance sampling efficiency by simultaneously running simulations at multiple temperatures and exchanging configurations to facilitate exploration of high-energy regions.[71]

Despite their advantages, MC simulations have limitations. The efficiency of sampling depends on the choice of move sets, which determine how new conformations are proposed. Optimized move sets and hybrid approaches that combine MC with other techniques can improve the robustness and coverage of conformational space exploration.

While MD and MC simulations each have strengths and limitations, integrating these approaches offers a powerful strategy for exploring peptide conformations comprehensively. MD provides detailed kinetic and dynamic information, while MC excels at sampling diverse conformations across the energy landscape. Hybrid methods, such as MD simulations guided by MC sampling, can leverage the advantages of both techniques to achieve both extensive coverage and accurate modeling of conformational transitions.

Advanced Computational Techniques and the Role of AlphaFold

In addition to MD and MC simulations, other computational approaches play a significant role in peptide conformational analysis. Quantum mechanical (QM) calculations, known for their precision, provide highly accurate representations of potential energy surfaces by considering electronic interactions at an atomic level. However, these calculations are computationally prohibitive for larger peptide systems due to the extensive number of variables involved and the high processing power required. This limitation restricts their

application primarily to small peptides or specific regions of larger peptides where high accuracy is crucial.

To address the limitations of purely quantum or classical methods, hybrid quantum mechanics/molecular mechanics (QM/MM) approaches have been developed. These methods offer a compromise by combining the accuracy of quantum mechanics for regions of interest, such as active sites or specific peptide bonds, with the efficiency of classical mechanics for the rest of the peptide. This enables detailed energy calculations in localized areas while maintaining feasible computational costs for the entire system. QM/MM methods are particularly beneficial when studying reactions, metal-peptide interactions, or specific conformational changes that demand quantum-level detail but do not require it for the entire structure.

AlphaFold represents a major breakthrough in the field of protein and peptide structure prediction [14, 75]. Developed by DeepMind, AlphaFold leverages deep learning algorithms and massive amounts of training data to predict the three-dimensional structures of proteins and peptides based solely on their amino acid sequences. Unlike traditional methods that rely heavily on energy landscape exploration and sampling, AlphaFold bypasses this process by using trained neural networks to infer structural features directly from sequence data. The result is a rapid and highly accurate prediction of static structures, which has set new benchmarks in the field of computational biology.

Although AlphaFold's primary focus has been on predicting protein structures, it has shown remarkable success in handling peptides as well. Its ability to provide quick and accurate static structure predictions makes it an invaluable tool for initial structural analysis. This is particularly useful for researchers who need a reliable starting model for further analysis or experimental validation. However, despite its strengths, AlphaFold has limitations. It does not provide insights into the dynamic behavior of peptides over time, a crucial aspect of understanding peptide functionality in a biological context[76–78]. The dynamic nature of peptides, including their conformational changes, folding pathways, and interactions with other molecules, is essential for comprehensive modelling and analysis. This is where MD and MC simulations remain indispensable, as they can track the temporal evolution of peptide conformations and capture the range of states a peptide may adopt under different conditions.

The integration of these diverse computational methods—ranging from high-accuracy QM calculations to efficient deep learning algorithms like AlphaFold—provides a robust toolkit for researchers studying peptide conformations. Each approach contributes unique strengths, from the detailed electronic interactions captured by QM and QM/MM methods to the fast and precise structure predictions made possible by AlphaFold, complemented by the dynamic insights gained from MD and MC simulations. Together, these methods offer a comprehensive framework for peptide analysis, enhancing our understanding of their structural properties and biological functions.

1.4 Data-Based Methods for Peptide Analysis

An alternative method for studying peptide conformations leverages data from the Protein Data Bank (PDB), a rich repository of experimentally derived protein structures. [5, 6, 20, 79–83]By mining this database, researchers can analyze the frequency and occurrence of specific amino acid sequences and represent these findings as probability density functions (PDFs) of backbone torsional angles (such as ϕ and ψ angles in the Ramachandran plot). These PDFs provide an empirical representation of the conformational space that peptides occupy, revealing their intrinsic preferences for certain structural arrangements. This data-driven approach allows researchers to discern patterns and tendencies in peptide behavior, which can be used to predict conformational states and inform computational models.

However, despite the advantages of using PDB data, significant challenges arise as the length of the peptide chain increases. The occurrence data for longer peptide sequences becomes sparse due to the combinatorial explosion of possible conformations and sequences.[7, 20] This sparsity limits the statistical power of the generated PDFs, making it difficult to accurately capture conformational preferences for longer peptides or rare amino acid sequences. Consequently, while this method can provide valuable insights for short peptide fragments or common motifs, it struggles to extend to more complex, longer peptides without substantial data augmentation or alternative analytical approaches.

Research into peptide conformational preferences has highlighted that even short peptide fragments have inherent tendencies to adopt certain secondary structures. These preferences are significant in the context of protein structure prediction algorithms, where accu-

rate modeling of peptide fragments can greatly influence the overall predicted structure of a protein. Studies using techniques like Nuclear Magnetic Resonance (NMR) spectroscopy have shown that peptide fragments in solution often exhibit a bias toward conformations similar to those found in their native protein structures, such as alpha helices, beta sheets, and beta turns[30, 84, 85]. These findings support the idea that peptide conformation is not solely determined by sequence but also by intrinsic propensities governed by the backbone's torsional constraints and interactions.

1.5 Data-Driven Approaches for Peptide Conformation Analysis

Building on the discussion of computational models such as MD and MC simulations, data-driven methods offer an alternative approach to studying peptide conformations. By leveraging experimentally derived structures from the Protein Data Bank (PDB), researchers can extract valuable insights into the conformational tendencies of peptides without relying solely on simulations. The PDB serves as a vast repository of protein structures, capturing the spatial arrangements of amino acids in diverse biological systems. This information can be analyzed to identify patterns and preferences in peptide conformations, providing a foundation for empirical modelling.

Using PDB data, researchers represent the frequency and distribution of backbone torsional angles, such as ϕ and ψ angles, as probability density functions (PDFs). These PDFs reflect the observed conformational space occupied by peptides, offering insights into intrinsic structural preferences. Visualized through tools like the Ramachandran plot, this analysis reveals how specific amino acid sequences tend toward particular secondary structures, such as alpha helices, beta sheets, or beta turns. Such empirical representations can complement computational models by narrowing down conformational possibilities and guiding predictions.

However, the utility of this approach diminishes as peptide chain length increases. The combinatorial explosion of potential conformations and sequences creates data sparsity, making it challenging to generate robust PDFs for longer peptides or rare sequences. This limitation restricts the statistical power of data-driven models and hinders their

ability to accurately predict conformations for complex peptides. While effective for short fragments or common motifs, the approach becomes less reliable for extended sequences unless augmented with alternative methods or enriched datasets.

Experimental studies, such as those using Nuclear Magnetic Resonance (NMR) spectroscopy, reinforce the observation that even short peptide fragments exhibit intrinsic preferences for adopting secondary structures similar to those found in native proteins. These findings highlight that peptide conformations are not solely sequence-dependent but are also influenced by backbone torsional constraints and environmental interactions. Such tendencies play a critical role in protein folding models, where the accurate representation of peptide fragments can significantly affect the overall predicted structure of a protein.

In addition to their importance in protein structure prediction, insights from PDB-based analysis have direct applications in drug discovery. Peptides are often employed as scaffolds for designing therapeutics due to their ability to interact with target proteins and modulate biological functions. Understanding the conformational preferences of a peptide allows researchers to design analogues with improved stability, specificity, and bioactivity. This knowledge helps reduce off-target effects and enhances the therapeutic potential of peptide-based drugs.

While the limitations of data sparsity and sequence variability pose challenges, PDB-based methods remain a valuable resource for peptide conformational analysis. By integrating these empirical insights with computational simulations and experimental techniques, researchers can better understand the structural behaviour of peptides, advancing fields like structural biology and pharmaceutical design.

1.6 Proposed Method: Multi-Point Probability Distributions via Optimal Transport Theory

This thesis addresses the challenge of data sparsity in determining the conformations of larger peptides by introducing a novel method for deriving multi-point probability distributions (MPD) using optimal transport theory and data from the Protein Data Bank (PDB)[16, 86–90]. The objective is to accurately represent the conformational space

that peptides can occupy. Traditional approaches often rely on single-residue probability distributions, which fail to capture the intricate interdependencies between residues in a peptide chain. By incorporating correlations between backbone torsional angles across the sequence, this method provides a more comprehensive understanding of the peptide's conformational landscape.

The foundation of this method lies in generating high-quality triplet distributions, which encode critical short-range correlations among torsional angles. These distributions serve as inputs for an optimization process based on a potential energy function. The approach minimizes the energy landscape while preserving essential conformational properties, ensuring both accuracy and computational efficiency. The methodology has been applied to tetrapeptides composed of alanine (Ala) and glycine (Gly). These amino acids were chosen for their simplicity and lack of bulky side chains, which make them ideal candidates for demonstrating the effectiveness of the method without unnecessary structural complexity.

Tetrapeptides as Building Blocks for Longer Peptides

This thesis proposes that tetrapeptides act as fundamental building blocks for larger peptides. To address the challenge of data sparsity in longer peptide chains, the method begins by determining distributions for tetrapeptides based on PDB data. By leveraging these distributions, it becomes possible to generate multi-point probability distributions for longer peptides, addressing data limitations and creating a scalable approach for modelling peptide conformations. This ability to construct longer peptide distributions from tetrapeptide data forms the central idea of this research.

The computation of MPD for tetrapeptides is achieved using the multi-marginal optimal transport (MOT) framework. This method extends classical optimal transport by considering multiple distributions simultaneously and determining an optimal joint distribution that minimizes a given cost function, specifically the potential energy of the peptide. This ensures that the resulting distributions satisfy the input triplet distributions as marginals while achieving a low-energy state, making it a robust and efficient tool for conformational analysis.

The MOT framework is noteworthy for its versatility, as it has found applications in fields as diverse as mathematics, physics, and economics. Its capacity to handle complex

distributional problems with multiple constraints makes it particularly suited to the challenges of peptide conformational modelling. By applying MOT to the study of peptide structures, this research introduces an innovative methodology for understanding peptide conformations.

Implications and Applications of the Approach

Understanding peptide conformations using this method has far-reaching implications. The integration of energy landscapes, computational simulations, and data-driven insights offers a unified view of peptide behaviour. This enriched understanding reveals how peptides adopt functional conformations, interact with other molecules, and respond to environmental changes. Such insights are invaluable in drug discovery, where the identification of stable and transitional peptide states can guide the design of therapeutics with enhanced efficacy and fewer side effects.

The computational precision achieved through the use of optimal transport theory also improves predictive models in structural biology. These models are critical for designing peptide-based drugs, where a clear understanding of conformational preferences can inform the creation of molecules tailored to specific biological functions. The methodologies developed in this research open pathways for extending these insights to longer and more complex peptides, with significant implications for understanding protein folding and advancing biotechnology.

Organization of the Thesis

The thesis is organized into the following chapters to systematically present the development and application of the proposed methods:

- **Chapter 2**: Explores the geometry of peptides and the development of the potential energy function used as the cost function in the optimal transport framework.
- **Chapter 3**: Provides a detailed overview of optimal transport methods, focusing on their adaptation to peptide conformational analysis.
- **Chapter 4**: Presents the results of the tetrapeptide analysis, showcasing the application and validation of the method.

- **Chapter 5**: Expands the approach to address the conformations of longer peptides, building on the findings from the tetrapeptide analysis.
- **Chapter 6**: Concludes with a summary of the contributions, potential limitations, and a discussion of future directions in peptide modelling and its applications.

Through this structure, the thesis constructs a cohesive narrative, linking the theoretical development of the method to its practical applications. By addressing the limitations of traditional approaches and demonstrating the potential of optimal transport theory, this research establishes a foundation for advancing peptide modelling and related fields.

Chapter 2

Geometry of the Peptide

Following the introduction and motivation chapter, this chapter establishes the elementary framework necessary for understanding peptide geometry by defining key elements such as coordinates, bond lengths, bond angles, and dihedral angles. Grasping the geometry of peptides is essential to understanding their structural and functional roles within biological systems. Peptides, composed of amino acids linked by peptide bonds, form the continuous backbone structures of proteins and exhibit the capacity to adopt diverse three-dimensional (3D) conformations.

The internal geometry of peptides—encompassing bond lengths, bond angles, and dihedral angles—plays a pivotal role in determining their structural arrangement and flexibility. These geometric properties collectively influence how peptides fold, interact with other molecules, and perform critical biological functions such as binding, catalysis, and signal transduction. This chapter delves into the structural features that define peptide conformation, highlighting how their internal degrees of freedom shape both their 3D configuration and their functional behavior in biological systems.

Each amino acid in a peptide consists of a central alpha carbon (C_α)(See Fig.aa) that serves as an essential pivot point within the structure. The C_α is bonded to an amino group (N) on one side and a carbonyl group (C) on the other, forming a repetitive backbone sequence that supports the peptide chain. The 3D conformation of a peptide is determined by its unique sequence of amino acids and can be affected by external factors such as pH, temperature, and interactions with surrounding molecules. To fully characterize the 3D structure of a peptide, it is crucial to determine the Cartesian coordinates of each atom,

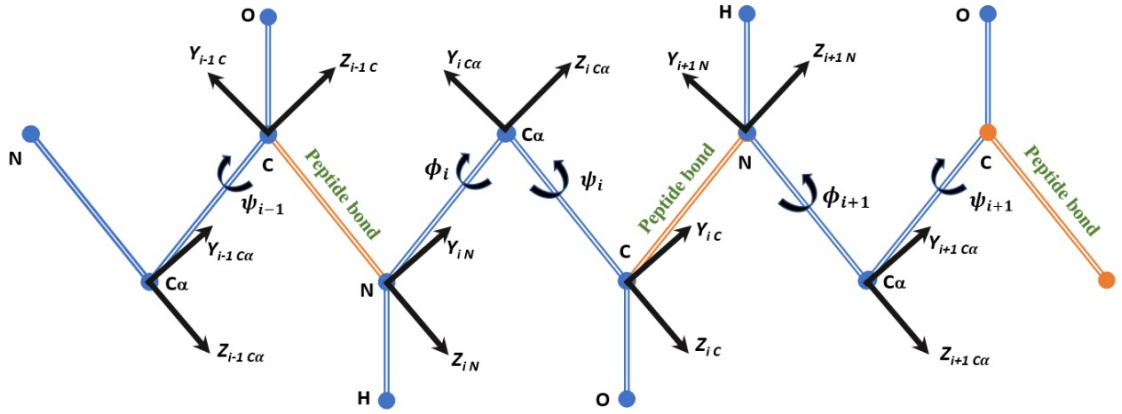


Figure 2.1: Peptide backbone is shown with local frames attached to every backbone atom. Dihedral rotation is restricted about peptide bond

which specify their exact positions in space. These coordinates are derived from the internal degrees of freedom, which dictate how atoms are spatially arranged and influence the overall conformation of the peptide. Understanding these internal coordinates is key to comprehending the structural diversity and functional potential of peptides in biological systems.

To fully characterize the 3D structure of a peptide, it is essential to determine the Cartesian coordinates of every atom in the molecule. These coordinates specify the precise location of each atom in space and can be derived from the internal degrees of freedom of the peptide. The internal degrees of freedom define how atoms within the peptide are arranged relative to one another and are vital for understanding its conformational behavior.

The internal degrees of freedom of a peptide can be classified into three main categories:

- **Bond lengths:** The first category consists of bond lengths, which represent the fixed distances between pairs of covalently bonded atoms. Common bond lengths in peptides include those between the $N-C_\alpha$, $C_\alpha-C$, and $C-O$ atoms. These lengths are generally stable and do not vary significantly under typical biological conditions. The constancy of bond lengths contributes to the rigidity of the peptide's structural framework.

- **Bond angles:** The second category encompasses bond angles, defined as the angles formed between two adjacent bonds that share a common atom. An example of a

bond angle in a peptide is the angle formed by the $N - C_\alpha$ and $C_\alpha - C$ bonds. Although bond angles are relatively consistent, slight variations can occur due to the local environment and interactions within the peptide or with other molecules.

- **Dihedral angles (or torsional angles):** The third and most significant category involves dihedral angles, which describe the rotation around bonds between atoms. A dihedral angle is defined by four atoms, with two consecutive bonds forming the intersection of two planes, and the angle between these planes measured as the torsional angle. In peptides, the most important dihedral angles are the ϕ (phi) and ψ (psi) angles. The ϕ angle represents the rotation around the $N - C_\alpha$ bond, while the ψ angle corresponds to the rotation around the $C_\alpha - C$ bond. Collectively known as the backbone Ramachandran angles, these dihedral angles are critical for defining the conformational flexibility of peptides and proteins.

Each of these internal degrees of freedom contributes to the overall structure of the peptide. While bond lengths and bond angles establish the peptide's foundational framework, they exhibit minimal variation and are often kept constant for simplicity in structural modeling. On the other hand, dihedral angles provide significant rotational freedom, allowing the peptide chain to adopt a wide array of conformations. The values of ϕ and ψ angles are particularly influential, governing the folding pattern of the peptide backbone and determining the formation of secondary structural elements such as alpha helices and beta sheets.

The variability of dihedral angles contributes to the peptide's conformational diversity, enabling it to assume different shapes and structural motifs. This flexibility is essential for the biological functions of peptides and proteins, as it impacts their ability to interact with other molecules, bind to receptors, and participate in complex biochemical processes. The specific conformational preferences of peptides are often visualized in a Ramachandran plot, which maps the permissible and restricted regions of ϕ and ψ angles based on steric hindrance and energetic constraints.

Understanding and defining the internal coordinates of peptides is fundamental for accurately modeling their 3D structures and predicting their functions in biological systems. By combining computational tools and experimental data, researchers can generate de-

tailed models of peptide conformations, contributing to advances in drug design, structural biology, and molecular engineering.

2.1 Computation of the Cost Function for Peptide Conformations

In the analysis of peptide conformations, developing an optimal transport framework requires a well-defined cost function and appropriate input marginal distributions. The cost function, which quantifies the interaction energy K, is a crucial element as it encapsulates the physical interactions between atoms within the peptide. This analysis considers both van der Waals and electrostatic interactions to model the forces that stabilize or destabilize various peptide conformations. The interaction energy function K depends on the pairwise distances between atoms, which can be described in terms of internal coordinates, including the Ramachandran angles (ϕ, ψ).

The interaction energy function is mathematically expressed as:

$$K(\{(\phi_p, \psi_p)\}) = \sum_{i<j} \frac{\epsilon_{ij}}{\lambda} \left[\left(\frac{r_{0_{ij}}}{|\vec{R}_i - \vec{R}_j|} \right)^6 - 1 \right]^2 + \sum_{i<j} \frac{q_i q_j}{D|\vec{R}_i - \vec{R}_j|}. \tag{2.1}$$

Explanation of the Components:

The components of the energy function are as follows: - $|\vec{R}_i - \vec{R}_j|$ represents the distance between atoms i and j, which can be derived from the internal coordinates, particularly the Ramachandran angles (ϕ, ψ). - The first term in the equation accounts for van der Waals interactions. These interactions are described using the parameters ϵ_{ij} (the depth of the potential well) and $r_{0_{ij}}$ (the distance at which the potential reaches its minimum), with λ acting as a scaling factor to adjust the overall contribution of the interaction. - The second term represents electrostatic interactions, where q_i and q_j denote the charges on atoms i and j, respectively, and D is the dielectric constant that accounts for the medium's effect on the interaction.

Deriving Cartesian Coordinates from Internal Coordinates:

To compute the interaction energy K, it is necessary to convert the peptide's internal coordinates, such as bond lengths, bond angles, and dihedral angles (ϕ, ψ), into Cartesian

coordinates for each atom. This step ensures that the precise pairwise distances $|\vec{R}_i - \vec{R}_j|$ can be evaluated. The bond lengths and bond angles are generally assumed to be standard values, while the dihedral angles provide flexibility in the 3D arrangement of the peptide backbone.

Importance of CHARMM Force Field Parameters:

The parameters used in calculating the interaction energies in this thesis are derived from the CHARMM (Chemistry at HARvard Macromolecular Mechanics) force field. [22] CHARMM is a well-established framework known for its accuracy and reliability in modeling the structural and dynamic properties of biological macromolecules. We chose CHARMM because of the fact that it captures dihedral conformations better specifically for flexible peptides with residues like Gly. The force field provides empirical data for potential energy functions, which are essential for simulating realistic peptide behavior.

In summary, defining a cost function that accurately captures the physical interactions within a peptide is essential for modeling its conformational space. The combination of van der Waals and electrostatic terms ensures that both short-range and long-range interactions are considered, providing a comprehensive representation of the energy landscape. The derived Cartesian coordinates from the peptide's internal geometry allow for precise calculations of these interactions, supporting detailed conformational analysis and contributing to a deeper understanding of peptide structure and function.

2.2 Transformation from Internal Coordinates to Cartesian Coordinates

Converting internal coordinates to Cartesian coordinates is a fundamental step in understanding and modeling the three-dimensional structure of peptides. This thesis employs the Denavit-Hartenberg (DH) method, a mathematical framework originally developed for robotics, to achieve this conversion. The DH method is an essential concept in forward kinematics, which involves determining the Cartesian coordinates of the end-point, or "end-effector," of a robotic arm based on its joint parameters. In the context of peptide modeling, the analogy extends to calculating the positions of atoms within a peptide chain using bond lengths, bond angles, and dihedral angles as the internal coordinates.

The Denavit-Hartenberg Method in Peptide Geometry

The DH method provides a systematic approach to relate internal coordinates to a global Cartesian coordinate system[91]. For peptides, each atom along the backbone can be considered analogous to a joint or link in a robotic arm. The internal coordinates include:

- *Bond lengths*: These define the distance between adjacent atoms.
- *Bond angles*: These describe the angles formed by three connected atoms.
- *Dihedral angles*: These specify the rotational angle between two intersecting planes formed by four consecutive atoms.

By applying the DH method, a sequence of transformations is created to map each local coordinate frame, attached to specific atoms, to a global coordinate frame. This recursive process ensures that all atoms in the peptide are represented in a unified Cartesian space.

Forward Kinematics in Peptide Modeling

Forward kinematics, the process of calculating the position of a system's components based on internal parameters, is particularly useful in modeling peptide structures. Each atom in the peptide chain serves as a node that can be positioned using transformations involving translations and rotations. The transformation from one local frame to the next relies on the bond length, bond angle, and dihedral angle between them.

- *Translation*: Moves the coordinate frame by the specified bond length between adjacent atoms.
- *Rotation*: Adjusts the coordinate frame based on the bond angle between atoms.
- *Torsional rotation*: Applies the dihedral angle to align the planes formed by consecutive sets of three atoms.

These transformations are applied sequentially from the start of the peptide chain to its end, effectively converting the peptide's internal coordinates into a set of Cartesian coordinates for all atoms.

Application of the DH Method for Peptide Backbones

In the DH framework, each local coordinate frame F_{ij} is defined with its origin at a specific atom A_{ij}. For peptides, the method involves creating a chain of transformations that map the local frame of each atom to the global reference frame. The algorithm calculates these positions step-by-step:

- Starting from a known reference frame for the first atom.
- Applying transformations defined by the internal coordinates (bond lengths, bond angles, and dihedral angles) to compute the positions of subsequent atoms.

Significance of Forward Kinematics in Structural Analysis

The forward kinematics approach facilitated by the DH method is critical for accurately modeling the 3D structure of peptides. By converting internal coordinates to Cartesian coordinates, researchers can visualize and analyze complex peptide conformations. This capability is essential for simulations, energy calculations, and understanding how peptides interact with other molecules. The DH method ensures that the geometric constraints of the peptide backbone are maintained, allowing for realistic modeling of its structure and flexibility.

In summary, the transformation from internal coordinates to Cartesian coordinates using the Denavit-Hartenberg method provides a reliable way to represent peptide structures in a 3D space. This process underpins many computational studies and contributes to advancements in peptide-based research, including structural biology and drug design.

Algorithm Based on the DH Method

Applying the DH method to peptides starts with the backbone, composed of repeating units of $-N - C_\alpha - C-$. Once the coordinates for these backbone atoms are generated, the method can be extended to include any additional atoms attached to the backbone. Each atom A_{ij} (where i represents the residue number and j specifies the atom within the residue, such as N, C_α, or C) is associated with a local coordinate frame. The

goal is to express the coordinates of A_{ij+1} in the frame attached to A_{ij} using a series of transformations.

The DH method involves translations and rotations defined by internal coordinates such as bond lengths d_{ij}, bond angles θ_{ij}, and dihedral angles ξ_{ij}. Each local frame F_{ij} has its origin at A_{ij}. The Z axis is aligned along the bond d_{ij-1}, denoted by the unit vector $\hat{z}_{ij}$.(see Fig.2.1) The X axis is defined as $\hat{x}_{ij} = \hat{z}_{ij} \times \hat{z}_{ij+1}$, while the Y axis is perpendicular to both, defined by $\hat{y}_{ij} = \hat{z}_{ij} \times \hat{x}_{ij}$.

The transformation matrix $R_{ij}(\theta_{ij}, \xi_{ij}, d_{ij})$ that converts the coordinates of a point P from frame F_{ij+1} to frame F_{ij} is:

$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = R_{ij}(\theta_{ij}, \xi_{ij}, d_{ij}) \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2.2}$$

This transformation includes:

- **Translation** along the bond length d_{ij}:

$$T(d_{ij}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d_{ij} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.3}$$

- **Rotation** to align the X axes by the dihedral angle ξ_{ij}:

$$R_z(\xi_{ij}) = \begin{pmatrix} \cos(\xi_{ij}) & \sin(\xi_{ij}) & 0 & 0 \\ -\sin(\xi_{ij}) & \cos(\xi_{ij}) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.4}$$

- **Rotation** to align the Z axes by the bond angle θ_{ij}:

$$R_x(\theta_{ij}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_{ij}) & \sin(\theta_{ij}) & 0 \\ 0 & -\sin(\theta_{ij}) & \cos(\theta_{ij}) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.5}$$

The combined transformation matrix is:

$$R_{ij} = R_x(\theta_{ij}) \cdot R_z(\xi_{ij}) \cdot T(d_{ij}) \tag{2.6}$$

By recursively applying this matrix from the initial global frame F_{11} (attached to the starting atom) to the final frame, the Cartesian coordinates of any atom in the peptide can be obtained:

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = R_{11} \cdot R_{12} \cdot R_{13} \cdot ... \cdot R_{n3} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.7}$$

Extending to Other Atoms Attached to the Backbone

Although the DH method is effective for backbone atoms, peptides often contain additional atoms connected to the backbone, such as side-chain carbons and hydrogens. For example, the C_α in alanine is bonded to a $C_\beta H_3$ group, whereas glycine has only a hydrogen atom. To include these atoms, their local Cartesian coordinates are first determined within the frame of the backbone atom they are attached to. The same transformation sequence can then be applied to project these local coordinates into the global frame.

For instance, if the oxygen atom O bonded to carbon C has local coordinates $(x'_{i3O}, y'_{i3O}, z'_{i3O})$, its global Cartesian coordinates can be found by:

$$\begin{pmatrix} x_{i3O} \\ y_{i3O} \\ z_{i3O} \\ 1 \end{pmatrix} = R_{11} \cdot R_{12} \cdot R_{13} \cdot ... \cdot R_{n3} \begin{pmatrix} x'_{i3O} \\ y'_{i3O} \\ z'_{i3O} \\ 1 \end{pmatrix} \tag{2.8}$$

To optimize computing time, the planarity of the peptide bond can be utilized by fixing the global frame at an atom located in the center of the peptide chain. The peptide plane at the midpoint of the chain is aligned to lie on the plane of the paper. Cartesian coordinates for the atoms on either side of this central plane can then be calculated separately relative to the global frame.

Using this method, precise Cartesian coordinates for all atoms in the peptide, including those connected to the backbone, can be determined efficiently. This systematic approach provides a robust foundation for calculating energy functions and analyzing peptide conformations, which are essential for understanding the structure and function of peptides in biological systems.

Chapter 3

Method: Optimal Transport

3.1 Introduction to Optimal Transport Theory

In this chapter, we explore the theory of optimal transport, which serves as a central tool in this thesis. The origins of optimal transport theory trace back to 1781 [92], when Gaspard Monge first posed a fundamental question regarding the minimization of transport costs. Monge's problem was simple in concept yet profound in its implications: how can one move a certain quantity of material, such as soil, from one location to another while minimizing the total transportation cost? To illustrate this, imagine a construction worker tasked with building a sandcastle at a designated location using sand from a separate pile. Each grain of sand must be transported from the pile to the construction site, incurring a cost proportional to the distance it travels. The worker's objective is to minimize the overall cost of moving the sand while ensuring that all the grains are utilized in forming the sandcastle. This situation exemplifies an *optimal transport problem*, where one distribution (the sand pile) is transformed into another (the sandcastle). Optimal transport problems can also be interpreted as resource allocation problems, where the goal is to allocate resources efficiently from one configuration to another.

Monge's Original Formulation:

Monge's problem, from a mathematical standpoint, involves finding a way to transform one distribution into another in the most cost-effective manner. This task involves defining a transport plan, or map, that minimizes the total transportation cost. Despite its intuitive formulation, Monge's problem presented significant mathematical challenges due

to its non-linear constraints. These constraints made finding an optimal transport map highly complex, rendering the problem difficult to solve both analytically and numerically. The inherent non-linearity of the optimization problem posed formidable obstacles, as non-linear optimization problems are notoriously challenging and require sophisticated methods for resolution.

The formulation laid by Monge laid the foundation for what would evolve into a rich and multidisciplinary area of study, influencing fields such as geometry, probability, and analysis. Despite the foundational nature of Monge's work, its mathematical complexity spurred the development of more tractable solutions.

The Kantorovich Approach and Linear Programming

The challenges associated with Monge's original problem led to significant advancements, particularly through the work of Leonid Kantorovich [93] in the 20th century. Kantorovich introduced a method that relaxed the non-linear constraints present in Monge's problem, effectively transforming it into a linear programming problem. By allowing for a broader set of transport plans, where mass could be split between destinations, Kantorovich's formulation provided a more flexible and solvable approach. This relaxation made it possible to represent the transport problem in terms of linear constraints, which are far easier to handle using existing numerical techniques.

In essence, Kantorovich reformulated Monge's problem by introducing a "relaxed" transport plan that could be expressed as a matrix, where entries represented the fraction of material moved from one point to another. This approach not only simplified the original problem but also allowed for the development of robust algorithms to find numerical solutions. The Kantorovich formulation thus marked the transformation of the optimal transport problem from a mathematically rigid challenge into a practical optimization task within the framework of linear programming.

Broader Implications and Applications

The development of optimal transport theory, from Monge's original conception to Kantorovich's practical reformulation, has had far-reaching implications. This field has since expanded into diverse branches of mathematics and has found applications [94–102] in geometry, probability theory, fluid dynamics, and economics. The concept of comparing and

transforming distributions is essential in many modern computational problems, including machine learning, computer vision, and resource management.

The Kantorovich approach provided the mathematical community with a powerful toolset for analyzing and solving problems involving the efficient allocation of resources and transformations between distributions. Its linear nature has made it a cornerstone for various algorithms and numerical methods, allowing for practical applications that range from logistics to data science.

Outline of the Chapter

This chapter begins by exploring Monge's original formulation of the optimal transport problem, providing a conceptual understanding of its foundational principles and challenges. Following this, we delve into the Kantorovich method, detailing how his linear programming approach addressed the difficulties of Monge's problem and paved the way for modern advancements in optimal transport theory.

3.2 Monge's Optimal Transport Problem

To understand Monge's optimal transport problem, we can start with a simple transport scenario before generalizing it to a more formal mathematical framework. Imagine there are n production sites located at coordinates $(x_1, x_2, \dots, x_n)$, each producing a uniform amount of product. Similarly, there are n consumption sites at coordinates $(y_1, y_2, \dots, y_n)$, each requiring an equal share of the product. The task is to transport the product from the production sites to the consumption sites in a way that minimizes the total cost. For simplicity, we assume that the total amount of production matches the total demand at the consumption sites.

Mathematically, this can be expressed as:

$$\begin{aligned} \mu(x_i) &= \nu(y_j) \quad \forall i, j, \\ \sum_{i=1}^{n} \mu(x_i) &= \sum_{j=1}^{n} \nu(y_j) = 1, \end{aligned} \tag{3.1}$$

where $\mu(x_i)$ and $\nu(y_j)$ represent the measures of production and consumption at each site, respectively.

If the cost of transporting a unit of product from production site x_i to consumption site y_j is denoted by $c(x_i, y_j)$, then the total cost for transporting the product can be expressed as:

$$C = \sum_{i=1}^{n} \sum_{j=1}^{n} c(x_i, y_j)\, \mu(x_i). \tag{3.2}$$

The essence of the optimal transport problem is to find an optimal map ϕ that pairs production sites $X = \{x_i\}$ with consumption sites $Y = \{y_j\}$ such that the total transportation cost is minimized. This minimum cost, referred to as the optimal transport cost, is represented as W_M and defined by:

$$W_M(\{x_i\}, \{y_j\}) = \inf_{\phi} \sum_{i=1}^{n} c(x_i, \phi(x_i))\, \mu(x_i). \tag{3.3}$$

Formal Formulation of Monge's Problem

To pose Monge's problem more rigorously, consider two non-negative measures μ and ν defined on spaces X and Y, respectively, with the constraint that the total mass of μ matches the total mass of ν:

$$\mu(X) = \nu(Y). \tag{3.4}$$

A cost function $c(x, y)$ is defined on the product space $X \times Y$, representing the cost of transporting a unit of mass from point $x \in X$ to point $y \in Y$. The task is to find a map ϕ that rearranges the measure μ on space X to match the measure ν on space Y. The map ϕ must be one-to-one and ensure that the transported mass aligns perfectly with the consumption distribution.

The total cost associated with a transport plan defined by ϕ is expressed as:

$$W_M[\phi] = \int_X c(x, \phi(x))\, d\mu(x). \tag{3.5}$$

The goal of Monge's optimal transport problem is to find an optimal mapping ϕ^* from the set of all admissible mappings such that the total transport cost W_M is minimized:

$$W_M[\phi^*] = \inf_{\phi} W_M[\phi]. \tag{3.6}$$

Monge's formulation provides an elegant way to conceptualize the optimal transport of resources. However, the problem's non-linear constraints make finding solutions analytically and numerically challenging, leading to the development of alternative methods such

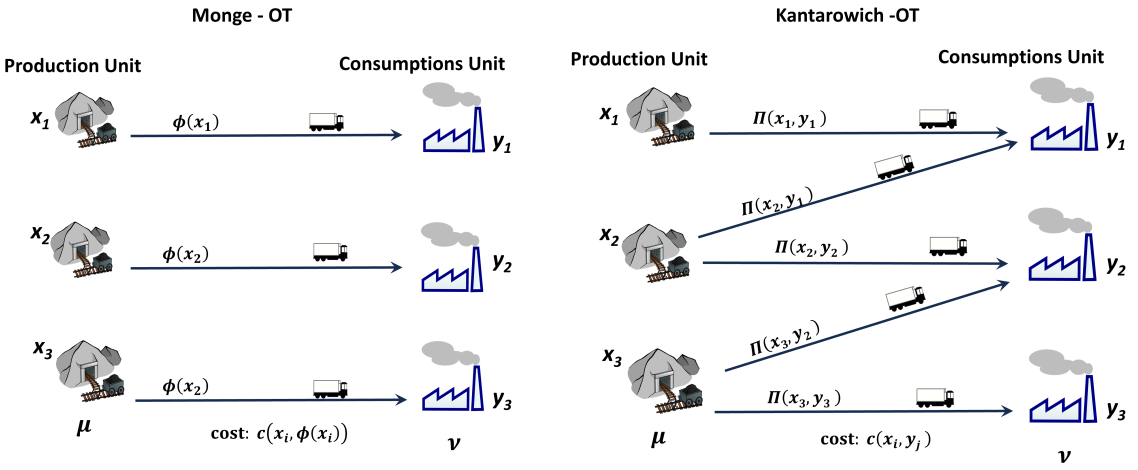


Figure 3.1: The figure illustrates the transport from the production unit to the consumption unit in both Monge Optimal Transport (left) and Kantorovich Optimal Transport (right). In Kantorovich OT (right), a single production unit located at x_2 can supply multiple consumption units at y_1 and y_2. Conversely, in Monge OT (left), each production unit is associated with a single target consumption unit.

as Kantorovich's relaxation of the problem, which will be discussed in the subsequent sections.

3.3 The Kantorovich Approach to Optimal Transport

Building upon the foundational work of Monge, the optimal transport problem presented significant challenges due to its non-linear constraints and the potential non-existence of an optimal transport map ϕ^*. To overcome these difficulties and broaden the scope of the problem, Leonid Kantorovich introduced a reformulation that relaxed the original constraints, transforming Monge's non-linear problem into a linear programming problem. This reformulation made the problem more tractable, ensured the existence of solutions, and enabled more practical numerical handling.

To better understand Kantorovich's approach, consider revisiting the simplified transport problem and modifying its constraints. In Monge's formulation, it was assumed that the production and consumption amounts were uniform, allowing for a one-to-one mapping. However, in real-world scenarios, the production output at different sites may vary, as may the demand at consumption sites (see Fig.3.1) . The only constraint is that the total production should match the total consumption. In this scenario, representing the transport plan as a simple map is no longer feasible. To handle this complexity, Kantorovich

introduced a variable to represent the amount of product transported between any two sites, known as the *coupling of measures*, denoted by $\Pi(x_i, y_j)$.

Defining the Coupling Matrix and its Properties

Consider a situation with m production sites and n consumption sites, where $\mu(x_i)$ indicates the amount of product produced at location x_i and $\nu(y_j)$ represents the demand at location y_j. The coupling $\Pi(x_i, y_j)$ must satisfy the following conditions:

$$\Pi : X \times Y \to \mathbb{R}\geq 0, \tag{3.7}$$

$$\sum_{i=1}^{m} \Pi(x_i, y_j) = \nu(y_j), \quad \forall j, \tag{3.8}$$

$$\sum_{j=1}^{n} \Pi(x_i, y_j) = \mu(x_i), \quad \forall i. \tag{3.9}$$

These constraints ensure that the total product delivered to each consumption site matches its demand and that the total product dispatched from each production site equals its output. Unlike Monge's original problem, which focused on finding a direct transport map, Kantorovich's approach allows for fractional allocation between sites, represented by the coupling $\Pi(x_i, y_j)$.

Calculating the Optimal Transport Cost

If the cost of transporting a unit of product from x_i to y_j is represented by $K(x_i, y_j)$, then the total transportation cost is expressed as:

$$W_K[\Pi] = \inf_{\Pi} \sum_{i=1}^{m} \sum_{j=1}^{n} K(x_i, y_j)\, \Pi(x_i, y_j), \tag{3.10}$$

subject to the constraints in Equations (3.8) and (3.9). The cost function K is assumed to be a positive, convex function. This reformulation turns the problem into a linear programming problem, which contrasts with the non-linear nature of Monge's original formulation.

Matrix Representation of the Transport Problem

To represent the problem in matrix form, we define Π, K, μ, and ν as column vectors:

$$\Pi = \begin{bmatrix} \Pi(x_1, y_1) \\ \Pi(x_2, y_2) \\ \vdots \\ \Pi(x_m, y_n) \end{bmatrix}, \tag{3.11}$$

$$K = \begin{bmatrix} K(x_1, y_1) \\ K(x_2, y_2) \\ \vdots \\ K(x_m, y_n) \end{bmatrix}, \tag{3.12}$$

$$\mu = \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_m) \end{bmatrix}, \tag{3.13}$$

$$\nu = \begin{bmatrix} \nu(y_1) \\ \nu(y_2) \\ \vdots \\ \nu(y_n) \end{bmatrix}. \tag{3.14}$$

In this thesis, both μ and ν are treated as probability distributions, meaning they sum to 1:

$$\sum_{i=1}^{m} \mu(x_i) = \sum_{j=1}^{n} \nu(y_j) = 1. \tag{3.15}$$

Advantages of Kantorovich's Formulation

Kantorovich's approach offers several significant advantages over Monge's original formulation:

- *Existence of Solutions*: The linear nature of the problem ensures that a solution exists, even when a direct transport map does not.
- *Numerical Solvability*: The problem can be effectively solved using linear programming algorithms, making it feasible for computational applications.
- *Flexibility in Transport Plans*: The coupling matrix Π allows for partial allocation of resources, providing a more realistic representation of practical transport scenarios.

The reformulation by Kantorovich not only simplified the mathematical complexity of the problem but also paved the way for applications across various domains, including economics, data science, and optimization. The properties and extensions of this method have made it a powerful tool for analyzing and solving optimal transport problems. The following sections will explore the mathematical details and applications of Kantorovich's approach in greater depth.

3.4 Dual Problem in Optimal Transport

In optimal transport (OT) theory, every Kantorovich formulation has a corresponding dual problem [103] that provides an alternative perspective for addressing the original optimization. While the primal problem focuses on minimizing transportation costs, the dual formulation converts it into a maximization problem under constraints defined by the cost function $K(x_i, y_j)$. This dual approach involves maximizing a function $J[\varphi, \psi]$ that depends on potential functions $\varphi(x_i)$ and $\psi(y_j)$, which are defined over the production and consumption sites x_i and y_j, respectively.

Mathematically, the dual formulation of the Kantorovich problem can be represented as:

$$J[\varphi, \psi] = \sup_{\varphi, \psi} \left\{ \sum_{i=1}^{m} \varphi(x_i)\, \mu(x_i) + \sum_{j=1}^{n} \psi(y_j)\, \nu(y_j) \right\}, \tag{3.16}$$

subject to the constraint:

$$\varphi(x_i) + \psi(y_j) \leq K(x_i, y_j), \tag{3.17}$$

which must hold for all i and j—that is, for each potential transport path between x_i and y_j.

Interpretation and Example of the Dual Problem

To gain an intuitive understanding of the dual problem, consider an illustrative example inspired by Cédric Villani's explanation. Suppose we need to transport goods from m production sites $\{x_i\}$ to n consumption sites $\{y_j\}$. In this context, $\varphi(x_i)$ can be interpreted as the loading cost at each production site x_i, while $\psi(y_j)$ represents the unloading cost at each consumption site y_j. The transportation cost per unit from x_i to y_j is given by $K(x_i, y_j)$.

In the dual objective function $J[\varphi, \psi]$ expressed in Equation (3.16), the terms $\varphi(x_i)\mu(x_i)$ and $\psi(y_j)\nu(y_j)$ represent the total loading and unloading profits, weighted by the amounts $\mu(x_i)$ and $\nu(y_j)$. The condition $\varphi(x_i)+\psi(y_j) \leq K(x_i, y_j)$ ensures that the combined loading and unloading charges do not exceed the actual transportation cost between any given pair (x_i, y_j). This constraint maintains the feasibility of the transport plan.

The Duality Theorem in Optimal Transport

One of the core results in optimal transport theory is the duality theorem, which states that the maximum value of the dual problem $J[\varphi, \psi]$ is equal to the minimum value of the primal Kantorovich transport cost $W_K[\Pi]$. This relationship can be formally expressed as:

$$J[\varphi, \psi] = W_K[\Pi]. \tag{3.18}$$

This equivalence implies that if an optimal solution exists for the primal OT problem, there is also an optimal solution for the dual problem, and their optimal values coincide. This principle of duality is fundamental not only for theoretical reasons but also for practical applications, as it provides an alternative approach to finding solutions to complex transport problems.

Significance of the Dual Formulation

The dual problem in optimal transport is more than just a theoretical construct; it offers several practical advantages:

- *Computational Efficiency*: In many cases, solving the dual problem can be simpler and more computationally efficient than solving the primal problem, especially when dealing with large-scale data.
- *Economic Interpretation*: The dual variables φ and ψ can be interpreted as potential functions representing the marginal value of transporting goods from production to consumption sites. This interpretation can provide insights into optimal pricing and resource allocation.
- *Robustness and Flexibility*: The dual formulation allows for the incorporation of additional constraints and regularization terms, making it adaptable for various applications in economics, logistics, and data science.

The dual approach to optimal transport complements the primal Kantorovich formulation by offering an alternative viewpoint that facilitates both the theoretical understanding and practical resolution of transport problems. This duality forms a cornerstone of modern optimal transport theory and continues to be a powerful tool in mathematical analysis and applied optimization. The following sections will explore further implications and applications of the dual problem in various contexts.

3.5 Linear Programming in Optimal Transport

With the reformulation of the discrete optimal transport (OT) problem by Kantorovich, the problem can be expressed as a linear programming (LP) problem, allowing it to be solved using established linear programming techniques. Linear programming is an optimization method used to maximize or minimize an objective function subject to linear constraints. The origins of linear programming date back to military logistics during World War II, with significant advancements made by George Dantzig in 1947 through the development of the simplex algorithm.

The primary objective of linear programming is to optimize (maximize or minimize) a linear objective function. The general form of such an objective function is:

$$\min \quad Z[t] = c_1 t_1 + c_2 t_2 + \cdots + c_n t_n, \tag{3.19}$$

where $\{t_i\}$ are known as decision variables and are represented as a vector $t = (t_1, t_2, \dots, t_n)$. The coefficients $\{c_i\}$ are similarly represented as a vector $c = (c_1, c_2, \dots, c_n)$. The objective function Z is linear in t_i, and the goal is to find a vector $t = t^*$ that minimizes (or maximizes) Z while satisfying certain linear constraints.

Linear Constraints in Linear Programming

The decision variables t_i must satisfy the following linear constraints:

$$a_{11} t_1 + a_{12} t_2 + \cdots + a_{1n} t_n = b_1, \tag{3.20}$$

$$a_{21} t_1 + a_{22} t_2 + \cdots + a_{2n} t_n = b_2, \tag{3.21}$$

$$\vdots \tag{3.22}$$

$$a_{m1} t_1 + a_{m2} t_2 + \cdots + a_{mn} t_n = b_m, \tag{3.23}$$

with the additional non-negativity condition:

$$t_1, t_2, \dots, t_n \geq 0. \tag{3.24}$$

These constraints define a feasible region, which is the set of all possible solutions t that satisfy the constraints. The optimal solution t^* lies within this feasible region.

Matrix Formulation of Linear Programming

The system of constraints and the objective function can be expressed in matrix form:

$$\text{minimize} \quad Z[t] = c^T \cdot t, \tag{3.25}$$

$$\text{subject to} \quad A \cdot t = b, \tag{3.26}$$

where:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix},$$

and c and t are column vectors.

Application to the Optimal Transport Problem

For the OT problem in the Kantorovich formulation, the cost matrix K and the coupling matrix Π are key to setting up the linear program. The objective function to be minimized is:

$$W_K[\Pi] = \min_{\Pi} [K^T \Pi], \tag{3.27}$$

subject to the constraints:

$$\tilde{A} \cdot \Pi = \eta, \tag{3.28}$$

where $\tilde{A}$ is the constraint matrix of size $(m+n) \times (mn)$. Each row in $\tilde{A}$ corresponds to a constraint ensuring that the transported mass matches the production and consumption requirements:

$$\sum_{i=1}^{m} \Pi(x_i, y_j) = \nu(y_j), \quad \forall j, \tag{3.29}$$

$$\sum_{j=1}^{n} \Pi(x_i, y_j) = \mu(x_i), \quad \forall i. \tag{3.30}$$

Representation of the Constraint Vector η

The vector η is a concatenation of the production and consumption distributions μ and ν:

$$\eta = \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_m) \\ \nu(y_1) \\ \nu(y_2) \\ \vdots \\ \nu(y_n) \end{bmatrix}. \tag{3.31}$$

Constructing the Linear Programming Matrix $\tilde{A}$

The matrix $\tilde{A}$ is constructed so that each entry indicates whether a particular decision variable $\Pi(x_i, y_j)$ contributes to the production or consumption constraints. The entries are either 0 or 1, ensuring that the constraints in Equations (3.8) and (3.9) are satisfied.

Solving the Linear Program

In the linear programming framework, each entry in Π is treated as a decision variable to be optimized. The objective is to minimize the product $K^T\Pi$ subject to the constraints defined by $\tilde{A}$ and η. The solution provides the optimal transport plan Π^*, minimizing the total cost while satisfying all production and consumption constraints.

Linear programming thus transforms the OT problem into a structured optimization task, enabling efficient computational solutions and facilitating deeper analysis of the problem's characteristics.

Chapter 4

Optimal Transport Technique to Understand tetrapeptide Conformations

4.1 Introduction

In this chapter, we present a method for determining multi-point probability distributions (MPD) of tetrapeptides, designed to capture the multi-point correlations among the backbone torsional angles of specific tetrapeptide sequences. Tetrapeptides with a given sequence can adopt numerous conformations due to the variability in backbone torsional angles at each residue within an allowable range, as observed in data from the Protein Data Bank (PDB). While distributions of individual residue angles provide useful information, they are not sufficient to fully define the conformational space of tetrapeptides. This is because they lack details on the multi-point correlation functions among the backbone torsional angles, which are critical for accurately characterizing the range of feasible conformations.

To address this gap, our proposed method computes MPD by utilizing a database of high-quality triplet distributions that encode all short-range correlations. The process involves minimizing the expectation value of the effective potential energy function for all atoms within the peptide, starting from an input tripeptide. To demonstrate the method's

practicality and effectiveness, we apply it to tetrapeptides composed of alanine (Ala) and glycine (Gly), selected for their structural simplicity and minimal side chains.

The technique used to compute MPD from a given input set of distributions is the multi-marginal optimal transport (MOT) method, which builds upon the optimal transport theory discussed in the previous chapter. This approach seeks an MPD that minimizes the expected value of the potential energy function (cost function), while ensuring that the obtained MPD has the given set of input distributions as its marginals.

4.2 Method

For the sake of completeness, we provide a brief summary of Optimal Transport Theory (OT), with more detailed discussions available in the previous chapter. The fundamental goal of OT is to find a positive matrix Π that represents the correspondence between points in the set $\Phi = \{\phi\}$ and points in the set $\Psi = \{\psi\}$. This matrix should minimize the expected value of a given cost function $K(\phi, \psi)$, defined as:

$$E[\Pi] = \min_{\{\Pi\}} \sum_{\phi,\psi} K(\phi, \psi)\Pi(\phi, \psi) \tag{4.1}$$

with constraints,

$$\sum_{\psi} \Pi(\phi, \psi) = \rho(\phi) \tag{4.2}$$

$$\sum_{\phi} \Pi(\phi, \psi) = \tilde{\rho}(\psi) \tag{4.3}$$

The summation extends to all ϕ in Φ and ψ in Ψ. The minimum of E is to be found for the values of $\Pi(\psi, \phi)$ that subject to the constraint that it admits $\rho(\phi)$ and $\tilde{\rho}(\psi)$ as marginals. Here $\rho(\phi)$ and $\tilde{\rho}(\psi)$ are two input distributions which are defined on the set Φ and Ψ. The solution to the OT problem provides an optimal plan Π_{opt} and the corresponding minimum expected cost $E_{min} = E[\Pi_{opt}]$. Remarkably, this equation has a unique minimizer despite the infinite number of possible transport plans that are consistent with $\rho(\phi)$ and $\tilde{\rho}(\psi)$ as marginals. While the roots of the minimization problem is traced back to late eighteenth century, in recent times, there is a great interest in exploring the theoretical and computational aspects of this minimization [16].

OT application to peptides needs a set of input distributions, as discussed above. We use the database [20] in this work, which we refer to as a PDB(D). It is a database of the distributions of dihedral angles of dipeptide amino acids collected from PDB. It provides the conditional distribution of dihedral angles (ϕ, ψ) of central amino acid C, when amino acid, $R(L)$, is on the right(left) of C. Fig 4.1 shows the sequence of three amino acids with C, L, and R. We represent this conditional distribution as $\hat{f}(\phi, \psi|C, R)((\hat{f}(\phi, \psi|C, L))$. It also contains neighbor independent probability distributions $\hat{f}(\phi, \psi|C)$ of the central amino acid C. Here, C, R, and L run over all the 20 amino acids and so there are 800 possible distributions with 400 each for $\hat{f}(\phi, \psi|C, L)$ and $\hat{f}(\phi, \psi|C, R)$. The conditional triplet probability distribution $P(\phi, \psi|C, L, R)$ is estimated, using the conditional and neighbor-independent probability distributions.

$$P(\phi, \psi|C, L, R) = \frac{\hat{f}(\phi, \psi|C, L)\hat{f}(\phi, \psi|C, R)}{\mathcal{N}\hat{f}(\phi, \psi|C)} \tag{4.4}$$

where, $\mathcal{N}$ is normalization constant. This prescribed estimate, as provided in [20], is based on the assumption that both residues "L" and "R" are independent of each other. Instead of denoting the peptide sequence as C, L and R, we represent it by $i, i-1$and $i+1$ for the sake of simplicity of notations. A set of Ramachandran angles $\{(\phi_i, \psi_i)\}$, fully describe the conformation of this sequence for a fixed bond lengths and angles. The set (ϕ_i, ψ_i) for a specific amino acid, 'i', takes different values in various native proteins which results in distributions over (ϕ_i, ψ_i). The triplet distribution estimate $P(\phi_i, \psi_i|i, i-1, i+1)$ of amino acid i depends on the type of neighbouring amino acid '$i-1$' and '$i+1$', where $i = 0, 1, 2, ..., N-1$ Our interest is here to find the probability distribution Π of a sequence of 'N' number of amino acids. This Π minimises the expectation value of interaction energy $K(\{(\phi_i, \psi_i)\})$ consisting the van der Waals and electrostatic interactions of all-atom of peptide of 'N' amino acids where, $K(\{(\phi_i, \psi_i)\})$ is

$$K(\{(\phi_i, \psi_i)\}) = \sum_{i<j} \frac{\epsilon_{ij}}{\lambda}\left[\left(\frac{r_{0_{ij}}}{|\vec{R}_i - \vec{R}_j|}\right)^6 - 1\right]^2 + \sum_{i<j} \frac{q_i q_j}{D|\vec{R}_i - \vec{R}_j|}. \tag{4.5}$$

where $\vec{R}_i$ and $\vec{R}_j$ are Cartesian coordinates of the atoms in each amino acid, and the distance between them, $|\vec{R}_i - \vec{R}_j|$, depends on the set of all $\{(\phi_i, \psi_i)\}$ angles. The parameters ϵ_{ij}, $r_{0_{ij}}$, q_i and D are depth of the potential, effective van der Waals radii, partial charge

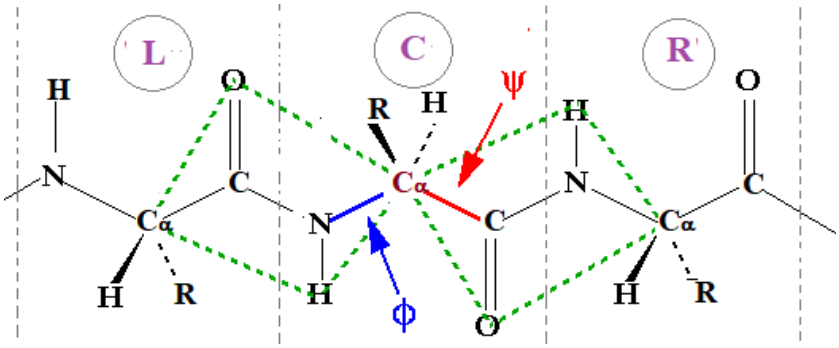


Figure 4.1: Sequence with three amino acids, where C, L and R, represent centre, left, and right amino acids respectively. (ϕ, ψ) angle corresponds to central acid C.

on each atom, and dielectric constant respectively [22]. The parameter λ is introduced to allow variation in the strength of the van der Waals term. For very large λ and the finite value of D, the interaction is purely electrostatic, while for very large D and $\lambda = 1$ the van der Waals interaction dominates. The cost function may be negative due to the electrostatic interactions present in the cost functions. To make it positive, a constant matrix can be added to the negative cost functions of size of K[1]. However, this modification does not affect the Π obtained from MOT

The input conditional triplet distributions $P(\phi_i, \psi_i | i, i-1, i+1)$ is converted into two distributions $\{\rho_i(\phi_i)\}$ and $\{\tilde{\rho}_i(\psi_i)\}$ as,

$$\sum_{\psi_i} P(\phi_i, \psi_i |\ i, i-1, i+1\ ) = \rho_i(\phi_i) \tag{4.6}$$

$$\sum_{\phi_i} P(\phi_i, \psi_i |\ i, i-1, i+1\ ) = \tilde{\rho}_i(\psi_i) \tag{4.7}$$

In the present situation, we have sets of $\{\Phi_i\}$ and $\{\Psi_i\}$, where $i = 0, 1, 2, \cdots, N-1$. Consequently, one needs $2N$ marginals, $\{\rho_i(\phi_i)\}$ and $\{\tilde{\rho}_i(\psi_i)\}$, as defined in Eqn(2). We search for $\Pi(\{\phi_i\}, \{\psi_i\})$ that minimizes the expected cost $E[\Pi]$ by solving a generalized version of OT, known as a multi-marginal transport theory (MOT). MOT has recently attracted more attention due to emerging applications in economics, mathematical finance, condensed matter physics, berry center, and image processing [86, 88, 89].

[1]In optimal transport, the cost function must be the lower semi-continuous function to have a unique value of multi-point distribution; however, K can be positive or negative, restricting its validity to the minimum distance between atoms,K can be bounded from the negative. therefore, we can make K the lower semi-continuous by a constant shift, as we did here

$$E[\Pi(\{\phi_i\},\{\psi_i\})] = \min_{\{\Pi\}} \sum_{\{\Phi_i\},\{\Psi_i\}} K(\{\phi_i\},\{\psi_i\})\Pi(\{\phi_i\},\{\psi_i\}) \tag{4.8}$$

where $\Pi(\{\phi_i\},\{\psi_i\})$ is MPD which satisfies the following constraints

$$\sum_{\{\psi_i,\phi_i\},\phi_i\neq\phi_k} \Pi(\{\phi_i\},\{\psi_i\}) = \rho_k(\phi_k) \tag{4.9}$$

$$\sum_{\{\phi_i,\psi_i\},\psi_i\neq\psi_k} \Pi(\{\phi_i\},\{\psi_i\}) = \tilde{\rho_k}(\psi_k) \tag{4.10}$$

where $k = 0, 1, 2, \cdots, N-1$.

We can cast the equations (8), (9), and (10) as a linear program (LP) in standard form, that is, a linear program with a linear objective (Eqn (8)); equality constraints (Eqns (9) and (10)) defined with a matrix and a constant vector; non-negative constraint on variables (Π). Solving LP, we obtain 'Π.' The most probable conformations of a given sequence of peptides corresponding to the peaks observed in Π. This method generates Π for peptides such as tetrapeptide, hexapeptide, and octapeptide. To demonstrate the effectiveness of this method, we consider only the tetrapeptides composed of Ala and Gly for their simple structure as it lacks side chains. We present the preliminary analysis and compare our results with data supplied by Sharmila A, which we would refer to as PDB(S) [2].

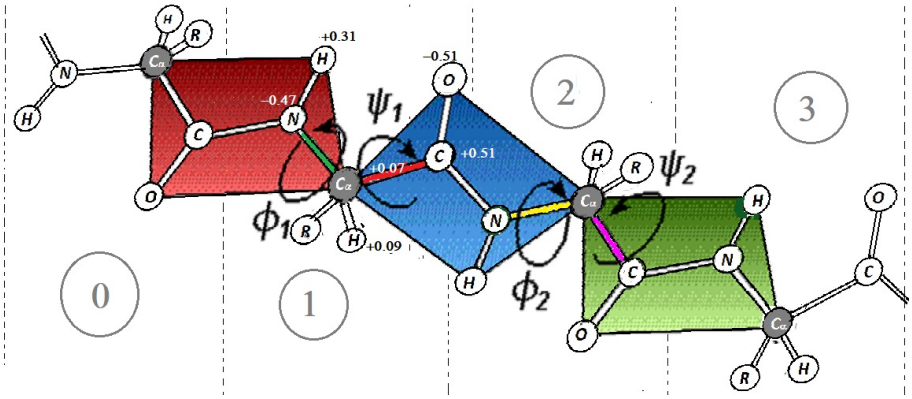
4.3 Results and Discussions

Tetrapeptide

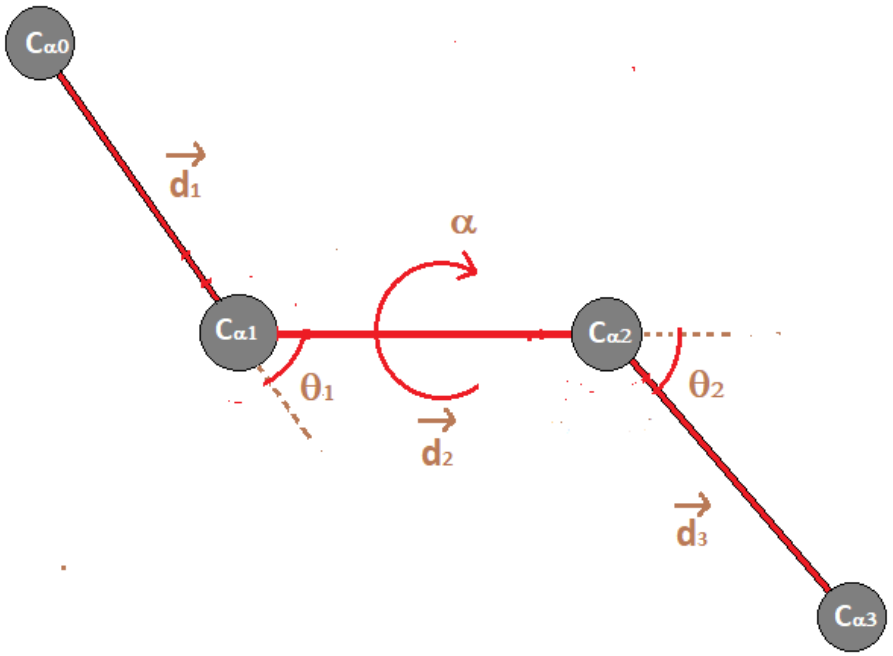
To describe PDB(S), the geometry of the tetrapeptide is shown in Fig 4.2a and 4.2b. In Fig 4.2a we show a sequence '0, 1, 2, 3' of amino acids of tetrapeptide in which bond lengths, bond angles, torsional angles (ϕ_0,ψ_0) and (ϕ_3,ψ_3) are fixed while the other two angles (ϕ_1,ψ_1) and (ϕ_2,ψ_2) vary. We fix the central peptide plane on the plane of the paper utilizing the translational and rotational invariance of the cost functions K. The other two peptide planes could rotate upon variation of (ϕ_1,ψ_1) and (ϕ_2,ψ_2) angles and the cost function $K(\phi_1,\psi_1,\phi_2,\psi_2)$ depends on four variables. The probability distribution $\Pi(\phi_1,\psi_1,\phi_2,\psi_2)$ that minimizes its expectation value has four marginals distribution

[2] Unpublished work on tetrapeptides by Sharmila Anishetty

$\rho_1(\phi_1)$, $\tilde{\rho}_1(\psi_1)$, $\rho_2(\phi_2)$ and, $\tilde{\rho}_2(\psi_2)$ which are determined using equations (5) and (6). We obtain $\Pi(\phi_1, \psi_1, \phi_2, \psi_2)$ by solving LP.



(a)



(b)

Figure 4.2: (a) Backbone sequence 0, 1, 2, and 3 depicts tetrapeptide with all atoms. The numbers -0.47, +0.31, +0.51, -0.51 are the partial charges of backbone atoms N, H, C and O respectively taken from Charmm force fields[22]. (b) Backbone of tetrapeptide with alpha carbon alone. The dihedral angle (α) rotation transforms the coordinates of the atoms to the right of the axis of rotation. The coordinates of the atoms to the left of the axis of rotation remain unchanged.

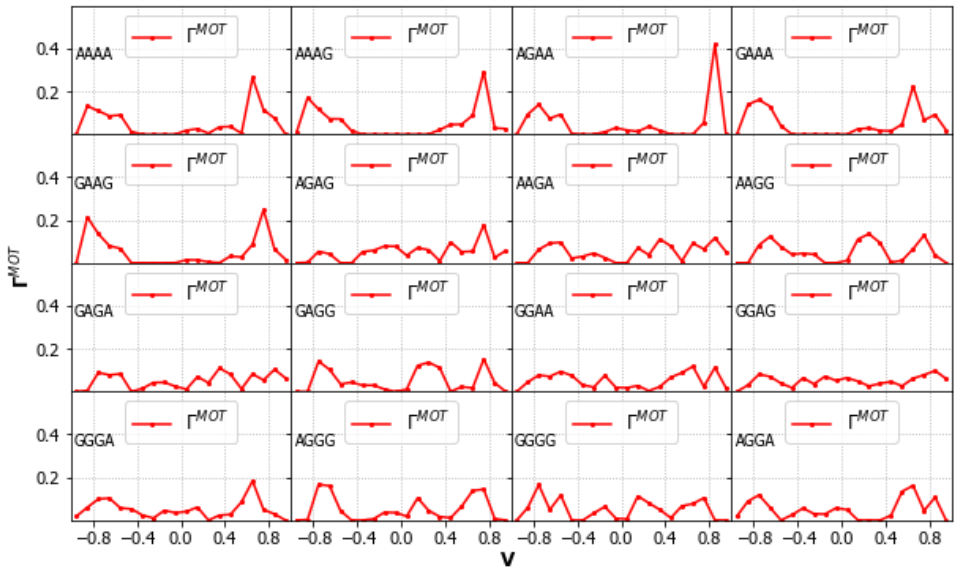
In Fig 4.2b, a backbone of the tetrapeptide with alpha carbons (C_α) is shown and the virtual bonds connecting $C_{\alpha_0} - C_{\alpha_1}$, $C_{\alpha_1} - C_{\alpha_2}$ and $C_{\alpha_2} - C_{\alpha_3}$ are defined by the normalized vectors $\vec{d}_1, \vec{d}_2$ and $\vec{d}_3$. $\theta_1 = \arccos(\vec{d}_1 \cdot \vec{d}_2)$ and $\theta_2 = \arccos(\vec{d}_2 \cdot \vec{d}_3)$ are the virtual bond angles. α is the angle between the planes formed by $C_{\alpha_0} - C_{\alpha_1} - C_{\alpha_2}$ and $C_{\alpha_1} - C_{\alpha_2} - C_{\alpha_3}$

planes. To generate the α rotations, we fix the $C_{\alpha_0} - C_{\alpha_1} - C_{\alpha_2}$ on the plane of the paper and C_{α_3} is allowed to rotate freely about the $C_{\alpha_1} - C_{\alpha_2}$ from 0 to 2π. Hence each α rotation in a given peptides always transforms the coordinates of the atoms of each amino acid on the right without disturbing the coordinates of atoms of amino acids on the left of α axis of rotation.

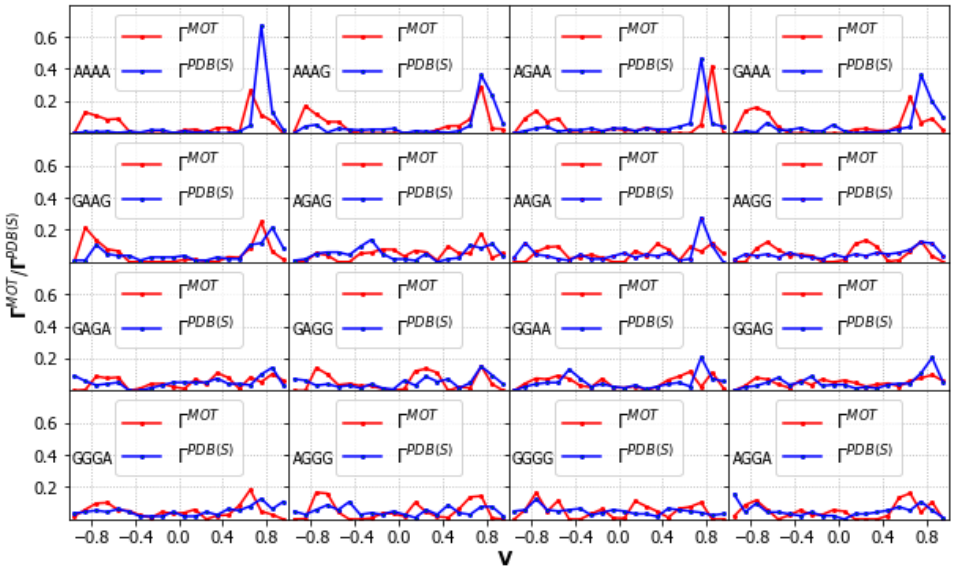
The PDB(S) database, which is used to compare results obtained from MOT, contains data for each tetrapeptide, including $\cos\theta_1, \cos\theta_2$ and V as well as references to the PDB file name and residue number. Here $V = (\vec{d}_1 \times \vec{d}_2) \cdot \vec{d}_3 = \sin\theta_1 \sin\theta_2 \sin\alpha$ is a signed tetrapeptide volume. The quartet tuple of torsional angles $(\phi_1, \psi_1, \phi_2, \psi_2)$ can be mapped to the triplet tuple of $(\theta_1, \theta_2, \alpha)$, thus, we can convert Π distributions to the distributions over (θ_1, θ_2, V). However, plotting three-dimensional distribution is not very practical and therefore we integrate the distribution over the variables θ_1 and θ_2 keeping V fixed and then obtain $\Gamma^{MOT}(V)$. Superscript is used to differentiate distributions obtained using MOT, PDB(S) and PDB(D) and so the distributions over volume for PDB(S) is referred as, $\Gamma^{PDB(S)}(V)$

We compare $\Gamma^{MOT}(V)$ with $\Gamma^{PDB(S)}(V)$ of all tetrapeptides consisting of Ala and Gly of PDB(S), which is shown in Fig 4.3. It can be seen in Fig 4.3a and 4.3b, there are two regions of volume V, where $\Gamma^{MOT}(V)$ is non zero, region "a" from $V = -0.8$ to -0.6 and region "b" from $V = 0.6$ to 0.8. Since $\Gamma^{PDB(S)}(V)$ and $\Gamma^{MOT}(V)$ are one-dimensional plots as a function of V, different $\phi_1, \psi_1, \phi_2, \psi_2$ points contribute to the same volume. To explicitly show the torsional angles which make up the peaks observed in Fig 4.3a and Fig 4.3b we put all the torsional angles correspond to the region "a" and region "b" in the Ramachandran scatter plot (Fig 4.4). The blue (red) dots in Fig 4.4 represent (ϕ_1, ψ_1) $((\phi_2, \psi_2))$ angles and we only show for GGGG and AAAA. The region "a" torsional angles for AAAA fall in the second while the same part for GGGG lies in the first quadrant, overlapping with the fourth quadrant. However, the region "b" torsional angles for GGGG, (ϕ_2, ψ_2) and (ϕ_1, ψ_1) fall in the second and the third quadrants, respectively.

In the positive region "a", we observe a peak with $\Gamma^{MOT}(V)$ greater than 0.2 in tetrapeptides involving three or more alanine, for example, AAAA, AAAG, AGAA, and GAAA and their respective torsional angles fall into the third quadrant of the scatter plot close to the alpha-helix region. But also a few points (ϕ_2, ψ_2) fall in the part of the boundary



(a)



(b)

Figure 4.3: (a) $\Gamma^{MOT}(V)$ obtained from MOT of all the tetrapeptides composed of Ala and Gly. (b) Comparison of $\Gamma^{MOT}(V)$ with that of $\Gamma^{PDB(S)}(V)$

between the second and third quadrants. We also see a peak in the plateau region "b" with $\Gamma^{MOT}(V)$ around 0.2, and its corresponding torsional angles are close to the beta-sheet part in the second quadrant. Hence, the peak at the positive volume observed in alanine dominating tetrapeptides suggests forming right-handed alpha helix conformations. A similar trend is seen in GAAG because central amino acids are alanine.

We see finite probability throughout the volume as the number of glycine increases. It

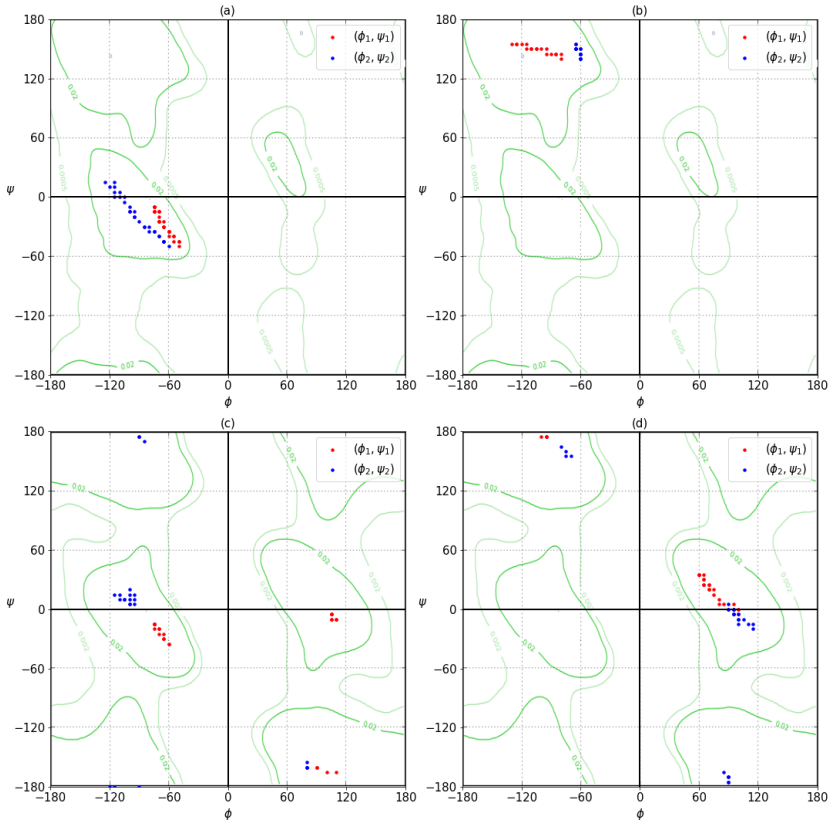


Figure 4.4: Ramachandran plots for the regions "a" and "b" for the volume ranging from 0.6 to 0.8 and -0.8 to -0.6: (a) and (d) for AAAA and (b) and (c) for GGGG.

is because glycine is more flexible than alanine and this behaviour of glycine is due to the lack of side chains, which enables it to have less restriction on conformations. On the contrary, alanine restricts its conformations because of a side-chain C_β atom in it.

The peak values near $V \sim 0.7$ with two glycines at the central positions, for example, GGGG and GGGA correspond to the conformations with (ϕ_1, ψ_1) falling near the boundary of the third quadrant while (ϕ_2, ψ_2) falling near the edge of the second and third quadrant of the scatter plot. (ϕ_1, ψ_1) and (ϕ_2, ψ_2) of the negative volume $(V \sim -0.7)$ conformations are in the first and the fourth quadrant because of the presence of glycine at the central locations which enables it to adopt conformations in these quadrants.

Our results also predict secondary structure elements such as beta turns in GGGG, GAGG, AAGG, and AAGA as can be seen in the Fig 4.5. The appearance of turns in our results is due to the fact that input distributions PDB(D) has been sampled in the loop region which involves turns as well. The observed peaks in GGGG at $V \sim 0.01, -0.68, 0.69$ and its corresponding torsional angles $(\phi_1 = 60, \psi_1 = -125, \phi_2 = -85, \psi_2 = 0\ )$, $(\phi_1 =$

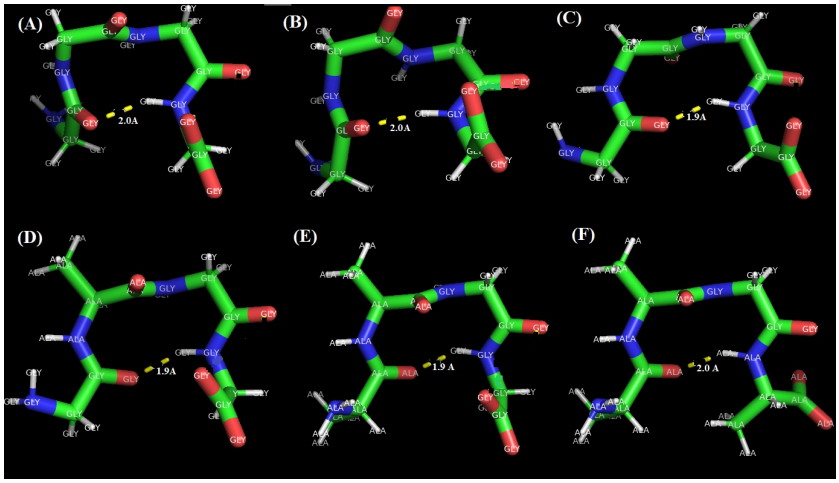


Figure 4.5: Type I turn, Type I' turn, and Type II' turn in GGGG are shown in (A), (B), and (C), respectively. While Type II turn observed in GAGG, AAGG, and AAGA are shown in (D), (E), and (F) respectively.

$60, \psi_1 = 35, \phi_2 = 90, \psi_2 = 0)$, and $(\phi_1 = -60, \psi_1 = -35, \phi_2 = -95, \psi_2 = 15\ )$ belong to primed turn II' and I' and turn I. Turn II is observed in the other three. The probability of type II' turn is ten times higher than I' and turn I in GGGG. The presence of glycine in the second position allows GGGG to form type II' and I' turns as glycine is allowed in the 4th and 1st quadrant of the scatter plot. The presence of AG sequence in the middle of the other three of the tetrapeptides, helps in the formations of turn type II, as A is allowed in the 2nd quadrant and G is populated at the boundary of the first and fourth quadrant of the Ramachandran plot. Though the torsional angles deviate slightly from the ideal angles of turns, yet it forms hydrogen bonds between amino acid 0 and 3 which is shown as a dotted(yellow) line in Fig 4.5.

Tripeptide

We obtain tripeptide distributions, $\tau_1$and τ_2 (Eq(11)), directly from II, since tetrapeptide is a concatenation of two tripeptides, for example, AAGA is a concatenation of AAG and AGA. The conformations of AAG and AGA are described by the torsional angles of the central amino acid. We obtain $\tau_1(\phi_1, \psi_1)$

$((\tau_2(\phi_2, \psi_2))$ by summing $\Pi(\phi_1, \psi_1, \phi_2, \psi_2)$ over $(\phi_2, \psi_2)((\phi_1, \psi_1))$.

$$\sum_{\phi_1, \psi_1} \Pi(\phi_1, \psi_1, \phi_2, \psi_2) = \tau_2(\phi_2, \psi_2)$$
$$\sum_{\phi_2, \psi_2} \Pi(\phi_1, \psi_1, \phi_2, \psi_2) = \tau_1(\phi_1, \psi_1) \tag{4.11}$$

Mapping ordered pairs (ϕ_1, ψ_1) and (ϕ_2, ψ_2) to uniquely θ_1 and θ_2, we can convert τ_1 and τ_2 to $\gamma_1^{MOT}(\cos\theta_1)$ and $\gamma_2^{MOT}(\cos\theta_2)$. $P(\phi_1, \psi_1|1, 0, 2)$ and $P(\phi_2, \psi_2|2, 1, 3)$ can also be converted to distributions $\gamma_1^{PDB(D)}(\cos\theta_1)$ and $\gamma_2^{PDB(D)}(\cos\theta_2)$. Similarly, we refer the distributions over $\cos\theta_1$ and $\cos\theta_2$ from PDB(S) as $\gamma_1^{PDB(S)}(\cos\theta_1)$ and $\gamma_2^{PDB(S)}(\cos\theta_2)$

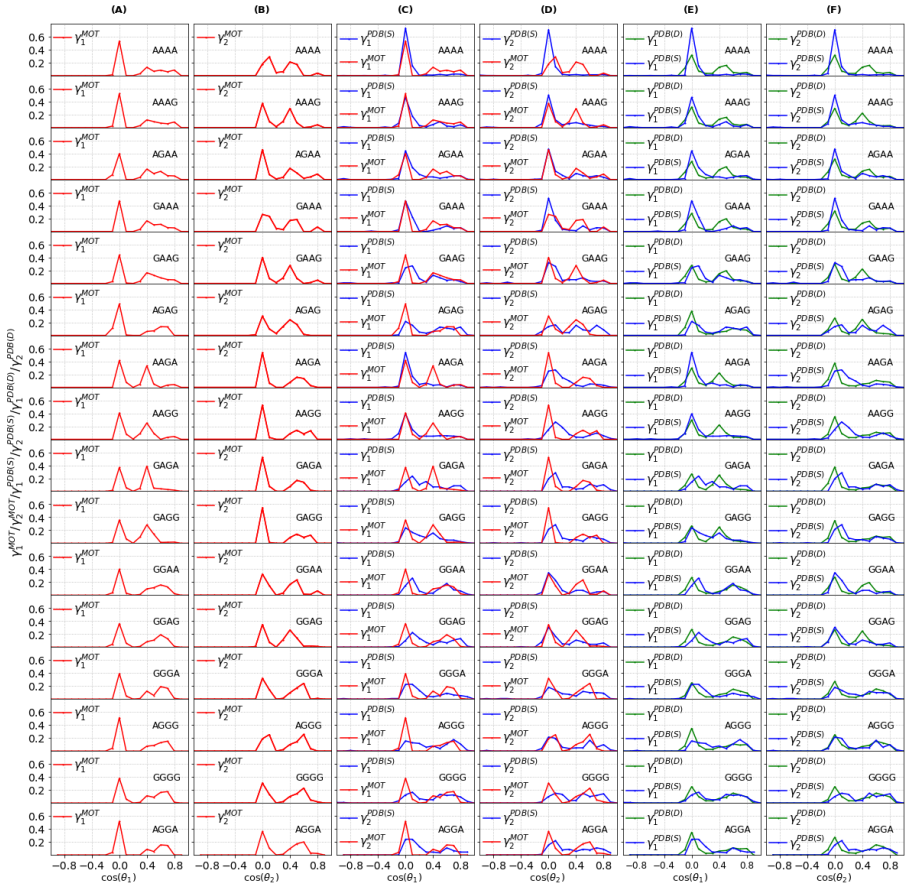


Figure 4.6: (A) and (B) are distributions, γ_1^{MOT} and γ_2^{MOT} as a function of $\cos(\theta_1)$ and $\cos(\theta_2)$ while (C) and (D) are its comparison with $\gamma_1^{PDB(S)}$ and $\gamma_2^{PDB(S)}$ respectively. Comparison between (E) $\gamma_1^{PDB(S)}$ and $\gamma_1^{PDB(D)}$ (F) $\gamma_2^{PDB(S)}$ and $\gamma_2^{PDB(D)}$.

In Fig 4.6(A) and Fig 4.6(B), $\gamma_1^{MOT}(\cos\theta_1)$ and $\gamma_2^{MOT}(\cos\theta_2)$ distributions are shown and these have a prominent peak at $\cos\theta_1 = \cos\theta_2 = 0.0$ belonging to alpha-helix region. In addition to these peaks, others are observed around 0.4, when the central amino acid is alanine. This peak shifts to right around 0.6 in the case of glycine being the central amino

acid and these peaks correspond to the extended beta-sheet regions in the Ramachandran scatter plot.

In Fig 4.6(C) and Fig 4.6(D), we now compare $\gamma_1^{MOT}(\cos\theta_1)$ and $\gamma_2^{MOT}(\cos\theta_2)$ with $\gamma_1^{PDB(S)}(\cos\theta_1)$ and $\gamma_2^{PDB(S)}(\cos\theta_2)$. As can be seen, the peak at $\cos(\theta_1) = 0.0$ in $\gamma_1^{MOT}(\cos\theta_1)$ match with AAAA, AAAG, AAGA, GAAA, AGAA and AAGG of $\gamma_1^{PDB(S)}(\cos\theta_1)$ while the peak at $\cos\theta_2 = 0.0$ in $\gamma_2^{MOT}(\cos\theta_2)$ match with GGAG, AGGG, AGAA, GGAA and AAAG $\gamma_2^{PDB(S)}(\cos\theta_2)$. However, the peaks at 0.4 and 0.6 of MOT do not match PDB(S). A similar mismatch can be seen in the comparison plot of $\Gamma^{MOT}(V)$ and $\Gamma^{PDB(S)}(V)$ in Fig 4.3b. The reason behind this mismatch is that output distributions obtained from MOT would depend on the choice of input distributions and here we use PDB(D) and PDB(S) input distributions which differ from each other, as PDB(D) sampling is specifically in loop regions. The comparison of these two input distributions are shown in Fig 4.6(E) and Fig 4.6(F),

Role of electrostatic interactions in cost functions

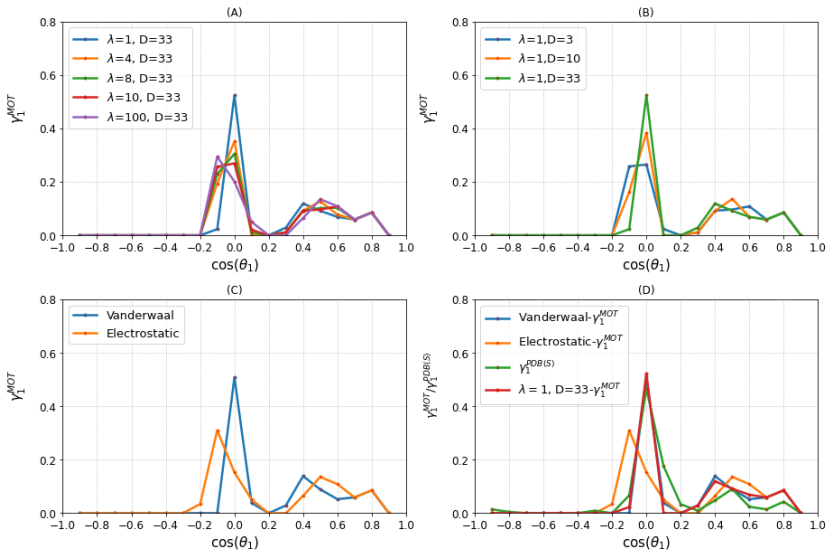


Figure 4.7: γ_1^{MOT} distributions of AAAG: (A) Various values of λ for $D = 33$. (B) Various values of D with $\lambda = 1$ (C) for very large λ (purely electrostatic interaction) and for very large D (purely Van der Waals interaction). (D) γ_1^{MOT} of AAAG of purely Van der Waals, electrostatic and $\lambda = 1$ for D=33 are compared with $\gamma_1^{PDB(S)}$

The MPD depends upon the input distributions containing the short-range correlation among backbone torsional angles and the potential energy function capturing the steric

and electrostatic interactions. The agreement of our results with data suggests the role of both van der Waals and electrostatic interactions.

In Fig 4.7, we show the effect of λ and D on γ_1^{MOT} of AAAG. The electrostatic interactions of atoms in Ala and Gly are weak since both are non-polar amino acids; the former has no side chains, and the latter has just C_β without a longer side chain.

The position of the peak ($\cos\theta_1 = 0.0$) for pure van der Waals interactions shifts towards the left with reduced amplitude when electrostatic interaction dominates (very large λ).

4.4 Conclusion

We presented the optimal transport method, which minimizes potential energy function with input tripeptide distributions of torsional angles as constraints. The most exciting aspect of this approach is to provide the multi-point correlations of the longer peptides from given short-range correlations encoded in the input distribution. Using this method, one can also verify whether the local nature of the interaction modeled by the potential energy function is appropriate by comparing its output data with experimental data. It is numerically less expensive compared to other approaches.

To demonstrate the effectiveness of this method, we studied the conformation of tetrapeptides composed of Ala and Gly and presented its detailed analysis. We showed the peak observed in distributions corresponds to the alpha-helix region matching quite closely with PDB(S) for AAAA, AGAA,AAAG, and GAAG. In addition, we pointed out why some parts do not match and discussed obtaining tripeptide distributions from the tetrapeptide distributions. Our results also predict beta turns in GGGG, GAGG, AAGG, and AAGA that are absent in PDB(S) which is owing to the fact that PDB(D) is sampled in the loop regions and turns. This method is numerically less expensive than other approaches. It can be applied to generate conformational distributions of longer peptides such as hexapeptides, octapeptides, etc.

The cost function parameters are taken from the Charmm force field parameters. The NumPy python generates the cost function. The scipy.optimize python module for Linear programming runs on the laptop with 8 GB RAM. The data points correspond to PDB(D) $\rho_i(\phi_i) \geq 0.009$ and $\tilde{\rho}_i(\psi_i) \geq 0.009$ are chosen as the input distributions to run linear

programming. The total time to get the MPD is 95.7 seconds.

Chapter 5

Tackling the Levinthal Problem with Recursive Optimal Transport

5.1 Introduction

In this chapter, we build upon the method introduced in Chapter 3, which outlined an approach to derive conformational distributions of tetrapeptides using dipeptide data from the Protein Data Bank (PDB). Here, we extend this methodology to explore the conformations of longer peptide sequences. As previously discussed, the limited availability of data for extended peptides in the PDB presents significant challenges in constructing optimized conformational landscapes. To address this issue, we enhance our approach by integrating backbone torsional angle distributions from tetrapeptides to construct the conformational landscapes of longer peptides.

This extended method employs a recursive framework that systematically combines the conformational distributions of shorter peptide fragments to build those of longer sequences. This strategy effectively manages the complex, high-dimensional nature of peptide structures despite the constraints posed by data limitations. One of the key innovations of our approach is its ability to substantially reduce computational complexity. While conventional methods face exponential growth in computational requirements as

peptide length increases, our approach scales more efficiently, allowing for the practical investigation of extended peptide conformations.

We validate this method by applying it to various peptide sequences composed solely of alanine and glycine, including hexapeptides, octapeptides, decapeptides, and an 18-residue sequence. These applications demonstrate the method's effectiveness and efficiency in navigating the intricate conformational space of extended peptides.

5.2 Method: Recursive Optimal Transport for Efficient Peptide Conformation Analysis

In the context of extended peptides, the conformational landscape of a peptide sequence $A_1A_2A_3...A_n$ of length n can be represented in a multidimensional space $X_n = \{\bar{x}_n\}$, where each tuple $\bar{x}_n = (x_1, x_2, x_3, ..., x_n)$ corresponds to the Ramachandran angles (ϕ_i, ψ_i) for each residue A_i. Each x_i fluctuates within specific bounds, leading to a distribution $\rho(x_i)$ for the torsional angles. However, individual distributions alone do not fully capture the overall conformational distribution of the peptide $\Pi_n(\bar{x}_n)$ because they lack the multi-point correlations between the backbone's Ramachandran angles, which are essential for accurately defining the conformational space. The PDB provides $\rho_i(x_i)$ as conditional probabilities for each residue A_i, serving as marginals for the overall distribution $\Pi_n(\bar{x}_n)$. The primary objective of the optimal transport problem in this context is to minimize the cost function:

$$E_n[\Pi_n] = \min_{\Pi} \sum_{\bar{x}_n} K_n(\bar{x}_n)\Pi_n(\bar{x}_n)$$

subject to the constraints:

$$\sum_{\bar{x}_n \setminus x_i} \Pi_n(\bar{x}_n) = \rho_i(x_i), \quad i = 1, 2, ..., n$$

These equations define a multimarginal optimal transport problem involving n marginals, making the computation particularly intensive for high-dimensional datasets.

While applying OT to fusion technique, each tuple $\bar{x}_n$ from the source and target distributions is divided into two components: $\bar{x}_{n_1}$ and $\bar{x}_{n_2}$. This modification reduces the complexity of the problem, transforming the multi-marginal transport into a more tractable two-marginal problem.

$$E_n[\Pi_n] = \min_{\Pi} \sum_{\bar{x}_{n_1}, \bar{x}_{n_2}} K_n(\bar{x}_{n_1}, \bar{x}_{n_2}) \Pi_n(\bar{x}_{n_1}, \bar{x}_{n_2})$$

subject to:

$$\sum_{\bar{x}_{n_1}} \Pi_n(\bar{x}_{n_1}, \bar{x}_{n_2}) = \Pi_{n_2}(\bar{x}_{n_2}), \quad \sum_{\bar{x}_{n_2}} \Pi_n(\bar{x}_{n_1}, \bar{x}_{n_2}) = \Pi_{n_1}(\bar{x}_{n_1})$$

In this framework, tuples represent sets of points that capture the multidimensional features and dependencies inherent in the data, thus aligning the transport problem more closely with the biological context. The source and target distributions, Π_{n_1} and Π_{n_2}, are linked through the cost function $K_n(\bar{x}_{n_1}, \bar{x}_{n_2})$, which quantifies the cost of transitioning between the configurations $\bar{x}_{n_1}$ and $\bar{x}_{n_2}$. The approach is depicted schematically in Fig. 5.1. This configuration ensures that the transport plan Π_n aligns with the marginal distributions of both the source and target, maintaining consistency across datasets.

This optimized approach captures the complex interdependencies within multidimensional data by reducing the original $(n+1)$-equation multi-marginal problem to a more manageable two-marginal distributional transport framework. It provides a robust tool for studying peptide conformations, addressing the complexities associated with high-dimensional peptide data.

Constructing the Potential Energy Function (K_n): The potential energy function integrates van der Waals and electrostatic interactions:

$$K_n(\bar{x}_n) = \sum_{i<j} \epsilon_{ij} \left[\left(\frac{r_{0_{ij}}}{|\vec{R}_i - \vec{R}_j|} \right)^6 - 1 \right]^2 + \sum_{i<j} \frac{q_i q_j}{D|\vec{R}_i - \vec{R}_j|}$$

where $\vec{R}_i$ and $\vec{R}_j$ denote atomic coordinates, $|\vec{R}_i - \vec{R}_j|$ depends on the torsional angles x_n, and $\epsilon_{ij}, r_{0_{ij}}, q_i$, and D represent potential depth, van der Waals radii, atomic charges, and the dielectric constant, respectively. Parameters from the CHARMM force field are used in these calculations [22].

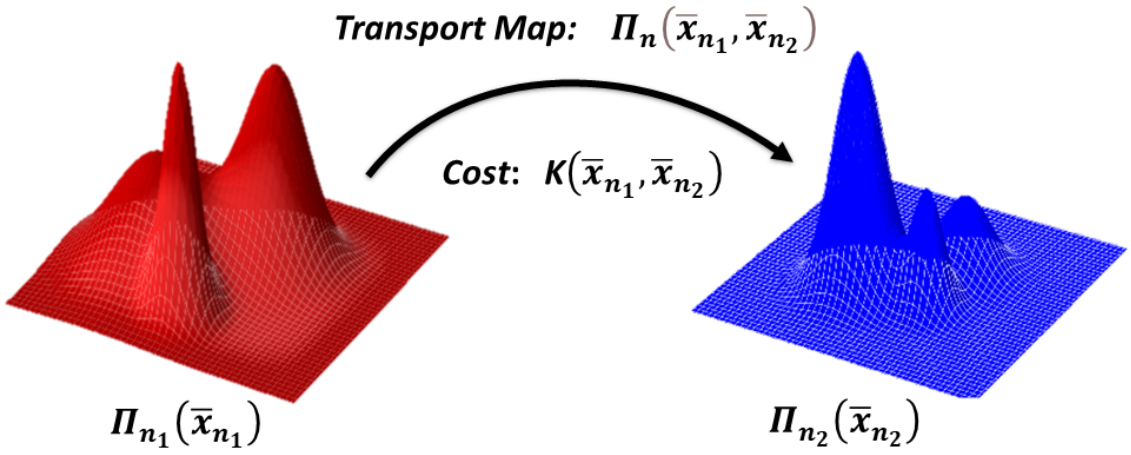


Figure 5.1: The optimal transport, functions between two multivariable distributions with marginals $\Pi(\bar{x}_{n_1})$ and $\Pi(\bar{x}_{n_2})$. By employing the cost function $K(\bar{x}_{n_1}, \bar{x}_{n_2})$, this method aims to establish the optimal transport map $\Pi_n(\bar{x}_{n_1}, \bar{x}_{n_2})$.

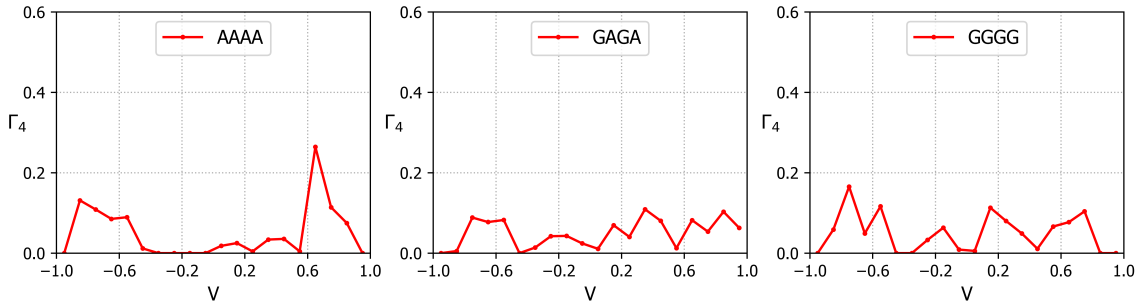


Figure 5.2: The figure shows tetrapeptide distributions (Γ_4) as a function of (V) volume for the tetrapeptides AAAA, GAGA, and GGGG, respectively, from the left [17].

Optimization of a Single Tetrapeptide: Following the methodology of Kannan et al. [17], the distribution Π_4 for a tetrapeptide $A_1A_2A_3A_4$ is optimized by adjusting the Ramachandran angles (ϕ, ψ) of the central residues while keeping the outer residues fixed. This optimization identifies several non-zero elements in the distribution, selecting m elements based on a cutoff criterion to emphasize the most significant configurations, forming a basis for higher-order distributions. Fig.5.2 illustrates the distribution Π_4 for tetrapeptides such as AAAA, GAGA, and GGGG as a function of tetrapeptide volume.

Fusion Process: To construct larger peptides, tetrapeptides are fused. For example, fusing $A_1A_2A_3A_4$ with $A'_1A'_2A_5A_6$ (aligning $A'_1 = A_3$ and $A'_2 = A_4$) creates a hexapeptide $A_1A_2A_3A_4A_5A_6$ as shown in Fig.5.3. The optimized angle sets from the initial tetrapeptides, $\bar{X}_4 = \{(x_2^{\text{opt}}, x_3^{\text{opt}})\}$ and $Y_4 = \{(x_{2'}^{\text{opt}}, x_5^{\text{opt}})\}$, are combined, and the cost function $K_6(\bar{x}_4, \bar{x'}_4)$ is calculated, forming an $m_1 \times m_2$ matrix.

Establishing Optimal Transport for Peptides: By utilizing the distributions $\Pi_4(\bar{x}_4)$, $\Pi'_4(\bar{x'}_4)$, and the cost function $K_6(\bar{x}_4, \bar{x'}_4)$, the distribution for the hexapeptide $\Pi_6(\bar{x}_4, \bar{x'}_4)$ is derived using two-marginal optimal transport, ensuring the resulting distribution retains the marginals of the original tetrapeptides, with the basic feasible solution encompassing $m_1 + m_2 - 1$ non-zero entries.

Creating Longer Peptide Chains: To form octapeptides, decapeptides, and beyond, this process is recursively repeated by fusing hexapeptides with additional tetrapeptides. In conventional multimarginal optimal transport, the computational complexity scales exponentially as $(m)^{4p}$ for p tetrapeptides. However, our approach reduces this complexity to a superlinear scaling of $(\sum_{i=1}^{p-1} m_i) \times m_p$.

Instead of experiencing full exponential growth, the recursive fusion significantly reduces calculation size. For example, rather than increasing at a rate of $(m)^{4p}$, our method scales more moderately by accumulating the sum of previously optimized configurations multiplied by the next set, rendering the process computationally feasible even for longer peptides. This pivotal realization transforms the problem from an otherwise intractable exponential scale to a manageable superlinear one.

We designate this scheme as "**Recursive Optimal Transport Theory (ROT)**," which builds upon the principles of optimal transport (OT) to recursively model the conformational landscapes of extended peptides. ROT provides a robust framework for navigating the complex, high-dimensional spaces inherent to peptide structures efficiently. By applying OT recursively, ROT streamlines the analysis of multidimensional peptide conformations, making it a powerful tool for studying peptide structures.

5.3 Structural Clustering and Data Visualization

Understanding the complex, multivariate distributions of peptide conformations obtained from the ROT method is essential for comprehending their structural properties. To analyze these multipoint distributions, we employ "band diagrams," which directly show the configurations and their corresponding probabilities based on the raw ROT data. These band diagrams deliver an immediate, intuitive depiction of the conformational landscape, providing a clear overview of the stability and variability of peptide structures without

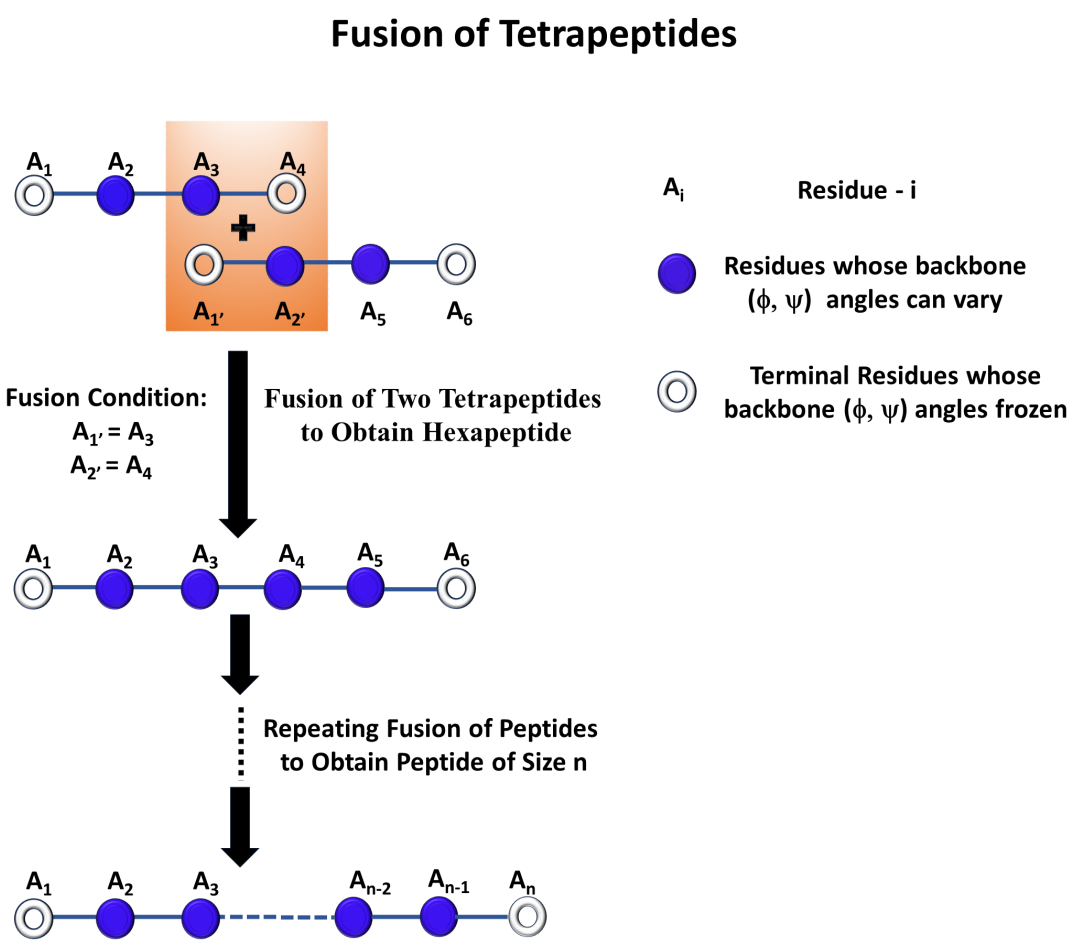


Figure 5.3: Illustration of the fusion process between two tetrapeptides to form a hexapeptide and further extend to longer peptides.

the need for extensive computations.

To conduct a deeper analysis, we employ clustering using the root mean square deviation (RMSD) of the C_α atoms, which assesses the structural resemblance between different conformations in 3D space. The clustering procedure begins with the most likely structure from the probability distribution Π_n as the starting reference. Structures are then categorized into clusters based on their RMSD values: those with RMSD less than 1 Å are grouped into one cluster, while those with RMSD of 1 Å or more are assigned to another cluster. This iterative procedure continues, using the next most probable ungrouped structure as the new reference at each subsequent clustering step until all structures have been classified.

This combined approach of band diagrams and clustering allows us to categorize peptide

sequences into three distinct structural classes:

Dominant Configurations: These are characterized by one, two, or three configurations with significantly higher Π_n values, indicating that these are the most likely stable conformations.

Disordered Peptides: In this category, a few distinct configurations have similar Π_n values, with all others much lower, suggesting the peptide lacks a single predominant conformation and exhibits disordered behavior.

No Unique Stabilizing Structure: This category is defined by a broad range of configurations with Π_n values that span from high to intermediate, suggesting that the peptide does not have a unique stabilizing structure but rather a continuum of possible conformations without clear dominance.

These classifications enable us to identify the structural properties of peptide sequences, determining whether they adopt stable conformations, display disordered behavior, or lack a defined structure entirely. Though clustering offers a precise and refined perspective by grouping probabilities into distinct clusters (Γ_n), band diagrams provide an immediate visual representation of peptide behavior across various conformational states. By combining these two methods—utilizing band diagrams for a quick overview and clustering for in-depth analysis—we obtain a thorough understanding of peptide conformational landscapes, gaining insights into their stability, variability, and the potential presence of continuum states within sequences.

5.4 Results and Discussion

We implemented our method across a spectrum of peptides composed solely of alanine and glycine, including 13 hexapeptides, 4 octapeptides, 10 decapeptides, and an 18-residue peptide. To maintain clarity and focus in the main text, we concentrated on three representative hexapeptides, each illustrating one of the three identified categories: dominant, disordered, and non-stabilizing structures, along with three decapeptides, since these result from the fusion of hexapeptides. The results for the remaining peptides are provided in the supplementary materials. In the core section, we display the band and bar plots for these representative types. Additionally, we offer detailed visualizations of two structures

each for the dominant and disordered categories, with further structural analyses presented in the supplementary section. This strategy enables us to underscore key findings without overloading the main discussion with excessive data.

Conformational Preferences of Alanine-Rich Peptides: Insights into Helical Structures and Stability

We start by examining the conformational preferences of sequences made up of repeating alanines, with the tetrapeptide AAAA serving as the primary unit. This tetrapeptide shows a non-zero probability distribution in relation to its volume V, as depicted by the plot of Γ_4 against the tetrapeptide volume V [17]. The visual data in Fig. 5.2 indicate a peak at a certain volume, indicating a preferred conformation. By extending this structure, combining two AAAA units forms a more intricate but still predominantly preferred arrangement, a pattern that holds true for both the hexapeptide AAAAAA and the decapeptide AAAAAAAAAA. These peptides display a repeating AA sequence that favours certain conformations. This repeating pattern not only enhances the stability of the peptides but also shows their inclination to retain a dominant conformation as the peptide length increases, which is vital for comprehending the stability and behaviour of longer alanine-rich peptides.

Fig. 5.4 presents the band and bar plots for the hexapeptide AAAAAA, labeled as (E) and (F), respectively. Similarly, Fig. 5.5 shows these plots for the decapeptide AAAAAAAAAA, labeled (D) and (E), respectively.

The hexapeptide AAAAAA mainly adopts a right-handed alpha-helical conformation, with its most likely structure being strongly favored over other forms. This preference is illustrated by a significant gap of 0.016 in the band plot between the second and third most probable structures, signifying a strong energetic bias towards the alpha-helix. In the corresponding bar plot, the primary conformation is categorized in cluster 0, with a Γ_6 value of 0.40. About 50% of the probability is associated with conformations similar to the 3_{10} helix, while the remaining are closely aligned with the right-handed alpha helix, underscoring the stability of this structure.

Similarly, the decapeptide AAAAAAAAAA, formed by fusing two hexapeptides, continues to exhibit a dominant right-handed alpha-helical structure, with a persistent gap of 0.011

in the Π_{10} band plot (D) between the second and third most probable structures. The most dominant conformation appears in cluster 0 of the bar plot (E), with a Γ_{10} value of 0.150. This consistent preference across varying lengths underscores the robust structural stability associated with right-handed alpha helices in alanine-rich peptides, affirming their role as reliable models for studying polypeptide conformations.

In addition to the dominant right-handed alpha helices, alanine-rich peptides also display minor alternative conformations, such as polyproline II (PPII) helices. For instance, in the hexapeptide, cluster 1 shows an alternative structure with $\Gamma_6 = 0.13$, suggesting that despite the prevalence of the alpha-helix, there is some degree of conformational flexibility within these sequences. Although these alternative structures are less frequent, they provide insight into the dynamic range of conformations that alanine-rich peptides can adopt, which may be relevant in more complex polypeptide and protein structures.

Overall, the alanine-rich sequences analyzed here demonstrate a pronounced conformational preference for stable, dominant structures, primarily right-handed alpha helices. This preference is highlighted by the significant probability gaps observed in the plots, emphasizing the strong energetic favorability for these conformations. The detailed visual representations of these configurations enhance our understanding of how structural motifs like the right-handed alpha helix contribute to the broader stability and folding properties of polypeptides.

In Fig. 5.4, the dominant configurations of AAAAAA with all atoms are shown in label (A), and the backbone shape of the structure alone is displayed in label (B). Labels (C) and (D) depict the aligned backbone structures of clusters 0 and 1 from the bar diagram (F), respectively. Similarly, in Fig. 5.5, the dominant configuration of AAAAAAAAAA with all atoms is shown in label (A), and the backbone shape of the structure alone is in label (B). Label (C) shows the aligned backbone structure of cluster 0 from the bar diagram (D), providing a comprehensive visualization of the conformational landscape of these peptides.

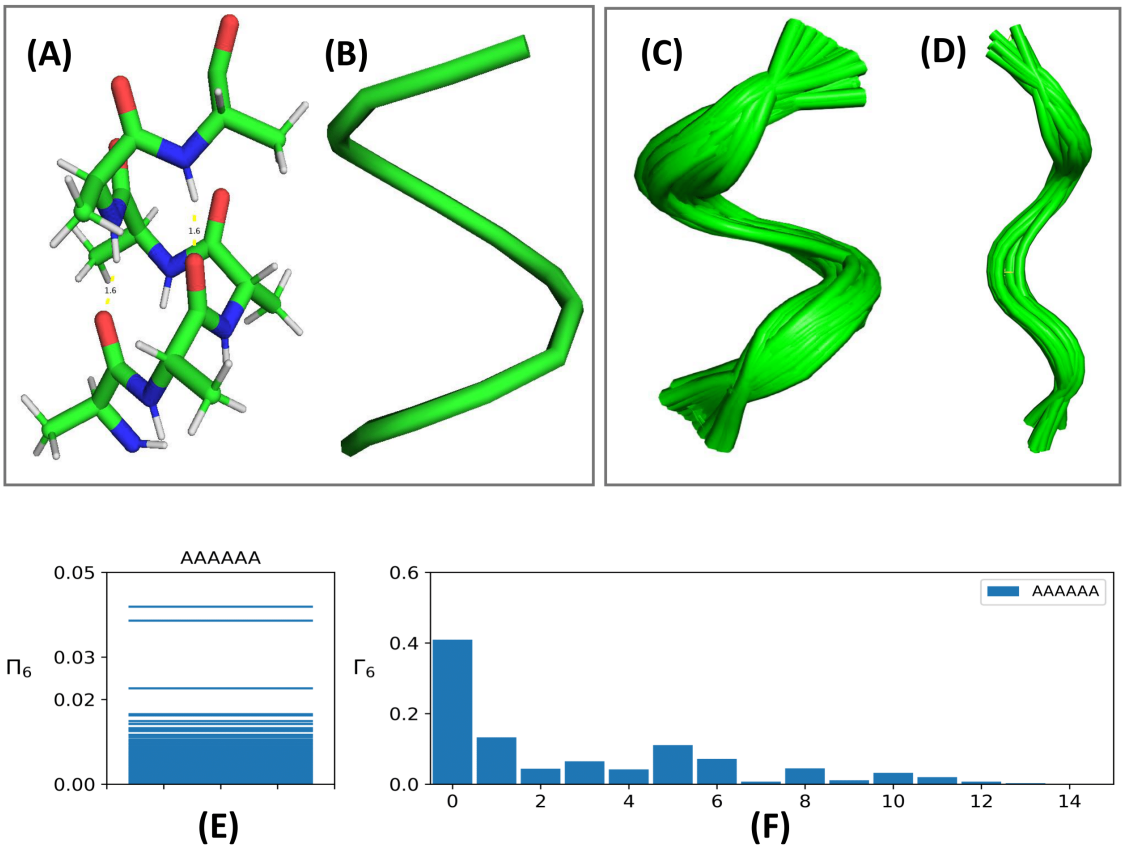


Figure 5.4: The figure presents the band diagram (E) and bar diagram (F) of the peptide AAAAAA. Label (A) displays the dominant configuration of AAAAAA with all atoms, while (B) shows the backbone shape of the structure alone. Labels (C) and (D) illustrate the aligned backbone structures of clusters 0 and 1 from the bar diagram (F), respectively.

Transition to Repeating GA Sequences: Exploring Disordered Structures

In contrast to the orderly stability found in alanine-rich peptides, sequences containing glycine and alanine (GA) motifs typically show disordered conformational tendencies. While alanine-rich sequences usually form stable, prominent alpha-helical structures, GA repeats do not exhibit such structural dominance. Instead, they adopt a range of conformations that are marked by considerable flexibility and lack strong stabilizing interactions. This transition from structured to disordered behavior underscores the significant influence that sequence composition exerts on peptide stability and conformation..

To investigate these patterns, we focus on the conformational preferences of repeating GA sequences, utilizing the tetrapeptide GAGA as the basic building block. The distribution for GAGA, depicted in the second plot of Fig. 2, indicates a non-zero probability, sug-

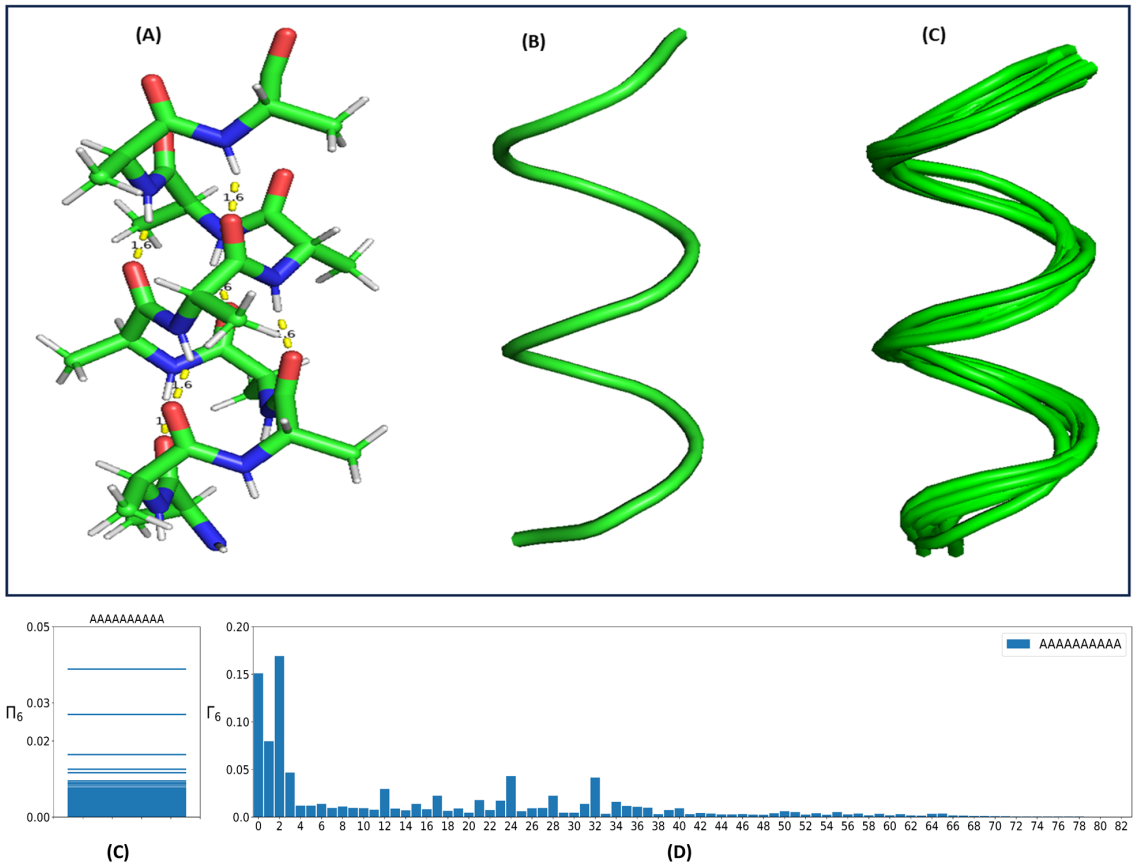


Figure 5.5: The figure displays the band diagram (C) and bar diagram (D) of the decapeptide AAAAAAAAAA. Label (A) shows the dominant configuration of AAAAAAAAAA with all atoms, while (B) depicts the backbone shape of the structure alone. Label (C) illustrates the aligned backbone structure of cluster 0 from the bar diagram (D).

gesting a variety of accessible conformations. Extending this pattern by connecting two such tetrapeptides results in a highly disordered state, as shown by the band plot in Fig. 3 (G). This disordered state continues in the hexapeptide GAGAGA and the decapeptide GAGAGAGAGA, both of which incorporate the repeating GA motif.

In the bar plot of the hexapeptide GAGAGA (Fig. 5.6 (H)), cluster 0, with $\Gamma_6 = 0.16$, primarily includes 3_{10} helices and repeated type I β turns, which are structurally similar due to their compact and helical nature. Additionally, GAGAGA forms repeated type II β turns, also known as beta bend ribbons, which are seen in cluster 2 with $\Gamma_6 = 0.095$. The alternating presence of glycine, especially in the third and fourth positions, facilitates the formation of type II β turns, as glycine uniquely adopts the overlapping dihedral angles in the I and IV quadrants of the Ramachandran plot, which are essential for these turns. Type II' β turns are also observed in the central four residues, represented in cluster 6

with $\Gamma_6 = 0.034$, supported by glycine's flexibility that allows access to these otherwise less common angles in quadrant IV of the Ramachandran plot.

The decapeptide GAGAGAGAGA further exemplifies this disordered nature, as depicted in the band plot (Fig. 5.7 (J)), where it consistently adopts a range of conformations without a dominant structure. Repeated type II β turns, or beta bend ribbons, are prominently featured in cluster 1 of Fig. 5.7 (K), while cluster 5 contains a mix of 3_{10} helices and repeated type I β turns, with Γ_{10} values of 0.02 and 0.06, respectively. The presence of these flexible structures underlines the inherent variability and lack of a singular, stable conformation in GA-rich sequences.

Overall, sequences with the repeating $(GA)_n$ motif consistently form repeated β turn structures across various lengths, such as hexapeptides, octapeptides, and decapeptides, without a clear dominant conformation emerging. This behavior starkly contrasts with alanine-rich peptides, which tend to stabilize into a predominant right-handed alpha-helical structure. The differences underscore the unique structural tendencies of GA motifs, which favor disordered and flexible conformations over the stabilized, structured forms observed in alanine-rich peptides. This flexibility may have implications for the functional properties of GA-containing peptides and proteins, offering a broader range of dynamic conformational states.

Fig.5.6 and 5.7 illustrate these findings in detail. In Fig. 5.6, the band diagram (G) and bar diagram (H) of the hexapeptide GAGAGA are shown. Label (A) presents a single conformation representing the 3_{10} helix from cluster 0, while (B) illustrates the aligned structures that form the 3_{10} helix within cluster 0. Label (C) shows a single conformation of the repeated type I β turn from cluster 0, and (D) displays the aligned structures of the type I β turn within cluster 0. The single conformation of the repeated type II β turn, or beta bend ribbon, is depicted in (E), with (F) showing the aligned structures of repeated type II β turns from cluster 2.

In Fig. 5.7, the band diagram (J) and bar diagram (K) of the decapeptide GAGAGAGAGA are presented. Label (A) shows a single conformation from cluster 1, and (B) depicts the aligned, repeated type II β turn structures characteristic of cluster 1. Label (C) displays a single conformation from cluster 5, with (D) illustrating the respective aligned 3_{10} helix conformations of cluster 5. In (E), the single conformation of the repeated type I β turn

is shown, while (F) presents the aligned structures of repeated type I β turns within this cluster. Label (H) features a single conformation with a type I turn followed by a 3_{10} helix, and (I) shows the aligned structures of various conformations from the same cluster 5.

Disordered configurations of GAGAGA and GAGAGAGAGA with all atoms are depicted in Fig. 5.6 (A) and Fig. 5.7 (A), respectively, with their backbone shapes alone shown in (B). Labels (C) and (D) in these figures display the aligned backbone structures of clusters 0 and 1 from the bar diagrams (H), respectively, providing a comprehensive visualization of the disordered and diverse conformational landscape of GA-rich sequences.

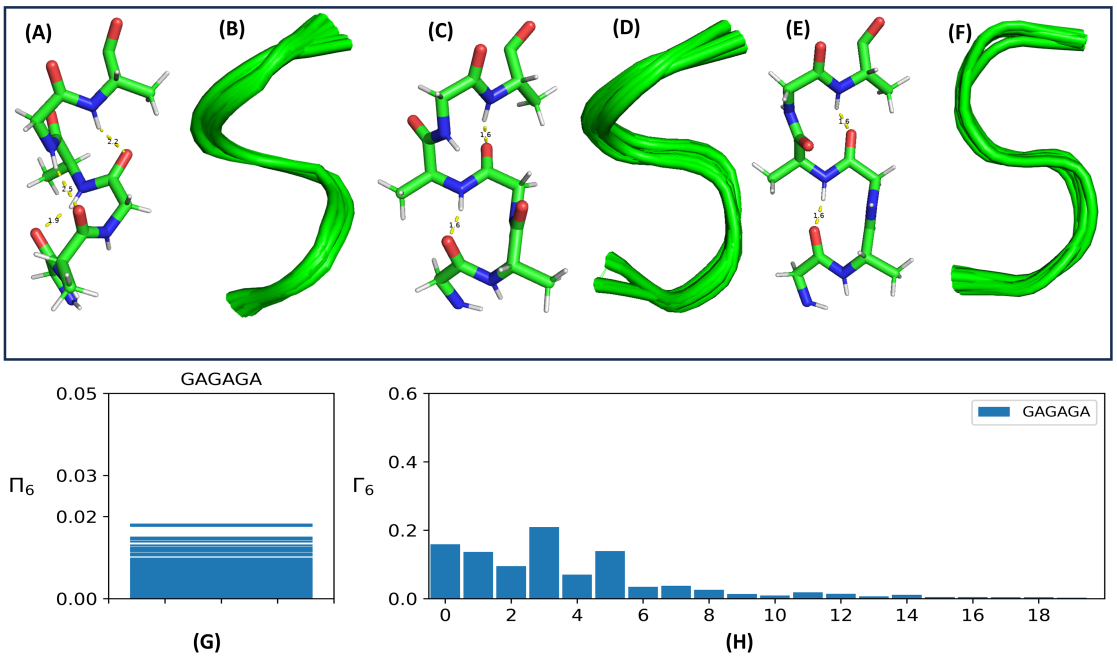


Figure 5.6: The figure displays the band diagram (G) and bar diagram (H) of the hexapeptide GAGAGA. Label (A) shows a single conformation representing the 3_{10} helix from cluster 0, with (B) illustrating the aligned structures that form the 3_{10} helix within cluster 0. Label (C) depicts a single conformation representing the repeated type I β turn from cluster 0, and (D) shows the aligned structures that form the type I β turn within cluster 0. The single conformation of the repeated type II β turn, also known as the beta bend ribbon, is shown in (E), while (F) displays the aligned repeated type II β turn structures of cluster 2.

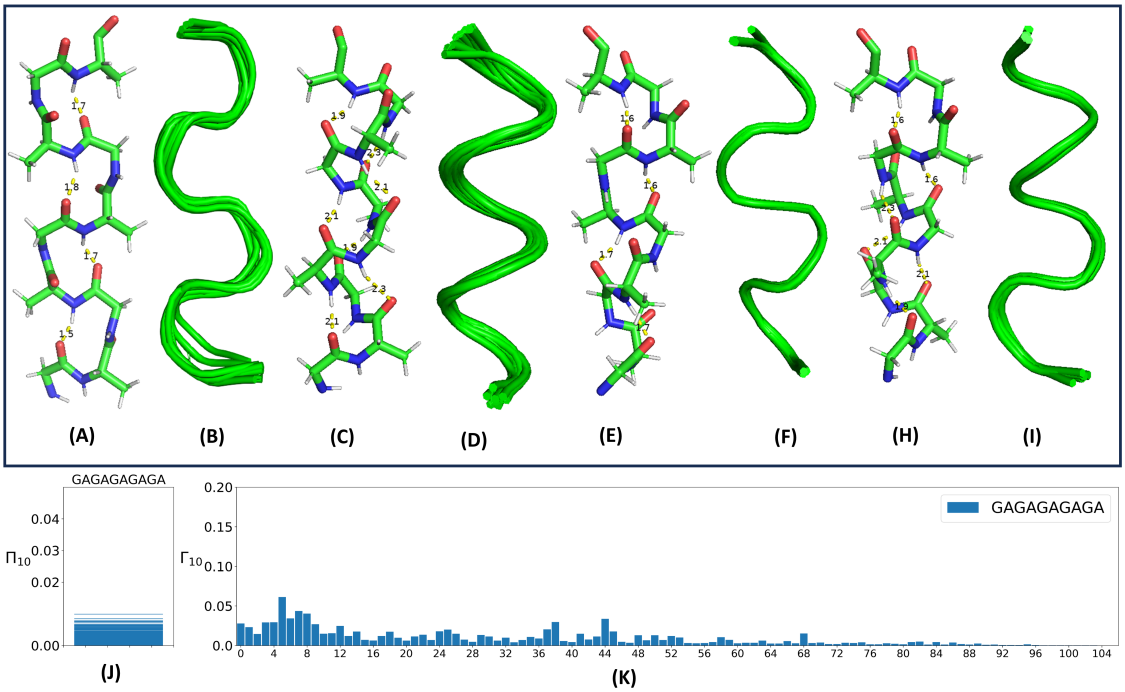


Figure 5.7: The figure includes the band diagram (G) and bar diagram (H) of the decapeptide GAGAGAGAGA. Label (A) shows a single type II β turn conformation from cluster 1, and (B) illustrates the aligned, repeated type II β turn structures of cluster 1. Cluster 5, which includes three sets of closely related conformations 3_{10} helices, repeated type I β turns, and structures with a 3_{10} helix followed by a type I turn along the backbone are shown in (C) through (I). (C) and (D) depict a single 3_{10} helix and aligned 3_{10} helices. (E) and (F) show the single and aligned repeated type I turns. Finally, (H) presents a conformation with a 3_{10} helix followed by a type I turn, and (I) shows its respective all aligned structures.

Conformational Analysis of Glycine-Rich Sequences: Lack of Stabilizing Structures

We now turn our attention to glycine-rich sequences, which are known for their exceptional flexibility. The hexapeptide GGGGGG, derived from the fusion of GGGG tetrapeptides—whose distribution is depicted in Fig. 5.2—exemplifies this flexibility, as evidenced by its distribution plot. Among the hexapeptides analyzed, GGGGGG is unique in displaying negligible or no gaps in its probability band plot (see Fig. 5.8), indicating a lack of structural preference or stabilization. The differences in probabilities between the most probable, second most probable, and subsequent structures are extremely small, often as low as 0.001, underscoring the absence of a dominant conformational state.

In the bar plot, the peptide shows no preferred cluster with a notably high probability; in-

stead, all conformations appear equally likely, reflecting its inherent flexibility. GGGGGG primarily adopts type II' and type I' β turns, which are distributed across the first, middle, and last four residues of the sequence, but these turns do not aggregate into a predominant secondary structure. For example, type II' and I' turns in cluster 0 contribute to a combined Γ_6 value of 0.13, evenly split between the two types of turns. This uniform distribution of minor structural motifs illustrates that GGGGGG lacks a distinctive conformation and does not possess significant stabilizing forces to favor any one structure over another.

The distinct conformational behavior of glycine-rich sequences, characterized by their inherent flexibility and resistance to adopting a dominant structure, stands in stark contrast to the more ordered alanine-rich peptides and even the disordered but somewhat structured GA sequences. This observation is crucial for understanding the broad structural landscape of peptides, highlighting how sequence composition directly influences conformational preferences. Glycine-rich peptides exemplify the extreme end of the flexibility spectrum, lacking any stabilizing elements that might otherwise encourage the formation of specific secondary structures.

A similar trend is observed in decapeptides derived from glycine motifs, such as GGGGGGGGGG. These longer glycine-rich sequences continue to exhibit a lack of dominant structural preferences, maintaining a diverse array of minor conformations without significant stabilization of any particular form. This behavior further emphasizes the role of glycine in promoting structural diversity and flexibility within peptide chains, which is in stark contrast to the more structured and stable configurations seen in peptides dominated by alanine. Understanding these differences provides valuable insight into the factors that govern peptide stability and folding, as well as the potential functional implications of such flexibility in biological contexts.

Our investigation into conformational preferences, detailed through precise probability distributions, uncovers insights that are often beyond the reach of conventional experimental or simulation techniques. While methods such as NMR or CD spectroscopy are effective for identifying prominent structures, they generally fall short in accurately quantifying the probabilities of various conformations across the full peptide state ensemble. Similarly, molecular dynamics (MD) simulations can explore the conformational landscape

over time but frequently encounter computational limitations, making it difficult to sample rare conformations, particularly in highly flexible sequences like glycine-rich peptides. In contrast, our approach directly evaluates the probabilities of different conformational states, offering a comprehensive view that encompasses both prevalent and less common structures, thereby providing a deeper understanding of peptide behavior. This probabilistic framework enables clear differentiation between dominant structures, disordered states, and highly flexible sequences, yielding insights that surpass the resolution typically available with traditional methods. Consequently, our findings broaden our understanding of the peptide conformational landscape by capturing states that might be overlooked or underrepresented in conventional experimental or simulation approaches. This method highlights not only the variability and flexibility inherent to peptide structures but also the importance of considering the full ensemble of conformations when studying peptide dynamics and stability.

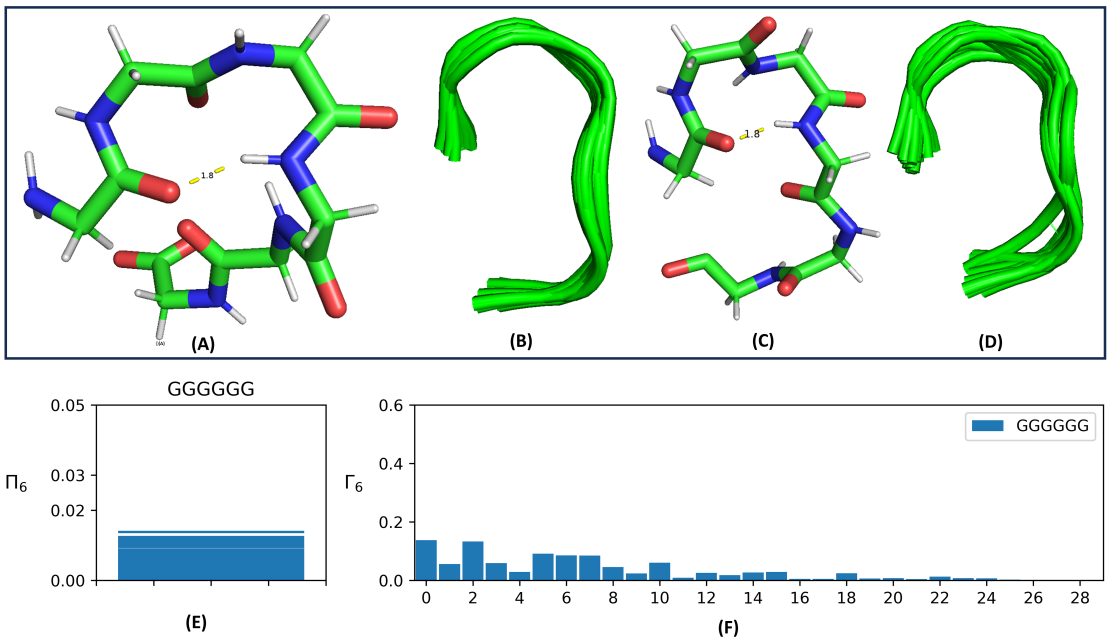


Figure 5.8: The figure displays the band diagram (E) and bar diagram (F) of the hexapeptide GGGGGG. Label (B) shows the aligned structures of the type II' β turn in cluster 0, with (A) depicting a single conformation from these aligned type II' β turn structures. Similarly, label (D) shows the aligned structures of the type I' β turn in cluster 0, with (C) presenting a single conformation from the aligned type I' β turn structures.

5.5 Conclusion and Future Directions

We have developed the Recursive Optimal Transport (ROT) method, a novel approach designed to tackle the challenges of exploring the vast configurational spaces of extended peptides—a significant barrier in computational simulations. Unlike methods such as AlphaFold, which utilize deep learning models to predict protein structures from sequence information, ROT leverages optimal transport theory to systematically and efficiently navigate these extensive spaces. ROT distinguishes itself by breaking down complex peptide structures into smaller, manageable segments, like dipeptides and tetrapeptides, which are sequentially assembled into larger peptide chains. This approach transforms a high-dimensional problem into a series of lower-dimensional tasks, greatly enhancing computational efficiency.

At each step, smaller configurations are integrated using a sequence of optimal transport calculations, where the cost function reflects both physical interactions and probabilistic constraints. This recursive strategy not only streamlines the modeling process but also improves predictive accuracy by focusing on the most relevant configurations from a constrained set of possibilities. ROT has been successfully applied to simulate a variety of peptides, ranging from hexapeptides and octapeptides to decapeptides and even an 18-residue sequence, as detailed in the Supplementary Information. This demonstrates its capability to accurately capture diverse peptide conformations, providing insights that traditional experimental or simulation methods often fail to achieve. However, the scarcity and fragmented nature of experimental data for larger peptides make direct comparisons challenging.

ROT can be extended to McCann interpolation for dynamic systems, broadening its application to complex biological processes, such as protein-protein interactions and cellular signaling pathways. This expansion enhances the adaptability and effectiveness of ROT, offering deeper insights into the mechanisms underlying these processes and disease pathways. Its ability to generate detailed energy landscapes also makes it a valuable tool in drug discovery and design, where identifying energetically favorable conformations and transition states is crucial for developing targeted therapeutic solutions.

Beyond biological systems, ROT can be applied in materials science, including the study

of polymers and glassy systems, where understanding structural properties is key. Its capacity to handle extensive configurational spaces and provide detailed insights into conformational preferences makes ROT a versatile tool across multiple disciplines. Ongoing efforts to optimize fusion techniques aim to further enhance the method's stability and efficiency, expanding its applicability even further. This continuous development underscores ROT's potential as a powerful computational framework for exploring complex structural landscapes in peptides and beyond.

Our findings on conformational preferences, quantified through detailed probability distributions, reveal insights that are often elusive with conventional experimental or simulation techniques. Methods such as NMR and CD spectroscopy are effective at identifying dominant structures, but they typically struggle to accurately quantify the probabilities of multiple conformations within the full peptide ensemble. Similarly, molecular dynamics (MD) simulations can explore conformational landscapes over time but are frequently limited by computational constraints, making it difficult to sample low-probability conformations, particularly in highly flexible sequences like glycine-rich peptides.

In contrast, ROT directly measures the probabilities of various conformational states, providing a comprehensive view that includes both major and minor structures, thereby offering a more nuanced understanding of peptide behavior. This probabilistic framework allows for clear differentiation between dominant structures, disordered states, and highly flexible sequences, delivering insights beyond the resolution typically achievable by traditional methods. As a result, our findings enhance our understanding of the conformational landscape of peptides, capturing configurations that might be missed or underrepresented in conventional studies. This approach not only highlights the variability and flexibility inherent in peptide structures but also underscores the importance of considering the entire conformational ensemble when studying peptide dynamics and stability.

In summary, the ROT method represents a transformative approach to studying peptide conformations by directly addressing the limitations of traditional experimental and simulation techniques. By quantifying the probability distributions of various conformational states, ROT offers a detailed and comprehensive perspective on peptide behavior, including both dominant and minor structures. This enhanced understanding of the conformational landscape is critical for advancing peptide research, from fundamental studies of protein

folding to practical applications in drug discovery and materials science. ROT's ability to systematically explore and manage large configurational spaces positions it as a powerful tool for uncovering the complex structural dynamics of peptides, emphasizing the critical role of probabilistic analysis in capturing the full diversity of peptide conformations.

The cost function parameters are taken from CHARMM force field parameters. [22] The Fragbuilder Python module is used to generate peptide structures.[104] The NumPy Python module is used to compute the cost function and the Gurobi Python module is used for linear programming computations.[105] Whole calculations are carried out on a desktop with 16 GB RAM with 8 core Intel processor, Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

This thesis explores the conformations of peptides, emphasizing the creation and use of advanced computational techniques to comprehend the conformational landscapes of peptides. It combines essential peptide biochemistry principles with state-of-the-art computational modeling to make notable progress in the field..

Summary of Key Chapters

Chapter 1: Introduction

The introduction elucidated the pivotal role of peptides within biological systems, underscoring their structural and functional diversity. Peptides, characterized by their intrinsic conformational flexibility, are integral to numerous biochemical processes, including cellular communication and enzymatic activities. This chapter establishes the foundation for the research by highlighting the complexities involved in comprehending the conformational landscape of peptides, attributed to their extensive configurational space and the impact of both intrinsic and extrinsic influences. Furthermore, it elucidated the limitations of relying solely on experimental data, such as that provided by the Protein Data Bank (PDB), which, despite its invaluable contributions, frequently proves inadequate in capturing exhaustive data for longer peptides due to data sparsity and the combinatorial explosion of potential conformations. This underscored the imperative for the development

of sophisticated computational tools to address these deficiencies in peptide modeling.

Chapter 2: Peptide Geometry

This chapter concentrated on the structural fundamentals of peptides, elucidating the internal coordinates that govern their three-dimensional conformations. By scrutinizing the backbone architecture of peptides, the chapter offered a thorough analysis of the essential geometric properties, including bond lengths, bond angles, and dihedral (torsional) angles. These parameters are crucial as they collectively influence the conformational behavior of peptides, determining their folding, interactions, and functions within biological systems. An understanding of these elements is not simply of academic interest; it directly affects the capability to model peptides with precision, as subtle variations in these coordinates can result in significant alterations in the overall structure.

The chapter also elaborated on the methods used to transform internal coordinates into Cartesian coordinates—a process that bridges theoretical geometric descriptions with practical computational modeling. This transformation is crucial because Cartesian coordinates represent the precise spatial locations of atoms and serve as the input for energy calculations and simulations. By converting internal properties into a global coordinate system, researchers can create detailed three-dimensional models of peptides, which are essential for visualizing and predicting conformational states.

This chapter sets the stage for the use of advanced computational techniques detailed in the following chapters. In Chapter 4, for example, an accurate representation of internal coordinates was crucial for the analysis of peptide conformations using multi-marginal optimal transport (MOT) to ensure that multi-point probability distributions (MPD) accurately depicted actual peptide structures. Furthermore, precise internal geometry modeling aided the development of recursive optimal transport (ROT) methods in Chapter 5, where these coordinates were crucial for extending the analysis to longer peptide chains..

The geometric principles of peptides outlined in this chapter form the basis for energy calculations that help identify low-energy conformations and determine the stability of different peptide structures. This topic is especially relevant in Chapter 3's study of optimal transport theory, as accurately calculating interaction energy requires knowledge of the geometric constraints and transformations of peptide structures. Additionally, this

chapter's in-depth analysis of dihedral angles and their influence on peptide folding is revisited in subsequent chapters, highlighting their significance in multi-marginal and recursive distributional analyses.

Chapter 3: Optimal Transport

This chapter delved into the mathematical framework underlying optimal transport, chronicling its evolution from Monge's classical formulation to Kantorovich's contemporary methodology. Monge's initial problem focused on minimizing transportation costs through the mapping of resources between distributions, thereby introducing the crucial concept of efficient resource allocation. However, it was constrained by non-linear challenges and analytical difficulties. Kantorovich's reformulation significantly transformed this field by converting the problem into a linear programming model, enabling a more adaptable approach with partial transport plans. This adjustment not only made the problem more manageable but also broadened its applicability to more intricate, high-dimensional systems.

In this chapter, optimal transport was introduced as an effective analytical method for examining the transformation of one distribution to another while minimizing a specific cost function. This approach was creatively applied to peptide conformational analysis to model peptide transitions between different states within their energy landscapes. By using a cost function that mirrors the energy shifts during these transitions, optimal transport enabled the exploration of the relationships between various conformations in terms of stability and transition likelihood.

The Kantorovich approach's capacity to manage linear constraints and distributional couplings proved exceptionally beneficial for peptide studies, offering a comprehensive framework that effectively accounted for intricate interactions among various components of the peptide structure. In contrast to traditional methods, which may simplify conformational changes or overlook interdependencies, this approach facilitated the mapping and comparison of conformational distributions with enhanced precision and profundity.

Covering the foundation for utilizing optimal transport to explore the conformational landscape of peptides, this chapter addressed a theoretical gap and prepared the way for more sophisticated computational approaches discussed in upcoming chapters. The theoretical

concepts presented were further elaborated with practical applications in multi-marginal optimal transport (MOT) in the following chapter, which investigated multiple distributions together to offer a more comprehensive analysis of peptide structures. Moreover, the discussion on optimal transport laid the groundwork for recursive methods that were developed later to effectively model longer peptide chains and their conformational intricacies.

Employing optimal transport theory for peptide modeling marked a notable advancement in the analysis of peptide conformation, offering a method to evaluate and forecast peptide conformational changes. This technique improved comprehension of static conformational tendencies and also paved the way for examining dynamic transitions crucial to peptide functionality. The chapter concluded by illustrating how the Kantorovich framework could facilitate more extensive applications in structural analysis, providing knowledge that goes beyond static modeling into the predictive and exploratory aspects of peptide study.

Chapter 4: Multi-Marginal Optimal Transport for Tetrapeptide Analysis

This chapter explored and elaborated on the multi-marginal optimal transport (MOT) framework, with a particular focus on its application to the study of tetrapeptides. The MOT method extends classical optimal transport by allowing for the joint analysis of multiple probability distributions. This innovation was essential in achieving a more sophisticated and thorough comprehension of the peptide's conformational landscape, given that peptides are fundamentally multi-dimensional entities with intricate interdependencies among their torsional angles..

In this chapter, the use of MOT was developed from the core concepts of optimal transport discussed earlier, extending to capture interactions that go beyond simple pairwise distributions. By incorporating high-quality multi-point probability distributions (MPD), this method enabled the examination of short-range correlations between adjacent torsional angles in the peptide backbone. This represented a noteworthy enhancement over single-point or pairwise approaches, which often overlook the collective dynamics of peptide structures and may miss key conformational subtleties. The MPD approach offered a more comprehensive depiction of the conformational space, accounting for how alterations in one section of the peptide could affect neighboring segments, thereby providing a deeper

understanding of the overall structure.

The findings described in this chapter show that the MOT framework effectively forecasts the conformational preferences of tetrapeptides, underscoring the usefulness of multi-point analysis in structural biology. Enhanced accuracy in mapping the conformational landscape allowed the MOT method to uncover patterns and associations among torsional angles that simpler analyses might miss. This was particularly clear in peptide studies containing alanine and glycine, where MOT elucidated the interactions guiding their structural tendencies and flexibilities.

The importance of this approach went beyond the particular instance of tetrapeptides. The MOT framework demonstrated its capacity for applicability in modeling longer and more intricate peptide sequences by effectively capturing torsional correlations. This ability paved the way for the recursive techniques explored in later chapters, where the approach was adapted for longer peptide chains like hexapeptides and decapeptides. The recursive optimal transport (ROT) method subsequently expanded upon the concepts introduced here, facilitating more extensive use of these computational strategies for modeling complex peptide conformations.

This chapter explored MOT to provide practical insights for overcoming typical challenges in peptide modeling, particularly data sparsity and the vastness of conformational space dimensions. By exploiting multi-point correlations and the cost-minimizing attributes of optimal transport, this approach achieved a balance between computational practicality and biological precision. Its capacity to manage the complexity of various distributions simultaneously represented a significant advancement in computational tools for structural biology. Overall, this chapter positioned the MOT framework as an influential tool for peptide conformational analysis. Its precise capture and prediction of conformational preferences emphasized its potential for wider uses, such as enhancing peptide-based drug development and predictive models of protein folding. The method's development and successful application to tetrapeptides not only showcased its immediate advantages but also hinted at its future integration and expansion in subsequent chapters, which will explore more intricate peptide systems with equally rigorous methodologies.

Chapter 5: Recursive Optimal Transport (ROT) – A Computational Approach for Extended Peptides

The ROT (Recursive Optimal Transport) approach discussed in this chapter represents a major step forward in analyzing longer peptide chains. Expanding upon the fundamental ideas outlined in the multi-marginal optimal transport (MOT) framework from the previous chapter, ROT takes these concepts further by employing a recursive method to effectively model the intricate conformational dynamics of extended peptide sequences. This recursive strategy is crucial for managing the exponential expansion of conformational space that occurs with increasing peptide length, a difficulty that has historically constrained conventional modeling methods.

The ROT method capitalized on the advantages of MOT to manage multi-point distributions and incorporated recursive techniques to enable iterative examination of peptide segments. By segmenting longer peptide chains into smaller, more manageable parts and applying distributional optimal transport to these components, ROT successfully represented the entire conformational landscape while retaining essential details. This recursive approach guaranteed that local conformational dynamics were maintained and that overall structural integrity was preserved, facilitating an accurate representation of the extended peptide behavior.

This chapter emphasizes a major innovation: ROT's ability to alleviate the computational difficulties involved in modeling extended peptide chains. As peptides grow longer, the possible conformations multiply exponentially, resulting in daunting computational and data issues. ROT tackled these by employing a systematic, incremental methodology that amalgamated findings from shorter segments, thereby creating an integrated model for longer chains. This approach not only simplified the computational burden but also preserved the accuracy of the conformational evaluation, offering a more feasible and scalable solution.

The application of structural clustering and data visualization techniques was pivotal in deciphering the results of ROT. These methods categorized conformational data into clusters according to similarity, offering significant insights into the variety of structures that extended peptides can form. Visualization tools were instrumental in recognizing predominant conformations, structural patterns, and motifs that could be associated with certain sequence characteristics or functional outcomes. This feature was especially apparent in

examining hexapeptides, octapeptides, and decapeptides, where ROT demonstrated its proficiency in managing longer peptide chains while maintaining high accuracy.

The findings illustrated that ROT could unveil the intricate equilibrium between stability and flexibility in elongated peptides. For example, in sequences featuring both alanine and glycine, the technique identified how structural rigidity and flexibility were distributed across the chain, providing insights into how different residues impacted the overall conformational dynamics. This comprehensive understanding bears significant implications for fields such as protein folding research, where insights into peptide segment behavior can inform broader theoretical frameworks and models. Moreover, the recursive method laid a foundation for potential future advancements. Employing ROT not only tackled the immediate challenge of expanding peptide conformational analysis but also paved the way for incorporating dynamic models that consider time-dependent alterations in peptide structures. This might involve adapting ROT to investigate conformational transitions, folding pathways, or interactions with other biomolecules, thereby further augmenting its usefulness in structural biology.

Overall, the ROT methodology detailed in this chapter represented a significant advancement in the analysis of peptide conformations. By expanding upon the MOT framework with a recursive application approach, ROT offered a robust method for modeling and understanding the varied structures of longer peptides. The implementation of structural clustering and visualization enhanced result interpretation, providing a clear insight into the conformational diversity of extended peptide chains. This method showed it is possible to scale up computational peptide analyses while maintaining accuracy, paving the way for more intricate studies and practical applications in drug discovery, protein engineering, and advanced peptide research.

Comparison with Existing Methods

For the hexapeptide AGAGAG, Recursive Optimal Transport (ROT) completes the ensemble construction in approximately **320 minutes on a single processor**, compared to a standard molecular dynamics (MD) simulation (80 ns) which requires roughly **80 hours on 140 processors**. This highlights that ROT is orders of magnitude faster and

far less demanding in computational resources.

Traditional Monte Carlo approaches face similar challenges: they must explore vast numbers of random moves to build conformational ensembles, often requiring large computational clusters and extended runtimes, and may become trapped in low-probability regions, potentially missing important rare conformers. By contrast, ROT efficiently fuses local backbone-angle distributions via optimal transport and minimizes interaction energy in a direct step.

A key limitation of ROT lies in its dependence on the quality of the input marginal distributions; sparsity or bias in these inputs propagates to the final ensemble. In contrast, MD simulations generate their own Boltzmann-weighted ensembles and provide full time-series dynamics, which ROT currently lacks, though kinetic extensions are possible. Deep learning-based methods such as AlphaFold produce rapid predictions of one or several best structures but do not estimate full conformational ensembles.

Overall, ROT combines the strengths of physics-based energy minimization and data-driven empirical distributions to efficiently generate comprehensive probability distributions of peptide structures with minimal computational cost. Its unique balance of speed and probabilistic coverage presents a compelling complement to existing methods.

Synthesis of Findings

The integration of MOT and ROT methodologies provided a robust computational toolkit for studying peptide conformations. The analysis showed that:

- **Peptide-specific insights**: Alanine-rich peptides were observed to favor stable, helical structures, while glycine-rich sequences exhibited considerable conformational flexibility, lacking consistent secondary motifs.

- **Scalability and Limitations**: While the ROT approach successfully extended conformational modeling to longer peptides, challenges related to computational efficiency and data sparsity for even more complex sequences were identified, highlighting areas for further methodological refinement.

6.2 Outlook

The methodologies developed and applied in this thesis pave the way for future advancements in peptide modeling and related fields. The potential extensions and applications include:

Applications and Future Research Directions

1. Expanding to Dynamic Systems

In this chapter, the ROT (Recursive Optimal Transport) method has been primarily applied to static conformational analysis. This approach has been instrumental in revealing the variety of structural forms that peptides may assume, effectively modeling both localized and extensive conformational preferences across diverse peptide sequences. However, ROT's potential extends into dynamic realms; by incorporating techniques like Macaan interpolation, ROT can be adapted to investigate the temporal dynamics of peptide conformations. Macaan interpolation, which allows for smooth transitions between conformational states over time, enables researchers to chart the continuous pathways that peptides traverse during folding and unfolding processes.Integrating this method with ROT would transform the approach into a dynamic analysis tool, facilitating the study of the kinetics associated with conformational changes. This expansion would be crucial for examining peptide transitions through intermediate states, highlighting the pathways from unfolded to folded structures and vice versa. Such insights could reveal the energy barriers encountered during folding, providing a deeper understanding of transition mechanisms.

The expansion of ROT to embrace dynamic behavior holds significant implications. This advancement would bridge the gap between static structure prediction and the real-time behavior of peptides, capturing the kinetic aspects of structural transformations. Grasping these dynamics is vital in various biological scenarios, such as peptide interactions with other molecules, receptor binding, or environmental responses. It would also illuminate the folding pathways leading to secondary and tertiary structure formation, offering a comprehensive view of peptide behavior in physiological contexts.

Incorporating Macaan interpolation within the ROT framework would enable the modeling of intricate, time-sensitive conformational rearrangements. Enhanced with this technique, researchers could simulate folding pathways and identify critical transitional states, providing potential predictive power for folding kinetics and stability. This would be particularly advantageous for studying peptides linked to diseases where misfolding or aggregation carries significant biological repercussions. By capturing peptide transitions between native and non-native states, scientists might uncover factors contributing to misfolding disorders, such as amyloid diseases.,Additionally, this enhancement would pave the way for exploring peptide flexibility and adaptability in response to external changes, such as variations in pH, temperature, or interactions with other biomolecules. Dynamic modeling would offer a more realistic portrayal of how peptides behave in natural environments, adding predictive capabilities that static models alone lack. Such modeling could also be used to examine the effects of mutations on peptide dynamics, exploring essential structure-function relationships vital for understanding numerous biological processes.

The dynamic ROT framework, combined with Macaan interpolation, could also advance computational methods used in drug discovery and peptide design. By simulating real-time folding and binding processes, researchers could pinpoint potential binding sites and anticipate the kinetics of peptide-based drug interactions with greater precision. This would enhance the rational design of peptides and peptide-mimetics that must maintain specific structural features over time to be effective. In conclusion, while the current ROT approach is superb for static conformational analysis, extending it to dynamic behavior via methods like Macaan interpolation opens a promising new chapter for peptide research. This development would enable a thorough modeling of conformational transitions, offering insights into folding pathways, structural rearrangement kinetics, and the dynamic interactions of peptides within complex biological systems. Integrating these dynamic features would vastly increase ROT's application scope, enhancing our understanding of peptide behavior and boosting the predictive power of computational models in structural biology and beyond.

2. Drug Discovery and Peptide Design

The comprehensive modeling capabilities outlined in this thesis have direct implications for drug discovery, particularly in designing peptide-based therapeutics. By understanding

conformational preferences and stability, researchers can develop peptides with enhanced bioactivity and reduced off-target interactions. The ROT framework, combined with machine learning algorithms, could be employed to screen peptide libraries efficiently and predict high-affinity candidates for specific targets.

3. Integration with Experimental Techniques

Coupling computational models with experimental data, such as NMR spectroscopy and X-ray crystallography, can further validate and refine predictions. This integration would enhance the accuracy of conformational models, making them more reliable for practical applications in structural biology.

4. Adaptation for Complex Biomolecules

The methodologies developed here could be adapted for use with larger biomolecular assemblies, such as proteins or protein-peptide complexes. Modifying the ROT approach to accommodate the complexity of these systems would enable comprehensive modeling of interactions at the molecular level.

General Perspective

The integration of computational techniques outlined in this thesis represents a significant step forward in understanding the complex nature of peptide conformations. The methodologies employed, particularly the multi-marginal optimal transport (MOT) and recursive optimal transport (ROT) frameworks, have set a new benchmark for analyzing peptide behavior across varying chain lengths. By utilizing advanced modeling techniques that incorporate multi-point correlations and recursive assessments, this work enhances our understanding of peptide conformational landscapes. These computational tools effectively uncover patterns and interactions that are challenging to identify through experimental methods alone, demonstrating their critical role in peptide research.

Future expansions of these techniques hold great promise for theoretical and applied sciences. Adapting these methods to account for dynamic and interactive biological systems could transform peptide and protein research by enabling the observation and prediction of biomolecular behavior over time and in diverse environments. Incorporating dynamic analyses, such as Macaan interpolation for simulating conformational transitions, would facilitate a shift from static to dynamic models. This approach could provide deeper in-

sights into folding pathways, stability under stress conditions, and the kinetics of binding interactions, offering a more comprehensive understanding of complex biological processes such as enzyme activity, molecular recognition, and cellular communication.

These advancements could lead to groundbreaking contributions in fields like molecular biology and biomedicine. In molecular biology, the ability to predict dynamic conformational changes could illuminate how peptides interact with biomolecules and adapt to environmental fluctuations. This capability would enhance research into protein-protein interactions, peptide-based inhibitors, and the development of novel biomolecules with targeted functions. In biomedicine, improved computational tools could optimize the development of peptide therapeutics by better predicting stable and bioactive conformations critical for drug efficacy. Dynamic ROT analysis could also significantly advance the study of peptide misfolding pathways, which are central to diseases such as Alzheimer's and Parkinson's, potentially offering new therapeutic strategies.

The integration of machine learning and artificial intelligence (AI) could further expand the applications of these computational approaches. Machine learning algorithms could enhance pattern recognition and predictive capabilities within the conformational data generated by MOT and ROT, enabling the creation of adaptable models for diverse peptide sequences and experimental setups. This integration would allow for real-time analysis, providing immediate feedback on conformational predictions, which would be particularly advantageous in fast-paced areas such as drug discovery and personalized medicine.

In summary, while this thesis has made substantial progress in advancing computational techniques for studying peptide conformations, the potential for future developments is immense. The methodologies developed and validated here establish a strong foundation for further research. The ongoing refinement and evolution of these approaches will undoubtedly lead to deeper insights into peptide structures, interactions, and functions. This progress has the potential to expand fundamental biological knowledge and revolut

Chapter 7

Supplementary

7.1 Hexapeptide

In this section of supplementary , we provide a detailed analysis of the remaining 10 hexapeptides: AAAAAG, AAAGGA, AAAGGG, AAGGAA, AGAAAA, AGAGAG, AGGAAA, GAAAAA, GGAAGG, and GGGAAA. The conformations of AAAAAA (dominant structure), GAGAGA (disordered peptide), and GGGGGG (lacking a unique stabilizing structure) have already been discussed in the manuscript.

Dominant Structure

Among the 13 hexapeptides analyzed, dominant configurations were observed for AAAAAA, AAAAAG, AAAGGG, and AGAAAA. As illustrated in Fig.7.1, the first column presents a band plot depicting the probability distributions (Π_6) of these peptides, while the second column shows bar graphs representing the probability distributions of each cluster (Γ_6). The bar and band plot of the hexapeptide AAAAAA is presented in the main article or the manuscript.(Fig.5.4)

A significant probability gap was observed either after the second most dominant structure, as in AAAGGG, or after the most dominant structure, as in AAAAAG and AGAAAA. Specifically, the most dominant structure in AAAAAG corresponds to a right-handed alpha helix, with a Π_6 gap of approximately 0.01. Similarly, in AGAAAA, the most dominant structure is also a right-handed alpha-helix with a Π_6 gap of approximately 0.01. This dominant structure appear in cluster 0 of the AGAAAA bar plot with a corresponding

value Γ_6 value close to 0.23 In many instances, the hydrogen bonds observed in the helical

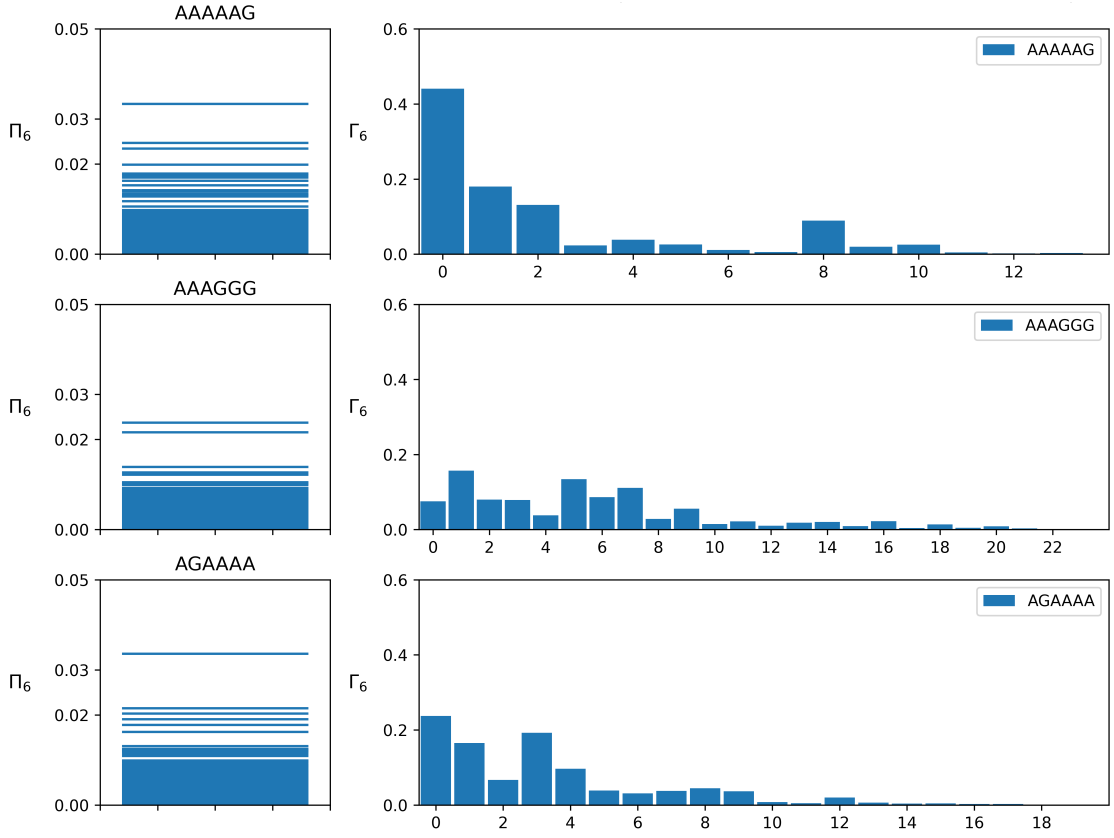


Figure 7.1: The figure shows hexapeptides with a few dominant configurations that have large band gaps. Each row corresponds to a particular peptide. The first and second columns display the band plot and bar plot, respectively.

structures of peptides within cluster 0 are bicentered. Consequently, it is beneficial to compare these peptides to the reference α-helix and 3_{10}-helix structures using root mean square deviation (RMSD) measures. The reference helices are generated with backbone ϕ, ψ angles of (-60, -45) for the α-helix and (-49, -26) for the 3_{10}-helix.

Upon conducting RMSD comparisons, it is evident that some peptides in cluster 0 are structurally closer to the 3_{10} helix. In the AGAAAA hexapeptide, the total Γ_6 value of 0.23 consists primarily of structures closer to the 3_{10} helix (0.19), with the remaining 0.04 contributed by structures closer to the α-helix. In AAAGGG, the most probable conformation is a 3_{10} helix, the second most probable structure being a type II' β turn, located in the four middle residues of the peptide. Notably, in AAAGGG, the probability gap between the second and third most probable structures is approximately 0.01. A similar trend is observed in peptide AAAAAA that has been presented in the manuscript.

In the corresponding bar plots, the two dominant structures of AAAGGG fall into clusters 0 and 1, with Γ_6 values of 0.07 and 0.15, respectively. It is noteworthy that in AAAGGG, the most probable structure corresponds to right handed alpha helical structure followed closely by a structure with type II β turn in the middle four residues. For AGAAAA and AAAAAG, the dominant structures in the band plot are located within cluster 0 of the bar plots, with Γ_6 values of 0.23 and 0.44, respectively, both corresponding to right-handed alpha helical structures.

In addition, other structures such as PPII helices, type I and II' β turns are observed in these four hexapeptides. For example, PPII helices appear in cluster 1 AAAAAG, with Γ_6 values of approximately 0.17. Type I turns are detected in the last four residues of AAAAAA (cluster 8, $\Gamma_6 = 0.04$) and AAAAAG (cluster 6, $\Gamma_6 = 0.01$).

Similarly to AAAAAA, certain peptides in cluster 0 of AAAAAG are structurally closer to the 3_{10} helix. Of the total Γ_6 value of 0.44 for cluster 0 of AAAAAG, approximately 0.21 is attributed to structures closer to the 3_{10} helix, while 0.23 corresponds to structures closer to the alpha helix. In AAAGGG, type II β turns are observed in clusters 1 and 8, involving a four-residue subsequence in the middle of the peptide, with Γ_6 values of 0.15 and 0.02, respectively. The mirror image of the type I turn, known as the type I' turn, is observed in cluster 6 of AAAGGG.

Disordered Structure

Fig 7.2 illustrates that 8 of the 13 hexapeptides (AAAGGA, AAGGAA, AGAGAG, AGGAAA, GAAAAA, GAGAGA, GGAAGG, and GGGAAA) fall into the category of disordered peptides, characterized by a very small probability band gap. The peptide GAGAGA is discussed in the manuscript.

In AAAGGA, the most probable structure is a random configuration without any secondary structure, closely followed by a 3_{10} helix, with a narrow gap of 0.002 in their Π_6 values. The observed 3_{10} helix appears distorted, with weak hydrogen bonds of lengths greater than 2 Å. In the bar plot of AAAGGA, these two most probable structures are represented in clusters 0 and 1.

In AAGGAA, the first two most probable structures are a 3_{10} helix and a type I turn in the first four residues, with a Π_6 gap of 0.001. This 3_{10} helix is also distorted due to weak

hydrogen bonds. These structures are depicted in clusters 0 and 1 in the bar plot.

For AGAGAG, the most probable and second most probable structures are a repeat type II β turn and a 3_{10} helix, respectively, with minimal differences in the Π_6 values. The first and second most probable structures are represented in clusters 0 and 1 of the bar plot.

In AGGAAA, the most probable structure is a type I β turn, followed by a random structure, with a minimal gap of 0.0002 in the values Π_6. These structures are shown in clusters 0 and 1 in the bar plot.

In GAAAAA, the first three most probable structures are right-handed alpha helices, with very small differences in the values of Π_6 (close to 0.001). Cluster 0 includes the three most probable structures. Similarly to AAAAAA, AAAAAG and AGAAAA, a comparison of structures in group 0 of GAAAAA reveals that peptides closer to the 3_{10} helix contribute 0.22 to the total Γ_6 value of 0.38, while those closer to the alpha helix contribute 0.16.

In GGAAGG, the first two most probable structures, both types II β in the last four residues, have a gap of 0.001 in their Π_6 values. These structures are depicted in cluster 0 of the bar plot.

In GGGAAA, the first and second most probable structures are a PPII helix and a type II β turn in the middle four residues, respectively, with a Π_6 gap of 0.001. The PPII helix is observed in cluster 0, while the type II β turn is observed in cluster 1.

In addition, other structures are observed in all eight hexapeptides and are grouped into clusters in the bar plot based on structural similarity. For example, PPII helices are also observed in hexapeptides such as cluster 14 of AGGAAA, and cluster 16 of AGAGAG, with values of Γ_6 of approximately 0.05, and 0.008 respectively. In particular, alanine-dominated peptides exhibit higher Γ_6 probabilities compared to other peptides that form PPII helices.

Other structures, such as type I, type II, type I', and type II' β turns, are also observed in various hexapeptides. However, many of these peptides deviate from the typical ϕ, ψ angles of the middle two residues of the tetrapeptide subsequence involved in turn formation, resulting in weak hydrogen bonds. These turns can occur in the first four, last four, or middle residues of the hexapeptide.

Type I turns are observed in the last four residues of hexapeptides in cluster 8 of AAAAAA, cluster 6 of AAAAAG and GAAAAA, cluster 4 of AGGAAA, and cluster 10 of AAGGAA,

with Γ_6 values of 0.04, 0.01, 0.01, 0.09, and 0.01, respectively. Type I turns are also seen in the last four residues of cluster 8 and 6 in GAAAAA, with Γ_6 values of 0.02 and 0.01.

Type II β turns are additionally observed in clusters 2, 5, and 8 of AAAGGA; clusters 4, 8, and 11 of AAGGAA; and clusters 5 and 6 of AGAGAG. These type II β turns involve a four-residue subsequence positioned in the middle of the hexapeptides. The associated Γ_6 values are 0.11, 0.04, and 0.05 for AAAGGA; 0.06, 0.008, and 0.03 for AAGGAA; and 0.10 and 0.01 for AGAGAG.

Some peptides that are not the most probable structures also form type-II turns in the first four residues. These turns are observed in clusters 2 and 14 of AAGGAA, clusters 4 and 8 of GGGAAA. The Γ_6 values for these clusters are 0.14 and 0.02 for AAGGAA, 0.11 and 0.09 for GGGAAA. Cluster 4 of GGGAAA is also associated with type II turns. The mirror image of the type I turn, known as the type I' turn, is observed in cluster 4 of GGGAAA, with a Γ_6 value of 0.037.

Type II' β turns are observed in the last four residues of AAAGGA (cluster 6, $\Gamma_6 = 0.07$). Cluster 6 of AGGAAA corresponds to type II' turns, formed by four-residue subsequences present in the middle, with Γ_6 values of 0.13. In GGGAAA, type II' turns are observed in the first four residues (cluster 2) and the middle four residues (cluster 4), with Γ_6 values of 0.07 and 0.144, respectively. In AGAGAG , type II' turns are observed in the last four and middle four residues, with Γ_6 values of 0.05 in cluster 10 of AGAGAG.

Furthermore, similar to the peptide GAGAGA, repeated turns, where one turn follows the other, are also observed in AGAGAG hexapeptide. Repeated types I and II are observed in AGAGAG (cluster 20, $\Gamma_6 = 0.004$).

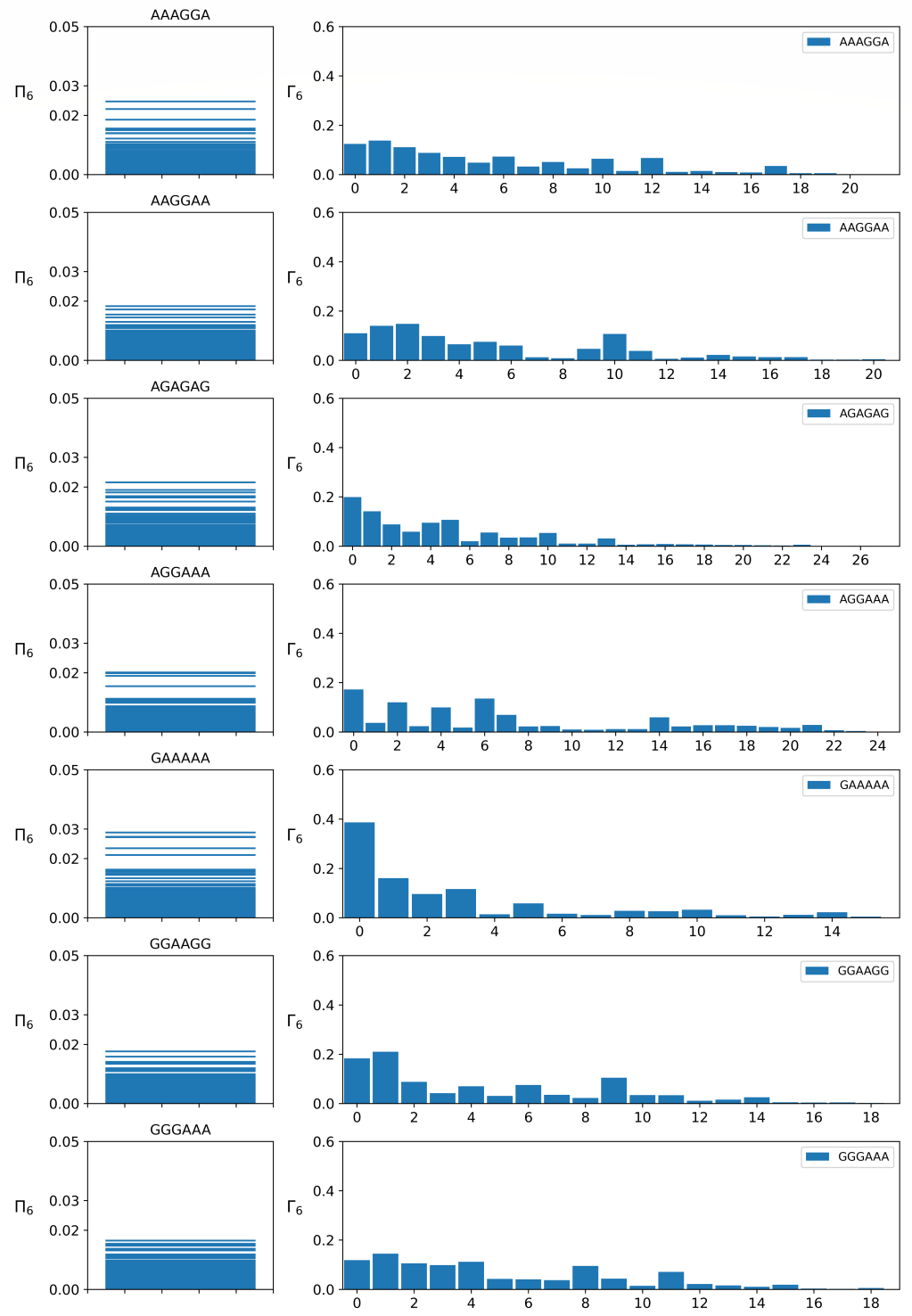


Figure 7.2: Figure shows 7 disordered hexapeptides with small probability gap between peptide structures. First column shows band plot and second column shows bar plot of the corresponding hexapeptides on the left

7.2 Analysis of Octapeptides

The octapeptide distributions for the peptides AAAAAAAA, AGAGAGAG, GAGAGAGA, and AGAAGAGG are obtained by concatenating the hexapeptide distributions composed of the first 6 residues and the last 4 residues of each respective octapeptides.

Dominant Structure in Octapeptides

Fig.7.3 display the bar plot and band plot for three octapeptides. In each figure, the first column presents the band diagram, where each horizontal line represents Π_8 obtained from ROT. The second column shows the bar plot, illustrating the probability of occurrence (Γ_8) of similar structures grouped into clusters.

Among the four peptides, three peptides AAAAAAA, AGAGAGAG, and AGAAGAGG exhibit notable differences in their structural configurations. For AAAAAAAA, there is a significant gap between the first four successive structures. In the cases of AGAGAGAG and AGAAGAGG, a significant probability gap is present after the most probable structure.

AAAAAAAA

The most probable structure of AAAAAAAA is a right-handed alpha helical structure with a probability of $\Pi_8 = 0.0419$. The second and third most probable structures are also right-handed alpha helices. The gap in Π_8 between the most probable and the second most probable structure is 0.007, and the gap between the second and third most probable structures is 0.012. The most dominant structure adopts $(\phi, \psi) = (-60, -40)$ angles, which are very close to the standard right-handed alpha helical reference angles ($(\phi, \psi) = (-60, -45)$). The ϕ, ψ angles of the second and third most probable structures deviate slightly from these standard values. The first three most probable structures fall into cluster 0 with $\Gamma_8 = 0.26$. **AGAGAGAG**

The most probable structure of AGAGAGAG corresponds to a repeated type II β turn along its backbone, also called a "beta bend" structure, with $\Pi_8 = 0.023$. The difference in Π_8 between the most probable and second most probable structures is 0.013. The most probable structure belongs to cluster 0 with $\Gamma_8 = 0.099$.

AGAAGAGG

The most probable structure of AGAAGAGG involves two repeated β type II turns starting from residues 3 to 8, with the remaining part forming a random structure. This structure has a probability of $\Pi_8 = 0.021$. The gap between the first and second most probable structures in AGAAGAGG is 0.006. The most probable structure falls into cluster 0 with $\Gamma_8 = 0.092$.

A structural comparison is made for the peptides in cluster 0 ($\Gamma_8 = 0.26$) of AAAAAAAA based on RMSD measures with respect to the reference alpha helix structure and the 3_{10} helix structure. Reference helical structures are generated with standard backbone angles $(\phi, \psi) = (-60, -45)$ and $(-49, -26)$, respectively, for the alpha helix and the 3_{10} helix. Comparing the reference alpha helix structure and the 3_{10} helix with all the peptides in cluster 0 shows that nearly 60 percent ($\Gamma_8 = 0.153$) of the peptides are closer to the alpha helix, while the remaining 40 percent ($\Gamma_8 = 0.106$) are closer to the 3_{10} helical structure. Apart from the above structures, we also observe various other structures such as 3_{10} and PPII helices in both AAAAAAAA and AGAGAGAG. In AAAAAAAA, PPII helices are observed with $\Gamma_8 = 0.0617$ in cluster 1. 3_{10} helices are exhibited by peptides in clusters 3 and 26 of AAAAAAAA and AGAAGAGG, respectively. The values of Γ_8 corresponding to their clusters 3 and 26 are 0.1408 and 0.0233 respectively.

Beta turns such as Type I and Type II are observed in peptides, and some parts adopt a random structure. A single Type I turn with a single hydrogen bond is observed in clusters 14 and 16 in residues from the third to the sixth position, with $\Gamma_8 = 0.0202$ and 0.0045, respectively, in AAAAAAAA, while the rest of the peptide exhibits a random structure.

In AGAGAGAG peptide, Type II turns are observed in peptides corresponding to clusters 1, 12, 19, 27, and 30 with $\Gamma_8 = 0.0299$, 0.0179, 0.00571, 0.0232, and 0.0129, respectively. Peptides in clusters 1, 12, 19, and 30 show Type II turns in residues starting from the fourth to the seventh, while peptides in cluster 27 adopt Type II turns in residues starting from the 2nd to the 5th residue.

Type II turns with more than one turn (two or three turns) are also observed in cluster 4 of AGAAGAGG, spanning residues from the third position to the sixth position, with Γ_8 values of 0.0929 and 0.0269, respectively. In cluster 19, peptides exhibit a Type I turn in the first 4 residues, followed by repeated Type II turns from the 3rd to the 8th residue, with a value of Γ_8 of 0.0126. The Type I turn-in peptides observed from cluster 19 of

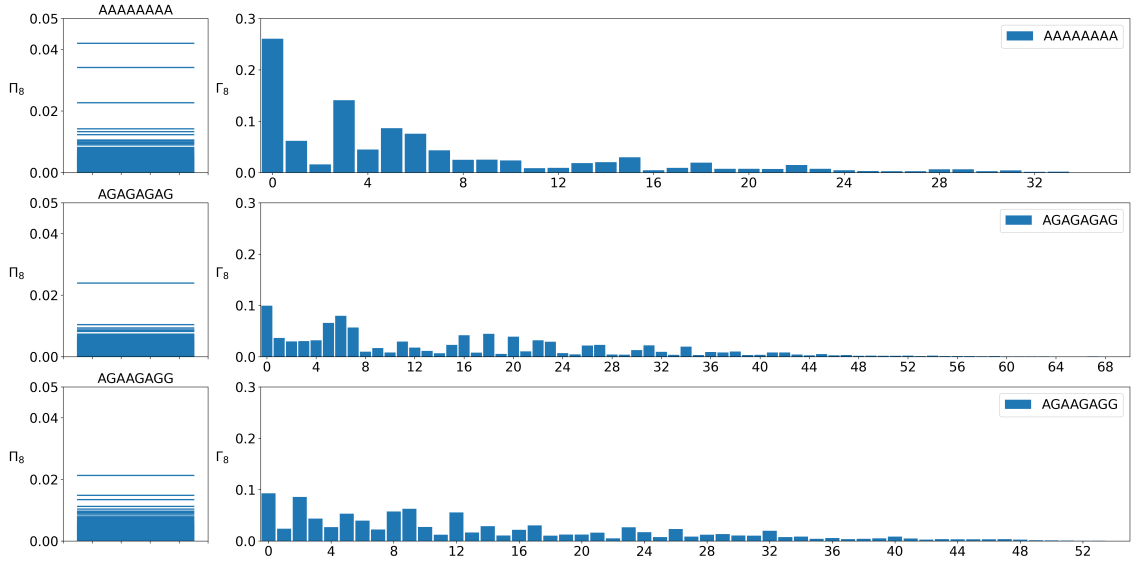


Figure 7.3: Figure shows 3 of the 4 octapeptides that falls under the category of dominant configuration. Column 1 corresponds to the band plot (Π_8 along the y-axis) and column 2 corresponds to bar plot after applying the clustering procedure where Γ_8 corresponds to each cluster. X-axis in bar plot shows the cluster numbers

AGAAGAGG demonstrate weak hydrogen bonding.

Similarly, repeated Type II turns are observed in peptides corresponding to group 0 of AGAGAGAG with a Γ_8 value of 0.0995. Furthermore, repeated Type II turns are found in clusters 3, 13, and 33 of AGAGAGAG within the middle 6 residues, with Γ_8 values of 0.0306, 0.0114, and 0.00379, respectively.

Disordered Structure in Octapeptides

Out of the four octapeptides, only the octapeptide GAGAGAGA fits into this category. Fig .7.4 presents both the band plot and the bar plot. The horizontal lines indicate the Π_8 values derived from the ROT, while the vertical bars in the bar plot represent the probability of each cluster (Γ_8). The most likely structure corresponds to a 3_{10} helix with $\Pi_8 = 0.012$. The difference between the most probable and the second most probable structures is smaller than that of the dominant structures and is 0.001. The most probable structure is located within cluster 0 of the bar plot, with $\Gamma_8 = 0.143$. The 3_{10} helices observed in peptides like GAGAGAGA are distorted, featuring weak hydrogen bonds with a bond length exceeding 2 Å in most regions..

In GAGAGAGA, Type II turns are found in peptides from clusters 1, 3, 4, 7, 8, 12,

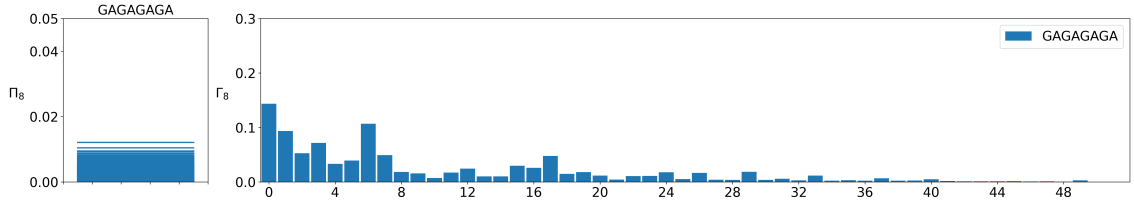


Figure 7.4: Figure shows remaining 1 of the 4 octapeptides that falls under the category of disordered configuration. Column 1 corresponds to band plot (Π_8 along y-axis) and column 2 corresponds to bar plot after applying the clustering procedure with Γ_8 corresponds to each cluster. With X axis in bar plot shows the cluster numbers

and 33 with $\Gamma_8 = 0.0935, 0.0718, 0.03342, 0.0494, 0.0183, 0.0245,$ and 0.0117, respectively. GAGAGAGA peptides in cluster 1 show a Type II

beta turn in the central 4 residues. Peptides from clusters 3 and 4 have Type II turns in the first four residues, while peptides from clusters 7 and 8 display Type II turns in the last four residues. In clusters 12 and 33, residues from the third to the sixth adopt Type II turns

In GAGAGAGA, clusters 2 and 26 show repeated Type II turns, also known as "beta bends," across their entire backbone, with each cluster displaying three repeated turns. The Γ_8 values for these clusters are 0.0528 and 0.0166, respectively. Furthermore, repeated Type II turns with two turns are found within the last six residues of the peptides in cluster 6, with a Γ_8 value of 0.1068."

7.3 Analysis of Decapeptides

Following our examination of hexapeptides and octapeptides, we now focus on decapeptides. Fig .7.5 illustrate band and bar graphs of 8 out of the 10 decapeptides: AAGGAAGGAG, AAGGAGAAGG, AGGAGAAGGA, GAGAAGGAAG, AGAAGGAAGG, GGAGAAGGAA, AGAGAGAGAG, and GGGGGGGGGG. Except for AAAAAAAAAA, which falls into the dominant configuration category, the remaining peptides are categorized as disordered, indicated by moderate gaps in their band plots. The conformations of AAAAAAAAAA (dominant structure), GAGAGAGAGA (disordered peptide) have already been discussed in the main manuscript (Fig.5.5 and 5.7).

Disordered Peptides in Decapeptides

Detailed analysis of the eight decapeptides is presented below.

AAGGAAGGAG: The most probable structure is a Type II β turn in the first four residues ($\Pi_{10} = 0.0097$). The second most probable structure, with $\Pi_{10} = 0.0096$, also forms a Type II β turn in the first four residues. These structures are found in clusters 0 and 1, with Γ_{10} values of 0.0334 and 0.0374, respectively.

AAGGAGAAGG: Both the most probable and second most probable structures exhibit multiple Type II β turns in various segments, with Π_{10} values of 0.012 and 0.011, respectively. These structures fall into cluster 0, with a Γ_{10} value of 0.055.

AGGAGAAGGA: The most probable structure ($\Pi_{10} = 0.01$) falls into cluster 0 with $\Gamma_{10} = 0.012$. The second most probable structure shows Type II turns in two segments (residues 3-6 and 6-9) with $\Pi_{10} = 0.007$, falling into cluster 1 with $\Gamma_{10} = 0.02$. The gap in Π_{10} between the two structures is 0.003.

GAGAAGGAAG: The most probable structure starts with a 3_{10} helix followed by a Type I' turn, forming weak hydrogen bonds ($\Pi_{10} = 0.013$), and falls into cluster 0 with $\Gamma_{10} = 0.07$. The second most probable structure forms a 3_{10} helix ($\Pi_{10} = 0.010$) and falls into cluster 1 with $\Gamma_{10} = 0.0799$. The gap in Π_{10} is 0.003.

AGAAGGAAGG: The most probable structure exhibits a Type II β turn in residues 3-6 ($\Pi_{10} = 0.01$). The second most probable structure also shows a Type II β turn in the same residues ($\Pi_{10} = 0.009$). The gap in Π_{10} between the two structures is 0.001. Both structures fall into cluster 0 with $\Gamma_{10} = 0.049$.

GGAGAAGGAA: The most probable structure corresponds to a Type I β turn in the first four residues. The second most probable structure shows Type II β turns in two segments (residues 2-5 and 5-8). The Π_{10} values are 0.0124 and 0.0109, with a small gap. These structures fall into clusters 0 and 2 with Γ_{10} values of 0.0124 and 0.0165, respectively.

AGAGAGAGAG: The most probable structure shows repeated Type II β turns with $\Pi_{10} = 0.012$. The second most probable structure has Type II β turns in two segments (residues 3-6 and 6-9) with $\Pi_{10} = 0.011$. These structures fall into clusters 0 and 1 with Γ_{10} values of 0.05 and 0.01, respectively.

GAGAGAGAGA: The most probable structure shows Type II β turns in multiple segments (residues 1-4, 3-6, and 7-10) with $\Pi_{10} = 0.009$. The second most probable structure, known as a β bend, has $\Pi_{10} = 0.008$. These structures fall into clusters 0 and 1 with Γ_{10} values of 0.027 and 0.023, respectively.

GGGGGGGGGG: The most probable structure is a random coil, while the second most probable structure has repeated Type II' β turns with $\Pi_{10} = 0.0086$. The second structure falls into cluster 1 with $\Gamma_{10} = 0.00826$.

Additional Structural Observations in Decapeptides

Beyond the most probable structures, additional configurations such as 3_{10} helices and Type II turns are observed across different clusters:

AAGGAAGGAG: Clusters 2 and 64 show 3_{10} helical structures with Γ_{10} values of 0.0242 and 0.0202, respectively.

AAGGAGAAGG: Clusters 18 and 27 display 3_{10} helices that span residues 4-9 and 3-8 with Γ_{10} values of 0.0281 and 0.0263, respectively.

AGGAGAAGGA: Clusters 6, 11, and 31 exhibit 3_{10} helices in various segments, with other parts forming random structures.

GAGAAGGAAG: Cluster 10 forms 3_{10} helices with Γ_{10} values of 0.0799 and 0.0419, highlighting different helical patterns compared to AAAAAAAAA.

Type II turns also appear in various subsequences, contributing to the diverse configurations: **AAGGAAGGAG**: Type II turns in cluster 41 with Γ_{10} of 0.0163, spanning residues 2-5.

AAGGAGAAGG: Type II turns in clusters 7, 9, 20, 49, and 56 with Γ_{10} values ranging from 0.0164 to 0.0389, spanning various residues.

AGAAGGAAGG: Type II turns in clusters 14, 22, and 32 with Γ_{10} values from 0.0278

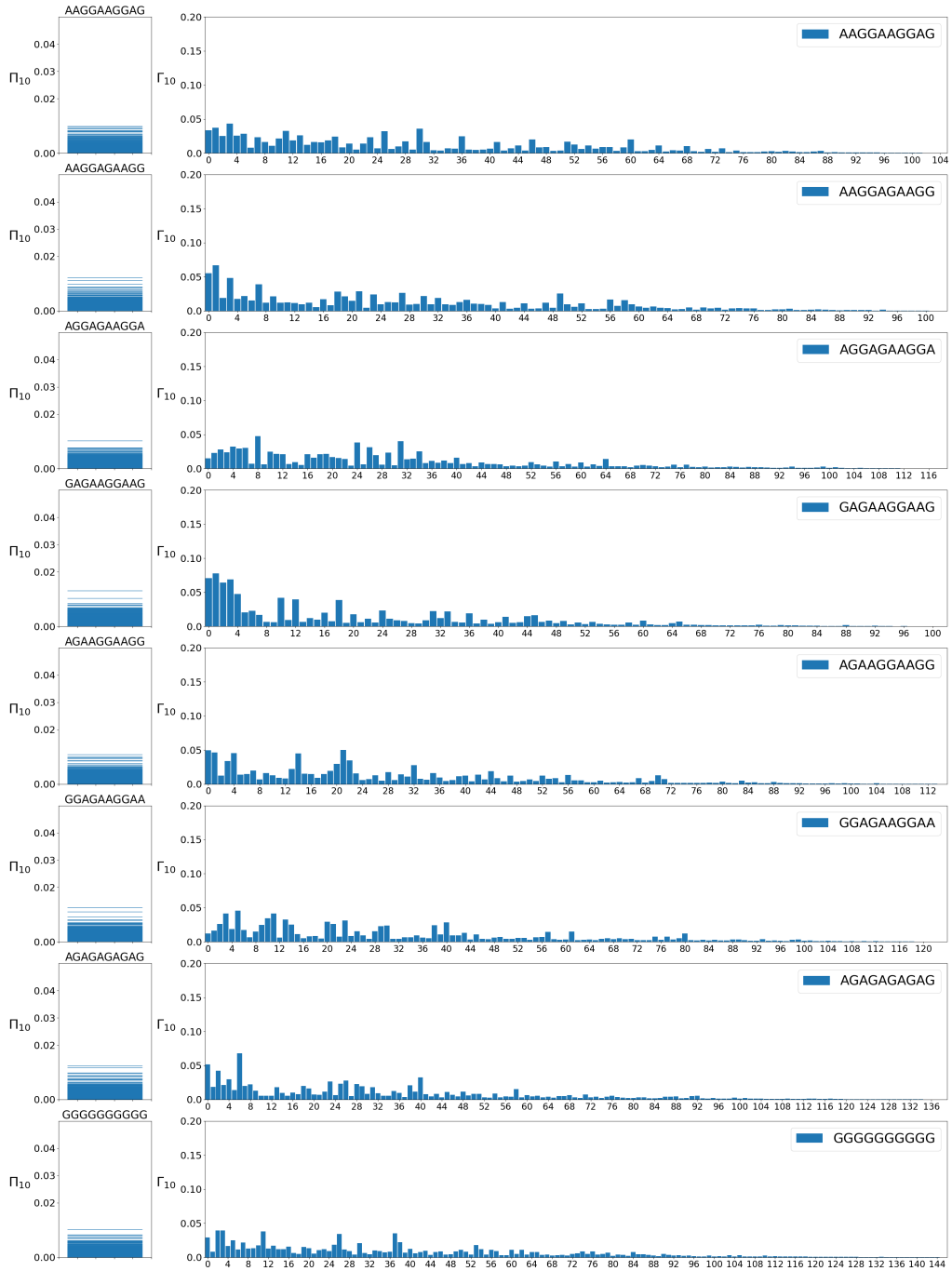


Figure 7.5: Figure shows the remaining 8 decapeptides out of the 10 decapeptide that falls in disordered configurations with not very large probability gap. First column shows the band plot and probability of each structure (Π_{10} - along the y-axis) and the second column shows the clustered peptide based on structural similarity with Γ_{10} along y axis indicating the probability of each structural clusters and x axis indicating the clusters

to 0.0450.

AGGAGAAGGA: Type II turns in clusters 3 and 27 with Γ_{10} values of 0.0196 and 0.0237, covering different residue spans.

GGAGAAGGAA: Cluster 38 shows a Type II turn in the first four residues with Γ_{10} of 0.0243.

GGGGGGGGGG: Type II turns in clusters 2, 8, and 37 with Γ_{10} values from 0.0126 to 0.0396, spanning various residues.

Single Type I and Type II' turns are also observed: **AGAAGGAAGG**: Cluster 13 forms a Type I turn from residues 3-6 with $\Gamma_{10} = 0.0222$.

GGGGGGGGGG: Cluster 112 shows a Type I turn from residues 5-8 with $\Gamma_{10} = 0.001830$.

AAGGAAGGAG: Cluster 3 displays a Type II' turn from residues 2-5 with $\Gamma 10 = 0.0433$.

GGGGGGGGGG: Clusters 4 and 5 show Type II' turns in different segments with Γ_{10} values of 0.0164 and 0.0250.

Furthermore, several decapeptides demonstrate a combination of structural motifs: **AAGGAAGGAG**: 3_{10} helices and Type II' turns in various clusters with values of Γ_{10} up to 0.0325.

AAGGAGAAGG: Cluster 1 shows a 3_{10} helix and Type II turns with $\Gamma_{10} = 0.0668$.

AGAAGGAAGG: Cluster 2 shows similar patterns with $\Gamma_{10} = 0.0125$.

AGGAGAAGGA: Cluster 0 shows 3_{10} helices and Type I/I' turns with $\Gamma_{10} = 0.01503$.

GAGAAGGAAG: Cluster 24 exhibits a $3,10$ helix and Type II' turn with $\Gamma_{10} = 0.0231$.

GGAGAAGGAA: Clusters 4, 10, 11, and 13 display both 3_{10} structures and β-turns with Γ_{10} values up to 0.041.

Lastly, multiple β turns are observed at different segments in decapeptides such as AAGGAAGGAG, AAGGAGAAGG, AGAAGGAAGG, AGGAGAAGGA, GAGAAGGAAG, and GGAGAAGGAA, with the remaining parts adopting random structures:

AAGGAAGGAG: Clusters 5, 7, 8, and 44 show Type II/II' turns with Γ_{10} values up to 0.0286.

AAGGAGAAGG: Clusters 12, 16, and 23 exhibit Type II turns in different segments with Γ_{10} values up to 0.0234.

AGAAGGAAGG: Clusters 3 and 8 show Type II turns with Γ_{10} values of 0.0336 and 0.00674.

AGGAGAAGGA: Clusters 2, 20, and 26 exhibit Type II turns with Γ_{10} values up to 0.03116.

GAGAAGGAAG: Clusters 2, 4, and 12 show Type II/Type I turns with Γ_{10} values up to 0.064.

GGAGAAGGAA: Clusters 30 and 97 show Type II turns with Γ_{10} values of 0.0238 and 0.00117.

GGGGGGGGGG: Clusters 7, 12, 13, and 26 show Type II/II' turns with Γ_{10} values up to 0.0343.

7.4 Structural Analysis of an 18-Residue Peptide

The ROT scheme can, in principle, generate probability configurations for peptides of any size. However, to test our method, we applied it to an 18 residue peptide and analyzed the structural distributions using two decapeptide distributions as input marginals. The sequence studied features a single glycine at the 9th position, represented as AAAAAAAA-GAAAAAAAAAAA (A_8GA_9). Fig. 7.6 displays the clusters of structures observed in the peptide (A_8GA_9). Due to the number of clusters being 465, they are presented in four figures. Similarly, the bar plot for this peptide, shown in Fig 7.6, is divided into four subplots.

Short Helical Structures: Clusters 0 and 1 each have different types of helices. Cluster 0 exhibits a short 3_{10} helix from residues 1 to 6, transitioning into a longer 3_{10} helix, linked by a random coil, with a Γ_{18} value of 0.0303. Cluster 1 contains a long helical structure in which residues 1 to 8 form an α helix, followed by a 3_{10} helix, with a Γ_{18} of 0.0196. Cluster 5 presents a continuous α helix with a Γ_{18} of 0.00536.

Helix-Coil Combinations: Clusters 104, 105, and 106 display combinations of helices and coils. Cluster 104 features a PPII helix that transitions to a α helix across the first and last nine residues, with a Γ_{18} of 0.00292. Cluster 105 shows a helix-coil-helix structure featuring a Type II turn in the subsequence AAGA (residues 7-9), with a Γ_{18} of 0.0020635. Cluster 106 contains a PPII helix followed by a Type II turn, with a Γ_{18} of 0.00205.

Helices Connected by Turns: Clusters 127, 117, 207, and 266 consist of helices connected by various types of turns. Clusters 127, 117, and 207 include peptides with two helices connected by a turn, with respective Γ_{18} values of 0.0026, 0.00265, and 0.0012. Clusters 117 and 207 include Type II' turns, while Cluster 117 also contains a Type I turn. Cluster 266 has two α helices connected by a random structure, with a Γ_{18} of 0.00134.

Different type of Helices and Random Structures: Clusters 141, 142, and 200 demonstrate different transitions between helices and random structures. Cluster 141 has two 3_{10} helices linked by a Type II' turn in the middle, with a Γ_{18} of 0.00169. Cluster 142 features an α helix that transitions through a random structure to a 3_{10} helix, with a Γ_{18} of 0.00169. Cluster 200 contains a Type II turn that connects a 3_{10} helix to a PPII helix, with a Γ_{18} of 0.00127.

Beta Turns and Random Structures: Clusters 217, 208, 129, 131, 133, and 135 are characterized by various combinations of beta turns and random structures. Cluster 217 shows two PPII helices linked by a Type II turn from residues 7 to 10, with a Γ_{18} of 0.00114. In Cluster 208, a Type I turn links a PPII helix to a random coil, with a Γ_{18} of 0.00116. Clusters 129, 131, 133, and 135 each contain single beta turns, with the remaining peptide structures being random. Their respective Γ_8 values are 0.0017, 0.00181, 0.00175, and 0.00173.

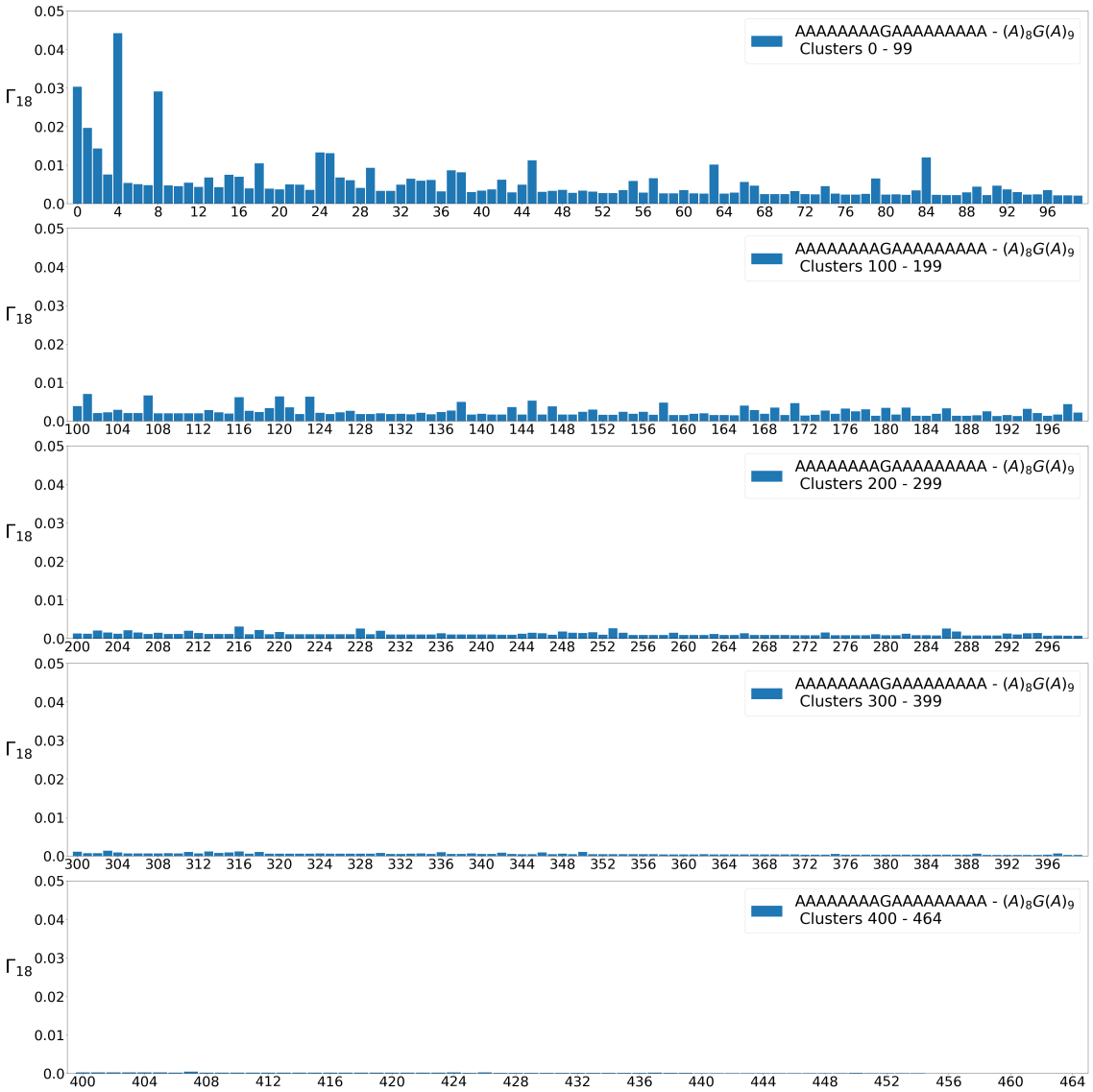


Figure 7.6: Γ_{18} is plotted against each cluster. In total, there are 465 clusters. To display all the bars for each cluster, the plot is divided into four subplots. The first subplot, starting from the top, encompasses clusters 0 to 99; the second one includes clusters 100 to 199; the third one covers clusters 200 to 299; the fourth comprises clusters 300 to 399, and the last one represents clusters 400 to 465.

7.5 Aligned Structures in Hexa,Octa, Deca and 18 Residue Peptide

To recognize the 3D structures within each cluster, we display the aligned structures of selected clusters of some of the specific peptides. Specifically, we present the first 12 clusters for the hexapeptides AAAAAA, GAGAGA, and GGGGGG (Fig. 7.7), and for the decapeptides AAAAAAAAAA, GAGAGAGAGA, and AGAGAGAGAG (Fig 7.8). For

the octapeptide and 18-residue peptide, we show 10 selected clusters (Fig.7.9). In each figure, the structural clusters are marked as "Cluster -i" to differentiate between clusters. Here, i represents the cluster number that corresponds to the respective cluster in the bar plot of the given peptide. Here While aligned structural plots were generated for all peptides and clusters, we present only a subset for clarity and brevity. The complete set of plots and detailed cluster information is available upon request.

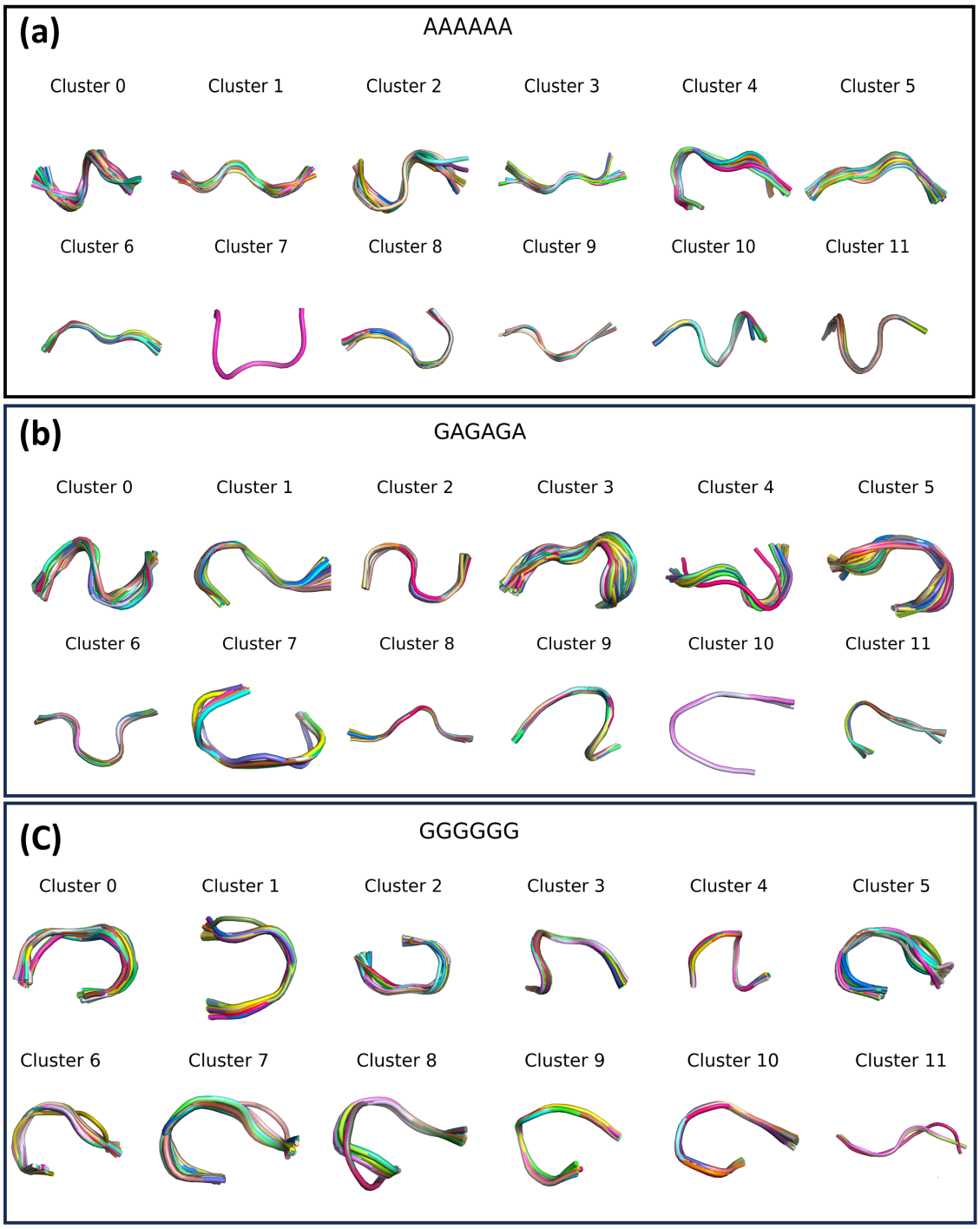


Figure 7.7: Aligned structures of the first 12 clusters observed in the hexapeptides: (a) AAAAAA, (b) GAGAGA, (c)GGGGGG

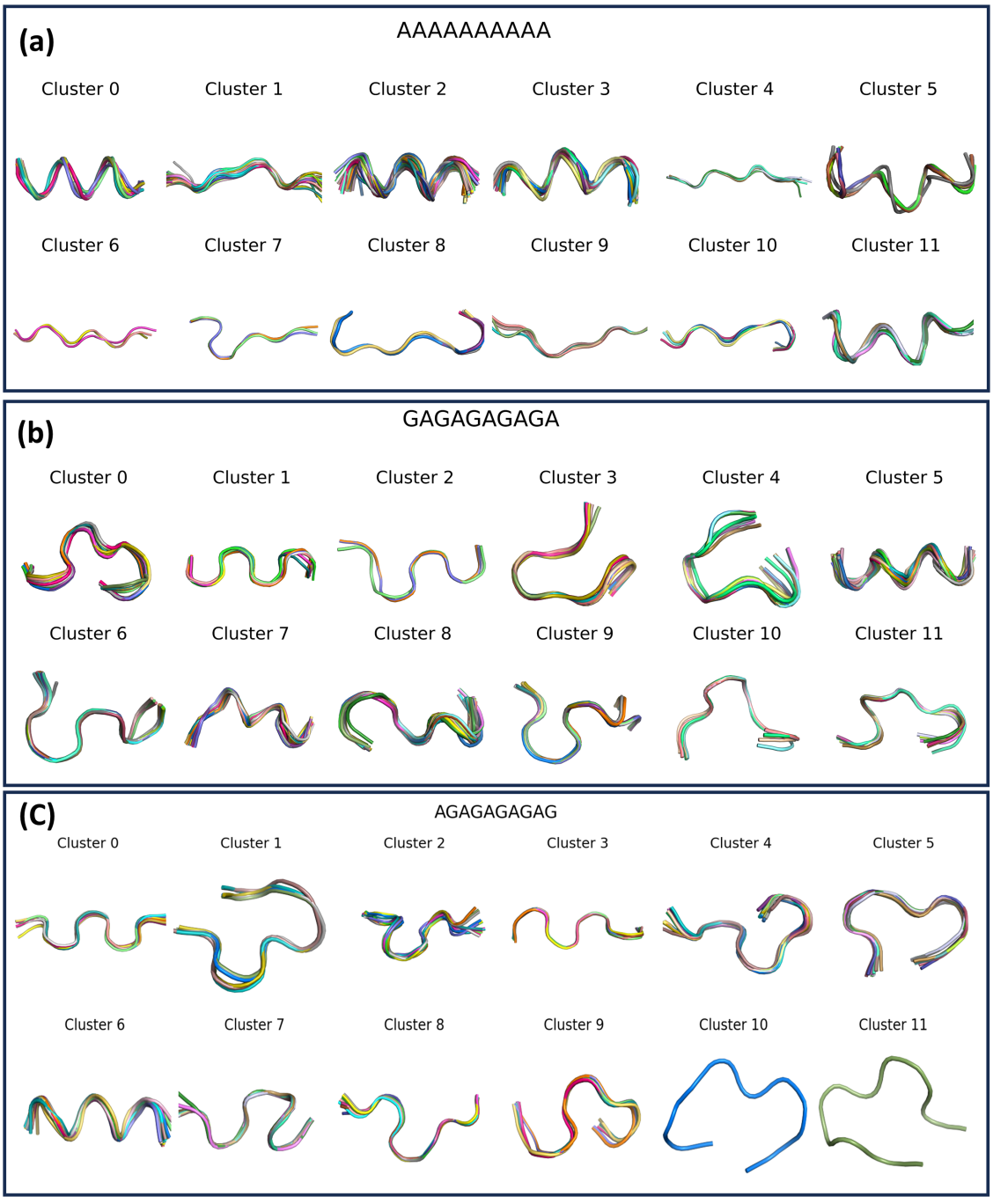


Figure 7.8: Aligned structures of the first 12 clusters observed in the decapeptides: (a) AAAAAAAAAA, (b) GAGAGAGAGA, and (c) AGAGAGAGAG.

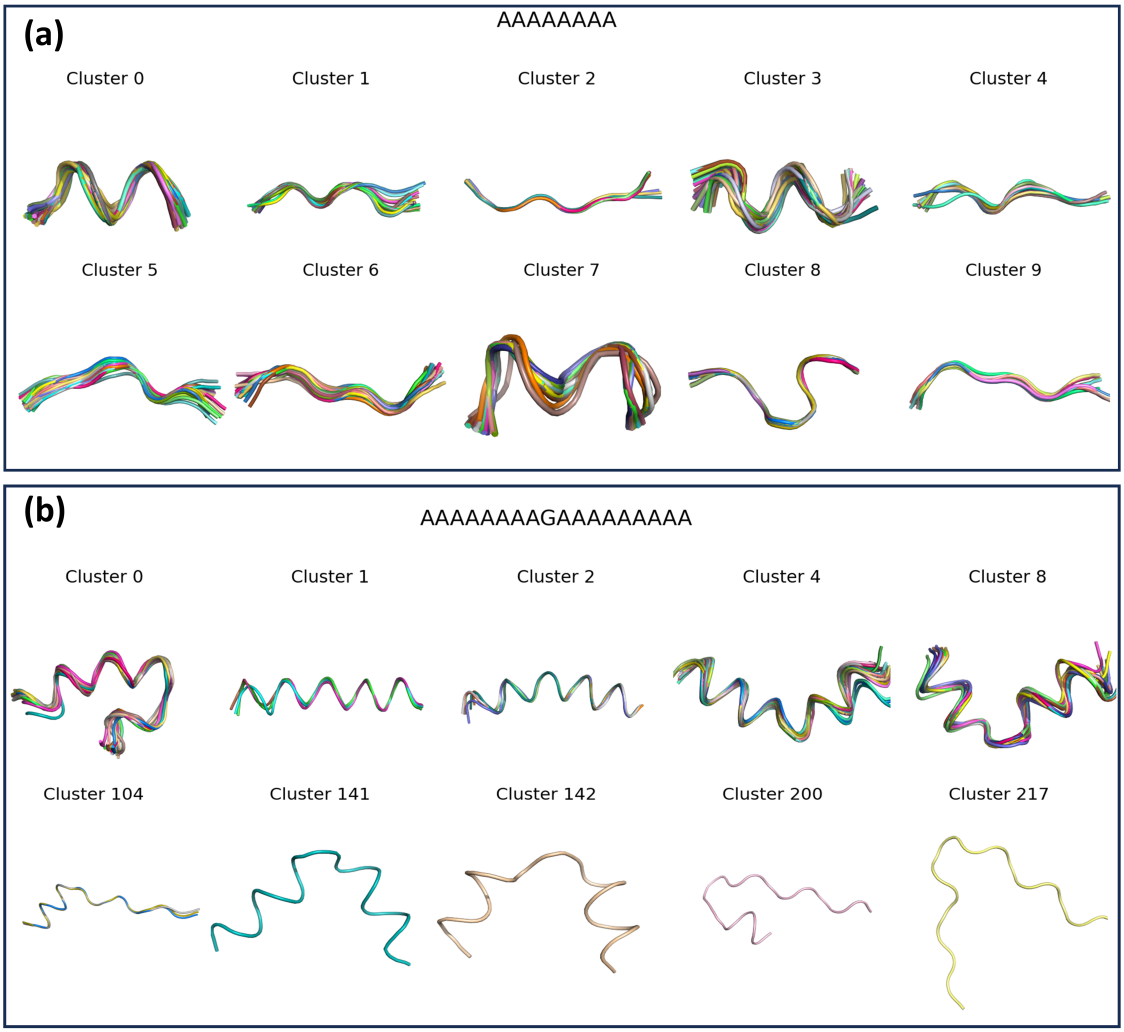


Figure 7.9: Aligned structures showing (a) the first 10 clusters of the octapeptide AAAAAAAA and (b) 10 selected clusters from the 18-residue peptide sequence A_8GA_9

Bibliography

[1] Michael R Yeaman and Nannette Y Yount. Mechanisms of antimicrobial peptide action and resistance. *Pharmacological reviews*, 55(1):27–55, 2003.

[2] Yuki Hirakawa and Shinichiro Sawa. Diverse function of plant peptide hormones in local signaling and development. *Current Opinion in Plant Biology*, 51:81–87, 2019. Cell signalling and gene regulation.

[3] Vasso Apostolopoulos, Joanna Bojarska, Tsun-Thai Chai, Sherif Elnagdy, Krzysztof Kaczmarek, John Matsoukas, Roger New, Keykavous Parang, Octavio Paredes Lopez, Hamideh Parhiz, et al. A global review on short peptides: frontiers and perspectives. *Molecules*, 26(2):430, 2021.

[4] Reinhard Schweitzer-Stenner. The relevance of short peptides for an understanding of unfolded and intrinsically disordered proteins. *Physical Chemistry Chemical Physics*, 25(17):11908–11933, 2023.

[5] Sharmila Anishetty, Gautam Pennathur, and Ramesh Anishetty. Tripeptide analysis of protein structures. *BMC structural biology*, 2(1):9, 2002.

[6] Roman Dallüge, Jan Oschmann, Olaf Birkenmeier, Christian Lücke, Hauke Lilie, Rainer Rudolph, and Christian Lange. A tetrapeptide fragment-based design method results in highly stable artificial proteins. *Proteins: Structure, Function, and Bioinformatics*, 68(4):839–849, 2007.

[7] Lerzan Ormeci, Attila Gursoy, Guzin Tunca, and Burak Erman. Computational basis of knowledge-based conformational probabilities derived from local-and long-

range interactions in proteins. *Proteins: Structure, Function, And Bioinformatics*, 66(1):29–40, 2007.

[8] El-Ad David Amir, Nir Kalisman, and Chen Keasar. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function, and Bioinformatics*, 72(1):62–73, 2008.

[9] Bosco K Ho and Ken A Dill. Folding very short peptides using molecular dynamics. *PLoS computational biology*, 2(4):e27, 2006.

[10] Vincent A Voelz, M Scott Shell, and Ken A Dill. Predicting peptide structures in native proteins from physical simulations of fragments. *PLoS computational biology*, 5(2):e1000281, 2009.

[11] S Gnanakaran, Hugh Nymeyer, John Portman, Kevin Y Sanbonmatsu, and Angel E Garcıa. Peptide folding simulations. *Current opinion in structural biology*, 13(2):168–174, 2003.

[12] Karissa Y Sanbonmatsu and Angel E García. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 46(2):225–234, 2002.

[13] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Peptide Science: Original Research on Biomolecules*, 60(2):96–123, 2001.

[14] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[15] Eli Fritz McDonald, Taylor Jones, Lars Plate, Jens Meiler, and Alican Gulsevin. Benchmarking alphafold2 on peptide structure prediction. *Structure*, 31(1):111–119, 2023.

[16] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[17] Vigneshwaran Kannan, Ramesh Anishetty, and SR Hassan. Optimal transport technique to understand peptide conformations. *Computational Biology and Chemistry*, 98:107684, 2022.

[18] Ulrich Schollwöck. The density-matrix renormalization group. *Reviews of modern physics*, 77(1):259, 2005.

[19] Steven R White. Density matrix formulation for quantum renormalization groups. *Physical review letters*, 69(19):2863, 1992.

[20] Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I Jordan, and Roland L Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS computational biology*, 6(4):e1000763, 2010.

[21] Oliviero Carugo and Kristina Djinović-Carugo. Half a century of ramachandran plots. *Acta Crystallographica Section D: Biological Crystallography*, 69(8):1333–1341, 2013.

[22] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[23] H BIELKA GDR, N Sharon, and EW Australia. Nomenclature and symbolism for amino acids and peptides. *Pure and Applied Chemistry*, 56:595–624, 1984.

[24] Modi Wetzler and Paris Hamilton. 8 - peptides as therapeutics. In Sotirios Koutsopoulos, editor, *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*, pages 215–230. Woodhead Publishing, 2018.

[25] Laszlo Otvos Jr and John D Wade. Current challenges in peptide-based drug discovery, 2014.

[26] Andy Chi-Lung Lee, Janelle Louise Harris, Kum Kum Khanna, and Ji-Hong Hong. A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences*, 20(10):2383, 2019.

[27] Louic S Vermeer, Yun Lan, Vincenzo Abbate, Emrah Ruh, Tam T Bui, Louise J Wilkinson, Tokuwa Kanno, Elmira Jumagulova, Justyna Kozlowska, Jayneil Patel, et al. Conformational flexibility determines selectivity and antibacterial, antiplasmodial, and anticancer potency of cationic α-helical peptides. *Journal of Biological Chemistry*, 287(41):34120–34133, 2012.

[28] Athanassios Stavrakoudis. Conformational studies of the 313-320 and 313-332 peptide fragments derived from the α iib subunit of integrin receptor with molecular dynamics simulations. *International Journal of Peptide Research and Therapeutics*, 15:263–272, 2009.

[29] Danilo Roccatano. A molecular dynamics simulation study of glycine/serine octapeptides labeled with 2, 3-diazabicyclo [2.2. 2] oct-2-ene fluorophore. *The Journal of Chemical Physics*, 160(14), 2024.

[30] Francisco J Blanco, German Rivas, and Luis Serrano. A short linear peptide that folds into a native stable β-hairpin in aqueous solution. *Nature structural biology*, 1(9):584–590, 1994.

[31] Mark S Searle, Dudley H Williams, and Leonard C Packman. A short linear peptide derived from the n-terminal sequence of ubiquitin folds into a water-stable non-native β-hairpin. *Nature Structural Biology*, 2(11):999–1006, 1995.

[32] Carol E Stotz and Elizabeth M Topp. Applications of model β-hairpin peptides. *Journal of pharmaceutical sciences*, 93(12):2881–2894, 2004.

[33] D. Voet, J.G. Voet, and C.W. Pratt. *Fundamentals of Biochemistry: Life at the Molecular Level*. Wiley, 2016.

[34] A.V. Finkelstein, O.B. Ptit□s□yn, and O.B. Ptitsyn. *Protein Physics: A Course of Lectures*. Series in soft condensed matter. Elsevier Science, 2002.

[35] Prasun Kumar, Neil G Paterson, Jonathan Clayden, and Derek N Woolfson. De novo design of discrete, stable 310-helix peptide assemblies. *Nature*, 607(7918):387–392, 2022.

[36] Diego Núñez-Villanueva. Revisiting 310-helices: biological relevance, mimetics and applications. 2024.

[37] Kang Chen, Zhigang Liu, and Neville R Kallenbach. The polyproline ii conformation in short alanine peptides is noncooperative. *Proceedings of the National Academy of Sciences*, 101(43):15352–15357, 2004.

[38] Zhengshuang Shi, Robert W Woody, and Neville R Kallenbach. Is polyproline ii a major backbone conformation in unfolded proteins? *Advances in protein chemistry*, 62:163–240, 2002.

[39] R Brian Dyer, Shelia J Maness, Eric S Peterson, Stefan Franzen, R Matthew Fesinmeyer, and Niels H Andersen. The mechanism of β-hairpin formation. *Biochemistry*, 43(36):11560–11566, 2004.

[40] Marina Ramírez-Alvarado, Tanja Kortemme, Francisco J Blanco, and Luis Serrano. β-hairpin and β-sheet formation in designed linear peptides. *Bioorganic & medicinal chemistry*, 7(1):93–103, 1999.

[41] Maciej Ciemny, Mateusz Kurcinski, Karol Kamel, Andrzej Kolinski, Nawsad Alam, Ora Schueler-Furman, and Sebastian Kmiecik. Protein–peptide docking: opportunities and challenges. *Drug discovery today*, 23(8):1530–1537, 2018.

[42] Karim M ElSawy. Energy landscape of pentapeptides in a higher-order (ϕ, ψ) conformational subspace. *Advances in Physical Chemistry*, 2016(1):3240674, 2016.

[43] David A Evans and David J Wales. Free energy landscapes of model peptides and proteins. *The Journal of chemical physics*, 118(8):3891–3897, 2003.

[44] Debayan Chakraborty, Yassmine Chebaro, and David J Wales. A multifunnel energy landscape encodes the competing α-helix and β-hairpin conformations for a designed peptide. *Physical Chemistry Chemical Physics*, 22(3):1359–1370, 2020.

[45] GN T Ramachandran and V Sasisekharan. Conformation of polypeptides and proteins. *Advances in protein chemistry*, 23:283–437, 1968.

[46] H Jane Dyson and Peter E Wright. Equilibrium nmr studies of unfolded and partially folded proteins. *nature structural biology*, 5(7):499–503, 1998.

[47] Ron Unger, David Harel, Scot Wherland, and Joel L Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins: Structure, Function, and Bioinformatics*, 5(4):355–373, 1989.

[48] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[49] Gregory B Quinn, Chunxiao Bi, Cole H Christie, Kyle Pang, Andreas Prlić, Takanori Nakane, Christine Zardecki, Maria Voigt, Helen M Berman, Philip E Bourne, et al. Rcsb pdb mobile: ios and android mobile apps to provide data access and visualization to the rcsb protein data bank. *Bioinformatics*, 31(1):126–127, 2015.

[50] Annick Thomas, Sébastien Deshayes, Marc Decaffmeyer, Marie Hélène Van Eyck, Benoit Charloteaux, and Robert Brasseur. Prediction of peptide structure: how far are we? *Proteins: Structure, Function, and Bioinformatics*, 65(4):889–897, 2006.

[51] Vibin Ramakrishnan, Kirti Patel, and Ruchika Goyal. *De Novo Peptide Design: Principles and Applications*. Academic Press, 2022.

[52] Handan Arkin and Tarik Çelİk. Structure of energy landscape of short peptides. *International Journal of Modern Physics C*, 14(01):113–120, 2003.

[53] Junichi Higo, Nobutoshi Ito, Masataka Kuroda, Satoshi Ono, Nobuyuki Nakajima, and Haruki Nakamura. Energy landscape of a peptide consisting of α-helix, 310-helix, β-turn, β-hairpin, and other disordered conformations. *Protein Science*, 10(6):1160–1171, 2001.

[54] Mahmoud Moradi, Volodymyr Babin, Christopher Roland, Thomas A Darden, and Celeste Sagui. Conformations and free energy landscapes of polyproline peptides. *Proceedings of the National Academy of Sciences*, 106(49):20746–20751, 2009.

[55] Rudi Podgornik. Energy landscapes: Applications to clusters, biomolecules and glasses (cambridge molecular science), 2007.

[56] E. Clementi and S. Chin. *Structure and Dynamics of Nucleic Acids, Proteins, and Membranes*. Springer US, 2012.

[57] Z Li and H A Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[58] Harold A. Scheraga. The multiple-minima problem in protein folding. *AIP Conference Proceedings*, 239(1):97–107, 10 1991.

[59] Cyrus Levinthal. Are there pathways for protein folding? *Journal de chimie physique*, 65:44–45, 1968.

[60] Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[61] Ken A Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nature structural biology*, 4(1):10–19, 1997.

[62] Robert Zwanzig. Simple model of protein folding kinetics. *Proceedings of the National Academy of Sciences*, 92(21):9801–9804, 1995.

[63] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of sciences*, 84(21):7524–7528, 1987.

[64] Leandro Martínez. Introducing the levinthal's protein folding paradox and its solution. *Journal of Chemical Education*, 91(11):1918–1923, 2014.

[65] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.

[66] Hao Geng, Fangfang Chen, Jing Ye, and Fan Jiang. Applications of molecular dynamics simulation in structure prediction of peptides and proteins. *Computational and structural biotechnology journal*, 17:1162–1170, 2019.

[67] Vivek P Raut, Madhuri A Agashe, Steven J Stuart, and Robert A Latour. Molecular dynamics simulations of peptide- surface interactions. *Langmuir*, 21(4):1629–1639, 2005.

[68] L Ramya and N Gautham. Conformational space exploration of met-and leu-enkephalin using the mols method, molecular dynamics, and monte carlo simulation—a comparative study. *Biopolymers*, 97(3):165–176, 2012.

[69] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science. Academic Press, 2001.

[70] Yuji Sugita and Yuko Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chemical Physics Letters*, 329(3-4):261–270, 2000.

[71] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. i. formulation and benchmark test. *The Journal of chemical physics*, 118(14):6664–6675, 2003.

[72] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. ii. application to a more complex system. *The Journal of chemical physics*, 118(14):6676–6688, 2003.

[73] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[74] Nana Heilmann, Moritz Wolf, Mariana Kozlowska, Elaheh Sedghamiz, Julia Setzler, Martin Brieg, and Wolfgang Wenzel. Sampling of the conformational landscape of small proteins with monte carlo methods. *Scientific reports*, 10(1):18211, 2020.

[75] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex

Bridgland, et al. Alphafold 2. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 2020.

[76] Letícia MF Bertoline, Angélica N Lima, Jose E Krieger, and Samantha K Teixeira. Before and after alphafold2: An overview of protein structure prediction. *Frontiers in bioinformatics*, 3:1120370, 2023.

[77] Lei Wang, Zehua Wen, Shi-Wei Liu, Lihong Zhang, Cierra Finley, Ho-Jin Lee, and Hua-Jun Shawn Fan. Overview of alphafold2 and breakthroughs in overcoming its limitations. *Computers in Biology and Medicine*, page 108620, 2024.

[78] Anastassis Perrakis and Titia K Sixma. Ai revolutions in biology: The joys and perils of alphafold. *EMBO reports*, 22(11):e54046, 2021.

[79] Abhishek K Jha, Andres Colubri, Muhammad H Zaman, Shohei Koide, Tobin R Sosnick, and Karl F Freed. Helix, sheet, and polyproline ii frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28):9691–9702, 2005.

[80] Christopher Bystroff and Shekhar Garde. Helix propensities of short peptides: molecular dynamics versus bioinformatics. *Proteins: Structure, Function, and Bioinformatics*, 50(4):552–562, 2003.

[81] Robert I Cukier. Generating intrinsically disordered protein conformational ensembles from a database of ramachandran space pair residue probabilities using a markov chain. *The Journal of Physical Chemistry B*, 122(39):9087–9101, 2018.

[82] Thomas Hamelryck, Kanti Mardia, and Jesper Ferkinghoff-Borg. *Bayesian methods in structural bioinformatics*. Springer, 2012.

[83] Wolfgang Kabsch and Christian Sander. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proceedings of the National Academy of Sciences*, 81(4):1075–1078, 1984.

[84] Diego Caballero, Jukka Määttä, Alice Qinhua Zhou, Maria Sammalkorpi, Corey S O'Hern, and Lynne Regan. Intrinsic α-helical and β-sheet conformational preferences: a computational case study of alanine. *Protein Science*, 23(7):970–980, 2014.

[85] Silvia Pizzanelli, Claudia Forte, Susanna Monti, Giorgia Zandomeneghi, Andrew Hagarman, Thomas J Measey, and Reinhard Schweitzer-Stenner. Conformations of phenylalanine in the tripeptides afa and gfg probed by combining md simulations with nmr, ftir, polarized raman, and vcd spectroscopy. *The Journal of Physical Chemistry B*, 114(11):3965–3978, 2010.

[86] Brendan Pass. Multi-marginal optimal transport and multi-agent matching problems: uniqueness and structure of solutions. *arXiv preprint arXiv:1210.7372*, 2012.

[87] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

[88] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. A numerical method to solve multi-marginal optimal transport problems with coulomb cost. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 577–601. Springer, 2016.

[89] Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic theory*, 42(2):397–418, 2010.

[90] Patrice Koehl, Marc Delarue, and Henri Orland. Statistical physics approach to the optimal transport problem. *Physical review letters*, 123(4):040603, 2019.

[91] J.J. Craig. *Introduction To Robotics: Mechanics And Control, 3/E.* Pearson Education, 2009.

[92] G. Monge. *Mémoire sur la théorie des déblais et des remblais.* Imprimerie royale, 1781.

[93] L. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

[94] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018.

[95] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Now Foundations and Trends, 2019.

[96] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[97] François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

[98] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 13–18 Jul 2020.

[99] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 29, 2016.

[100] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[101] Giuseppe Buttazzo, G Carlier, et al. Optimal spatial pricing strategies with transportation costs. *Contemp. Math*, 514:105–121, 2010.

[102] Giuseppe Buttazzo, Aldo Pratelli, and Eugene Stepanov. Optimal pricing policies for public transportation networks. *Siam Journal on Optimization*, 16(3):826–853, 2006.

[103] C. Villani and American Mathematical Society. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[104] Anders S Christensen, Thomas Hamelryck, and Jan H Jensen. Fragbuilder: an efficient python library to setup quantum chemistry calculations on peptides models. *PeerJ*, 2:e277, 2014.

[105] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.