Statistical methods for the structural analysis of highly flexible proteins

Javier González-Delgado

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *11/10/2023* **par :**

Javier GONZÁLEZ DELGADO

Statistical methods for the structural analysis of highly flexible proteins

JURY

Yohann DE CASTRO	Professeur des Universités	Rapporteur
Elodie LAINE	Maîtresse de Conférences	Rapporteure
Philippe BERTHET	Professeur des Universités	Examinateur
Nathalie SIBILLE	Directrice de Recherche	Examinatrice
Pau BERNADÓ	Directeur de Recherche	Membre invité
Juan CORTÉS	Directeur de Recherche	Directeur de thèse
Pierre NEUVIAL	Directeur de Recherche	Directeur de thèse

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut de Mathématiques de Toulouse, LAAS-CNRS
Directeur(s) de Thèse :
Juan Cortés et *Pierre Neuvial*
Rapporteurs :
Yohann De Castro et *Elodie Laine*

Remerciements

Tout d'abord, je tiens à remercier les rapporteurs Yohann de Castro et Elodie Laine pour avoir consacré leur temps et leurs efforts à la lecture et à l'évaluation de cette thèse. Je tiens également à exprimer ma gratitude envers les examinateurs Nathalie Sibille et Philippe Berthet pour avoir accepté de faire partie du jury. Un remerciement spécial à Pau Bernadó, qui dans la pratique a été mon troisième directeur de thèse. Merci beaucoup pour ton intérêt, tes efforts et ton implication remarquable.

Je tiens à remercier les personnes qui ont collaboré avec moi sur les projets qui composent cette thèse : Kresten Lindorff-Larsen, Pablo Mier, Amin Sagar et Christophe Zanon. J'exprime ma gratitude particulière à Alberto González-Sanz, mon voisin de bureau pendant ces trois années, avec qui j'ai également eu le plaisir de travailler. Je salue tous les autres chercheurs m'ayant fait découvrir la communauté des mathématiques et de la biologie structurale sous ses plus belles couleurs.

J'exprime ma gratitude envers toutes les personnes de l'Institut de Mathématiques de Toulouse qui m'ont accompagné pendant ces trois années, en particulier mes collègues et amis doctorants. Avoir un environnement de confiance, de soutien et de sécurité est essentiel pendant une thèse, et c'est quelque chose que j'ai eu la chance d'avoir dès le premier jour. Je vous remercie pour votre affection et votre accueil et je vous souhaite à toutes et à tous le meilleur pour l'avenir.

Je tiens également à remercier les amis que j'ai eu le plaisir de rencontrer lors de mon expérience à Toulouse, ainsi que ceux qui étaient déjà présents. Ces années ont été exceptionnelles grâce à vous. Quiero dar las gracias especialmente a mis padres, por vuestro apoyo y cariño, por estar siempre.

Par dessus tout, je remercie les deux personnes qui m'ont rendu très heureux en travaillant sur ce projet pendant trois ans: mes directeurs de thèse Pierre Neuvial et Juan Cortés. En plus de vos qualités scientifiques évidentes, votre exceptionnelle qualité humaine a fait de cette expérience un véritable privilège. Pour m'avoir fait profiter du chemin, je vous remercie.

"Without leaps of imagination or dreaming, we lose the excitement of possibilities. Dreaming, after all is a form of planning."

Gloria Steinem

Contents

Résumé

La reconnaissance de la pertinence fonctionnelle des protéines désordonnées a entraîné un changement de paradigme en biologie structurale. Avec les progrès des méthodes de simulation et des modèles génératifs, la communauté scientifique a désormais accès à des ensembles conformationnels à résolution atomique d'un grand nombre de systèmes. Cependant, l'analyse structurale de ces objets ne peut pas être réalisée en utilisant les mêmes techniques que celles employées dans l'étude des protéines rigides ou globulaires. Leur nature intrinsèquement probabiliste exige l'adoption d'une perspective plaçant la statistique comme un prisme fondamental pour comprendre la relation séquence-structure. Dans cette thèse, nous présentons des outils statistiques pour la caractérisation et la comparaison, à la fois à l'échelle locale et globale, d'ensembles de protéines hautement flexibles. La stratégie générale consiste à définir des distributions de probabilité qui capturent avec précision la variabilité structurale des ensembles, puis à utiliser des techniques statistiques avancées pour caractériser et comparer de manière appropriée ces descripteurs. Dans certains cas, l'absence d'outils bien adaptés au problème nous amènera à définir de nouvelles méthodes statistiques qui seront utiles d'un point de vue plus général. La première partie de la thèse se concentre sur l'analyse structurale locale. Dans le chapitre 2, nous démontrons l'interdépendance des influences des acides aminés voisins sur la structure protéique locale. Ensuite, dans le chapitre 3, nous utilisons la théorie du Transport Optimal pour définir des tests d'homogénéité à deux échantillons pour des mesures sur le tore plat bidimensionnel, où sont supportées les distributions de probabilité décrivant la structure locale des protéines. Ces outils sont appliqués dans le chapitre 4 pour évaluer l'effet du codon traduit sur la conformation locale. La deuxième partie du manuscrit aborde l'analyse structurale globale. Dans le chapitre 5, nous présentons WASCO, un outil pour comparer des ensembles de protéines désordonnées basé sur la distance de Wasserstein. Dans le chapitre 6, nous fournissons des garanties statistiques pour des méthodes classiques de clustering conformationnel couramment utilisées pour caractériser des ensembles. Plus précisément, nous étendons la théorie de l'inférence après clustering lorsque les observations et les variables présentent des structures de dépendance arbitraires. Enfin, nous concluons en introduisant WARIO dans le chapitre 7, une méthode de caractérisation des ensembles qui généralise les cartes de contact au cadre des protéines flexibles, en incorporant des techniques de clustering avancées qui dévoilent la variabilité des interactions résidu-résidu. Les méthodes présentées dans cette thèse ont été rendues disponibles à la communauté sous forme de logiciel open-source, assurant également la reproductibilité des résultats présentés.

Mots-clés: Protéines intrinsèquement désordonnées, Statistique, Bioinformatique.

Abstract

The recognition of the functional relevance of disordered proteins has brought about a paradigm shift in Structural Biology. With the advancement of simulation methods and generative models, the scientific community now has access to atomic-resolution conformational ensembles of a large number of systems. However, the structural analysis of these objects cannot be carried out using the same techniques employed in the study of rigid/globular proteins. Their intrinsically probabilistic nature demands the move to a perspective that places statistics as a fundamental prism for understanding the sequence-structure relationship. In this thesis, we present statistical tools for the characterization and comparison, at both local and global scales, of ensembles of highly flexible proteins. The general strategy consists of defining probability distributions that accurately capture the structural variability of the ensembles, and then employing advanced statistical techniques to appropriately characterize and compare these descriptors. In some cases, the absence of tools well-adapted to the problem will lead us to the definition of new statistical methods that will be useful from a more general standpoint. The first part of the thesis focuses on the local structural analysis. In Chapter 2, we demonstrate the interdependence of the influences of neighboring amino acids on the local protein structure. Then, in Chapter 3, we use Optimal Transport theory to define two-sample goodness-of-fit tests for measures on the two-dimensional flat torus, where the probability distributions that describe the protein local structure are supported. These tools are applied in Chapter 4 to assess the effect of the translated codon on the local backbone conformation. The second part of the manuscript addresses global structural analysis. In Chapter 5, we present WASCO, a tool for comparing ensembles of disordered proteins based on the Wasserstein distance. In Chapter 6, we provide statistical guarantees for classical conformational clustering methods commonly used to characterize ensembles. More precisely, we extend the theory of post-clustering inference when observations and variables exhibit arbitrary dependency structures. Finally, we conclude by introducing WARIO in Chapter 7, an ensemble characterization method that generalizes contact maps to the framework of flexible proteins, incorporating advanced clustering techniques that unravel the variability of residue-residue interactions. The methods presented in this thesis are developed with mathematical rigor and aim to provide statistical guarantees whenever possible. Their implementation has been made available to the community through open-source software, also ensuring the reproducibility of the presented results.

Keywords: Intrinsically disordered proteins, Statistics, Bioinformatics.

Chapter 1

Introduction

1.1 Intrinsic disorder in proteins

Proteins are essential molecules in all living organisms. They play a central role in the majority of biological processes, operating at the molecular, cellular, and organismal levels. The term *protein* was first introduced by the Swedish chemist Jöns Jacob Berzelius in a letter to the Dutch chemist Gerardus Johannes Mulder in 1838 [118]:

"Or je présume que l'oxyde organique, qui est la base de la fibrine et de l'albumine (et auquel il faut donner un nom particulier p. ex. protéine) est composé d'un radical ternaire, combiné avec de l'oxygène dans quelqu'un de ses rapports simples que la nature inorganique nous présente."

This letter marked the beginning of a long journey that Structural Biology embarked on to understand the structure of these macromolecules and connect them to their crucial functions at the higher levels of the living world. Of course, this trip went hand in hand with the technological advancements that enabled the experimental determination of protein structure. Following the early X-ray crystallographic techniques, cryo-electron microscopy (cryo-EM) and Nuclear Magnetic Resonance (NMR) emerged as a major breakthrough for single-particle reconstruction, solving the three-dimensional structure of a macromolecule at atomic scale [175]. These advancements have continuously pushed the boundaries of achievable resolution and enabled the observation of structures with increasing size and complexity. The more and better we can observe, the richer the approaches and perspectives that allow deciphering what we see. Thanks to Structural Biology, we are able to make objects at the subatomic scale visible and embrace the "seeing is believing". However, *understanding* what we see needs the involvement of a diverse family of areas of knowledge, in which, with the recent recognition of the importance of disorder, Mathematics must take part.

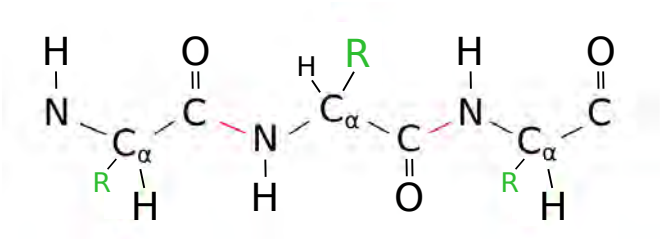


Figure 1.1: Simplified representation of a polypeptide. Peptide bonds, backbone atoms and side chains are marked in red, black and green respectively.

1.1.1 Protein structure and function

A protein is a macromolecule constituted of amino acid residues linked by peptide bonds. This type of polymer molecule is also called a polypeptide. An amino acid is a molecule composed of a carbon atom (α-carbon) attached to a carboxyl group (-COOH), an amine group (-NH_2), one hydrogen and a variable side chain, also called radical. A peptide bond is a double bond between the carbon of the carboxyl group of one residue and the nitrogen of the following amine group. A simplified representation of a polypeptide is presented in Figure 1.1. Note that the formation of peptide bonds allows to distinguish two main parts in the protein. On one hand, the sequence of nitrogen, α-carbon, hydrogen, carbon and oxygen atoms that is referred to as *backbone*, depicted in black in Figure 1.1. One the other hand, the *side chains*, that is, the family of different radicals bonded to each α-carbon, illustrated in green in Figure 1.1. The side chains determine the physico-chemical properties of amino acids and constitute the fingerprint of the protein.

The sequence of amino acid residues is called the *primary structure* (Figure 1.2a). For simplicity, we will also refer to primary structure as *sequence*. While there are approximately 500 naturally occurring amino acids known, only 20 of them are found in proteins. This already hints at the complexity of the world we are delving into, since up to 20^L proteins with a sequence length L are conceivable. For proteins with 100 amino acids, this implies envisioning up to 10^{130} possible sequences in a universe containing 10^{82} atoms. The impossibility of knowing all proteins highlights the need for intelligent strategies to understand their behavior based on the available information. Structural Biology seeks to attain so by deciphering the mechanisms that govern the transformation of the primary structure into the protein three-dimensional form, that we will generally refer to as *structure*. This process is known as *folding*. During the folding process, some parts of the sequence adopt relatively stable and well-defined *secondary structure* elements, being α-helices, β-sheets the most representative ones Figure 1.2b). The spatial arrangement of these elements, which are connected by turns and coil region, forms the *tertiary structure* (Figure 1.2c). For a in-depth introduction to protein structure, we refer to [175, 271, 1, 56].

Proteins perform numerous functions that are closely related to their structural and dynamic properties [175]. For instance, enzymes catalyze various types of chemical re-

(a) (b) (c)

Figure 1.2: Primary (a), secondary (b) and tertiary (c) structure of the protein. In (b,c), α-helices are depicted in red and β-sheets are marked in yellow.

actions. Other proteins function as nutrient and storage proteins, crucial for the growth and survival of seeds in many plants. Others enable cell contraction, bind and transport substances, act as structural proteins to give cells a defined shape, and regulate various cellular processes. All these functions are often directly dependent on the folded structure of the protein, also known as *native state*. The native state is not a fixed conformation but rather a collection of accessible states that the protein can adopt, depending on factors such as solvent conditions and temperature. The energies of the native state of proteins that fold into a well-defined 3D structure present stable global minima. However, many proteins do not fit this description, presenting relatively flat energy landscapes with multiple local minima. These proteins, known as *Intrinsically Disordered Proteins* (IDP), are in a constant shape-changing and transitioning between different states [139]. The collection of all these conformations is referred to as protein *ensemble*. The absence of an equilibrium state requires the classical techniques traditionally used to study the structure-function relationship to be readapted, opening up to new paradigms that allow for the understanding of the functional richness conferred by their structural variability.

1.1.2 Intrinsically disordered proteins: falling of the structure-function paradigm

Until the end of the 20th century, the vast majority of the scientific community supported the so-called *structure-function paradigm*: a functional protein requires a stable and well-defined structure. In addition, protein-protein interactions depend on the precise complementation of surfaces. Classic models such as the "lock-key" proposed by the Nobel laureate Emil Fischer in 1894 can be found within this framework [95]. Stating that unstructured proteins are denatured makes intrinsically disordered proteins challenge that paradigm [282]. Indeed, although they are devoid of a stable secondary and tertiary structure in isolation, IDP perform a large diversity of biological functions by exploiting their intrinsic flexibility [293]. Moreover, IDP can malfunction under certain circumstances, such as mutations or inconvenient environmental conditions. This phenomenon may induce severe diseases including cancer, cardiovascular or neurodegenerative diseases [288]. All of this justifies the functional relevance of disordered proteins and the need to readapt the structure-function paradigm to incorporate structural variability.

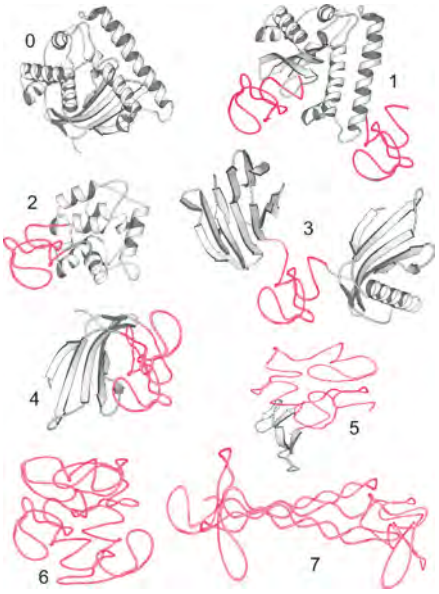


Figure 1.3: Figure 1 in [290]. Different levels of order (gray) and disorder (red): (0) no disorder, (1) disordered termini, (2) disordered linker, (3) disordered loop, (4) disordered domain, (5) disordered protein with some residual structure, (6) wholly disordered, mostly collapsed protein and (7) wholly disordered, extended protein.

During the past 20 years, Structural Biology has been incorporating disorder into the study of proteins. This has led to a shift from the rigid vs. unstructured dichotomy to considering disorder as a continuum. Indeed, most proteins are neither fully ordered nor fully disordered but contain ordered and disordered regions at different ratios [290]. This is illustrated in Figure 1.3, extracted from [290]. The passage to continuum also impacted the study of energy landscapes, that now consider weakly funneled profiles as a transition between deep energy minima and the rugged landscapes of very disordered systems [221]. As we can see, disorder manifests with varying intensity and sequence placement. Furthermore, the failure to fold is encoded by the primary structure, as IDP exhibit unique sequential properties. Some of them are the compositional bias [289], manifesting through residues with low hydrophobicity (for low compaction) and high net charge (for high charge-charge repulsion), low content of predicted secondary structure [180] or high sequence variability (low conservation) [113]. This highlights how the folding process is strongly related to the protein sequence. Deciphering this relationship is essential for the development of structure prediction methods from protein sequence, which have experienced a boom with the arrival of deep learning. This is the case of the well-known AlphaFold algorithm [147], that can predict protein structure with atomic-level accuracy. The AlphaFold Protein Structure Database [296] contains a predicted structure for nearly every protein in the proteome of numerous organisms that have been totally or partially sequenced, including proteins with intrinsically disordered regions (IDRs). However, IDRs

present low values of the AlphaFold confidence metric, called predicted local distance difference test (pLDDT), which means low confidence in the structural predictions, and thus inaccurate descriptions of these (*a priori*) disordered regions. Therefore, the study of disordered proteins requires alternative approaches that intelligently combine experimental, theoretical and computational methods. An overview of the state-of-the-art is presented in the following section.

1.1.3 Existing approaches to model IDP

Arguably, the most relevant feature that sets apart the study of ordered proteins and IDP is the access to experimental data. There is a stark contrast in the quantity of experimentally known structures between both worlds. The Protein Data Bank (PDB) [20] is a freely accessible data base containing more than 200000 experimental structures of folded proteins. Its counterpart for disordered systems is the Protein Ensemble Database [168], an open access repository that includes IDP data, but contains 280 entries so far. In this context, experimental data cannot provide accurate information of each of the individual conformations in the ensemble, but only average measurements. This makes experimental IDP data useful as a *restraint to simulate*. Indeed, the study of disordered proteins as conformational ensembles is largely governed by simulation and modeling techniques often calibrated with experimental data. In the following section, we present a brief overview of the two major families that integrate the ensemble generation methods. As these techniques are not the focus of this thesis, the outline presents only some of the most relevant contributions within a broad and diverse field of study. For a more extensive picture of the existing literature, we refer to the reviews [58, 26, 252, 28, 264, 156].

Ensemble Generation Methods

The first category of computational methods aims at generating representative sets of conformations through an effective exploration of the conformational space. The term "effective" arises from the computational infeasibility of randomly inspecting the complete state space. In fact, an efficient exploration incorporates information derived from experimentally determined structures, optimizing the computational procedure. The most distinctive knowledge-based method is Flexible-Meccano (FM) [219], that builds each conformation by sequentially assembling peptide plane units using a residue-specific coil library obtained from crystallographic structures. Together with Flexible-Meccano, TraDES [93] is another popular stochastic sampling technique. The conformations produced by these methods are validated through their adjustment to experimental data, using computational tools as ENSEMBLE [162], ASTEROIDS [214] or EOM [22, 285]. These techniques make use of NMR measurable parameters or small-angle scattering of X-rays (SAXS) data. Although lower in resolution, SAXS is capable of retrieving overall structural and dynamic information about biological macromolecules, including those that cannot crystallize, like IDP [94, 136, 160, 234, 273]. However, approaches like Flexible-Meccano fail at capturing

secondary structure elements that involve multiple consecutive residues in IDP [21, 142]. This limitation was overcome in [88] by refining the experimental calibration using an extensive coil library of three-residue fragments.

The second big family of methods uses physical models to sample the conformational space, simulating the dynamic behavior of IDP. The preeminent technique within this context are Molecular Dynamic simulations (MD), that solve Newton's equations of motion to recreate the time evolution of the system [155, 231]. Although being capable of suitably representing the state space of IDP, MD present a major drawback that lies in their excessive computational cost when applied to large molecules. Indeed, the substantial radius of gyration exhibited by IDP in comparison to folded proteins, as well as their inherent fluctuations, make considerably increase the size of the simulation box containing the protein and water molecules. A solution to deal with this type of systems is the use of coarse-grained models, that provide a more simplistic representation of the protein but allow a wider investigation of the state space [158, 68, 159]. Besides, the accuracy of MD-based techniques is strongly dependent on the force-fields and solvation models that are employed, whose determination for flexible proteins is a very active area of research [140, 301]. The performance of MD methods can also be reinforced by integrating experimental data to narrow down the exploration of the conformational space [71, 179, 309]. Remarkable hybrid approaches have also been proposed, performing MD simulations with Machine-Learning-derived potentials, such as CALVADOS [277, 276]. A physical-based alternative to MD are Monte Carlo methods (MC), among which we might stand out the Markov chain Metropolis scheme [202], its variant Hamiltonian Switch Metropolis Monte Carlo [206] adapted to the study of IDR, or ABSINTH [301], an intermediate MC approach between coarse-grained and all-atom models.

Ensemble Characterization and Comparison Methods

The methods presented below aim at producing ensemble representations of disordered proteins. Indeed, the vast majority of methodological contributions in the study of flexible proteins focus on compensating for the lack of experimental data by simulating conformational ensembles. Here, a natural reflection arises: once we are capable of generating IDP ensembles with indeterminate size, what is next? What do we do with all this data? How do we transform the output of generative models into concise and interpretable representations that allow to understand the sequence-structure relationship in IDP? More succinctly: how can we *characterize* and *compare* ensembles of highly flexible proteins? To these questions, the only possible answer is to resort to techniques conceived for handling the variability of inherently disordered objects: Probability and Statistics. Random systems must be described as probabilistic objects, and samples drawn from such systems must be analyzed using statistical techniques. This idea is schematized in Figure 1.4, that situates the contribution of this thesis (in purple) with respect to the state-of-the-art previously described (in blue).

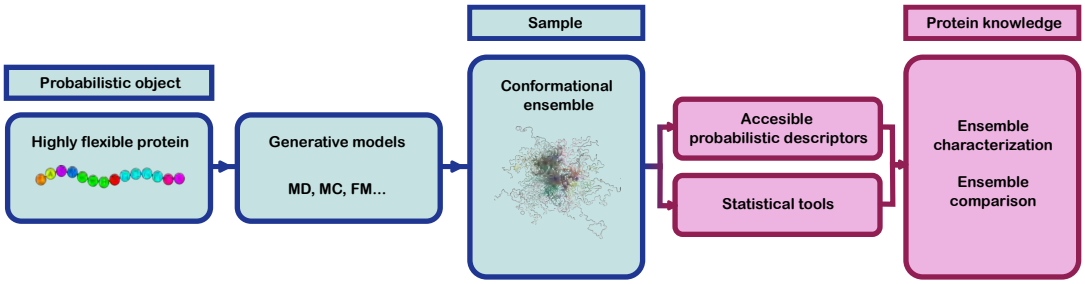


Figure 1.4: Conformational ensembles conceived as samples of IDP states produced by knowledge-based and physic-based generative models. This data is featured by probabilistic descriptors and analyzed with statistical tools, aiming at accurately characterizing and comparing the retrieved ensembles.

The main objective of this thesis is to provide the natural subsequent step to IDP modeling techniques, making the outputs of generative models interpretable with statistical guarantees. To achieve so, it is essential to define methods that allow for a compact and interpretable *characterization* and *comparison* of IDP ensembles. The applications that motivate these objectives are numerous and diverse. Among them we might highlight the posterior analysis of how generative models perform, including their comparison with experimental data, the relative comparison of force-fields and solvation models or the assessment of the effect of experimental restrains. Some other remarkable applications are the evaluation of the effect of sequence mutations, that may lead to the incorporation of IDP to the field of protein design, or the definition of loss functions and compact descriptors required for the development of Machine Learning algorithms. This last point would be determinant for the extension of structure prediction methods to sequences with IDR.

The development of methods for the characterization and comparison of disordered ensembles is gaining increasing relevance in Structural Biology, with numerous remarkable contributions in recent years [167, 7, 59, 57]. These studies, that will be discussed in more detail in the chapters composing this manuscript, provide interesting and innovative contributions. However, we believe that they have not yet fully integrated the probabilistic nature of flexible proteins in a productive manner. Here, we propose to tackle our objective by placing structural variability at the core and conceiving IDP as inherently probabilistic objects that need to be analyzed using the most suitable statistical techniques. We detail this strategy in the following section.

1.2 The inherent probabilistic nature of flexible proteins

The strategy that we present to capture the intrinsic probabilistic nature of flexible proteins consists on (*i*) defining structural descriptors through probability distributions supported on well-suited spaces and (*ii*) characterizing and comparing such distributions with appropriate statistical techniques, providing statistical guarantees about the behavior of

the population when possible. We propose to perform such a strategy both at the local (amino acid scale) and global (entire sequence) levels. Local and global structural descriptors are defined in Section 1.2.2. Then, in Section 1.2.3, we introduce the main statistical tools that will be used to characterize and compare them. First, we recall some essential concepts from probability theory and set the notation that will be used throughout the manuscript. This is presented in the following section, which may be skipped by readers less interested in mathematical aspects.

1.2.1 Background and notation

This section gathers key notation and definitions that will be assumed throughout the manuscript. Further specific notation will be presented within each chapter. We start by defining probability spaces, that is, measure spaces where the measure of the whole space equals to one [29].

Definition 1.2.1 (Probability space)**.** *Let Ω be a non-empty set and Σ a σ-algebra, i.e. a set of subsets of Ω such that*

(i) $\Omega \in \Sigma$,

(ii) *Every countable union of elements of Σ is also in Σ,*

(iii) *The complement of every element of Σ is in Σ.*

If $\mathbb{P} : \Sigma \longrightarrow [0,1]$ is such that $\mathbb{P}(\Omega) = 1$ and countably additive, that is, $\mathbb{P}(\cup_{i\in\mathbb{N}} E_i) = \sum_{i\in\mathbb{N}} \mathbb{P}(E_i)$ for every countable collection of pairwise disjoint sets $\{E_i\}_{i\in\mathbb{N}} \subset \Sigma$, then $\mathbb{P}$ is called a probability measure on Σ and the triplet $(\Omega, \Sigma, \mathbb{P})$ is called a probability space.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, $\mathcal{E}$ a topological space [210] and $\mathcal{T}$ the σ-algebra generated by the topology of $\mathcal{E}$ [29]. Let $f : \Omega \longrightarrow \mathcal{E}$ be a measurable function, i.e. such that $f^{-1}(O) \in \Sigma$ for all $O \in \mathcal{T}$. The *push-forward measure* of $\mathbb{P}$ by f is the mapping $f_{\#}\mathbb{P} : \mathcal{T} \longrightarrow [0,1]$ such that $f_{\#}\mathbb{P}(O) = (\mathbb{P} \circ f^{-1})(O)$ for all $O \in \mathcal{T}$. This transformation is the key to define the probability distribution of a random variable.

Definition 1.2.2 (Random variable, distribution)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\mathcal{E}$ a topological space. A random variable is a measurable function $X : \Omega \longrightarrow \mathcal{E}$. The push-forward measure of $\mathbb{P}$ by X, denoted as $P_X := X_{\#}\mathbb{P}$, is called the (probability) distribution of X, or the law of X.*

For any random variable X defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in a topological space $\mathcal{E}$, we define the *support* of its distribution as the closed set

$$\text{supp}(P_X) = \{x \in \mathcal{E} \,:\, P_X(U_x) > 0 \text{ for all } U_x \text{ neighborhood of x}\}.$$

We will denote as $\mathcal{P}(\mathcal{E})$ the set of all probability distributions supported on $\mathcal{E}$, that is, whose support is a subset of $\mathcal{E}$. Note that the term *distribution* refers to a random

variable and the term *measure* operates directly on a probability space. However, when Ω is a topological space, we can take $\Omega = \mathcal{E}$, $\Sigma = \mathcal{T}$, X the identity mapping and speak directly of $\mathbb{P}$ as a *distribution*. As all the spaces considered here will be provided with a topology, we will use the terms *distribution* or *measure* interchangeably throughout the manuscript, omitting the push-forward of the random variable when the context is clear. Thus, we will also refer to $\mathcal{P}(\mathcal{E})$ as the set of probability *measures* supported on $\mathcal{E}$.

The concept of random variable can be extended to the case where its image space is a Cartesian product of topological spaces equipped with the product topology [210]. This construction can be considered from the combination of two random variables defined on the same probability space.

Definition 1.2.3 (Joint distribution, marginals)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, $\mathcal{E}$ a topological space and $X, Y : \Omega \longrightarrow \mathcal{E}$ two random variables. The joint distribution of X and Y is the probability distribution of the random variable*

$$\begin{aligned}(X, Y)\colon \Omega &\longrightarrow \mathcal{E} \times \mathcal{E} \\ \omega &\longmapsto (X(\omega), Y(\omega)),\end{aligned}$$

that is, the measure $P_{XY} = (X, Y)_{\#}\mathbb{P} \in \mathcal{P}(\mathcal{E} \times \mathcal{E})$. The measures P_X and P_Y are called the marginal distributions of P_{XY}.

Making implicit the push-forward of random variables, for a pair of measures $P, Q \in \mathcal{P}(\mathcal{E})$, we denote by $\Pi(P, Q)$ the set of probability distributions having P and Q as marginals. We can write $\Pi(P, Q)$ more precisely as follows

$$\Pi(P, Q) = \{\gamma \in \mathcal{P}(\mathcal{E} \times \mathcal{E}) \,:\, p^x_{\#}\gamma = P, \quad p^y_{\#}\gamma = Q\}, \quad \forall\, P, Q \in \mathcal{P}(\mathcal{E}),$$

where $p^x, p^y : \mathcal{E} \times \mathcal{E} \longrightarrow \mathcal{E}$ are such that $p^x(x, y) = x$ and $p^y(x, y) = y$ for all $x, y \in \mathcal{E}$. The elements of $\Pi(P, Q)$ are also referred to as *couplings*.

In practical problems, knowing the true underlying distribution P of a random variable X is often impossible. Instead, we usually have access to a *sample* of X. More precisely, we define a *sample* of X as a family of independent random variables $X_1, \ldots, X_n$ identically distributed as X (i.e. whose probability distribution is P). In practice, we observe a *realization* of $X_1, \ldots, X_n$, that is, the image of the random variables at n points $\omega_1, \ldots, \omega_n \in \Omega$. Realizations are commonly denoted in lower case letters as $x_1, \ldots, x_n$, where $x_i = X_i(\omega_i)$ for all $i \in \{1, \ldots, n\}$. Samples allow to obtain statistically meaningful information about the population through the *empirical measure* of P, defined below.

Definition 1.2.4. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, $\mathcal{E}$ a topological space and $\mathcal{T}$ the σ-algebra generated by the topology of $\mathcal{E}$. Let $X : \Omega \longrightarrow \mathcal{E}$ be a random variable with distribution P and $X_1, \ldots, X_n$ a sample of X, for $n \in \mathbb{N}$. The empirical measure of P is the probability measure $P_n : \mathcal{T} \longrightarrow [0, 1]$ satisfying*

$$P_n(E) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in E\} \quad \forall\, E \in \mathcal{T}, \tag{1.1}$$

where $\mathbb{1}\{A\}$ *denotes the indicator function of the event* A.

Note that writing $X_i \in E$ as in (1.1) is an abuse of notation, and formally we should replace it by $\{\omega \in \Omega : X_i(\omega) \in E\}$. For the sake of clarity, the widely accepted notation in (1.1) will be kept. The empirical measures are the main tool to *infer* the behavior of the population measures by only exploiting the information provided by a sample. This is further described in Section 1.2.3. Besides, the study of empirical measures provides a well-founded analysis of how the underlying distributions behave as P_n converges almost-surely [29, Theorem 6.1] and uniformly [29, Theorem 20.6] to P as n grows up to infinity. We refer to [29] for a deep introduction to Measure theory and Probability. To get familiar with the main concepts of Topology theory, that we will also be using, we refer to [210].

1.2.2 Probabilistic structural descriptors

The first step of our strategy involves defining suitable structural descriptors that integrate as much information as possible about the conformational variability of flexible proteins. We will do this by considering probability distributions well-adapted to the local and global structure of the system and, above all, whose corresponding random variables provide *accessible realizations*. In the words of a physicist, we aim to define random observables that can be *measured* in practice on protein models.

Local structural descriptors

The investigation of protein structural and dynamic properties at the local level primarily involves analyzing the backbone dihedral angles, ϕ and ψ, of individual amino acid residues along the sequence [39, 175]. An illustration for three consecutive amino acids is presented in Figure 1.5. The examination of the allowed values and the statistical distribution of (ϕ, ψ) has been a subject of study for over half a century, starting with the seminal work by Ramachandran *et al.* [237, 238]. The analysis of (ϕ, ψ) angles in polypeptide chains has numerous applications, such as the validation and refinement of structures determined from biophysical techniques [208, 182], the development of models or scoring functions for protein structure prediction and design [106, 153, 27, 35, 244, 280] or the investigation of denatured states of globular [267, 142] and intrinsically disordered proteins [265, 88]. While the values of (ϕ, ψ) are physically restricted for proteins that fold into a stable three-dimensional structure, they exhibit a high variability for IDP. Consequently, we are led to consider the pair (ϕ, ψ) as a random variable taking values on the two-dimensional flat torus $\mathbb{T}^2$, which is the Cartesian product of a pair of unit circles. A technical definition of $\mathbb{T}^2$ is presented in Chapter 3, where we also analyze its fundamental geometric and topological properties. As, indeed, $\mathbb{T}^2$ can be equipped with a topology, we are able to consider the set $\mathcal{P}(\mathbb{T}^2)$ of probability distributions supported on $\mathbb{T}^2$. Consequently, for a fixed amino acid, its dihedral angles (ϕ, ψ) will be associated with an element of $\mathcal{P}(\mathbb{T}^2)$, that we will define as the *local structural descriptor* of the amino acid residue.

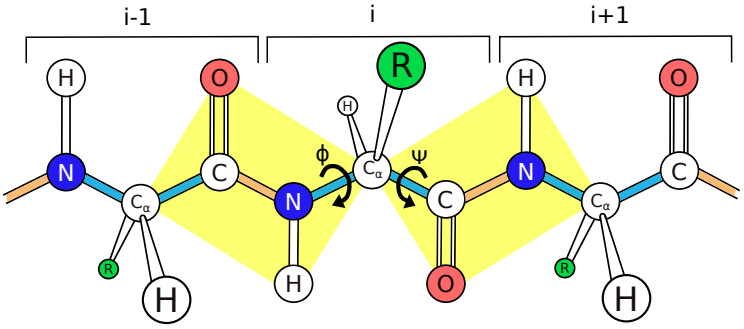


Figure 1.5: The determination of ϕ and ψ torsion angles. Each peptide bond (in orange) holds six atoms in a plane (in yellow), that fully describe the conformation of the i-th amino acid. The angle ϕ (resp. ψ) determines the rotation of the polypeptide backbone around the N-C_α (resp. C_α-C) bond.

Definition 1.2.5 (Local structural descriptor)**.** *Let (ϕ, ψ) be the random dihedral angles of an amino acid residue. Its local structural descriptor is defined as the probability distribution of (ϕ, ψ), which is an element of $\mathcal{P}(\mathbb{T}^2)$.*

As previously mentioned, we seek to consider structural descriptors that can be "measured" or, in other words, whose empirical probability distributions are easy to calculate. This is the case of (ϕ, ψ) angles, that can be experimentally determined with high resolution for rigid proteins or known when conformations are simulated with the methods presented in Section 1.1.3. This leads us to define the *empirical local structural descriptor* of an amino acid residue as the empirical probability distribution of its local structural descriptor.

Global structural descriptors

Structurally describing an entire sequence is a more complex task. Although some experimental methods such and X-ray crystallography and cryo-EM, as well as generative models, are capable of returning the coordinates of all atoms in the protein (for structured proteins), these coordinates cannot be compared between different conformations since they do not refer to an absolute reference system in which all states can be expressed. Furthermore, the structure of a state is invariant under rigid body transformations or, equivalently, under change of basis in the Euclidean vector space. Therefore, describing the global structure by using directly the all-atom coordinates would lead us to resort to the following equivalence relation[1]. If n_a denotes the number of atoms in the sequence, two elements $x, y \in \mathbb{R}^{3n_a}$ are equivalent, denoted as $x \sim y$, if and only if they are equal up to

[1]A binary relation $\sim$ on a set $\mathcal{X}$ is said to be an *equivalence relation* if it is reflexive, symmetric and transitive. The set of all elements in $\mathcal{X}$ that are equivalent to $x \in \mathcal{X}$ is called the *equivalence class* of x. The set of the equivalence classes of all the elements of $\mathcal{X}$ is called the *quotient set* of $\mathcal{X}$ by $\sim$, denoted by $\mathcal{X}/\sim$ [307].

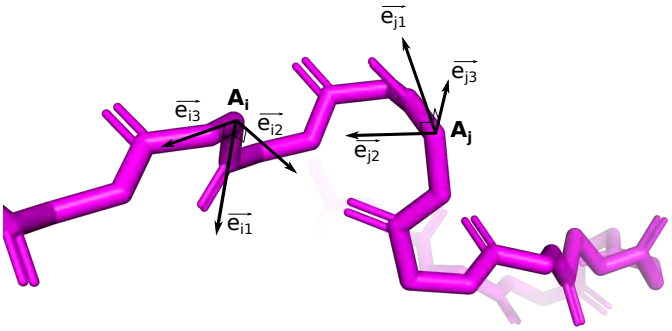


Figure 1.6: Illustration of the reference systems $\mathcal{F}_i$, $\mathcal{F}_j$ built at the i-th and j-th residues.

a rigid body transformation. Note that, indeed, $\sim$ is an equivalence relation as the space of rigid body motions is -among other things- a group. Such space is called the special Euclidean group of three dimensions and it is usually denoted as $SE(3)$ [135]. Then, we might define a global structural descriptor for the entire sequence as a probability distribution supported on the quotient space $\mathbb{R}^{3n_a}/\sim$. Although mathematically stimulating, resorting to $\mathcal{P}(\mathbb{R}^{3n_a}/\sim)$ to define structural descriptors is unnecessarily complicated and verges on mathematical pedantry. To capture the structure of the entire sequence, we propose to build a reference frame at every amino acid residue using the backbone atoms. An illustration for a pair of residues is presented in Figure 1.6.

Let L denote the sequence length and A_i the i-th amino acid, for $i \in \{1, \dots, L\}$. Using the coordinates of the i-th C, C_α and N atoms we are able to define a reference system that accounts for the geometrical configuration of the backbone at the i-th residue. The origin of the reference frame is set at the coordinates of the C_β atom, i.e. the first atom of the side chain (recall Figure 1.1) for non-glycine residues. For glycines, we place the origin at the coordinates of C_α. If we denote as $\mathcal{F}_i = \{\overrightarrow{e_{i1}}, \overrightarrow{e_{i2}}, \overrightarrow{e_{i3}}\}$ the i-th reference system, the global structure of the ensemble is described by L reference frames $\mathcal{F}_1, \dots, \mathcal{F}_L$. Note that each $\mathcal{F}_i$ can be formalized as an element of $SE(3)$. Besides the intricacy of comparing frames across different conformations, that can be solved in some cases, it should be remarked that relying on $SE(3)$ is extremely complex and requires handling Riemannian geometry. While this space is widely used in robotics [17, 315, 223], its application in this setting remains an excessively complicated and impractical task, due, for example, to the non-uniqueness of its geodesic curves, which hinders the computation of distances [226]. Although some remarkable contributions dealing with probability distributions and statistics in $SE(3)$ have been recently proposed [54, 205], we opt to define Euclidean descriptors of the family $\{\mathcal{F}_1, \dots, \mathcal{F}_L\}$ that allow their direct comparison across states and efficient computation.

Two different strategies will be followed according to whether ensembles must be compared or characterized. The former will rely on mapping the family $\{\mathcal{F}_1, \dots, \mathcal{F}_L\}$ to a Cartesian product of three-dimensional Euclidean spaces. This transformation will yield the definition of the *three-dimensional global structural descriptor* of the ensemble.

Definition 1.2.6 (Three-dimensional global structural descriptor)**.** *Let L denote the protein sequence length and $\{F_1 \dots, \mathcal{F}_L\}$ the family of reference frames built at every amino acid residue. Let*

$$T_{\mathbb{R}^3} \colon SE(3) \times \overset{L}{\cdots} \times SE(3) \longrightarrow \mathbb{R}^3 \times \overset{L(L-1)/2}{\cdots} \times \mathbb{R}^3 \\ (\mathcal{F}_1, \dots \mathcal{F}_L) \longmapsto \left(\overrightarrow{R}_{11}, \dots, \overrightarrow{R}_{(L-1)L}\right) \tag{1.2}$$

be the transformation that maps the family of reference frames to all the relative positions of every residue pair along the sequence. More precisely, $\overrightarrow{R}_{ij}$ denotes the coordinates of the origin of $\mathcal{F}_j$ with respect to $\mathcal{F}_i$, for every $i < j$. The three-dimensional global structural descriptor of the ensemble is defined as the $L(L-1)/2$-tuple

$$\left(P_{11}, \dots, P_{(L-1)L}\right) \in \mathcal{P}(\mathbb{R}^3) \times \overset{L(L-1)/2}{\cdots} \times \mathcal{P}(\mathbb{R}^3), \tag{1.3}$$

where $P_{ij} \in \mathcal{P}(\mathbb{R}^3)$ is the probability distribution of $\overrightarrow{R}_{ij}$, for every $i < j$.

Indeed, the definition of a reference frame at every amino acid allows the determination of the relative position of every residue pair. These positions will be random variables taking values on $\mathbb{R}^3$ and their probability distributions (1.3) will act as global structural descriptors of the protein ensemble. Note also that the realizations of every $\overrightarrow{R}_{ij}$ are comparable across conformations. Certainly, the mapping (1.2) transforms the structural configuration of the entire sequence into a set of three-dimensional Euclidean descriptors that do not depend on the absolute coordinates that were given as input. In other words, the realizations of $\overrightarrow{R}_{ij}$ are accessible and comparable, allowing the definition of the *empirical three-dimensional global structural descriptor* of the ensemble as the family of empirical counterparts of (1.3).

A different approach will be chosen when aiming at characterizing protein ensembles. In that case, the family of reference frames will be mapped to a Cartesian product of real intervals. Instead of analyzing all the relative positions of amino acid pairs, we will now account for their residue-residue interactions.

Definition 1.2.7 (One-dimensional global structural descriptor)**.** *Let L denote the protein sequence length and $\{\mathcal{F}_1, \dots, \mathcal{F}_L\}$ the family of reference frames built at every amino acid residue. Let*

$$T_{[0,1]} \colon SE(3) \times \overset{L}{\cdots} \times SE(3) \longrightarrow [0,1] \times \overset{L(L-1)/2}{\cdots} \times [0,1] \\ (\mathcal{F}_1, \dots \mathcal{F}_L) \longmapsto \left(\omega^C_{11}, \dots, \omega^C_{(L-1)L}\right) \tag{1.4}$$

be the transformation that maps the family of reference frames to a vector of elements in $[0,1]$ acting as a proxy for the interaction between the residues i and $j > i$. The one-dimensional global structural descriptor of the ensemble is defined as the $L(L-1)/2$-tuple

$$\left(P^C_{11}, \dots, P^C_{(L-1)L}\right) \in \mathcal{P}([0,1]) \times \overset{L(L-1)/2}{\cdots} \times \mathcal{P}([0,1]), \tag{1.5}$$

where $P^C_{ij} \in \mathcal{P}([0,1])$ is the probability distribution of ω^C_{ij}, for every $i < j$.

The quantities ω_{ij}^C account for the contact between amino acids at positions i and $j > i$. These variables will be conceived as an extension of the classical binary notion of contact, that is based on a universal threshold for the Euclidean distance [213, 275, 229]. In this case, the mapping (1.4) will transform the family $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ into a tuple of values that will vary continuously in $[0, 1]$ and will depend not only on the identity of the interacting amino acids and their positions in the sequence but also on the relative orientation of $\mathcal{F}_i$ and $\mathcal{F}_j$. Once again, these quantities are comparable between conformations and allow the definition of the *empirical one-dimensional global structural descriptor* as the vector of empirical probability distributions of (1.5).

1.2.3 Statistical tools to compare and characterize ensembles

The structure of protein ensembles will be described by the probability distributions presented in Definitions 1.2.5, 1.2.6 and 1.2.7. The next step is to find statistical tools that capture the variability of these distributions as faithfully as possible and provide compact, clear, and interpretable outcomes accounting for the conformational variability of flexible proteins and the changes in their secondary structure. Sometimes, suitable tools for this purpose were already found in the literature, corresponding even to standard methods used in Biostatistics and applied Mathematics in a wide manner. However, the problems of comparing and characterizing ensembles naturally give rise to questions for which there had not yet been a methodological answer. In this case, theoretical contributions that are applicable in a more general context have been proposed. It should be noted that, rather than the study of statistical problems, the motivation of this thesis is to provide answers to open problems in Structural Biology. Nevertheless, this has naturally led to the development of some mathematical techniques that may be of interest from a broader perspective. The families of statistical methods employed in this thesis are outlined in the following sections. We will omit the details about the first group as it comprises standard techniques commonly used in Structural Biology and Biostatistics. More attention will be paid to the subsequent ones, whose application in Structural Biology is more novel and where our methodological contributions have been made.

Clustering on a low-dimensional embedding (used in Chapter 7)

Dimension reduction is a widely employed technique in Biostatistics due to the intrinsic high dimensionality of biological data. Most of the applications of such theory are related to the very active fields of neuroimaging [211], single-cell [15, 272] or genetics [83, 81], among others. Here, we will focus on non-linear dimensionality reduction algorithms, that have shown efficient empirical performances when identifying underlying structures in complex data [80, 81, 15, 174, 232, 83]. In particular, we will make use of the Uniform Manifold Approximation and Projection (UMAP) algorithm [199]. This choice is motivated by its ability to preserve the high-dimensional topology of data and efficiently reveal population structure [81, 83, 15]. For some time now, the combination of non-linear

dimension reduction algorithms with clustering techniques is becoming a standard procedure to detect the structures unraveled by the low-dimensional projection and classify them into well-defined groups. The use of that strategy is supported by its successful empirical efficiency [82, 2, 15, 83]. In this thesis, we propose to project high-dimensional data to a low-dimensional UMAP space and perform the HDBSCAN [46] algorithm on the embedding, which we believe to be one of the most sophisticated density-based techniques. The main principles of UMAP and HDBSCAN algorithms are explained in Appendix E.1.

Optimal transport (used in Chapters 3-5)

Optimal Transport (OT) is a mathematical theory that has been gaining considerable relevance in recent years due to its efficient and versatile applicability. In particular, the popularity of OT has raised through its integration into Machine Learning techniques, notably in the framework of generative networks [9], robustness [262] or fairness [76, 69, 31], among others. With some notable exceptions [47, 19, 248, 67, 107], OT has not been widely used in Structural Biology. Here, we propose to rely on OT to account for differences between global and local structural descriptors. Let us first introduce the main concepts of this theory.

Optimal Transport is a specific case of *mass transportation*, which is the general problem of matching two probability distributions P, Q supported on a Polish space $\mathcal{X}$, that is, a separable and completely metrizable topological space [210]. Note that the Euclidean space of arbitrary dimension is Polish, as well as the two-dimensional flat torus $\mathbb{T}^2$, as shown in Chapter 3. Consequently, this theory is applicable to the distributions that compose the local and global structural descriptors[2] defined in Section 1.2.2. The problem of mass transportation aims at selecting a coupling in $\Pi(P, Q)$, that is, a joint probability distribution having P and Q as marginals.

A coupling can be seen as a random mapping, matching every instance in the support of P to possibly several counterparts in the support of Q with probability weights. This transformation can be also understood as a reconfiguration of the *probability mass* of P to recover the one of Q. More visually, we might think of each marginal distribution as a sand pile on $\mathcal{X}$. A coupling is a transportation plan transforming one pile into the other, that specifies how to move each elementary sand mass from the first distribution to recover the second one. A coupling is said to be *deterministic* if each instance from P is matched to a unique instance from Q. In that case, the coupling is localized on the graph of a (P-almost surely unique[3]) mapping $T : \mathcal{E} \longrightarrow \mathcal{E}$ that pushes forward P to Q, i.e. such that $T_{\#}P = Q$. We denote by $\mathcal{T}(P, Q)$ the set of measurable mappings pushing forward P to Q.

Optimal Transport has become a popular tool to define such couplings by selecting the

[2]As closed subsets of Polish spaces are Polish, this also applies for the distributions composing the one-dimensional local structural descriptors introduced in Definition 1.2.7.

[3]That is, if there exists another mapping $T' \neq T$ whose graph localizes the same coupling, it only differs from T on some set O with $P(O) = 0$.

ones that are optimal in some sense. This theory dates back to Monge [207] who in 1781 defined OT maps as functions that transform P into Q with minimal effort according to a positive cost function $c : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^+$. Formally, transport maps are defined as the solutions of

$$\inf_{T \in \mathcal{T}(P,Q)} \int_{\mathcal{X}} c(x, T(x))\, dP(x). \tag{Monge}$$

One mathematical challenge arises from the push-forward constraint, which makes the problem infeasible in many general scenarios, especially when distributions P and Q are not absolutely continuous with respect to the Lebesgue measure [29] or have an imbalanced number of atoms. This complication motivated the so-called Kantorovich relaxation of the OT problem introduced by Kantorovich and Rubinshtein in 1958 [154]:

$$\inf_{\gamma \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y)\, d\gamma(x, y). \tag{Kantorovich}$$

Solutions to (Kantorovich) are optimal couplings (non deterministic in general) between P and Q with respect to the cost c. Contrary to OT maps, they exist under very mild assumptions, like the non-negativeness of the cost [299]. Note that, since a push-forward operator can be identified with a coupling, the set of admissible solutions of (Monge) is included in the set of admissible solutions of (Kantorovich).

The solutions to (Kantorovich) are of particular interest to us as they define a distance in $\mathcal{P}(\mathcal{X})$ [299]. More precisely, for $p \geq 1$, the optimal value

$$\mathcal{W}_p(P, Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c^p(x, y)\, d\gamma(x, y) \right)^{\frac{1}{p}} \tag{1.6}$$

is called the *p-Wasserstein distance* between P and Q, and it represents the minimum transportation cost needed to reconfigure the mass of P to recover the mass of Q. Note that the Wasserstein distance is able to integrate the geometry of the underlying space $\mathcal{X}$ if the cost function is chosen, for instance, as the geodesic distance on $\mathcal{X}$. That makes it a well-adapted metric to capture the variability of the conformational space and appropriately compare a pair of structural descriptors.

We conclude by presenting how to solve (Kantorovich) when, in practice, we only have access to the empirical counterparts of P and Q, denoted by P_n, Q_m for $n, m \in \mathbb{N}$. This scenario corresponds to the *discrete* version of the Kantorovich problem, where the points of the sample drawn from P are sent to the points of the sample drawn from Q with probabilities given by a $n \times m$ matrix, that we identify with the coupling in (Kantorovich). Let X, Y be two random variables having P and Q as probability distributions respectively, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ two samples of X and Y and $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ two realizations of such samples. The discrete version of (Kantorovich) corresponds to solving

$$\inf_{M \in U(P_n, Q_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j)\, M_{ij}, \tag{1.7}$$

where $U(P_n, Q_m)$ is the set of real $n \times m$ matrices $M = (M_{ij})_{ij}$ such that $\sum_{i=1}^{n} M_{ij} = m^{-1}$ and $\sum_{j=1}^{m} M_{ij} = n^{-1}$. Once again, the *p-Wasserstein distance* between the empirical measures P_n and Q_m is given by

$$\mathcal{W}_p(P_n, Q_m) = \left(\inf_{M \in U(P_n, Q_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} c^p(X_i, Y_j)\, M_{ij} \right)^{\frac{1}{p}}. \tag{1.8}$$

Note that (1.6) is a positive real number whereas (1.8) is a random variable, as it is a function of the samples drawn from P and Q. Fortunately, the so-called empirical Wasserstein distance (1.8) presents strong statistical guarantees. In particular, it converges weakly to the distance between the true measures (1.6) as n and m grow up to infinity under mild assumptions [299, Corollary 6.9]. This justifies the use of (1.8) to account for the differences between the local and global structural descriptors by computing the Wasserstein distance between their empirical counterparts. For these practical applications, we will set $p = 2$ due to the well-known statistical properties associated to the quadratic cost, in particular the uniqueness of the solution to (Kantorovich) under mild assumptions [298, Theorem 2.12]. For a comprehensive understanding of the mathematical properties of the Wasserstein distance and the Optimal Transport problem, we refer to [299].

The resolution of the optimization problem (1.7) has become another extensive area of research. As the objective function and the constraints are linear in the variables of interest, the discrete formulation of the Kantorovich problem is a linear program. Consequently, it can be solved with a large family of algorithmic tools from linear programming and combinatorial optimization. Among them, we may highlight the classical Network Simplex algorithm [24], that is implemented in the more common OT solvers [98, 259]. Another popular strategy are Dual Ascent methods [144], notably the well-known Hungarian algorithm [25]. The primary challenge encountered when dealing with empirical optimal-transport solutions lies in their high computational complexity and memory requirements. Solving (1.7) typically demands $O((n+m)nm\log(n+m))$ computer operations. Besides, for non-standard cost functions, a $n \times m$ matrix of coefficients $C_{ij} = c(x_i, y_j)$ needs to be stored. In practical problems, notably in Machine-Learning applications, it is usual to consider entropic regularization schemes, that can reduce the computational complexity to $O(nm)$ operations [64]. However, these approximations do not overcome memory issues and loose the mathematical and statistical properties that motivate the use of Wasserstein distance in the inferential context. For a less technical introduction to OT and a thorough analysis of the computational aspects discussed below, we point to [228].

Hypothesis testing (used in Chapters 2-6)

Hypothesis testing constitute, along with estimation, the other fundamental pillar of statistical inference. The objective of such tests is to determine, based on the collected sample information, whether a certain hypothesis about a population characteristic should be rejected or not. Succinctly, to test a hypothesis is to determine whether it is "compatible"

with what is observed in the sample. More precisely, it involves comparing the validity of two complementary statements about the population. One of them is called the *null hypothesis* (H_0), while the other is called the *alternative hypothesis* (H_1). Note that statistical tests are not symmetric towards H_0 and H_1 in the sense that they do not choose the most plausible hypothesis based on the sample. Instead, they just aim to determine whether there is sufficient evidence to reject what H_0 claims. Consequently, the test never concludes that the null hypothesis is true, but rather that there is no evidence to reject it. Formally, we can define a statistical test as a measurable partition [61, Definition 15.1] of the sample space.

Definition 1.2.8 (Hypothesis test)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. A hypothesis test is a measurable partition of Ω in two regions, namely the critical -or rejection- region (C.R.), that includes the instances that yield the rejection of H_0, and its complementary set, the alternative region (A.R.), composed by the outcomes that do not yield rejection of H_0. The test is characterized by the indicator function of the critical region, also referred to as test function, $\pi \colon \Omega \longrightarrow \{0, 1\}$ where*

$$\pi(\omega) = \begin{cases} 1 & \text{if } \omega \in \text{C.R.} \\ 0 & \text{if } \omega \in \text{A.R.} \end{cases} \tag{1.9}$$

The potential of hypothesis tests comes in terms of the statistical guarantees that they provide regarding the partition (1.9). More precisely, test functions are built to ensure the control of the so-called *type I error*, that is, the probability of rejecting H_0 when it is true.

Definition 1.2.9 (Type I error)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\pi : \Omega \longrightarrow \{0, 1\}$ a test function. The type I error of π is defined as the probability of rejecting H_0 when it holds, that is,*

$$\mathbb{P}\left(\{\omega \in \Omega \,:\, \pi(\omega) = 1 \,|\, H_0\}\right) = \mathbb{P}_{H_0}\left(C.R.\right).$$

The procedure for constructing a hypothesis test begins by setting an upper bound on the type I error -the so-called *level of significance*- and selecting, among all tests that control it, the one that more efficiently detects the false null hypotheses or, in other words, the most *powerful* one.

Definition 1.2.10 (Power)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\pi : \Omega \longrightarrow \{0, 1\}$ a test function. The power of π is defined as the probability of rejecting H_0 when it is false, that is,*

$$\mathbb{P}\left(\{\omega \in \Omega \,:\, \pi(\omega) = 1 \,|\, H_1\}\right) = \mathbb{P}_{H_1}\left(C.R.\right).$$

Note that the problem of building a hypothesis test relies on the choice of a "good" critical region. To achieve so, the following considerations must be taken into account:

(*i*) Discrepancies with the null hypothesis are sought, so the critical region must include sample values that are unlikely to occur under H_0, even though they are possible,

(*ii*) Generally, the critical region is determined before analyzing the experimental results (although this is not always true, as discussed in the next section),

(*iii*) The critical region is usually expressed in terms of a statistic, that is, a real-valued measurable function of the sample, known as the *test statistic*. It measures the discrepancies between the samples in the critical region and the null hypothesis. The distribution of the test statistic under H_0 is used to ensure the level of significance.

Note that the test function (1.9) provides a binary output "rejection vs. non-rejection" about H_0. However, whether the null hypothesis is rejected or not, it is usually interesting to "measure the distance" between the sample result and H_0. This gives rise to the concept of p-value.

Definition 1.2.11 (p-value)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, π a test function defined on Ω whose test statistic is a function of a sample $X_1, \dots, X_n$. The p-value associated with a realization $(x_1, \dots, x_n)$ is the smallest significance level at which the null hypothesis H_0 is rejected by π.*

This value acts as a proxy for the plausibility of the sample under the null hypothesis. If the p-value is large, it means that we are working with samples that have a high probability of occurring if H_0 holds. In this case, there is insufficient evidence against the null and H_0 should not be rejected. However, rejection should be chosen if the p-value is small. We want to emphasize that the p-value can -and should- be considered as a quantitative indicator of the "credibility" of the null hypothesis. Accordingly, with appropriate precautions, p-values calculated under the same conditions (e.g. equality of sample sizes) are quantitatively comparable and provide a correct indicator of which samples contradict more the validity of H_0 within a family of realizations. This point will be key in the works presented here. Furthermore, it is worth noting that the p-value can also be used to effectively verify the proper definition of a hypothesis test. Following its definition, a p-value is statistically valid if and only it is Super-Uniform under H_0. A real-valued random variable X is said to be Super-Uniform if its cumulative distribution function (CDF) is upper bounded by that of the Uniform distribution, that is:

$$\mathbb{P}(X \leq x) \leq x \text{ for all } x \text{ in } [0, 1]$$

(see e.g. [172, Section 3.3]). Moreover, the closer the p-value distribution under the null hypothesis is to $\mathrm{U}[0, 1]$, the more powerful the corresponding test is. Checking for super-uniformity of p-values under H_0 is essential to ensure the suitability of the corresponding statistical test.

In this thesis, we mainly focus on a particular case of statistical test, known as *two-sample goodness-of-fit* test. In short, it aims at assessing whether two probability distributions are the same. More precisely, for two measures P, Q supported on a Polish space $\mathcal{X}$, the goal is to test for the following null and alternative hypotheses:

$$H_0 : P = Q \qquad \text{vs.} \qquad H_1 : P \neq Q. \tag{1.10}$$

Note that, under this framework, we are only testing for the equality of P and Q, independently of the identities of these distributions. The key issue here is the choice of a suitable test statistic that appropriately accounts for the differences between P and Q and whose null distribution is known[4]. The most commonly used approaches to test for (1.10) are mainly defined for measures supported on the real line (e.g. Kolmogorov-Smirnov and Wilcoxon statistics). However, testing the equality of distributions supported on more general spaces is a much less studied problem, and it is crucial in our objective of properly comparing local structural descriptors. The probability distributions accounting for the conformational variability of the protein at the amino acid scale (recall Definition 1.2.5) are supported on the two-dimensional flat torus, which is a non-Euclidean space. In particular, a test statistic defined to compare distributions in $\mathcal{P}(\mathbb{T}^2)$ needs to be adapted to the periodicity of their support. This is why the Wasserstein distance (1.6) turns out to be a well-suited metric to compare measures on $\mathbb{T}^2$ if the geodesic distance on such space is chosen as cost function. In this thesis, we propose two approaches to define two-sample goodness-of-fit tests in $\mathcal{P}(\mathbb{T}^2)$ using the Wasserstein distance as a test statistic. This will provide *statistical evidence* about the discrepancies between local structural descriptors or, in other words, the statistical significance of changes on local protein structure.

Post-selection inference (used in Chapter 6)

When performing a statistical investigation, a model for the data needs to be previously specified. This model might be the underlying distribution of the sample, the variables that explain a given outcome or a hypothesis to be tested. In a more classical -or academic- context, the model is set prior to collecting the data. This might be the case if, for instance, observations follow a known physical law. However, in more realistic applications, inference is performed on a model that is chosen *from the data*. A straightforward example is testing for the significance of the features selected by a regression model whose coefficients have been obtained from data. In this case, the null hypotheses to be tested, that is, the *questions* that should be answered through inference, *depend on the data*. If the same data are used for the subsequent testing step, the statistical guarantees are not ensured. This phenomenon is akin to overfitting in prediction tasks. Adapting inferential statistics to the framework where the model choice is guided by the data is the goal of *post-selection inference*. The relevance of this field has greatly increased in recent years, due to its usefulness in a wide family of areas like causal inference [16], high-dimensional linear regression [171] or neural networks [300], among others. Here, we focus on *selective hypothesis testing*, that addresses the definition of statistical tests when null hypotheses are chosen from the data. The basis of this theory was rigorously introduced in [97].

In this thesis, we aim at performing inference after clustering by testing for the difference between cluster means. Clustering algorithms aim to classify observations into a number of classes, which is usually directly or indirectly predetermined. The outcome

[4]We commonly refer to the distribution of any random variable under H_0 as its *null distribution*.

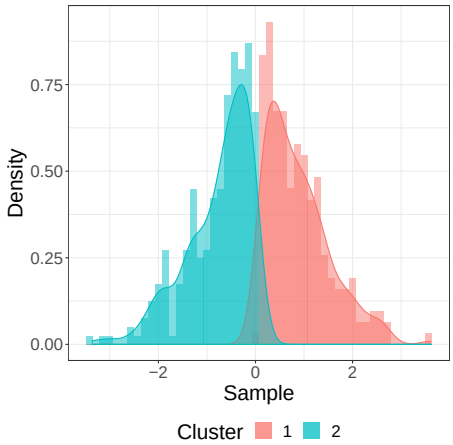


Figure 1.7: Empirical distributions of two groups of observations found after simulating one sample of size $n = 500$ drawn from the univariate distribution $\mathcal{N}(0, 1) + \mathcal{U}(-0.2, 0.2)$ and classifying it into two groups through a k-means algorithm. Colors indicate the classes retrieved by the pipeline. This Figure is adapted from [121, Figure 1].

of these methods is highly sensitive to the parameters required by each technique, and the resulting partition of the space can strongly vary for a given dataset. Although there are criteria to optimize the calibration of these parameters [278, 251, 119], it should be emphasized that there is generally no underlying true classification [90]. Clustering algorithms attempt to find classes that compactly represent the distribution of the dataset, but these groups are not inherent to the population in general and act only as a description of the sample. Post-clustering inference seeks to shed some light on this issue by providing statistical evidence of the true differences between clusters. The relevance of this problem can be easily illustrated by simulating a one-dimensional Gaussian random variable with a uniform noise and asking the classic k-means algorithm [117] to find two clusters. The result is presented in Figure 1.7. If, as it happens in practice, the underlying distribution is unknown, it is difficult to visually assess whether both groups correspond to two different populations. If we try to answer that question by omitting that the null hypothesis has been chosen by looking at the data, i.e. following the results of a clustering algorithm, the conclusion will be misleading. Indeed, a classical Z-test retrieves a p-value below the machine precision ($5.87 \cdot 10^{-67}$), yielding a strong rejection of the hypothesis of equal group means. If we perform selective inference, the approach presented in [51] to test for the difference of cluster means retrieves a p-value equal to 0.84, which is coherent with the true setting (here known, as the data have been simulated).

Our motivation here is to provide statistical guarantees about the clustering algorithms that are commonly used to characterize protein ensembles. Indeed, defining a partition of the conformational space through featuring the states by classical descriptors has become a standard technique [7, 59, 170, 169]. However, these characterizations lack statistical

evidence about the true differences between groups, which we believe to be essential to correctly interpret their output.

1.3 Outline of the thesis

The main objective of this thesis is to define statistical methods for the appropriate characterization and comparison of conformational ensembles of flexible proteins. In Section 1.2.2, we have introduced the probabilistic descriptors of the structural variability of these systems and, in Section 1.2.3, we have detailed the statistical methods we will use to analyze them. We are left with only one question to address, which is: how? The answer to this question is developed throughout the chapters that constitute this thesis. The manuscript is divided into two main parts, dedicated to the structural analysis at local and global scales, respectively. Within each part, we outline the strategies constructed to deploy the methods from Section 1.2.3 on the descriptors from Section 1.2.2, and present the resulting techniques for characterization and comparison. Now, we provide a brief overview of these methods through a plan of this manuscript, where the main results and contributions are outlined.

Software availability and reproducibility

The characterization and comparison methods presented in this thesis have been made freely available to the community. To ensure reproducibility, the code implementing all the statistical analyses and the data have been equally shared. Links are specified within each chapter.

1.3.1 Local structural analysis of protein ensembles (Part I)

The first part of the thesis is devoted to the structural analysis of ensembles at the local level. More precisely, we will study the probability distributions of the dihedral angles (ϕ, ψ) that define the local structural descriptors. This part is composed of three chapters:

- In Chapter 2, we assess the effect of the neighboring amino acids on (ϕ, ψ) distribution, showing that the identities of left and right residues must be *simultaneously* taken into account to describe local structures. This sets fragments of three amino acids (tripeptides) as the unit brick to analyze protein conformation locally.

- Chapter 3 introduces two approaches to perform two-sample goodness-of-fit tests on $\mathcal{P}(\mathbb{T}^2)$, using Optimal Transport theory and Wasserstein distance. These methods will be the main tool to account for statistically significant differences between local structural descriptors. We also illustrate their utility to reject Flory's isolated pair hypothesis [99].

- Finally, in Chapter 4, we present a less trivial application of the methods introduced in Chapter 3, that is, the assessment of the translated codon effect on (ϕ, ψ) distributions. This chapter arises in response to the work of Rosenberg *et al.* [249], where the same problem was analyzed but using an inadequate methodology.

Interdependence between nearest neighbor effects (Chapter 2)

The structural analysis of conformational ensembles, both at the local and the global level, must be built upon a solid foundation of how sequence influences the amino acid structure. Flory's isolated-pair hypothesis [99], that states that the (ϕ, ψ) angles of a given residue are independent of the identity of its neighbors, has already been refuted by the community in numerous studies [40, 225, 66, 260, 149, 204] (although, as shown in Chapter 3, none of these approaches provide statistical evidence for its rejection). However, an important question remains unclear: are the effects of the left and right neighbors independent? In other words, can the local structure of a protein be described based on two amino acid fragments (dipeptides), or should the unit block be the tripeptide? Contradictory answers have been proposed regarding this issue [109, 27, 129, 244], but none of them has been based on a well-grounded methodology that provides statistical guarantees about the distribution of (ϕ, ψ). Here, we aim to *test* for the independence of neighbors effects.

Let C denote the identity of an amino acid residue and L, R the identities of its left and right neighbor in the sequence, respectively. From the works [280, 244], we can show that the local structural descriptor given the whole tripeptide, denoted by $P(\phi, \psi \,|\, L, C, R)$, can be obtained from the ones given the left and right dipeptides as

$$P(\phi, \psi \,|\, L, C, R) = \frac{P(\phi, \psi \,|\, L, C)\, P(\phi, \psi \,|\, C, R)}{S\, P(\phi, \psi \,|\, C)}, \tag{1.11}$$

where S is a normalization constant, if and only if the following hypothesis holds

$$L \text{ and } R \text{ independent given } C \text{ and } (\phi, \psi). \tag{1.12}$$

Note that (1.11) corresponds to stating that the influences of the left and right neighbors on (ϕ, ψ) distribution can be independently considered to reconstruct it. As it is an equivalent statement, we propose to test for (1.12) using a classical χ^2 independence test [172]. Methodologically, a suitable approach to condition on $\{C, \phi, \psi\}$ is proposed, mainly consisting on intelligently discretizing $\mathbb{T}^2$ and performing one test per subdivision and value of C. Then, all p-values are corrected for multiplicity [125] and stratified by residue identity. Our results unequivocally demonstrate the coupled effects of the left and right neighbors, indicating that they cannot be considered independently of each other. Besides, we show that the magnitude of the interdependence, measured in terms of p-values, is affected by the physicochemical properties of the nearest neighbors and the structural origin of the data. These observations represent a fundamental step towards understanding sequence-structure relationships in peptides and proteins.

Two-sample tests to compare local structures (Chapter 3)

The structural investigation of (ϕ, ψ) angles involves quantifying the expected magnitude of structural effects associated with local changes in the sequence. In this context, the definition of a suitable distance between distributions on $\mathbb{T}^2$ whose statistical significance can be assessed is essential. In this chapter, we aim to test the hypotheses

$$H_0 : P = Q \qquad \text{vs.} \qquad H_1 : P \neq Q, \tag{1.13}$$

for a pair of local structural descriptors $P, Q \in \mathcal{P}(\mathbb{T}^2)$. As previously stated, distributions will be compared using the 2-Wasserstein distance (1.6), that integrates the underlying geometry of the conformational space at the local level. As the study of Optimal Transport in $\mathbb{T}^2$ has not completely been addressed, we start by extending the main results of this theory to the flat torus of arbitrary dimension, denoted by $\mathbb{T}^d$. In particular, we show the uniqueness under mild assumptions of the solution of (Kantorovich) in $\mathcal{P}(\mathbb{T}^d)$, and derive a Central Limit Theorem (CLT) for the fluctuations of the empirical transportation cost. We justify why the proposed CLT is not suitable to define an asymptotic goodness-of-fit test for (1.13), motivating the exploration of alternative approaches.

The strategy to define a two-sample test for (1.13) should rely on rejecting the null hypothesis when the Wasserstein distance between the empirical counterparts of P and Q is "too big" or, in other words, too improbable under H_0. If we denote as $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ two independent samples identically distributed as P and Q respectively, this yields the definition of the following critical region

$$C.R. = \{(x_1, \dots, x_n; y_1, \dots, y_m) : \mathcal{W}_2^2(P_n, Q_m) \geq c_{nm}(\alpha)\}, \tag{1.14}$$

where x_i (resp. y_j) denotes a realization of X_i (resp. Y_j) for $i = 1, \dots, n$ (resp. $j = 1, \dots, m$). The threshold $c_{nm}(\alpha) > 0$ must be chosen to ensure the type I error control at level $\alpha \in [0, 1]$ through the null distribution of the test statistic. However, knowing the distribution of $W_2^2(P_n, Q_m)$ under H_0 remains still an open and non-trivial problem, especially when the ground space has dimension higher than one. That intrinsic difficulty led us to search for alternative approaches that exploit the scenarios where a Wasserstein-based test statistic with known distribution can be defined. We propose two strategies in that regard, that we outline in the following paragraphs.

The first approach consists on bypassing the dimension problem by projecting P_n and Q_m to the one-dimensional closed geodesics of $\mathbb{T}^2$, that are closed spirals isomorphic to the circle S^1 [36]. Then, we define a test statistic based on the 2-Wasserstein distance to compare measures on S^1 whose null distribution we can derive. We show that this distribution does not depend on the identities of P and Q under H_0 or, in other words, that the defined statistic is *distribution-free* under the null. This allows the definition of a two-sample goodness-of-fit test for a pair of geodesic projections of P_n and Q_m. The strategy to define a p-value for the two-dimensional test is to (i) choose a family of randomly-selected closed geodesics on $\mathbb{T}^2$, (ii) for each geodesic, project P_n and Q_m and

retrieve a p-value for the equality of their projections and, finally, (*iii*) aggregate all the p-values by Bonferroni aggregation [32], obtaining a well-defined p-value for (1.13). We conclude by showing the consistency of the test under fixed alternatives, i.e. that its power tends to one as the sample sizes n,m grow up to infinity.

The second approach aims to directly compare the structural descriptors in the two-dimensional space. Due to the inability to construct an exact or asymptotic test based on (1.14), we propose to find an upper bound for its associated p-value. Note that upper bounding a p-value yields statistically valid hypothesis testing. Indeed, if the upper bound is smaller than the significance level, so will be the true -and unknown- p-value and rejection with type I error control will be ensured. In other words, this corresponds to have Super-Uniform p-values. However, a price to pay comes in terms of power loss. Succinctly, the idea is to first upper bound the deviations of the statistic from its expectation and then show that, under the null, this expectation presents a fast convergence rate to zero. This yields the definition of a two-sample test that is *asymptotically consistent at level* α [292, Definition 14.2]. That means that the statistical guarantees are ensured at the limit $n, m \to \infty$ so, in practice, that the test can be performed for large sample sizes. However, the finite sample conservativeness of this test becomes advantageous in our application context, making it complementary to the geodesic projection approach.

A numerical analysis is performed to illustrate the relative efficiency of both approaches and compare them to other alternative techniques from the literature. Besides, we demonstrate their suitability to detect differences on local structural descriptors, using a structural database of three-residue fragments extracted from experimentally-determined high-resolution protein structures [88] to reject the Flory's isolated pair hypothesis.

The codon effect on local structure (Chapter 4)

We conclude the first part of the manuscript by presenting a practical application of the methods introduced in Chapter 3. Numerous biological processes, such as mRNA splicing, translational rates and protein folding, have demonstrated the relevance of synonymous codon usage [220, 44]. Although the relation between synonymous codons and secondary structure in translated proteins has been extensively investigated [218, 258], Rosenberg *et al.* [249] took a more detailed approach by evaluating the impact of codon identity on the distribution of (ϕ, ψ) dihedral angles within secondary structure elements. Their work aimed to determine whether there are significant differences when synonymous codons are used, through the implementation of a statistical test for measures on $\mathcal{P}(\mathbb{T}^2)$. However, their statistical methodology is formally incorrect, casting doubt on the obtained results.

In this Chapter, we first demonstrate that the p-values defined in [249] are statistically invalid by proving that their distribution is not Super-Uniform under the null hypothesis (recall Section 1.2.3). Besides, we show that these p-values are highly conservative for large values of the statistic, yielding an important number of false negatives and thus ignoring substantial differences that might appear between local structural descriptors. Moreover, the multiple testing procedure used in [249] fails to control the False Discovery

Rate (FDR) as it needs the p-values to be Super-Uniform under the null [247]. The technical inaccuracies in this study prompted us to investigate the codon effect with appropriate statistical tools. This was the motivation to implement the methods presented in Chapter 3 to properly test for significant differences between codon-specific local structural descriptors.

Our results confirm the influence of the codon on (ϕ, ψ) distributions, but differ from those of [249] in the strength of significance of the differences depending on the secondary structure type. Besides, we evaluated the impact of structural classification and local sequence context on these findings. The results revealed that codon-specific effects exhibit similar levels of significance across different regions of $\mathbb{T}^2$. However, these effects may be more pronounced for specific secondary structure types, such as β-strands compared to α-helices. Furthermore, the results suggest that synonymous codon effects are amplified when considering the context of the local sequence, following the conclusions of Chapter 2.

1.3.2 Global structural analysis of protein ensembles (Part II)

The second part of the manuscript is devoted to the structural analysis of flexible proteins at the global level. We make use of the global structural descriptors defined in Section 1.2.2 to define statistical tools to compare and characterize conformational ensembles, and provide the classical clustering techniques with statistical guarantees. This part includes three chapters, outlined below.

- Chapter 5 presents WASCO, a Wasserstein-based statistical tool to compare conformational ensembles of highly flexible proteins. The main idea of WASCO is to use Wasserstein distance to compare the three-dimensional global structural descriptors (Definition 1.2.6), also integrating the local-level information through the techniques presented in Chapter 3. We show the usefulness of the method to compare different force-fields within MD simulations or to assess the effect of refinement with experimental data.

- Chapter 6 is devoted to the study of post-clustering inference when data present arbitrary dependence structures between features and observations. This work, that is the natural extension of the framework in [104, 51], provides the classical clustering techniques for ensemble characterization with statistical guarantees about the true differences between the retrieved groups of conformations.

- Chapter 7 presents WARIO, a contact-based characterization of conformational ensembles. The method adapts the classical contact maps that characterize folded structures to the ensemble framework, by characterizing a flexible protein through a weighted family of contact maps, built from the one-dimensional global structural descriptors (Definition 1.2.7). The applicability of WARIO is illustrated with the characterization of several conformational ensembles of IDP.

A Wasserstein-based tool to compare protein ensembles (Chapter 5)

The comparison of conformational ensembles is an essential problem in Structural Biology. When dealing with ensembles of highly-flexible proteins, the existing tools proposed in the literature are based on averaged descriptors across the set of conformations [167, 131]. However, reducing complex distributions to their mean usually entails considerable loss of information and hides relevant features that might distinguish the systems. In this Chapter, we present a comparison technique that integrates the whole variability of the conformational space and makes use of the Wasserstein distance to account for the differences between the entire probabilistic descriptors.

The main idea of WASCO is to compute the Wasserstein distance between the three-dimensional global structural descriptors of a pair of ensembles (Definition 1.2.6). More precisely, for every pair of residues at positions $i < j$ on the sequence, the method computes the distance $\mathcal{W}_{ij} = \mathcal{W}_2(P^A_{ij;n}, P^B_{ij;m})$, where $P^A_{ij;n}$ (resp. $P^B_{ij;m}$) denotes the ij component of the empirical global structural descriptor for ensemble A (resp. B). The quantity $\mathcal{W}_{ij}$ is the distance between the relative position distribution of residues i, j of both ensembles. The same idea is carried out to compare all the local structural descriptors (Definition 1.2.5) for every residue along the sequence. If $P^A_{i;n}$ (resp. $P^B_{i;m}$) denotes the i-th component of the empirical local structural descriptor for ensemble A (resp. B), WASCO computes the quantities $\mathcal{W}_i = \mathcal{W}_2(P^A_{i;n}, P^B_{i;m})$, to which we associate the p-value upper bound introduced in Chapter 3. Note that this formulation yields a clear and compact representation of the results through a triangular matrix, having as entries the quantities $\mathcal{W}_{ij}$ in the lower triangle and the distances $\mathcal{W}_i$ along the diagonal. Combining all the structural discrepancies at the local and the global level in the same representation allows to clearly stand out the more relevant residue-specific differences and assess the relation between changes on (ϕ, ψ) distributions at the amino acid level and structural disagreements at the entire sequence scale.

The variability in experimental and simulated structures causes uncertainties and statistical noise that may substantially bias the distance estimation. Therefore, the computed differences between global and local structural descriptors are corrected to filter such noise and highlight the relevant discrepancies between the ensembles. This filtering is carried out by estimating and removing the so-called *intra-ensemble* differences, that is, the counterparts of the quantities $\mathcal{W}_{ij}$ and $\mathcal{W}_i$ calculated between independent samples of the same ensemble. We also use the intra-ensemble differences to define a final score that allows to quantitatively interpret the computed Wasserstein distances using the noise as a reference. Besides, we define an overall distance that accounts for the difference between all the global and local structural descriptors (i.e. a distance in their product space), by properly aggregating the quantities $\mathcal{W}_{ij}$ and $\mathcal{W}_i$ after the noise correction.

We demonstrate the usefulness of WASCO to compare conformational ensembles (i) produced from MD simulations using different force fields, and (ii) before and after refinement with experimental SAXS data. We also show the applicability of the method to assess the convergence of MD simulations, and discuss further potential applications such

as in machine-learning-based approaches. One of the assets of this tool is its easy-to-use implementation as a Jupyter notebook, that has been made available to the community.

Post-clustering inference under dependence (Chapter 6)

A common strategy to characterize protein ensembles is to partition the conformational space through the implementation of clustering algorithms [7, 59, 170, 169]. However, as discussed in Section 1.2.3, the output of clustering techniques lacks of interpretability due to their high sensitivity to the pipeline parameters and the non-availability of an underlying true classification. This issue can be overcome by resorting to the theory of post-clustering inference, which provides statistical guarantees about the differences between cluster means. The mathematical techniques that allow such a selective testing are highly dependent on the clustering algorithm and the distribution of the data. Recently, the seminal work of Gao *et al.* [104] introduced the framework to perform inference after hierarchical clustering when observations are independent and identically distributed as p-dimensional Gaussian random variables with a spherical covariance matrix. This corresponds to the following matrix normal model [127]:

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \sigma^2\mathbf{I}_p), \tag{1.15}$$

where $\mathbf{X}$ is a $n \times p$ matrix whose rows are vectors of features in $\mathbb{R}^p$, the means of the p-dimensional Gaussian vectors are given by the rows of the $n\times p$ matrix $\boldsymbol{\mu}$, the $n\times n$ identity matrix accounts for the independence of observations and $\sigma\mathbf{I}_p$ is the covariance matrix of features for every row. Under (1.15), the authors in [104] defined a p-value that controls the selective type I error, that is, the probability of rejecting equality of cluster means *conditionally to the fact that these clusters have been found*. Besides, the authors showed that asymptotically over-estimating σ yields asymptotic control of the selective type I error, providing a suitable estimator that can be used in practice. Dealing with parameter estimation compatible with selective type I error control is a very remarkable and novel contribution, that had been overlooked in previous relevant works in the field [173, 243]. Recently, this approach was adapted to k-means clustering in [51] and to the feature-level framework, that is, the identification of the variables that contain a true signal, in [121].

Although the contributions [104, 51] are highly relevant from a statistical perspective, their application to realistic problems remains limited. Indeed, the model (1.15) requires the variables to be independent and to have equal variance, which is very unlikely to hold in practice. In particular, the descriptors commonly used in clustering techniques applied to protein structures are usually Euclidean distances between all the C_α atoms along the sequence. Even if these quantities can be considered as p-dimensional Gaussian random variables, their strong correlation prevents the assumption of (1.15). Besides, conformations may exhibit temporal dependence when generated with physical-based approaches, such as MD simulations. Consequently, a model admitting dependence structures between variables and observations is required in this framework. In this Chapter, we extend the

framework presented in [104, 51] to the general matrix normal model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}), \tag{1.16}$$

where $\mathbf{U}$ encodes the dependency between observations and $\boldsymbol{\Sigma}$ the covariance between features. We define a p-value that controls the selective type I error under (1.16) for hierarchical and k-means clustering algorithms. Furthermore, we generalize the over-estimation of σ to the matrix framework, showing that there is a partial order -the so-called Loewner partial order [127]- in the space of Hermitian matrices for which asymptotically over-estimating $\boldsymbol{\Sigma}$ ensures the asymptotic control of the selective type I error. We also provide an estimator of $\boldsymbol{\Sigma}$ that can be used in practice under some assumptions, that we show to be satisfied for several common models of dependence between observations. In addition to illustrating the numerical performance of the presented test with simulations on synthetic data, we demonstrate how the method provides statistical guarantees after clustering real data from protein ensembles whose conformations are featured with Gaussian descriptors.

A contact-based characterization of protein ensembles (Chapter 7)

The last contribution of the thesis is a method for characterizing conformational ensembles of highly-flexible proteins. Existing methods in the literature can be classified into two main families: clustering-based approaches and averaging-based approaches. The first ones propose an interesting idea that integrates the probabilistic nature of disordered ensembles. Furthermore, following the work proposed in Chapter 6, these methods can be provided with statistical guarantees about the differences between clusters. Together with using C_α-C_α distances to feature conformations, states are often compared using root-mean-square deviation (RMSD) [241, 185]. Classifying states in that way tends to group conformations based on good alignments, i.e. whose structures are globally similar. This approach, which makes sense for conformational ensembles of ordered/structured proteins, does not provide suitable characterizations here due to the high conformational variability of the system: IDP conformations do not align well. Consequently, forcing such an alignment does not yield clear and interpretable partitions of the conformational space. On the other hand, averaging-based approaches considerably reduce spatial variability and mask relevant but infrequent structural features of the system. The unsuitability of all these methods is illustrated in Chapter 7.

We propose to address the question of how to properly characterize a disordered ensemble by going back to the origins of structural characterization: contact maps. Contact and distance maps have served as one of the main tools for characterizing the structure of rigid proteins [229, 213, 275], demonstrating their suitability to detect structural domains [250, 164, 255, 137]. They consist on a binary triangular matrix $(C_{ij})_{ij}$, where $C_{ij} = 1$ if the Euclidean distance between the i-th and j-th C_α atoms is below a given threshold, and $C_{ij} = 0$ otherwise. Although they reveal as very useful tools to characterize rigid structures, their naive extension to conformational ensembles, consisting on estimating contact probabilities by averaging binary contacts across every conformation, loses

the contact patterns outside the diagonal that appear for sets of conformations with low occupancy. We believe that contact information should remain the key to characterizing the conformational variability of flexible proteins, but the average-based extension must be replaced by a more intelligent approach to unravel the structural complexity of IDP. The message we propose is clear: use contacts, *but cluster first*.

Chapter 7 introduces WARIO, a contact-based characterization of conformational ensembles of highly flexible proteins. This method exploits the potential of contact maps by wisely integrating the statistical behavior of disordered systems. This is done by first performing a well-adapted clustering algorithm that unravels how residue-residue interactions manifest across the protein dynamic. To do so, we feature conformations by the one-dimensional global structural descriptors (Definition 1.2.7), that is, by a continuous function taking values in $[0, 1]$ that acts as a proxy for the interaction between every residue pair. This function integrates sequence information and the relative orientation between the interacting amino acids, which we show to be crucial to correctly account for the formation of local structural motifs. Then, every group of conformations is described by its representative contact configuration. In short, a conformational ensemble is characterized by a *weighted family of contact maps*, accounting for its structural diversity through a set of contact patterns that appear with a given frequency along the conformational fluctuations of the protein. We illustrate the usefulness of WARIO through four examples of flexible proteins, and we compare it with the classical clustering approaches that use distances to feature conformations.

Part I

Local structural analysis of highly flexible proteins

Chapter 2

Statistical proofs of the interdependence between nearest neighbor effects on local backbone conformations

Backbone dihedral angles ϕ and ψ are the main structural descriptors of proteins and peptides. The distribution of these angles has been investigated over decades as they are essential for the validation and refinement of experimental measurements, as well as for structure prediction and design methods. The dependence of these distributions, not only on the nature of each amino acid but also on that of the closest neighbors, has been the subject of numerous studies. Although neighbor-dependent distributions are nowadays generally accepted as a good model, there is still some controversy about the combined effects of left and right neighbors. We have investigated this question using rigorous methods based on techniques from inferential statistics. Our results unambiguously demonstrate that the influence of left and right neighbors cannot be considered independently. Consequently, three-residue fragments should be considered as the minimal building blocks to investigate polypeptide sequence-structure relationships.

This work has been published in *Journal of Structural Biology, 214(4): 107907, 2022*, with Pau Bernadó, Pierre Neuvial and Juan Cortés. It is presented here with minor changes for the sake of coherence in the manuscript.

Contents

2.1 Introduction

The main variables to locally investigate protein structural and dynamic properties are the backbone ϕ and ψ dihedral angles of each of the amino acid residues along the sequence [39, 175] (see Figure 2.1 for an illustration). The allowed values of this pair of angles and its statistical distribution have been studied over half a century, since the seminal work by Ramachandran et al. [237, 238]. Several applications have motivated the detailed analysis of ϕ and ψ angles in polypeptide chains, such as the validation and refinement of structures determined from biophysical techniques [208, 182] and the development of models or scoring functions for protein structure prediction and design [106, 153, 27, 35, 244, 280]. The study of local structural preferences of polypeptides is also essential for the investigation of denatured states of globular proteins [267, 142] and intrinsically disordered proteins (IDPs) [265, 88].

Each amino acid type has a particular distribution of the ϕ and ψ angles [274, 261, 70, 128, 6]. These distributions are relatively similar for all natural amino acids, with the exception of glycine and proline. While glycine lacks a side chain, thus providing enhanced conformational variability, proline has a cyclic side chain, which severely restricts the accessible ϕ and ψ values [122]. Some early work assumed that the distribution depends only on the amino acid type, independently of the context, which is usually referred to as Flory's isolated-pair hypothesis [99, 321]. Despite its simplicity, Flory's isolated-pair hypothesis has been very useful to interpret Small Angle Scattering data reporting on the overall size of disordered and denatured proteins [161]. However, the availability of residue-specific information provided my Nuclear Magnetic Resonance (NMR) measurements, such as residual/scalar couplings and chemical shifts, evidenced that the conformational preferences of individual amino acid residues is influenced by their nearest neighbors [40, 225]. A wide variety of short peptides have been used in order to rationalize and quantify the effects exerted by the nearest neighbors [66, 216, 217, 149, 281, 260]. The ensemble of these studies identified aromatic and β-branched amino acids as having the strongest influence on the structure of their neighbors due to their steric hindrance [225, 149], although the role of solvation has been also pointed out by some authors [10]. Nearest neighbor ef-

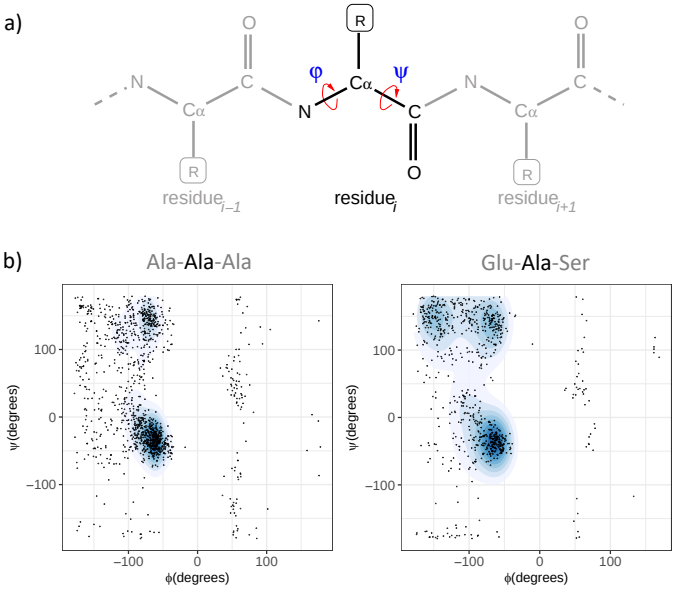


Figure 2.1: (a) Illustration of a three-residue fragment indicating the ϕ and ψ angles of the central residue. Only heavy atoms are represented, and the R corresponds to each amino acid side chain. (b) Distributions of the ϕ-ψ angles for an alanine residue with different neighbors.

fects were found particularly significant in several cases of repeated amino acids along the sequence [204, 203].

Various theoretical/computational approaches, building on experimentally-determined protein structures, have also been developed to investigate sequence-dependent structural preferences and to integrate them within predictive methods [150, 105, 27, 35, 244, 280]. In addition, the inability of simple single-residue-based coil models to recapitulate NMR data supports the influence of the sequence context in defining conformational ensembles of disordered and denatured proteins [266, 222, 142, 143, 21, 129, 88]. Finally mention that molecular dynamics simulations of simple tripeptides showed that the nearest neighbors influence the relative population between the regions of the Ramachandran space and the transitions rates between them [314]. Overall, these experimental and computational studies provide strong evidence for the effect of the nearest neighbors in defining the conformational preferences of a given amino acid residue.

An additional question arises when investigating nearest neighbor dependence: is the influence of left and right neighbors interdependent? This is an important issue as it determines whether the influence exerted by both neighbors can be studied separately. Contradictory answers have been given to this question. For instance, Griffiths et al. [109] postulated that electrostatic interactions between the left and right neighbors significantly affect the conformation of the central residue. Betancourt and Skolnick [27] directly considered three-residue fragments for the analysis of neighbor dependence, thus implicitly

assuming that left and right neighbors cannot be dissociated for this analysis. Results by Huang et al. [129] also suggested that left and right residues have to be simultaneously taken into account in order to appropriately estimate NMR residual dipolar couplings (RDCs) measured in IDPs. Conversely, independence of adjacent residues has been asserted by other authors, although only descriptive or statistically vague methods have been applied in this regard. Rata et al. [244] stated independence after visual comparison of two density estimations, and Shen et al. [265] based their analysis on an amino acid clustering approach, valid only under the independence hypothesis.

We have implemented statistical hypothesis testing methods [172] to investigate the interdependence between nearest neighbor effects on backbone dihedral angles. These statistical tests make it possible to prove the interdependence of neighbor effects, by assessing the independence of two categorical variables. Data for our analyses were extracted from a non-redundant set of experimentally-determined high-resolution protein structures. We constructed two datasets from three-residue fragments (called tripeptides from now on) depending on the structural context: considering all tripeptides in all structures, and considering only fragments from coil regions (i.e. tripeptides not contained in α-helices or β-strands). In the following, we will refer to these datasets as *All* and *Coil*, respectively. We would like to clarify here that although it is well known that statistical models to investigate disordered or unfolded proteins are in general more accurate when they are built from restricted structural datasets that do not contain secondary structure elements [274, 266, 142], we decided to perform the statistical tests for restricted and unrestricted datasets with the aim of analyzing differences.

The chapter is organized as follows:

- In Section 2.2.1, the collection of the tripeptide data is detailed.
- Section 2.2.2 presents the statistical methodology implemented to assess the interdependence of neighboring effects. First, the general testing procedure is detailed. Then, the strategy to evaluate the impact of polarity and size on the strength of interdependence is described.
- In Section 2.3, we present and discuss the results of our analyses, highlighting the statistical significance of the interdependence of neighboring effects. We show that both the physicochemical properties of the nearest neighbors and the structural origin of the data affect the magnitude of the interdependence.

2.2 Methods

2.2.1 Data collection

A database of three-residue fragments (tripeptides) was built from a curated database of experimentally-determined high-resolution protein structures. More precisely, we used protein domains from the SCOPe 2.07 release [49]. In order to remove highly-redundant

sequences, we used the 95% sequence-identity-filtered subset of these domains. In addition to structures determined by X-ray crystallography (with a resolution below 3Å), SCOPe also contains structures from NMR experiments. For each input file from NMR experiments containing more than one model, a distance filter was applied to corresponding tripeptides in each model to avoid repetitions in the database. A tripeptide structure was considered sufficiently different from another one already extracted from the same file, and was thus added to the database, if it met at least one of the two following criteria: the RMSD on ϕ and ψ angles was above 0.2 radians, or one of the dihedral angles differed by more than 0.6 radians. In total, 6,740,433 tripeptide structures were extracted. Tripeptides were classified by sequence (i.e. 8,000 tripeptide classes) and the backbone dihedral angles were collected in a dataset called *All*, since no additional structural criteria were considered for filtering.

A structurally filtered dataset, called *Coil*, was generated by removing tripeptides contained in α-helices or β-strands. For this, DSSP [151, 283] was employed to assign secondary structure labels to each amino acid residue in the structures extracted from the SCOPe database. A tripeptide was included in the filtered subset if none of its three residues had a DSSP code of H or E. Note that π-helices or 3-10-helices, which are relatively rare in our database, were not filtered out because they are usually small and can be observed inserted into coil regions. The secondary structure filtering reduced the number of tripeptide structures to 3,141,877, which is less than half the size of the *All* dataset.

For the analyses in this work, for both *All* and *Coil* datasets, we considered only tripeptides involving peptide bonds in *trans* conformation, which corresponds to the vast majority of the instances. Therefore, tripeptides involving at least one peptide bond in *cis* conformation were removed. We treated the cases of glycine and proline separately. We excluded from the datasets tripeptide sequences for which the number of available structures was very low, and thus not statistically interpretable. The number of required structures depends on each test, and is detailed in Section 2.2.2.

It should be noted that, in order to collect enough data for the analyses, we were less restrictive in the construction of the datasets compared to previous studies (e.g. [244, 280]). Nevertheless, this is acceptable in our case since our aim is not to develop a (differentiable) statistical potential, but to perform statistical tests, and because our implementation of these tests is reasonably resilient to noise. For the sake of rigor, we generated more restrictive (supposedly higher-quality) datasets considering only structures determined by X-ray crystallography with a resolution below 2Å. We performed the same analyses using these datasets, but considering only tripeptides for which the amount of data allowed a correct implementation of the statistical tests. Overall, the analyses (restricted to a small number or tripeptides sufficiently represented in the so filtered datasets) led to the same conclusions regarding the interdependent effects of left and right neighbors. These results are not presented here.

2.2.2 Statistical methodology

Assessing interdependence between left and right neighbors

We aimed at assessing whether the distribution of (ϕ, ψ), which depends on the three amino acids L, R and C, can be separately inferred from the information of L-C and C-R dipeptides, or the information on the tripeptide L-C-R is unavoidably required. Ting et al. [280] stated that, under the hypothesis

$$L \text{ and } R \text{ independent given } C \text{ and } (\phi, \psi), \tag{indep}$$

the probability density of (ϕ, ψ) given the whole tripeptide, $f(\phi, \psi \,|\, L, C, R)$, can be obtained from the information of the densities given by the left and right dipeptides as

$$f(\phi, \psi \,|\, L, C, R) = \frac{f(\phi, \psi \,|\, L, C)\, f(\phi, \psi \,|\, C, R)}{S\, f(\phi, \psi \,|\, C)}, \tag{2.1}$$

where S is a normalization constant. Moreover, Rata et al. [244] proved the reciprocal implication. We have thus the following equivalence:

$$\begin{gathered} L \text{ and } R \text{ independent given } C \text{ and } (\phi, \psi) \\ \Leftrightarrow \\ f(\phi, \psi \,|\, L, C, R) = \frac{f(\phi, \psi \,|\, L, C)\, f(\phi, \psi \,|\, C, R)}{S\, f(\phi, \psi \,|\, C)} \end{gathered} \tag{2.2}$$

which is proved as follows. Letting

$$\frac{1}{S} = \frac{P(L, C)\, \mathrm{P}(C, R)}{\mathrm{P}(C)\, \mathrm{P}(L, C, R)}, \tag{2.3}$$

we have

$$\begin{aligned} \frac{\mathrm{P}(\varphi, \psi \,|\, L, C, R)\, \mathrm{P}(\varphi, \psi \,|\, C)}{\mathrm{P}(\varphi, \psi \,|\, L, C)\, \mathrm{P}(\varphi, \psi \,|\, C, R)} &= \frac{\mathrm{P}(\varphi, \psi, L, C, R)\, \mathrm{P}(\varphi, \psi, C)}{\mathrm{P}(\varphi, \psi, L, C)\, \mathrm{P}(\varphi, \psi, C, R)} \frac{\mathrm{P}(L, C)\, \mathrm{P}(C, R)}{\mathrm{P}(L, C, R)\, \mathrm{P}(C)} \\ = \frac{\mathrm{P}(L, R, \varphi, \psi\,, C)}{\mathrm{P}(\varphi, \psi, C)} \frac{\mathrm{P}(\varphi, \psi, C)}{\mathrm{P}(L, \varphi, \psi, C)} \frac{\mathrm{P}(\varphi, \psi, C)}{\mathrm{P}(R, \varphi, \psi, C)} \frac{1}{S} &= \frac{\mathrm{P}(L, R \,|\, \varphi, \psi, C)}{S\, \mathrm{P}(L \,|\, \varphi, \psi, C)\, \mathrm{P}(R \,|\, \varphi, \psi, C)}\,, \end{aligned}$$

so that the conditional independence of L and R given C and (φ, ψ) is indeed equivalent to (2.1).

If Equation (2.1) is false, then, the probability density of (ϕ, ψ) of a central residue for a given tripeptide cannot be inferred from the information on the corresponding dipeptides (at least via the functional form stated in [280]). According to the equivalence (2.2), disproving hypothesis (indep) is enough to disprove (2.1). In order to test (indep), one can perform a χ^2 independence test between the categorical variables L and R for each fixed value of C and (ϕ, ψ). This requires a proper discretization of the space $\mathbb{T}^2$, in order

to obtain a set of values for (ϕ, ψ) that accurately represent the bi-dimensional random variable and that allow the implementation of the statistical test. Generally, a finer or coarser discretization entails a more or less faithful representation of the angular variable (ϕ, ψ), which ideally is continuous on $\mathbb{T}^2$. Consequently, an optimal discretization procedure will be the thinnest one allowing contingency matrices of the maximum dimension and with a number of points sufficiently large for the independence tests to be performed correctly. We propose three different discretization methods, whose parameters should be optimized. The three methods are based on:

(D1) The choice of a representative set

$$\mathcal{R} = \{(\phi_i, \psi_i)\}_{i \in 1,\dots,N_{\text{rep}}} \subset \mathbb{T}^2 \,.$$

(D2) For each representative value $(\phi_i, \psi_i) \in \mathcal{R}$, the choice of the set of points $\mathcal{R}_i = \{(\phi_{ij}, \psi_{ij})\}_{j \in 1,\dots,J_i}$ for which $(\phi_{ij}, \psi_{ij}) \equiv (\phi_i, \psi_i) \quad \forall\, j \in 1, \dots, J_i$, where $a \equiv b$ means that, in terms of the discretization, a and b belong to the same space subdivision.

The three proposed methods were built as follows and are illustrated in Figure (2.2):

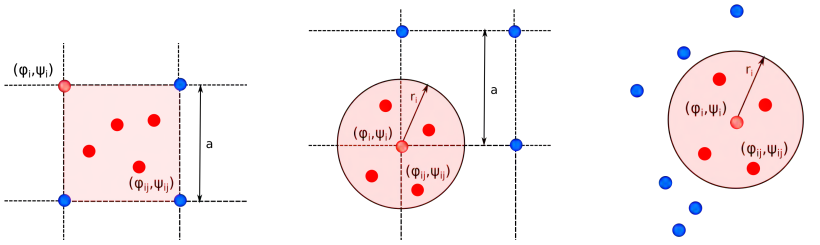
(I) $\mathcal{R}$ is a homogeneous square grid and $\mathcal{R}_i$ are the points belonging to the i-th cell.

(II) $\mathcal{R}$ is a homogeneous square grid and $\mathcal{R}_i$ are the points in the vertex-centered ball $B_{\mathbb{T}^2}\left((\phi_i, \psi_i), r_i\right)$.

(III) $\mathcal{R}$ is a subset of the data set sampled uniformly and without replacement, and the $\mathcal{R}_i$ are the points in the ball $B_{\mathbb{T}^2}\left((\phi_i, \psi_i), r_i\right)$.

For method I, the only parameter is the size $a = 2\pi/\sqrt{N_{\text{rep}}}$ of the square grid. It was chosen as the smallest value allowing maximum dimension contingency matrices with a large enough number of points. Due to physicochemical constraints, the whole $\mathbb{T}^2$ space is not accessible, and thus we limited ourselves to regions with non negligible density. To do so, a grid cell was kept only if it contained a minimum number of data points (i.e. if $J_i \geq N_{\min}$). For the analyses presented here, we chose $N_{\min} = 500$ and $N_{\text{rep}} = 30$.

For methods II and III, the radius r_i of each ball depends on (ϕ_i, ψ_i), and was determined in order to include a specific number $J_i = J$ of points in the ball, the same for all partitions. This allowed a discretization for which each subdivision had the same number of data points, and thus for which all the tests performed were comparable. In order to maintain a certain control on how (ϕ, ψ) values are identified together, a maximum radius R was established and only balls with $r_i < R$ were kept. The number of points J at each ball was chosen to guarantee contingency matrices with maximum dimension while providing a thin and reliable discretization. For method III, the number of representative points N_{rep} was also chosen according to the same considerations. Here, we chose $J = N_{\text{rep}} = 1000$ and $R = 0.1$.



(a) Method I. (b) Method II. (c) Method III.

Figure 2.2: The three proposed discretization methods.

It should be recalled that we do not need to perform tests over the whole $\mathbb{T}^2$ space. As the hypothesis (indep) is *conditional* to C and the continuous random variable (ϕ, ψ), it is equivalent to the hypothesis

$$L \text{ and } R \text{ independent given } C = c \text{ and } (\phi, \psi) = (\phi_0, \psi_0)$$
$$\text{for all amino-acids } c \text{ and all } \phi_0, \psi_0 \in [-\pi, \pi].$$

Therefore, rejecting (indep) means rejecting the independence of L and R for any fixed values of C and (ϕ, ψ). Consequently, implementing tests for a subset of the discretized space will properly answer our question if a significant result is retrieved.

The independence test was performed for the two aforementioned datasets, *All* and *Coil*, using the three proposed discretization methods, whose corresponding parameters were chosen according to the previously specified considerations. Given a central amino acid, one test was performed per point (ϕ_0, ψ_0) of the chosen grid, associating a distribution of p-values to each central residue. For methods II and III, sample sizes were fixed and therefore p-values can be quantitatively compared.As a large number of test was performed, a multiplicity adjustment was implemented [32]. Finally, an overall p-value for each amino acid was defined as the minimum adjusted p-value across the discretization.

Simulation of non-rejecting tests: The intrinsic randomness of discretization method III allows to simulate the proportion of non-rejecting tests for a given central amino acid. For $s = 1, \ldots, N_{\text{sim}} = 100$, we sample a representative set $\mathcal{R}_s$, perform the independence test across $\mathcal{R}_s$ and compute the proportion $\tilde{p}_s$ of p-values higher than a fixed threshold $\alpha = 0.05$. The set of all $\tilde{p}_1, \ldots, \tilde{p}_{N_{\text{sim}}}$ constitute a sample of the proportion of non-rejecting tests for the given amino acid. As p-values are quantitatively comparable, so are the proportions $\tilde{p}_s$. This corresponds to comparisons presented in Figure 2.6(b,d,f), between the two databases, in Section 2.3.

Polarity and size effect on interdependence: AUC score

We assessed whether properties such as the polarity and the size of the neighbors affect the strength of the interdependence. Note that some previous studies on nearest neighbor

effects (not dealing with interdependence) divided amino acid types into only two classes [225]: those involving aromatic and beta-branched side chains (FHITVWY) and the others, with the exception of glycine and proline. The large size of our databases enabled a finer classification. Consequently, we chose six representative amino acids for each of the following groups:

- Polar (P): Arg, Lys, Asp, Glu, Asn, Gln.
- Hydrophobic (H): Ala, Ile, Leu, Met, Phe, Val.
- Small (S): Ala, Ser, Thr, Asp, Asn, Cys.
- Large (L): Phe, Tyr, Trp, Arg, Ile, Lys.

The strategy was to repeat the independence test for all central amino acid types, but restricting the admissible settings of neighbors identities to those in these groups. For polarity (resp. size) we computed (indep) p-values when left and right neighbors belonged to the settings P-P, P-H, H-P and H-H (resp. L-L, L-S, S-L and S-S). However, reducing the number of classes that the categorical variables L and R induces a power loss. In other words, if the information about the variables whose independence we want to assess is trimmed-down, the test will have less information to state any result with the same evidence. Nevertheless, relative comparisons between two groups of p-values for the same number of classes are allowed, and statistically informative.

To facilitate a more direct comparison between settings, we defined a score representing the strength of the interdependence of neighbors in a given configuration. For a given setting C_L-X-C_R, where $C_L, C_R \in \{P, H\}$ (for polarity) or $C_L, C_R \in \{L, S\}$ (for size), let $F_{N_{\text{rep}}}^{C_L,C_R}$ denote the empirical cumulative distribution function (ECDF) of the p-values retrieved after testing hypothesis (indep) across a fixed discretization of size N_{rep}. Then, the Area Under the Curve (AUC) of $F_{N_{\text{rep}}}^{C_L,C_R}$ is defined as

$$\text{AUC}(C_L, C_R) = \sum_{i=1}^{N_{\text{rep}}} (p_{(i+1)} - p_{(i)})\, F_{N_{\text{rep}}}^{C_L,C_R}(p_{(i)}), \tag{2.4}$$

where $p_{(i)}$ is the i-th smallest p-value, for $i = 1, \ldots, N_{\text{rep}}$, and $p_{(N_{\text{rep}}+1)} = 1$. If the AUC for a given setting is close to 1, then the corresponding p-values are concentrated towards zero and, therefore, the statistical evidence that (indep) has to be rejected is high.

2.3 Results and discussion

2.3.1 Influence of the left and right neighbors are statistically interdependent

First, we assessed the significance of combined effects exerted by the nearest neighbors. As detailed in Section 2.2.2, we equivalently evaluated the independence of the left and right amino acid identities given a central residue in a defined conformation (ϕ,ψ). We

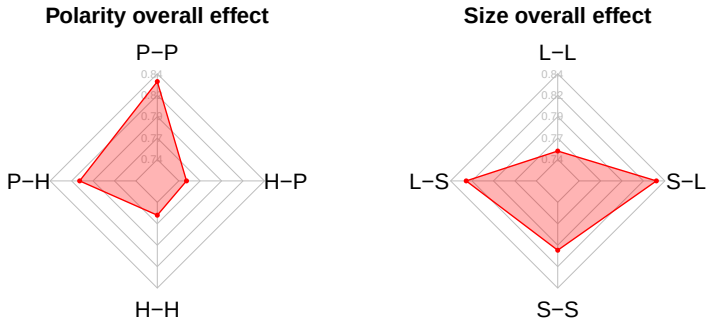


Figure 2.3: Radar plots showing the interdependence score (AUC) between neighbors with different physicochemical properties: (left) polarity/hydrophobicity, (right) size. P, H, L and S stand for polar, hydrophobic, large and small, respectively.

implemented one χ^2 (chi-square) test of independence per central amino acid residue and value of (ϕ,ψ), by discretizing the Ramachandran space and carrying out one test per subdivision. Then, the results of all tests across the discretization grid were summarized by a p-value, which quantifies the plausibility of the observed data assuming independence for each amino-acid type. Hence, p-values close to zero provide strong statistical evidence for the interdependence of the influence of the left and right nearest neighbors. The obtained p-values were lower than 10^{-10} for all amino acid types, from both *All* and *Coil* datasets. This implies the interdependence of the left and right nearest neighbors in determining the (ϕ, ψ) angles of the central residue.

2.3.2 The physicochemical properties of the nearest neighbors affect the magnitude of the interdependence

Figure 2.3 shows radar plots with AUC values for possible combinations of neighbors depending on their polarity (left plot) and size (right plot), and averaged for all amino acid types at the central position. General trends can be observed from this representation: (i) Interdependence is stronger when both neighbors are polar. This can be justified by the presence of attractive or repulsive electrostatic interactions between them, which may constrain the conformational space for the central residue, and that strongly depend on the specific pair of neighbors [204]. Adjacent charged amino acids can also modify the solvation energy and perturb the central residue [10]. (ii) Interdependence is weaker when both neighbors are large. This observation is less intuitive. A possible explanation would be that when both neighbors are large (regardless of their type), the conformational space of the central residue is more constrained [55], and thus, other effects due to the nature of each neighbor are less visible. The contrary occurs when at least one of the neighbors is small, as the central residue exhibits a less constrained conformational space.

Nevertheless, exceptions to these above-described general trends emerged when the

strength of the interdependence was analyzed individually for each amino acid. Case-by-case results comparing the four settings for each central residue are shown in Figures 2.4 (for polarity) and 2.5 (for size). The AUC values for each group were also compared with the "free" setting, for which no physicochemical properties are imposed to neighbors (i.e. considering all possible neighbors). They illustrate the very diverse degrees of interdependence depending on the central amino acid and the properties of the nearest neighbors, which highlights the need to take into account (at least) three-residue fragments to locally describe backbone conformational preferences. Due to how the score has been defined, one must not compare AUC values between different individual plots, but only inside each plot. All the differences between AUC scores were statistically significant.

Our analyses performed on the *Coil* dataset showed that, for 14 out of the 20 amino acids, dependence is stronger when both neighbors are polar than when they are both hydrophobic. With respect to size effect, for 16 out of 20 central residues dependence was found stronger when both neighbors were small than when they were large. No relationship was found between amino acids not following both general trends. However, more detailed analyses showed that amino acids that did not follow the general trend were among those for which the amount of data was more limited. This may suggest that with additional data, the general trend would probably be more widely satisfied. With respect to mixed neighbors settings, no clear general trend was found among all central amino acids. In all cases, all the corresponding AUC scores were significantly different to the "free" setting ones, showing that both polarity and size do affect interdependence also when neighbors have mixed properties. Moreover, all plots in Figures 2.4 and 2.5 were strongly asymmetrical with respect to the vertical axis, which evidences that polarity and size effects have a non-negligible directional component.

2.3.3 Combined neighbor effects are stronger in coil regions

We implemented two approaches to quantitatively assess whether neighbor interdependence is influenced by the structural origin of the datasets. The first approach lies in comparing the computed p-value distributions for each dataset (*All* and *Coil*). Here, a distribution of p-values is associated to each central amino acid (one test is performed at each point of its discretized Ramachandran space, recall Section 2.2.2 for details). Moreover, sample sizes are fixed for both datasets and thus p-values for *All* and *Coil* are now quantitatively comparable. Consequently, we can compare each pair of distributions and evaluate whether the interdependence of neighbors is stronger in one of the two data sets. Three representative examples of this comparison are shown in Figure 2.6(a,c,e) with alanine, glutamic acid and leucine as the central residue. p-value density estimates show that the independence hypothesis is more significantly rejected (i.e. interdependence is stronger) for the *Coil* dataset (Kolmogorov-Smirnov test states highly significant discrepancies). Note that the scales in Figure 2.6 vary between the three amino acids in order to better reflect the different behaviour of *All* and *Coil* distributions. Comparisons between

different plots are not really relevant.

The second approach simulates the proportion of statistically non-significant tests for both datasets (the lower this proportion, the more interdependent the left and right neighbors). Figure 2.6(b,d,f) exemplifies this approach using the same central residues. These figures shows that the simulated distribution corresponding to the *Coil* dataset is significantly closer to zero than that of the *All* dataset, substantiating our aforementioned conclusion. This observation contradicts the statements in [244], who suggested that the correlated effects of left and right neighbors were weak, especially in coil regions. Nevertheless, their statements about the lack of interdependence were based on vague statistical analyses compared to the rigorous statistical approach presented above.

2.4 Conclusions

We have investigated local sequence effects on the distribution of the ϕ-ψ angles, which are the main descriptors of polypeptide conformations, using rigorous statistical methods on datasets built form experimentally-determined high-resolution protein structures. Results of our analyses corroborate the large amount of experimental and computational studies describing the influence of the nearest neighbors, thus providing additional evidence for the rejection of Flory's isolated-pair hypothesis, including in disordered regions. Furthermore, our results unambiguously demonstrate coupled effects of the left and right neighbors, which cannot be considered independently of each other. This observation clarifies questions still open on this subject, and represents a fundamental step to understand sequence-structure relationships in peptides and proteins.

These results also have several direct implications for methodological developments in the context of molecular modeling and protein design. The most obvious one concerns sampling algorithms that use ϕ-ψ distributions to model flexible regions in proteins, such as loops or intrinsically disordered regions [21, 219, 88, 13]. More accurate conformational ensemble models will be obtained when explicitly considering coupled neighbor dependencies. The parameterization of the constants associated with backbone torsion angles in force-fields used for molecular dynamics simulations, and more particularly in the case of coarse-grained models, would also benefit from protocols that consider the local sequence context (i.e. going beyond residue-specific parameterization). Regarding structure prediction algorithms applied to globular proteins, although modern machine-learning-based algorithms mostly exploit evolutionary-conserved pairwise residue contacts, the incorporation of local structural constraints and preferences are crucial to obtain accurate solutions [147]. Thus, our observations suggest that the performance of these algorithms could be improved by explicitly considering triplets of consecutive amino acids for the conception of the neural network architecture.

This work focused on studying the interdependent effects of the nearest neighbors along the sequence (i.e. residues $i \pm 1$). It would be very interesting extend the analysis to more distant neighbors ($i \pm n$, with $n = 2, 3, 4, \dots$). Unfortunately, the amount of experimental

data currently available does not allow such an analysis in a general case. With the increase of available data from experimental techniques and/or high quality models generated by simulation or structural prediction methods, such an analysis seems feasible in the near future. It should be noted, however, that non-trivial mathematical challenges would also arise in addressing this question, which would require new methodological developments.

Software availability

The code implementing the statistical tests described in this chapter as well as the datasets are freely available:

- Software: https://gitlab.laas.fr/moma/STINA,
- Data: https://moma.laas.fr/static/data/tripeptide_angles_data.tar.

Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) through grant ANR-19-P3IA-0004, the LabEx CIMI (ANR-11-LABX-0040) and EpiGen-Med (ANR-10-LABX-12-01) within the French State Programme “Investissements d’Avenir”, and by the European Research Council under the European Union’s H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030] awarded to Pau Bernadó.

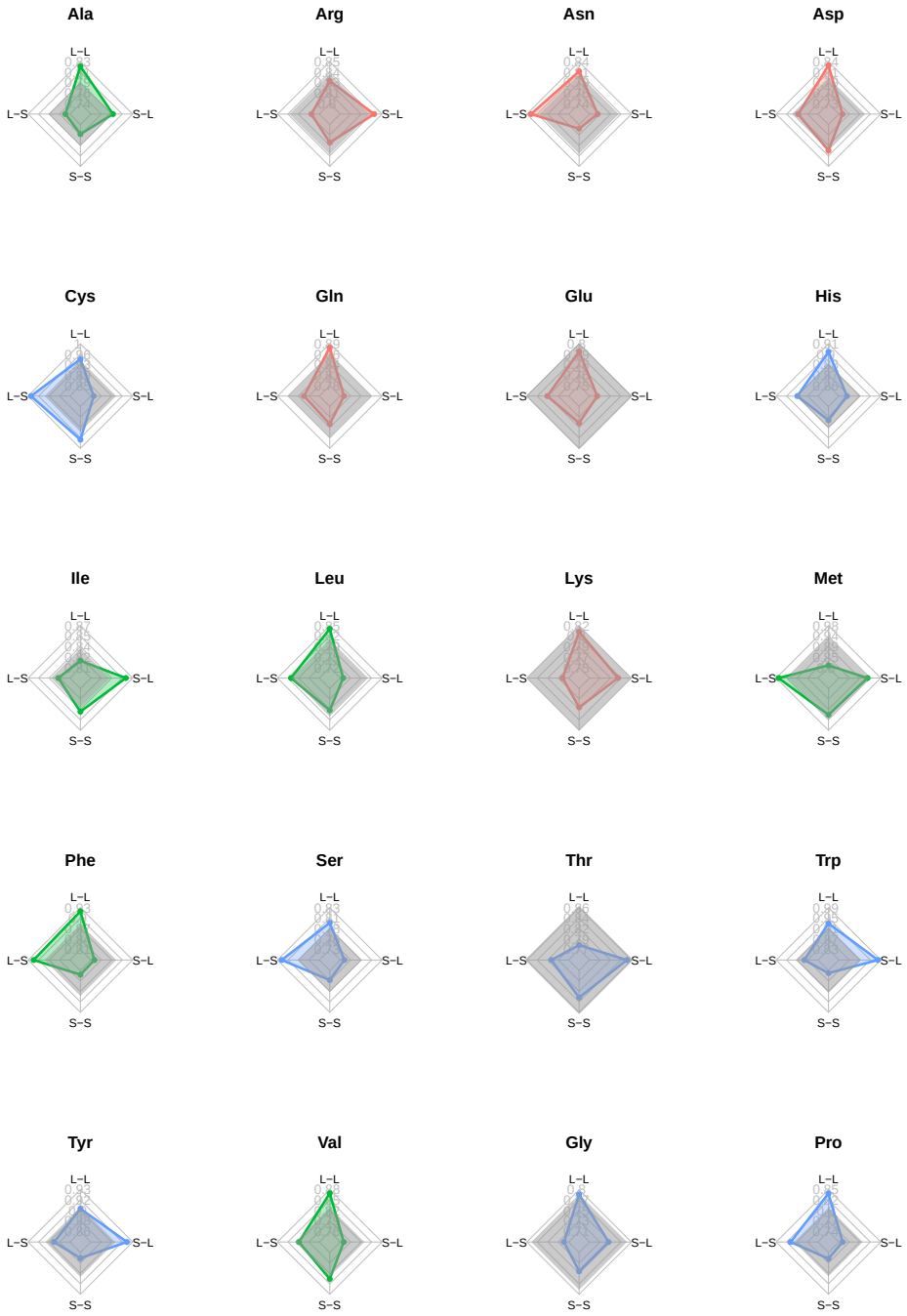


Figure 2.4: For each central amino acid type, comparison of the interdependence score (AUC) between the four possible polarity combinations for neighbors (where P stands for polar and H for hydrophobic). In gray, the same score when no physicochemical properties are imposed. Polar and hydrophobic central residues correspond to red and green plots respectively. Blue plots correspond to central residues not belonging to any of both categories.

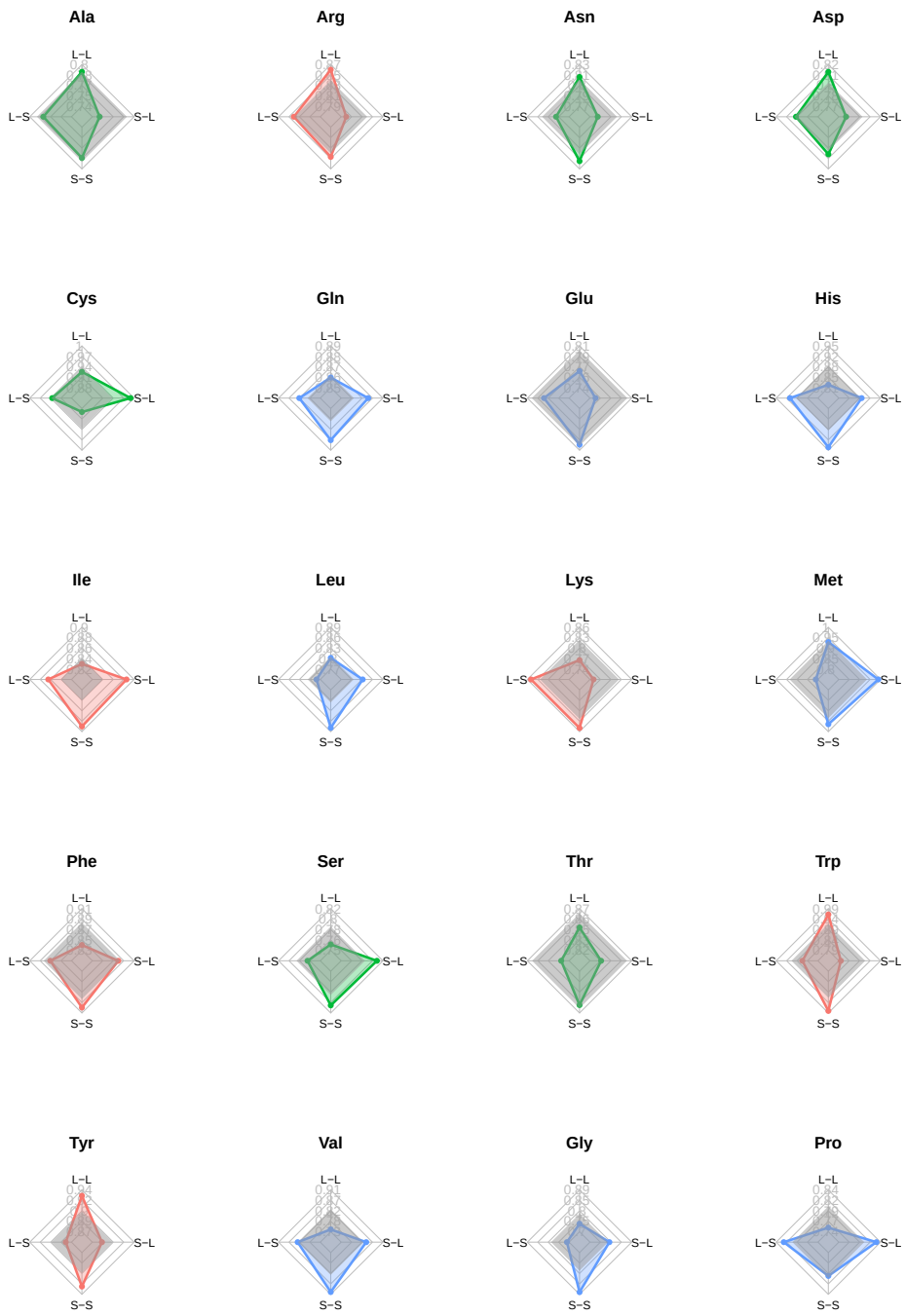


Figure 2.5: For each central amino acid type, comparison of the interdependence score (AUC) between the four possible size combinations for neighbors (where L stands for large and S for small). In gray, the same score when no physicochemical properties are imposed. Large and small central residues correspond to red and green plots respectively. Blue plots correspond to central residues not belonging to any of both categories.

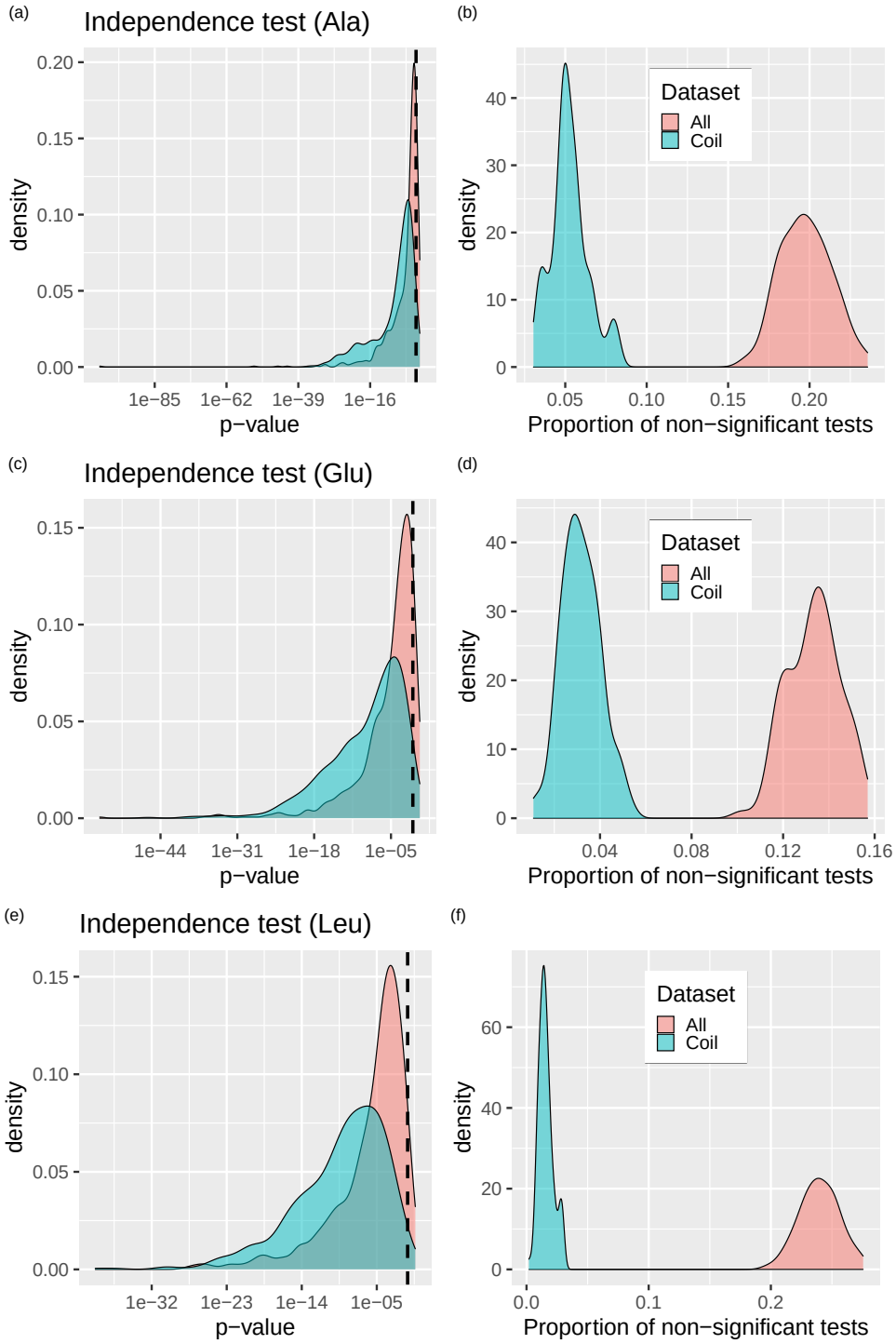


Figure 2.6: (a,c,e) Distribution of (-log10-scaled) p-values for the independence hypothesis tests performed on the *All* (red) and *Coil* (blue) datasets, for a fixed central residue. Dashed line indicates a level of significance of 0.05. (b,d,f) Distribution of the proportion of non-significant tests for a fixed central residue for the *All* and *Coil* datasets.

Chapter 3

Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance

Local protein structure is defined by two variable dihedral angles that take values from probability distributions on the flat torus. The goal of this chapter is to provide the space $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$ with a metric that quantifies local structural modifications due to changes in the protein sequence, and to define associated two-sample goodness-of-fit testing approaches. Due to its adaptability to the geometry of the underlying space, we focus on the Wasserstein distance as a metric between distributions. We extend existing results of the theory of Optimal Transport to the d-dimensional flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, in particular a Central Limit Theorem for the fluctuations of the empirical optimal transport cost. Moreover, we propose different approaches for two-sample goodness-of-fit testing for the one and two-dimensional case, based on the Wasserstein distance. We prove their validity and consistency and assess their performance by numerical experiments on synthetic data and protein structure data. The tests developped in this chapter are implemented in the R package `torustest`.

This work has been published in *Electron. J. Statist., 17(1): 1547–1586, 2023,* with Alberto González-Sanz, Juan Cortés and Pierre Neuvial. It is presented here with minor changes for the sake of coherence in the manuscript.

Contents

3.1 Introduction

When it comes to measure the distance between two probability distributions, the well known Wasserstein distance, derived from the theory of Optimal Transport (OT), provides both strong theoretical guarantees –it metrizes weak convergence [299]– and attractive empirical performance [228]. Most of the applications of such theory are related to the very active field of machine learning, notably in the framework of generative networks [9], robustness [262] or fairness [76], among others.

From a statistical point of view, one of the main caveats of the theory of OT comes from the *curse of dimensionality*: the rate of convergence of the empirical Wasserstein distance decreases as $n^{-1/d}$ with the dimension d [100]. Another important issue is the asymptotic behavior of the fluctuations of the empirical optimal transport cost. For probability measures supported in $\mathbb{R}^d$, it has been proved, using Efron–Stein's inequality that, for the cost L^2, the difference $\sqrt{n}(\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q))$ is asymptotically Gaussian [77]. Recently, the proofs have been extended to some general costs in $\mathbb{R}^d$, including the cost L^p, for $p > 1$ [74]. Concerning statistical goodness-of-fit tests based on Wasserstein distance, the one-sample case has already been addressed in [114] and, when the probability distributions are defined over $\mathbb{R}$, two-sample tests can be derived from [72, 209].

In this chapter, we focus on the d-dimensional flat torus $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ where, even from the purely theoretical point of view, OT has not been completely addressed, besides the work in [60], [193] or, more recently, in [187]. However, this space appears naturally when the probability measures are periodic (e.g. for distributions of angles). The main objective of this chapter is (1) to extend recent existing OT results to the space of probability measures on the flat torus $\mathcal{P}(\mathbb{T}^d)$, especially a Central Limit Theorem (CLT) for the fluctuations of the empirical optimal transport cost, and (2) to address in particular the two-dimensional case, by constructing two-sample goodness-of-fit tests based on the Wasserstein distance.

The analysis of (ϕ, ψ) distributions has several important applications, such as the validation or refinement of protein structures determined from biophysical techniques [208, 182], the prediction of some biophysical measurements to complement experiments

[265], and the development of potential energy models or scoring methods for protein structure modeling, prediction and design [27, 244, 280]. In this context, the definition of a suitable distance between distributions on $\mathbb{T}^2$ is essential. This would allow to quantify the expected magnitude of structural effects associated with local changes in the sequence, and therefore to develop improved versions of the aforementioned modeling and prediction techniques. Nevertheless, this has not been done satisfactorily in previous works. For example, significant differences between two laws are stated after visual comparison of two empirical distributions in [244] and [265], and the Hellinger distance is used to compare distributions on a non-periodic $[-\pi, \pi] \times [-\pi, \pi]$ in [280]. Powerful statistical tests remain to be defined and implemented in order to state such differences, being based on a metric that takes geometry into consideration. As many other commonly-used metrics, Hellinger distance ignores the underlying geometry of the space. Here, we propose to use the Wasserstein distance, whose advantageous geometrical and mathematical properties are described in [228], [298] and [299], to define goodness-of-fit testing techniques for two measures on $\mathbb{T}^2$, allowing a more accurate study of the distribution of protein local conformations.

This chapter is organized as follows:

- Section 3.2 starts by introducing the general framework of measures on the flat torus in general dimension, followed by the precise formulation of the optimal transport problem. Section 3.2.1 is devoted to the study of the shape of the solutions, recalling that they are the gradients of periodic convex functions and showing the uniqueness of the potential in Corollary 3.2.2. Section 3.2.2 proves through Theorem 3.2.5 that the optimal transport potentials converge, up to an additive constant, when the measures converge weakly. This result implies that the method of [77] based on Efron–Stein's inequality can be applied to derive a Central Limit Theorem, see Theorem 3.2.6 in Subsection 3.2.3. Finally, we show how the previously defined CLT does not allow the definition of an asymptotic test.
- Section 3.3 shows how Wasserstein distance can be used to define two-sample goodness-of-fit tests in the two-dimensional flat torus. We propose two testing approaches. The first one, introduced in Subsection 3.3.1, consists in testing the equality of two measures projected into a finite number of closed geodesics on $\mathbb{T}^2$. The second, presented in Subsection 3.3.2, is a conservative procedure based on upper-bounding the exact p-values. This is possible thanks to a concentration inequality given in Theorem 3.3.7, together with faster convergence rates for the expectation.
- Section 3.4 reports numerical experiments illustrating the relevance of these theoretical results, first with synthetic data and then with real data from protein structures, showing that our methods behave well in both cases.

To facilitate reading, the proofs are relegated to Appendix A, but in some cases the intuition behind the proof is provided in the main text for clarity.

3.2 Optimal transport in $\mathbb{R}^d/\mathbb{Z}^d$

Let $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ be defined as the quotient space derived from the equivalence relation $\mathbf{x}\mathcal{R}\mathbf{y}$ if $\mathbf{x}-\mathbf{y} \in \mathbb{Z}^d$. For each $\mathbf{x} \in \mathbb{R}^d$ we denote as $\boldsymbol{x} \in \mathbb{T}^d$ its equivalence class and reserve the notation τ for the canonical projection map $\mathbf{x} \mapsto \tau(\mathbf{x}) = \boldsymbol{x}$. The topology of the quotient space is defined as the finest one that makes τ continuous. With this topology, the space $\mathbb{T}^d$ is a Polish space with the distance derived from the Euclidean norm $\|\cdot\|$,

$$d(\boldsymbol{x}, \boldsymbol{y}) := \inf_{\mathbf{p}\in\mathbb{Z}^d} \|\mathbf{x}-\mathbf{y}+\mathbf{p}\|. \tag{3.1}$$

Note that the last claim is true since the projection map τ is in fact a metric identification, $(\mathbb{R}^d, \|\cdot\|)$ is a Banach space and $\mathbb{Z}^d$ is a closed subset, then it is complete, metrizable through d and separable.

Set $p > 1$. For two probability measures $P, Q \in \mathcal{P}(\mathbb{T}^d)$, a probability measure $\pi \in \mathcal{P}(\mathbb{T}^d \times \mathbb{T}^d)$ is said to be an *optimal transport plan for the cost* d^p between P and Q if it solves

$$\mathcal{T}_p(P, Q) := \inf_{\gamma\in\Pi(P,Q)} \int_{\mathbb{T}^d\times\mathbb{T}^d} d^p(\boldsymbol{x}, \boldsymbol{y}) d\gamma(\boldsymbol{x}, \boldsymbol{y}), \tag{3.2}$$

where $\Pi(P, Q)$ is the set of probability measures $\gamma \in \mathcal{P}(\mathbb{T}^d \times \mathbb{T}^d)$ such that $\gamma(A \times \mathbb{R}^d) = P(A)$ and $\gamma(\mathbb{T}^d \times B) = Q(B)$ for all Borel measurable subsets A, B of $\mathbb{T}^d$.

The Kantorovich problem (3.2) can be formulated in a dual form, as follows

$$\mathcal{T}_p(P, Q) = \sup_{(f,g)\in\Phi_p(P,Q)} \int_{\mathbb{T}^d} f(\boldsymbol{x}) dP(\boldsymbol{x}) + \int_{\mathbb{T}^d} g(\boldsymbol{y}) dQ(\boldsymbol{y}), \tag{3.3}$$

where

$$\Phi_p(P, Q) = \{(f, g) \in L_1(P) \times L_1(Q) : \ f(\boldsymbol{x}) + g(\boldsymbol{y}) \le d^p(\boldsymbol{x}, \boldsymbol{y}) \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{T}^d\}.$$

The element $\psi \in L_1(P)$ is said to be an *optimal transport potential from P to Q for the cost* d^p if there exists $\varphi \in L_1(Q)$ such that the pair (ψ, φ) solves (3.3). Recall from [299] that the solutions of (3.3) are pairs (f, g) of d^p-conjugate d^p-concave functions. This means that

$$f(\mathbf{x}) = \inf_{\boldsymbol{y}\in\mathbb{T}^d} \{d(\boldsymbol{x}, \boldsymbol{y})^p - g(\boldsymbol{y})\} \text{ and } \ g(\boldsymbol{y}) = f^{d^p}(\boldsymbol{y}) = \inf_{\boldsymbol{x}\in\mathbb{T}^d} \{d(\boldsymbol{x}, \boldsymbol{y})^p - f(\boldsymbol{x})\}. \tag{3.4}$$

Furthermore, since $\mathbb{T}^d$ is a Polish space, then Theorem 4.1 in [299] implies that there exists a solution π^* of (3.2). Additionally, Theorem 5.10 in [299] establishes that $\mathrm{supp}(\pi^*)$ is *d^p-cyclically monotone*. This means that for any finite sequence $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^n \subset \mathrm{supp}(\pi^*)$ and any bijection $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$, the following inequality holds:

$$\sum_{k=1}^n d^p(\boldsymbol{x}_k, \boldsymbol{y}_k) \le \sum_{k=1}^n d^p(\boldsymbol{x}_k, \boldsymbol{y}_{\sigma(k)}).$$

Note that, if Q is a probability measure in $\mathbb{T}^d$, its support is defined as the closed set $\text{supp}(Q) \subset \mathbb{T}^d$ composed by $\boldsymbol{x} \in \mathbb{T}^d$ such that for any neighborhood $\mathcal{U}_{\boldsymbol{x}}$ of $\boldsymbol{x}$ it holds that $Q(\mathcal{U}_{\boldsymbol{x}}) > 0$. The interior of the support is denoted by $\mathcal{X}_Q$.

With the same obvious notation we can define a $\|\cdot\|^p$-*cyclically monotone* set. Note that for $p = 2$, $\|\cdot\|^2$-cyclical monotonicity is equivalent to the concept of cyclical monotonicity in convex analysis, described in [246]. Recall that a set $A \subset \mathbb{R}^d \times \mathbb{R}^d$ is *cyclically monotone* if for every finite sequence $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^n \subset A$ and every bijection $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ it holds that

$$\sum_{k=1}^{n} \langle \mathbf{x}_k, \mathbf{y}_k \rangle \geq \sum_{k=1}^{n} \langle \mathbf{x}_k, \mathbf{y}_{\sigma(k)} \rangle.$$

Consequently, the concept of d^p (resp. $\|\cdot\|^p$) -cyclical monotonicity is the natural generalization, to other spaces and costs, of cyclical monotonicity.

In some cases, that we will study later on, there exists some measurable map T such that the optimal transport plan π satisfies $\pi = (I \times T)\#P$, where the symbol $T\#P$ denotes the *push forward measure* of P through T, which is defined by $T\#P(A) := P(T^{-1}(A))$, for all measurable $A \subset \mathbb{T}^d$, and I denotes the identity map. Therefore, the problem becomes equivalent to the following *Monge* formulation:

$$\mathcal{T}_p(P, Q) = \inf_{T\#P=Q} \int_{\mathbb{T}^d} d^p(\boldsymbol{x}, T(\boldsymbol{x}))dP(\boldsymbol{x}). \tag{3.5}$$

3.2.1 Existence of $\|\cdot\|^p$-cyclically monotone mappings

A cyclically monotone map is the natural generalization of a non decreasing function in the real line (as being the gradient of a convex function, see [246]). Cyclical monotonicity provides a powerful tool for statistical studies, see [114, 73, 53] among others. The existence of cyclically monotone maps between probability measures in $\mathbb{R}^d$ has been investigated, in parallel, by [63] and [41], with the restrictive assumption of finite second order moment, relaxed in [194]. For periodic measures, the celebrated result of [60] showed the existence of such maps. The concept of cyclically monotone map also appears naturally when solving an optimal transport problem with quadratic cost in $\mathbb{R}^d$. Therefore, for any potential cost $\|\cdot\|^p$, the natural generalization is the one of $\|\cdot\|^p$-cyclically monotone. In fact, the existence of a $\|\cdot\|^p$-cyclically monotone mapping between probability measures with finite moment of order $p > 1$ was proved in [102]. To the authors' knowledge, no previous work has dealt with the existence of $\|\cdot\|^p$-cyclically monotone mappings between periodic probability measures. Consequently, the main result of this section is Theorem 3.2.1, which shows the existence and uniqueness of a $\|\cdot\|^p$-cyclically monotone preserving map $\mathbf{S}_p$ between periodic measures, for $p > 1$, and relates it with the solution of (3.5). Then, Theorem 3.2.2 guarantees, under certain assumptions of regularity on the support of P, that the solution of (3.3) is unique up to an additive constant.

Note that, in practice, a probability $P \in \mathcal{P}(\mathbb{T}^d)$ defines a periodic measure $\mu_P \in \mathcal{M}(\mathbb{R}^d)$ w.r.t. any $\mathbf{p} \in \mathbb{Z}^d$. In other words, $T_{\mathbf{p}}\#\mu_P = \mu_P$, for all $\mathbf{p} \in \mathbb{Z}^d$, where $T_{\mathbf{p}} : \mathbb{R}^d \to \mathbb{R}^d$

is the shift operator $\mathbf{x} \mapsto \mathbf{x} + \mathbf{p}$. A measure μ_P is *periodic* if it is the natural extension of some probability measure $P \in \mathcal{P}(\mathbb{T}^d)$. As anticipated, the goal of this section is to show the existence of $\|\cdot\|^p$-cyclically monotone mappings between two periodic measures $\mu_P, \mu_Q \in \mathcal{M}(\mathbb{R}^d)$ absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, denoted as $\mu_P, \mu_Q \ll \ell_d$. As commented before, [60] established the existence of a $\|\cdot\|^2$-cyclically monotone map (which a.s. is the gradient of a convex function φ) such that $\nabla\varphi\#\mu_P = \mu_Q$. Theorem 1.25 in [256] entails that there is a unique solution of the Monge problem in the torus, described by the relation $T = \mathbf{x} - \nabla f(\mathbf{x})$, where the sum is to be intended modulo $\mathbb{Z}^d$ and f is an optimal transport potential for the quadratic cost. Note that this is a quite similar relation (between potentials and transport) to the one in the quadratic transport problem in $\mathbb{R}^d$.

The proof of Theorem 3.2.1 starts by realizing that since $\mathbb{T}^d$ is a Polish space, then Theorem 4.1 in [299] implies that there exists a solution π^* of (3.2). Furthermore, Theorem 5.10 in [299] establishes that $\text{supp}(\pi^*)$ is d^p-cyclically monotone, which implies that the set

$$\Gamma = \{(\mathbf{x}+p, \mathbf{y}+p) : \ (\boldsymbol{x}, \boldsymbol{y}) \in \text{supp}(\pi^*), \ \mathbf{x} \in [0,1]^d, \ d(\boldsymbol{x}, \boldsymbol{y}) = \|\mathbf{x}-\mathbf{y}\| \text{ and } p \in \mathbb{Z}^d\} \quad (3.6)$$

is cyclically monotone. Corollary 3.5 in [102] implies that this cyclically monotone set is contained in the graph of a $\|\cdot\|^p$-differential

$$\partial^{\|\cdot\|^p}\varphi_p(\mathbf{x}) = \{\boldsymbol{y} : \ \varphi_p(\mathbf{z}) \leq \varphi_p(\mathbf{x}) + \|\mathbf{z}-\mathbf{y}\|^p - \|\mathbf{x}-\mathbf{y}\|^p, \text{ for all } \mathbf{z} \in \mathbb{R}^d\}$$

of a $\|\cdot\|^p$-concave function φ_p (defined as in (3.4) but replacing d^p with $\|\cdot\|^p$). In conclusion, the a.s. uniqueness of this $\|\cdot\|^p$-differential ends the proof.

Theorem 3.2.1. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures such that $\mu_P \ll \ell_d$. Then, there exists a unique solution $\mathbf{T}_p$ of (3.5). Moreover, there exists a μ_P-a.e. defined $\|\cdot\|^p$-cyclically monotone map $\mathbf{S}_p$ such that*

- *the relation $\mathbf{T}_p \circ \tau = \tau \circ (\mathbf{S}_p)$ holds μ_P-almost surely,*
- *and $\mathbf{S}_p\#\mu_P = \mu_Q$.*

The following result gives the uniqueness, up to additive constants, of the optimal transport potential, where the assumptions are given with respect to its associated periodic measures. In particular, we need to have *negligible boundary* of μ_P which means that the boundary of its support has Lebesgue measure 0, $\ell_d(\partial\,\text{supp}(\mu_P)) = 0$. The proof investigates the intrinsic relation between the optimal transport potentials and the previously described $\mathbf{T}_p$, which allows the use of general results for the uniqueness of $\|\cdot\|^p$-concave functions (see [74]) which have the same gradient a.s. in a connected domain of $\mathbb{R}^d$.

Theorem 3.2.2. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures with connected support such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$ with negligible boundary. Then, there exists a unique, up to an additive constant, d^p-concave function f_p solution of (3.3).*

The assumption of connected support can be relaxed, via [270, Theorem 2], to the setting where both measures have disconnected support. If the supports of μ_P and μ_Q decompose into closures of connected open components

$$\operatorname{supp}(\mu_P) = \bigcup_{i\in\mathcal{I}} \mathcal{X}_{i,\mu_P}, \quad \operatorname{supp}(\mu_Q) = \bigcup_{j\in\mathcal{J}} \mathcal{X}_{j,\mu_Q}, \tag{3.7}$$

where $\mathcal{I}$ is finite index set and $\mathcal{J}$ is a countable index set, then, assuming for all non-empty proper $\mathcal{I}' \subset \mathcal{I}$ and $\mathcal{J}' \subset J$ that

$$\sum_{i\in\mathcal{I}'} \mu_P(\mathcal{X}_{i,\mu_P}) \neq \sum_{j\in\mathcal{J}'} \mu_Q(\mathcal{X}_{j,\mu_Q}), \tag{3.8}$$

it follows by [270, Lemma 5] that no degenerate transport plan exists. Hence, invoking Theorem 2 in [270] in conjunction with Theorem 3.2.2, yields an extension of the uniqueness result to measures with disconnected support.

Corollary 3.2.3. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$ with negligible boundary where* (3.7) *and* (3.8) *hold. Then, there exists a unique, up to an additive constant, d^p-concave function f_p solution of* (3.3).

The importance of Corollary 3.2.3 mainly lies in that it enables the study of the asymptotic behavior of the potential, allowing us to apply Arzelá-Ascoli like reasoning, as explained in the following section.

3.2.2 Asymptotic behaviour

This section deals with the asymptotic properties of the transport map and potentials. We consider two sequences of probability measures $\{\alpha_n\}_{n\in\mathbb{N}}, \{\beta_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{T}^d)$ converging weakly to P and Q respectively,

$$\alpha_n \xrightarrow{w} P \quad \text{and} \quad \beta_n \xrightarrow{w} Q.$$

Since $\mathbb{T}^d$ is compact, here the weak convergence is in the sense that for every continuous function $h \in \mathcal{C}(\mathbb{T}^d)$, $\int h(\boldsymbol{x}) d\alpha_n(\boldsymbol{x}) \to \int h(\boldsymbol{x}) dP(\boldsymbol{x})$. Once again, thanks to that compactness the existence of moments of any order is always fulfilled for $P \in \mathcal{P}(\mathbb{T}^d)$. As a consequence, Theorem 7.12 in [298] implies that $\alpha_n \xrightarrow{w} P$ if and only if the *p-Wasserstein distance* $\mathcal{W}_p(\alpha_n, P) := (\mathcal{T}_p(\alpha_n, P))^{\frac{1}{p}}$ tends to 0. An analogous reasoning implies the convergence $\mathcal{T}_p(\alpha_n, \beta_n) \to \mathcal{T}_p(P, Q)$ for the two-sample case.

The idea of this section is to take advantage of the fact that any d^p-concave function f is continuous whereby it is finite. Moreover, it has bounded continuity modulus, so we can apply Arzelá-Ascoli's Theorem by fixing the constants.

Lemma 3.2.4. *Every d^p-concave function f is Lipschitz (in its definition domain* $\operatorname{dom}(f)$*) with constant $L = 2\,p\,d^{\frac{p-1}{2}}$, with respect to the metric* (3.1).

The proof of the next Theorem first proceeds by choosing the sequence $\{a_n\}_{n\in\mathbb{N}}$ to guarantee the uniform boundedness of the sequence $\{(f_n, g_n)\}_{n\in\mathbb{N}}$ of solutions of (3.3). This, together with Lemma 3.2.4 and Arzelá-Ascoli's Theorem, implies that $\{(f_n, g_n)\}_{n\in\mathbb{N}}$ is relatively compact. The uniqueness of solutions of (3.3), described in Theorem 3.2.2, allows us to conclude.

Theorem 3.2.5. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures with connected supports whose associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$ with negligible boundary. Let $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\beta_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{T}^d)$ be two sequences of probability measures converging weakly to P and Q respectively. Denote by (f_n, g_n) (resp. (f, g)) the solution of the dual problem between α_n and β_n (resp. P and Q). Then there exists a sequence of real numbers $\{a_n\}_{n\in\mathbb{N}}$ such that $f_n + a_n \to f$ uniformly on the compact sets of $\mathcal{X}_P$.*

3.2.3 Asymptotic normality

This section is devoted a proof of a Central Limit Theorem (CLT) for the fluctuations of the empirical optimal transport cost. Recall that the previous section proves that, under certain regularity assumptions, there exists a unique optimal transport potential from P to Q. Let f_p be such a potential. We will use Efron-Stein's inequality to derive that

$$\sqrt{n}\left(\mathcal{T}_p(P_n, Q) - \mathbb{E}\mathcal{T}_p(P_n, Q)\right) \xrightarrow{w} N(0, \sigma_p^2(P, Q)),$$

with

$$\sigma_p^2(P, Q) = \mathrm{Var}(f_p(X)). \tag{3.9}$$

Then, we will see that the same holds in the two sample case. The idea is not new: it has already been used with the same goal in [77] for the quadratic cost in $\mathbb{R}^d$, and in its extension to general costs in [74]. Moreover, when using regularized optimal transport, [200] showed that the same technique can be applied. A similar result, but using the idea in [75] of differentiating the supremum in the functional sense by applying the general result of [65], yields also a CLT on the torus for $p \geq 2$, see [134].

Theorem 3.2.6. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures with connected supports such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$ with negligible boundary. Then, for any $p > 1$, we have*

$$\sqrt{n}\left(\mathcal{T}_p(P_n, Q) - \mathbb{E}\mathcal{T}_p(P_n, Q)\right) \xrightarrow{w} N(0, \sigma_p^2(P, Q)),$$

and, if $m = m(n)$ satisfies that $m \longrightarrow +\infty$ and $\frac{n}{n+m} \to \lambda \in (0, 1)$ as $n \to \infty$,

$$\sqrt{\tfrac{nm}{n+m}}\left(\mathcal{T}_p(P_n, Q_m) - \mathbb{E}\mathcal{T}_p(P_n, Q_m)\right) \xrightarrow{w} N\left(0, (1-\lambda)\sigma_p^2(P, Q) + \lambda\sigma_p^2(Q, P)\right),$$

where $\sigma_p^2(P, Q)$ and $\sigma_p^2(Q, P)$ are defined in (3.9) and satisfy

$$\sqrt{\tfrac{nm}{n+m}}\mathrm{Var}(\mathcal{T}_p(P_n, Q_m)) \longrightarrow (1-\lambda)\sigma_p^2(P, Q) + \lambda\sigma_p^2(Q, P). \tag{3.10}$$

It is clear that the limit of Theorem 3.2.6 degenerates to 0 when $P = Q$. Suppose now that $P \neq Q$ are satisfying the assumption of Theorem 3.2.6. The limit, in the one-sample case, is degenerate if and only if $\mathrm{Var}(f_p(X)) = 0$. Since the optimal transport potentials are unique up to additive constants, see Theorem 3.2.2, we can suppose that $E(f_p(X)) = 0$. Thus, the degeneracy is equivalent to $E(f_p(X)^2) = 0$, hence $f_p = 0$ P-a.s. and the same holds for f_p^{dp}. This implies, in particular, that $\mathcal{W}_p(P, Q) = 0$ which occurs only if $P = Q$.

Our initial motivation to prove Theorem 3.2.6 was to find an asymptotic distribution of $\mathcal{T}_p(P_n, Q_m)$ allowing the definition of a two-sample goodness-of-fit test. Even for measures supported on the real line, the only asymptotic results account for the case $P \neq Q$, providing the asymptotic behaviour of the Wasserstein statistic under the alternative hypothesis. The idea of switching H_0 and H_1 and testing for similarities has been studied in several previous works, all considering measures supported on $\mathbb{R}$. Gaussian deviations from the true distance $\mathcal{T}_2(P, Q)$ are proved in [76], which allows testing of $\mathcal{T}_2(P, Q) \geq \Delta_0$, for a given threshold Δ_0. In the same way, the earlier work [101] introduced such an asymptotic test for assessing similarities based on the trimmed Wasserstein distance, allowing sample dependency.

Unfortunately, the same strategy can not be applied in our case, as the derived CLT for measures supported on $\mathbb{T}^2$ (Theorem 3.2.6) only states Gaussian deviations from the mean. Indeed, if we use (3.10), we could consider the statistic

$$\frac{\mathcal{T}_p(P_n, Q_m) - \mathbb{E}\mathcal{T}_p(P_n, Q_m)}{\sqrt{\mathrm{Var}(\mathcal{T}_p(P_n, Q_m))}} \xrightarrow[P \neq Q]{w} N(0, 1), \tag{3.11}$$

where, in practice, the variance and expectation could be estimated by bootstrapping the given samples (as long as bootstrap consistency is ensured). The recent works of [133] and [187] show that, in small dimension –$d = 2, 3$ and at most 4–, the value $\mathbb{E}\mathcal{T}_p(P_n, Q_m)$ can be substituted by the population $\mathcal{T}_p(P, Q)$. That gives rise to

$$\frac{\mathcal{T}_p(P_n, Q_m) - \mathcal{T}_p(P, Q)}{\sqrt{\mathrm{Var}(\mathcal{T}_p(P_n, Q_m))}} \xrightarrow[P \neq Q]{w} N(0, 1), \tag{3.12}$$

see [133, Example 5.7] for general p or [187, Corollary 8] for $p = 2$. However, for dimension $d > 4$ and $p = 2$, this substitution is no longer valid [188, Proof of Proposition 21].

When $P \neq Q$, the statistics in (3.11), (3.12) converge in law to a standard Gaussian distribution. This is illustrated in Figure A.4. However, one would expect the statistic to be stochastically larger under $P \neq Q$ than under $P = Q$, allowing the distinction of the null and the alternative hypotheses. Nevertheless, due to the aforementioned degeneracy of Theorem 3.2.6 when $P = Q$, this condition fails to be satisfied and no asymptotic test can be implemented from this result. Further discussion about this issue can be found in Section 3.5. Therefore, the rest of this paper is devoted to alternative approaches to define suitable two-sample goodness-of-fit tests for measures supported on $\mathbb{T}^2$.

3.3 Two-sample goodness-of-fit tests

Let us first formulate the problem. Denote by $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$ two independent and identically distributed random samples of laws $P, Q \in \mathcal{P}(\mathbb{T}^2)$ respectively, and by P_n, Q_m their corresponding empirical probability measures. We aim to test

$$H_0 : P = Q \qquad \text{against} \qquad H_1 : P \neq Q \tag{3.13}$$

via the definition of a statistic $T_{nm} = T(P_n, Q_m)$, representing an estimate of discrepancy between P_n and Q_m, together with the critical region

$$R = \{(x_1, \dots, x_n; y_1, \dots, y_m) \, : \, T_{nm} \geq c_{nm}(\alpha)\}, \tag{3.14}$$

where x_i (resp. y_j) denotes a realization of X_i (resp. Y_j) for $i = 1, \dots, n$ (resp. $j = 1, \dots, m$). The critical value $c_{nm}(\alpha)$ in (3.14) is given for a fixed significance level α by

$$c_{nm}(\alpha) = \inf\{t > 0 \, : \, F_{nm}(t) \geq 1 - \alpha\}, \tag{3.15}$$

where F_{nm} is the distribution function of the statistic T_{nm} under H_0. We are therefore considering the test

$$\pi_{nm} = \begin{cases} 1 & \text{if} \quad T_{nm} \geq c_{nm}(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

Equivalently, a p-value for this test is $p_{nm} = 1 - F_{nm}(T_{nm})$. Ideally, we would like T_{nm} to be $\mathcal{T}_p(P_n, Q_m)$. However, knowing the distribution of the latter statistic under H_0 remains an open problem. The one-sample case in $\mathbb{R}^d$ has recently been addressed in [114], but approaches for two-sample testing in arbitrary dimension, and for measures on more general spaces, have not already been proposed to the best of our knowledge. The lack of solutions may be explained by the intrinsic difficulty of characterizing the distribution of $\mathcal{T}_p(P_n, Q_m)$ when $P = Q$ especially when the dimension is larger than one. In the next subsections, we propose two alternative approaches to define (3.16), both based on the 2-Wasserstein distance, that allow two-sample goodness-of-fit testing for measures on $\mathbb{T}^2$.

3.3.1 Geodesic projections into $\mathbb{R}/\mathbb{Z}$

Our first approach for testing the equality of two measures P, Q on $\mathbb{R}^2/\mathbb{Z}^2$ is to test the equality of their geodesic projections. This bypasses the dimension problem and allows the implementation of testing techniques based on Wasserstein distance for one-dimensional spaces. Geodesics on $\mathbb{T}^2$ are the images by the canonical projection τ of straight lines on $\mathbb{R}^2$ [36]. Lines with irrational slope map to geodesics which are dense on $\mathbb{T}^2$, and only lines with rational slope map to *closed* geodesics on the torus, which are closed spirals isomorphic to $\mathbb{R}/\mathbb{Z}$ (see [36, Figure VII.10] for an illustration).

The strategy is to project P_n and Q_m into N_g closed geodesics, and to test the equality of each pair of projected measures, which will be supported on $\mathbb{R}/\mathbb{Z}$. These geodesics can be chosen a priori by the practitioner, or sampled from the set of all closed geodesics on $\mathbb{T}^2$. We propose a sampling method in Appendix A.1.1. This method prioritizes simpler geodesics (that is, with a smaller number of revolutions over the torus) in order to ease computational implementation. The algorithm we used to project samples on $\mathbb{T}^2$ to a given geodesic is described in Appendix A.1.2. To avoid repetition of the same test, and to ensure independence between the computed p-values, we require all the N_g geodesics to be different.

In this section, we propose a two-sample Wasserstein test to assess the equality of two measures supported on the circle, and state how to combine the resulting N_g-tuple of p-values into a global p-value for the bi-dimensional problem. From now on, to simplify notation, we will denote by $\mathcal{T}_2$ any squared Wasserstein distance, the ground space being inferred from the corresponding measures.

Two-sample goodness-of-fit test on $\mathbb{R}/\mathbb{Z}$

Optimal Transport on the circle has been recently studied in detail in [132], where the limit laws of the one and two-sample empirical Wasserstein distance for measures on $\mathbb{R}/\mathbb{Z}$ are derived. However, the considered statistics are not distribution-free, so that only one-sample goodness-of-fit tests can be derived from these results. Still, the authors of [132] also propose a b-out-of-n bootstrap approach, for $b = o(n)$, to define a two-sample goodness-of-fit test. Unfortunately, type I error fails to be controlled since the bootstrapped p-value under the null hypothesis is (substantially) stochastically smaller than a uniform random variable. This can be observed by simple numerical experiments based on the implementation proposed by [132], for example by comparing two equally-sized samples from a Uniform distribution. We believe that this is due to a lack of consistency of the two-sample bootstrap for the proposed statistic. In order to bypass this issue, we now propose a convenient alternative approach based on a distribution-free two-sample statistic.

Let $P^c, Q^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$ and P_n^c, Q_m^c be their corresponding empirical probability measures. We aim to test

$$H_0 : P^c = Q^c \qquad \text{against} \qquad H_1 : P^c \neq Q^c.$$

If $\mathbb{R}/\mathbb{Z}$ is parameterized by the set $[0, 1)$ with the geodesic distance

$$d_{\mathbb{R}/\mathbb{Z}}(x, y) = \min\{|x - y|, 1 - |x - y|\},$$

the cumulative distribution functions of P^c, Q^c, denoted as F, G respectively, can be defined as in [132] as

$$F(t) = P^c([0, t]), \quad G(t) = Q^c([0, t]) \qquad \forall\, t \in [0, 1). \tag{3.17}$$

Then, we can write

$$\mathcal{T}_2(P^c, Q^c) = \inf_{\alpha \in \mathbb{R}} \int_0^1 \left(F^{-1}(t) - (G - \alpha)^{-1}(t) \right)^2 dt, \tag{3.18}$$

where the pseudo-inverse is defined as $H^{-1}(s) = \inf\{t : H(t) > s\}$, for any distribution function H. The formulation (3.18) was first proved in [236] for discrete measures, and extended to arbitrary measures in [78]. It shows how the Optimal Transport problem on the circle reduces to the same problem on $[0, 1) \subset \mathbb{R}$ if both measures are relocated on the real line choosing as origin the minimizing element α. This is well illustrated in [132]. We first remark that if one of the two measures is the uniform law on $\mathbb{R}/\mathbb{Z}$, the infimum on (3.18) has an explicit formulation.

Lemma 3.3.1. *Let $P^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$, and F be its cumulative distribution function. Let U be the uniform distribution on $\mathbb{R}/\mathbb{Z}$. Then,*

$$\mathcal{T}_2(P^c, U) = \int_0^1 \left(F^{-1}(t) - t - \alpha_0(F) \right)^2 dt,$$

where the optimal origin is given by

$$\alpha_0(F) = \int_0^1 (F^{-1}(t) - t)\, dt.$$

If we replace P^c and F by their empirical counterparts, P_n^c and F_n, Lemma 3.3.1 allows the definition of the statistic $\mathcal{T}_2(P_n^c, U)$, which is distribution-free when $P^c = U$.

Lemma 3.3.2. *Let $P^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$, P_n^c be its empirical probability measure, and U be the uniform distribution on $\mathbb{R}/\mathbb{Z}$. Then, if $P^c = U$,*

$$n\, \mathcal{T}_2(P_n^c, U) \xrightarrow[n]{w} \int_0^1 \mathbb{B}(t)^2\, dt - \left(\int_0^1 \mathbb{B}(t)\, dt \right)^2,$$

where $\mathbb{B}$ is a standard Brownian bridge, and the weak convergence is understood as convergence of probability measures on the space of right-continuous functions with left limits.

Lemma 3.3.2 can be used to define a one-sample goodness-of-fit uniformity test, based on the squared Wasserstein distance on the circle. This would complement the work in [132], where such a test was introduced for the 1-Wasserstein distance. As our aim here is to define a two-sample test, we adapt the idea of [239] to compare two measures on the circle, by considering the 2-Wasserstein distance between $G_m^{-1}(F_n)$ and the uniform distribution. We can therefore consider the statistic

$$T_{nm}^c = \frac{nm}{n+m} \mathcal{T}_2(G_m \# P_n^c, U) = \tag{3.19}$$
$$\frac{nm}{n+m} \int_0^1 \left(G_m(F_n^{-1}(t)) - t - \alpha_0(F_n^{-1}(G_m)) \right)^2 dt,$$

which is also distribution-free when $P^c = Q^c$. The following result is the counterpart of Lemma 3.3.2.

Proposition 3.3.3. *Let $P^c, Q^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$, having continuous and strictly increasing cumulative distribution functions. Let P_n^c, Q_m^c be their corresponding empirical probability measures, and F_n, G_m be their empirical cumulative distribution functions. If $\frac{n}{m} \to \lambda$ when $n, m \to \infty$ for some $\lambda \in [0, \infty)$ then, under $P^c = Q^c$, it holds that*

$$T_{nm}^c = \frac{nm}{n+m} \mathcal{T}_2(G_m \# P_n^c, U) \xrightarrow[n,m]{w} \int_0^1 \mathbb{B}(t)^2 \, dt - \left(\int_0^1 \mathbb{B}(t) \, dt \right)^2 .$$

Consequently, with the notation of the beginning of Section 3.3, we propose the test

$$\pi_{nm}^c = \begin{cases} 1 & \text{if} \quad T_{nm}^c \geq c_{nm}^c(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

where the critical value $c_{nm}^c(\alpha)$ is given by

$$c_{nm}^c(\alpha) = \inf \left\{ t > 0 \, : \, F_{nm}^c(t) \geq 1 - \alpha \right\} ,$$

with F_{nm}^c denoting the distribution function of T_{nm}^c under H_0. Equivalently, a p-value for this test is $p_{nm}^c = 1 - F_{nm}^c(T_{nm}^c)$. Following Proposition 3.3.3, the critical value or, equivalently, the p-value for a given sample, can be approximated with arbitrary precision using a Monte Carlo algorithm. The following result guarantees the consistency of (3.20).

Proposition 3.3.4 (Consistency)**.** *Let $P^c, Q^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$ having continuous and strictly increasing cumulative distribution functions. If $P^c \neq Q^c$, it holds*

$$\lim_{n,m \to \infty} \mathbb{P} \left(\pi_{nm}^c = 1 \right) = 1 \quad \textit{for any } \alpha > 0.$$

Combining a N_g-tuple of tests on $\mathbb{R}/\mathbb{Z}$

Consider the problem of testing the equality of N_g pairs of projections of P_n and Q_m into N_g different closed geodesics. Instead of a single statistic, we now have a sample $(T_{nm,1}^c, \ldots, T_{nm,N_g}^c)$ of statistics which, under the null hypothesis, are identically distributed as T_{nm}^c (by Proposition 3.3.3). Equivalently, one can think of a sample of p-values $(p_1, \ldots, p_{N_g})$ which, following (3.20), are given by

$$p_i = 1 - F_{nm}^c(T_{nm,i}^c) \quad i = 1, \ldots, N_g. \tag{3.21}$$

These individual p-values can be aggregated as follows:

$$p^{N_g} = N_g \min_{i=1}^{N_g} p_i. \tag{3.22}$$

This aggregation is akin to the Bonferroni correction for Family Wise Error Rate (FWER) control in multiple testing [32]. As such, p^{N_g} defined in (3.22) is a valid p-value for the

two-dimensional test, regardless of the possible dependencies between the N_g individual p-values. This implies that the two-dimensional test

$$\pi^g_{nm,N_g} = \begin{cases} 1 & \text{if} \quad p^{N_g} \leq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{N_g-geod}$$

controls the type I error for any $\alpha > 0$ (see Appendix A.2.2 for a proof). Regarding consistency under fixed alternatives, by construction, (N_g-geod) will fail to detect differences between two measures on $\mathbb{T}^2$ whose projected distributions are identical for all the N_g geodesics considered. Therefore, π^g_{nm,N_g} will not be consistent under such alternatives, which, are arguably very unlikely in practice if N_g is large enough. Otherwise, consistency is guaranteed.

Proposition 3.3.5 (Consistency)**.** *Let $P, Q \in \mathcal{P}(\mathbb{T}^2)$ such that $\mu_P, \mu_Q \ll \ell_2$ and P^c_i (resp. Q^c_i), $i = 1, \dots, N_g$, be the circular projected distributions of P (resp. Q) to N_g closed geodesics of $\mathbb{T}^2$. If $P^c_i \neq Q^c_i$ for at least one $i \in \{1, \dots, N_g\}$, it holds*

$$\lim_{n,m\to\infty} \mathbb{P}\left(\pi^g_{nm,N_g} = 1\right) \quad \textit{for any } \alpha > 0.$$

Remark 3.3.6. *The assumption in Proposition 3.3.3 that the projected measure $P^c \in \mathcal{P}(\mathbb{R}/\mathbb{Z})$ has continuous and strictly increasing cumulative distribution function is satisfied if the underlying measure $P \in \mathcal{P}(\mathbb{T}^2)$ satisfies $\mu_P \ll \ell_2$. See Appendix A.2.2 for a proof.*

The time complexity of (N_g-geod) is $\mathcal{O}(n+m)$. Indeed, $n+m$ operations are needed to compute $G_m(F_n^{-1}(t))$ and $F_n^{-1}(G_m(t))$ for a given t. Therefore, computing the test statistic (3.19) can be done in $\mathcal{O}(n+m)$ operations, where the complexity constant depends on the number of subdivisions of $[0,1]$ set by the numerical integration method chosen to compute (3.19). Moreover, the time complexity of the algorithm described in Appendix A.1.1 to sample closed geodesics is also $\mathcal{O}(n+m)$ in practice, as a consequence of the distribution from which the geodesics are drawn. This is empirically illustrated in Figure A.5.

3.3.2 p-value upper bounding

If we set $\mathcal{T}_2(P_n, Q_m)$ as the statistic T_{nm} for the test (3.16), the p-value for a given sample would be given by

$$\mathbb{P}_{H_0}(\mathcal{T}_2(P_n, Q_m) \geq t_{nm}), \tag{3.23}$$

where t_{nm} denotes the statistic realization. The goal of this section is to find an upper bound for (3.23), which will itself be a valid p-value for (3.16) if it controls type I error (that is, if it remains with probability $1-\alpha$ over a fixed significance level α under H_0). We will also require the power of the corresponding test to tend to 1 under fixed alternatives. We start by upper bounding the deviations of the statistic from the mean. Using McDiarmid's inequality [195], we obtain the following result, which extends to the two-sample case the inequality in [306, Proposition 20], for the quadratic cost.

Theorem 3.3.7. *Let $P, Q \in \mathcal{P}(\mathbb{T}^2)$ and P_n, Q_m be two empirical probability measures of laws P, Q respectively. Then, for all $t \in \mathbb{R}$, we have*

$$\mathbb{P}\left(\mathcal{T}_2(P_n, Q_m) - \mathbb{E}\mathcal{T}_2(P_n, Q_m) > t\right) \leq \exp\left(-\frac{nm}{n+m} 8t^2\right). \tag{3.24}$$

After that, we study the convergence speed of the expectation under the null hypothesis. Using directly the results exposed in [100], only bounds of order

$$\mathbb{E}\mathcal{T}_2(P_n, Q_m) = O\left(n^{-\frac{1}{2}} + m^{-\frac{1}{2}}\right) \tag{3.25}$$

can be expected. However, the recent work in [5] shows that the convergence of the mean (3.25) becomes faster under some regularity assumptions. On the one hand, we require the density of the induced periodic measure μ_P to be Hölder continuous[5] and absolutely continuous w.r.t. the Lebesgue measure ℓ_2 in $\mathbb{R}^2$. On the other hand, we require the set $\mathrm{supp}(P)$ to be connected and to have $\mathcal{C}^1$ boundary, in the sense that it can be locally parameterized by a $\mathcal{C}^1$ curve.

Assumption 1. *(1) $P \in \mathcal{P}(\mathbb{T}^2)$ is supported in a connected set with $\mathcal{C}^1$ boundary, with $\mu_P \ll \ell_2$. (2) Its probability density p is Hölder continuous and bounded from below in its support ($p(x) \geq \lambda > 0$ for all $x \in \mathrm{supp}(\mu_P)$).*

If Assumption 1 is satisfied, then from Lemma B.1 and Theorem 6.3. in [5] we can derive the following asymptotic bound for the two-sample null expectation.

Lemma 3.3.8. *Let $P = Q \in \mathcal{P}(\mathbb{T}^2)$ satisfy Assumption 1 and $m = m(n)$ be a sequence such that $m \xrightarrow[n\to\infty]{} \infty$ and $\frac{n}{m} \to \lambda \in (0, 1)$. Then, we have*

$$\limsup_{n\to\infty} \frac{n}{\log(n)} \mathbb{E}\mathcal{T}_2(P_n, Q_m) \leq \frac{1}{4\pi}\left(1 + \frac{1}{\lambda}\right). \tag{3.26}$$

Note that Assumption 1 is not especially restrictive. It is satisfied by any continuously differentiable density, whose connected support can be locally given by the graph of a continuously differentiable function. Examples include bivariate von Mises distributions or uniform distributions in connected smooth sets.

The idea to define the test is to combine Theorem 3.3.7 with Lemma 3.3.8 and upper bound (3.23) for sufficiently large sample sizes. If we take the limit for the expectation in (3.24) under the null, we have the following result.

Proposition 3.3.9. *Let $P, Q \in \mathcal{P}(\mathbb{T}^2)$ and P_n, Q_m be two empirical probability measures of laws P, Q respectively. For all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that for all $n, m \geq N_\varepsilon$, we have*

$$\mathbb{P}_{H_0}\left(\mathcal{T}_2(P_n, Q_m) > t\right) \leq \exp\left(-\frac{nm}{n+m} 8(t-\varepsilon)^2\right) =: \xi_{nm,\varepsilon}(t) \qquad \forall\, t > 0. \tag{3.27}$$

[5] A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be locally Hölder continuous in a compact set X for some $\alpha > 0$ if, for every $x \in X$, there exists some $\epsilon > 0$ such that $|f(x) - f(y)| \leq C\|x - y\|^\alpha$ if $y \in X$ and $\|y - x\| < \epsilon$.

For a fixed $\varepsilon > 0$, the bound (3.27) can be used to define a test (3.16) for any $\alpha > 0$ as follows:

$$\pi^{ub}_{nm,\varepsilon} = \begin{cases} 1 & \text{if} \quad \xi_{nm,\varepsilon}\left(\mathcal{T}_2(P_n, Q_m)\right) \leq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{UB}$$

By Proposition 3.3.9, the test (UB) will control type I error for all $n, m \geq N_\epsilon$. In practice, the threshold N_ε depends on the unspecified constant hidden in (3.26), which is dragged from the results in [5]. The following result shows that, nevertheless, asymptotic consistency at level α of (UB) is guaranteed.

Proposition 3.3.10 (Asymptotic consistency at level α)**.** *Let $P, Q \in \mathcal{P}(\mathbb{Z}^2)$. The test* (UB) *is asymptotically of level α. If $P = Q$, we have, for any $\varepsilon > 0$,*

$$\lim_{n,m\to\infty} \mathbb{P}\left(\pi^{ub}_{nm,\varepsilon} = 1\right) \leq \alpha \quad \textit{for any } \alpha > 0. \tag{3.28}$$

Under fixed alternatives, the test is consistent if $\mathcal{T}_2(P, Q) > \varepsilon$:

$$\lim_{n,m\to\infty} \mathbb{P}\left(\pi^{ub}_{nm,\varepsilon} = 1\right) = 1 \quad \textit{for any } \alpha > 0.$$

The last result ensures asymptotic consistency at level α if the two compared measures are further than ε in the squared 2-Wasserstein distance. This can be used to calibrate the sensibility of (UB) if the practitioner possesses some prior information about the differences that the test should accept. This would ensure smaller N_ε without implying a power decrease. For the simulation and case studies presented here, we will set ε to the machine precision $\varepsilon_m = 2.2 \cdot 10^{-16}$ (for a standard double-precision floating-point format). The corresponding N_ε should be affordable thanks to Lemma 3.3.8, responsible of the satisfactory power of (UB). Due to the improved convergence speed of the expectation, we will have sharp bounds (3.27) for reasonable sample sizes, allowing the detection of differences for our practical purposes. This is illustrated in Section 3.4.2.

The computational complexity of (UB) is given by the numerical algorithm solving the Optimal Transport problem. Here, we used the Fast Network Simplex for Optimal Transport [34], which has $\mathcal{O}((n+m)^2)$ time complexity and $\mathcal{O}((n+m)^2)$ memory cost, due to the cost matrix computation.

3.4 Numerical experiments

This section is devoted to assess the performance of the two-sample goodness-of-fit tests (N_g-geod) and (UB), and to show how they can be implemented to evaluate differences on protein structure data. In Section 3.4.1 and 3.4.2, we evaluate the relative efficiency of both tests, comparing their performance with other methods not based on Optimal Transport. Section 3.4.3 illustrates one possible application to protein structure investigations, by stating statistical evidence of nearest neighbors effects on local protein conformations.

3.4.1 Small-sample performance

To make an informative analysis of the performance of tests (N_g-geod) and (UB), we studied how their power function behaves for alternatives converging to the null hypothesis. We also assessed whether the proposed approach to define a Wasserstein test on the circle contributes to a better power. In particular, we compared the power function of (N_g-geod) with variations of the same test. On the one hand, to evaluate whether the choice of an optimal origin to relocate the measures on $[0, 1)$ is advantageous, we considered the same statistic (3.19) but with α_0 being random and uniformly chosen in $[0, 1]$. It is easy to check that the modified statistic is distribution-free under the null, by proceeding analogously to Proposition 3.3.3. On the other hand, to study whether the use of Wasserstein distance for the one-dimensional statistic contributes to a better power, we relocated the measures in $[0, 1)$ (again after choosing a random origin on $\mathbb{R}/\mathbb{Z}$) and compared them with the well-known Anderson-Darling two-sample statistic. To study the effect of the number N_g of geodesics, we performed the test (N_g-geod) for $N_g \in \{2, 3, 4, 5\}$. We also compared all the previous approaches with the two-dimensional extension of the Kolmogorov-Smirnov two-sample test proposed by Fasano and Franceschini [92], defined for measures supported on $\mathbb{R}^2$. This allows the assessment of whether taking into account the geometry of the underlying space contributes to a better performance.

For the small-sample case, we compared samples of size $n = m = 50$ drawn from a bivariate von Mises (bvM) distribution [189] of means $\mu = \nu = 0.5$, and concentration parameters $\kappa_1, \kappa_2, \kappa_3$ with equally-sized samples drawn from a uniform distribution on $\mathbb{T}^2$. The density of the bvM cosine model is given by

$$f(\varphi, \psi) = c(\kappa_1, \kappa_2, \kappa_3)(\exp(\kappa_1 \cos(\varphi - \mu) + \kappa_2 \cos(\phi - \nu) - \kappa_3 \cos(\varphi - \mu - \psi + \nu)),$$

where the explicit form of the normalization constant $c(\kappa_1, \kappa_2, \kappa_3)$ is stated in [189]. The null hypothesis corresponds to the case $\kappa_1 = \kappa_2 = \kappa_3 = 0$. For the converging alternatives, we distinguished two scenarios:

(a) No dependence structure: $\kappa_3 = 0$ and $\kappa_1 = \kappa_2 \in [0, 3]$ as varying parameter.

(b) Only dependence structure: $\kappa_1 = \kappa_2 = 0$ and $\kappa_3 \in [0, 3]$ as varying parameter. Here, the marginal laws are uniform distributions on $[0, 1]$ [189].

The rejection probability was estimated as the proportion of rejections at level $\alpha = 0.05$ among 5000 repetitions of each test for a fixed value of the corresponding varying parameter. Results for both scenarios are shown in Figure 3.1, where 'W-geodesic' stands for the test (N_g-geod), 'Naive W-geodesic' for its random origin variation, 'AD-geodesic' for the comparison with the Anderson-Darling two-sample statistic, and 'Upper bound' for the test (UB).

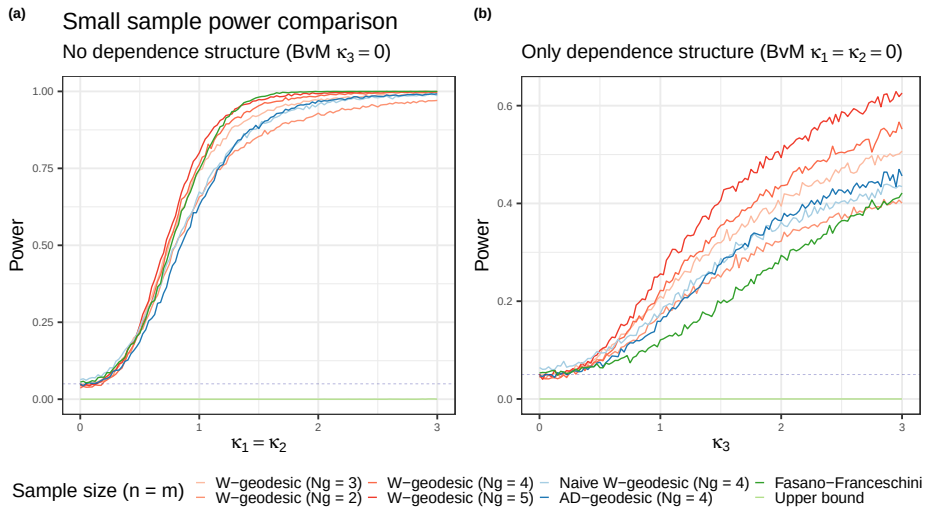


Figure 3.1: Empirical power of two-sample goodness-of-fit tests for measures supported on $\mathbb{T}^2$, under bivariate von Mises (BvM) alternatives with no dependence structure and different marginal laws (a) and with equal marginal laws and dependence structure (b). The simulated samples had sizes $n = m = 50$. The empirical power corresponds to the proportion of rejections at level $\alpha = 0.05$ (dashed line) among 5000 repetitions of the test for fixed concentration parameters.

The first conclusion that we can state after Figure 3.1 is that the test (UB) has zero power for small sample sizes. This was expected by Proposition 3.3.9, as large values of n, m are required to ensure sharp bounds. However, some interesting conclusions can be extracted regarding the other tests. First, the test (N_g-geod) has power α under H_0. Indeed, further simulations confirmed that the approach described in Section 3.3.1.0 ensures the uniformity of the combined p-value's null distribution. Together with the illustrated consistency of test (N_g-geod), we can observe the considerable gain in power when comparing measures with the Wasserstein statistic (3.19) by choosing an optimal origin on the circle. The choice of a random origin ('Naive W-geodesic' curve) or the use of techniques that do not rely on Optimal Transport ('AD-geodesic' or Fasano-Franceschini curve) notably reduce the test power, specially when differences are presented on the dependence structure (Figure 3.1b). Finally, the choice of the number N_g of geodesics seems to have an effect on power. As one could have expected, increasing the number of geodesic projections improves the test's ability to detect slighter differences. Consequently, the practitioner is entitled to indefinitely increase N_g, paying back on computation time (or implementation complexity, if geodesics are randomly chosen, see Appendix A.1.1).

3.4.2 Asymptotic performance

This section is devoted to assess the suitability of the upper bound testing technique (UB) when large sample sizes are available. Here, we studied the relative efficiency of

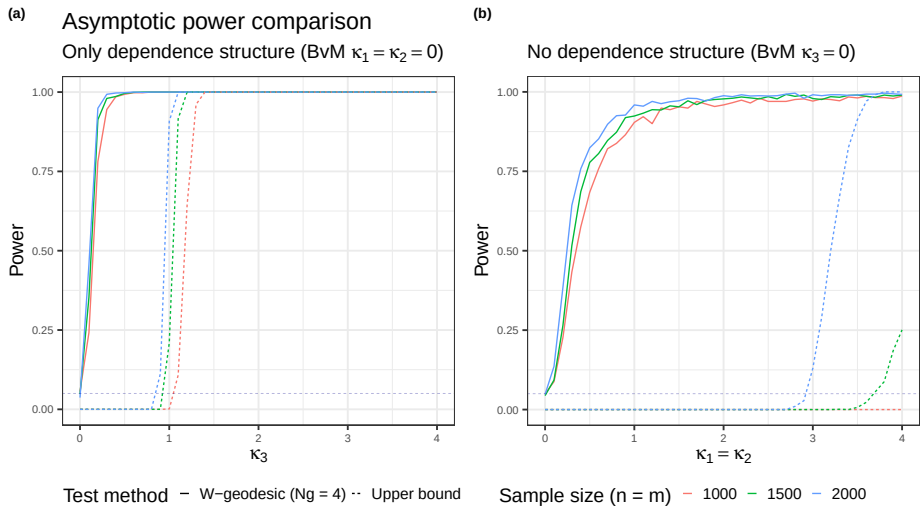


Figure 3.2: Empirical power of two-sample goodness-of-fit tests for measures supported on $\mathbb{T}^2$, under bivariate von Mises (BvM) alternatives with no dependence structure and different marginal laws (a) and with equal marginal laws and dependence structure (b). The empirical power corresponds to the proportion of rejections at level $\alpha = 0.05$ (dashed line) among 1000 repetitions of each test for fixed concentration parameters.

tests (UB) and (N_g-geod) for the same converging alternatives as in Section 3.4.1, with $n = m \in \{1000, 1500, 2000\}$. Results are shown in Figure 3.2, where the parameter of interest ($\kappa_1 = \kappa_2$ or κ_3) took values in $\{0.1, 0.2, \ldots, 4\}$, and the empirical power was estimated as the proportion of rejections at level $\alpha = 0.05$ among 1000 test repetitions.

Figure 3.2 shows that the test (UB) is powerful when sample sizes are large enough. As its corresponding p-value has been defined as an upper bound of the actual p-value (3.23), it will be quite a conservative test and, therefore, relatively less efficient than (N_g-geod). This is illustrated in both panels. In any case, the test (UB) can be useful in practice. Besides the detection of big differences, the practitioner may be interested in the acceptance of small and controlled discrepancies between samples, which may be due, for instance, to experimental inaccuracies. In scenarios where a less conservative method as (N_g-geod) may detect such differences, one might prefer to rely on a test method that allows slight dissimilarities and stands out only the more relevant ones. Consequently, even if the test (UB) is clearly less efficient than our first candidate (N_g-geod), we believe it can be of interest in some practical scenarios, such as several situations appearing in Structural Biology problems. This is further discussed in Section 3.5.

3.4.3 Application to protein structure analysis

A method to accurately compare local structural preferences in conformational ensemble models of proteins is useful to investigate sequence-structure-function relationships, allowing for instance to understand the effect of mutations. For most amino acid types (for

all excepting proline and glycine), the distribution of ϕ and ψ angles is supported on the same subset of $\mathbb{T}^2$, which, even if there exist some physically forbidden regions due to strong repulsive forces between non-bonded atoms at short distance, is connected and has a smooth boundary. We can also assume that density is continuously differentiable and strictly positive in its support, so that Assumption 1 is satisfied.

The aim of this section is to make use of the tests (N_g-geod) and (UB) to show that the distribution of (ϕ, ψ) does not depend only on the amino acid type, but also on the sequence context, and particularly on the closest neighbors. This corresponds to rejecting Flory's isolated-pair hypothesis [99]. Even if the importance of the closest neighbors effect is widely accepted in the Structural Biology community [150, 105, 27, 244, 280], only purely descriptive methods have been employed to state so, and no goodness-of-fit techniques have been used to the best of our knowledge. For a given amino acid C, we denote by P_C the distribution of (ϕ, ψ) supported on $\mathbb{T}^2$. If we take into account the identities L, R of C's left and right neighbors, the distribution of (ϕ, ψ) is now given by P_{LCR}. The objective is to test

$$H_0 : P_C = P_{LCR} \qquad \text{against} \qquad H_1 : P_C \neq P_{LCR} \tag{3.29}$$

to assess whether nearest neighbors significantly affect dihedral angles distributions. An example of two samples drawn from P_C and P_{LCR} is depicted in Figure 3.3. For the analysis presented here, we used the structural database of three-residue fragments (also called tripeptides) presented in Section 2.2.1, that were extracted from experimentally-determined high-resolution protein structures [88]. The large available sample sizes allow us to illustrate the asymptotic behaviour of (UB). We selected the 71 tripeptides *L-C-R* for which the database contained more than 3000 points. For each one, we compared the corresponding sample of (ϕ, ψ) values with an equally-sized sample drawn from P_C (sampled from the sub-database containing (ϕ, ψ) values from tripeptides having C as central amino-acid). The data were rescaled to $[0, 1] \times [0, 1]$ before applying the tests. As the p-values for the test (N_g-geod) are computed by Monte Carlo simulation, they are lower-bounded by $1/N_{MC}$, where N_{MC} is the number of Monte Carlo replicas [230]. This point is important here, as due to the large number of performed tests, we had to correct p-values for multiplicity [125]. The results are depicted in Figure 3.4, where we show the empirical cumulative distribution function of both tests' corrected p-values, for three increasing ranges of sample sizes.

From Figure 3.4, we can state that the geodesic projection test (N_g-geod) strongly rejects the null hypothesis at level $\alpha = 0.05$ for the three considered sample size ranges, being all p-values truncated to the Monte Carlo precision. Repeating the same analysis for $N_g = 3, 4$ did not change the shape of the (N_g-geod) p-values curves, which was expected as higher values of N_g yield a power increase. A clear asymptotic behaviour is observed for the upper bound technique (UB), as power at level α tends to one when sample sizes increase. Note that, for the largest range of sample sizes, (UB) is relatively more efficient

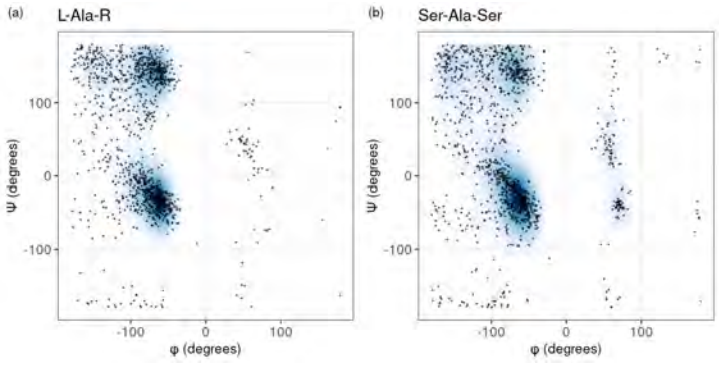


Figure 3.3: (a) Sample and kernel density estimate of alanine (ϕ, ψ) distribution when the identity of its left and right neighbor is not taken into account. (b) Sample and kernel density estimate of (ϕ, ψ) distribution corresponding to tripeptide Ser-Ala-Ser (a fragment of the three consecutive amino-acids serine, alanine, serine).

than (N_g-geod), due to the Monte Carlo truncation. Both procedures lead to rejection of the null hypothesis, and therefore to the statement that nearest neighbors effect on (ϕ, ψ) distributions is statistically significant. This analysis suggests that both (N_g-geod) and (UB) are suitable for assessing differences on local protein structures, as the available sample sizes (which may be up to $\sim 10^5$ in some practical scenarios) are large enough to state significant conclusions.

3.5 Discussion

The main goal of this work was to define suitable two-sample goodness-of-fit tests for measures on $\mathbb{T}^2$. This naturally led us to enrich the existing theoretical results [60, 187, 193] on Optimal Transport for periodic measures. In particular, we studied the shape of the solutions to the Monge problem (3.5), which allowed the extension of a Central Limit Theorem to $\mathbb{T}^d$, for any $p > 1$. Our original inspiration when first investigating these theoretical results was to use the Central Limit Theorem 3.2.6 to define a two-sample asymptotic test. However, the derived limit distribution degenerates when $P = Q$ and prevents such an application. Nevertheless, the Wasserstein distance on $\mathbb{T}^2$ for the quadratic cost was used to define two efficient testing techniques, which address our initial goals.

The first approach bypasses the dimension problem by projecting the measures to closed geodesics on $\mathbb{T}^2$ and subsequently test their equality. This required the investigation of how to project samples on closed geodesics and, moreover, how to conveniently sample closed geodesics. The answers we propose here, notably in Sections A.1.1 and A.1.2, together with their supplied practical implementations, may be of interest in fur-

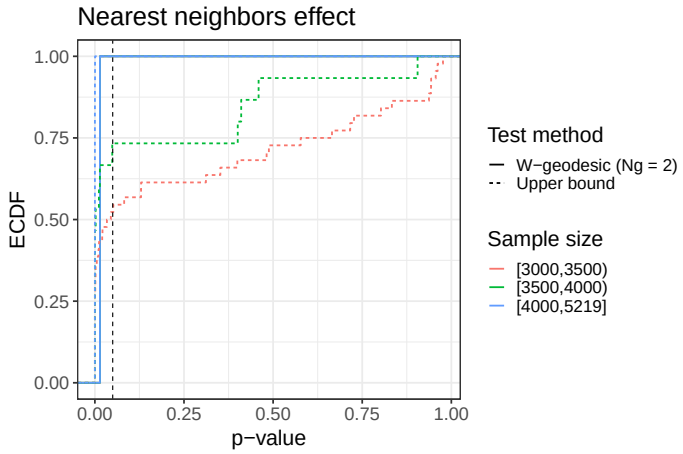


Figure 3.4: Empirical cumulative distribution function of p-values corresponding to test hypotheses (3.29) with (N_g-geod) ('W-marginal') and (UB) (Upper bound) testing methods, for 71 different combinations of L,C,R. To illustrate the asymptotic behaviour, p-values were classified in three ranges of sample sizes. For each test method, p-values were corrected for multiplicity using Holm-Bonferroni correction [125]. Marginal test p-values were computed with a Monte Carlo simulation of $N_{MC} = 5000$ replicas. The black dashed line indicates an arbitrary significance level of $\alpha = 0.05$.

ther practical situations. Furthermore, they suggest one possible extension of the Sliced Wasserstein distance [33] to the two-dimensional flat torus. As closed geodesics on $\mathbb{T}^2$ are isomorphic to $\mathbb{R}/\mathbb{Z}$, the equality of the projected measures is assessed through a two-sample Wasserstein test on the circle which, to the best of our knowledge, is the only efficient procedure proposed up to now.

The second proposed approach consists in upper-bounding the exact p-values (3.23). This is possible thanks to the derived concentration inequalities (3.24) for the two-sample empirical Wasserstein distance with the quadratic cost, and to the improved convergence speed of its expectation, as shown in Lemma 3.3.8. As with any upper-bounding technique, the corresponding test is conservative and only efficient for large sample sizes, which reduces its range of application. However, this test could be relevant in some practical scenarios. For example, Molecular Dynamics simulations (which simulate the temporal evolution of the structure of a protein using force-fields based on physical models), produce samples on $\mathbb{T}^2$ that may present small and meaningless differences when re-running simulations multiple times with slightly different initial conditions. In such a situation, we expect that the first technique (N_g-geod) will reject the equality of their corresponding distributions, while the conservative test (UB) will accept differences between independent replicas of the same simulation. Consequently, (UB) will only detect more important discrepancies, which are the only ones of interest for practical purposes.

Regarding the practical implementation of both tests, some differences appear with

respect to computing time. The main advantage of (N_g-geod) is the explicit formulation of Wasserstein distance on one-dimensional spaces, which avoids the use of any Optimal Transport solver. As a result, its time complexity is linear in the sample size. However, the statistic null-distribution must be simulated with the desired precision, which may slow down the procedure. Note that, in any case, this distribution can be simulated once and be tabulated for any further implementation. The time complexity of (UB) exclusively lies on the Optimal Transport solver chosen to compute Wasserstein distance. For very large sample sizes, this might lead to a substantially slower process.

The issue of two-sample goodness-of-fit testing studied in Section 3.3 remains largely open. Our contribution in this respect is to propose easily implementable goodness-of-fit testing approaches that are built on top of state-of-the-art tools in Optimal Transport. Finding the exact or asymptotic distribution of the Wasserstein statistic in general dimension remains one of the main unsolved problems of the theory of Optimal Transport, preventing the construction of more efficient two-sample goodness-of-fit tests. An asymptotic approach for measures supported on a finite set has been presented in [268] and, in the one-dimensional case, [23] have obtained a CLT under the null $P = Q$ for deviations of $W_p(P_n, Q_n)$ from the true distance $W_p(P, Q)$ (instead of $\mathbb{E}(W_p(P_n, Q_n))$). The results of [23] are already quite challenging mathematically, and extensions to higher dimensions are clearly beyond the scope of the present work. Altogether, we believe that the goodness-of-fit tests defined in this paper constitute a relevant building block for the study of the sequence-structure-function relationship in proteins, and in particular for Intrinsically Disordered Proteins (IDPs), allowing their structural investigation with mathematical guarantees. Furthermore, the interest of the techniques here presented may go beyond the Structural Biology community, as they allow solving the goodness-of-fit testing problem for two distributions lying in general periodic spaces, which appears in various application domains.

Software availability

The statistical tests presented in this chapter are implemented in the R package `torustest`, available at https://github.com/gonzalez-delgado/torustest. The package also includes the algorithms introduced in Appendix A.1. Empirical Wasserstein distances were computed using the R package `transport` [259].

Acknowledgements

This work was supported by the AI Interdisciplinary Institute ANITI, which is funded by the French “Investing for the Future – PIA3” program under the Grant agreement ANR-19-PI3A-0004, and by the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme “Investissements d’Avenir”.

Chapter 4

The translated codon effect on local backbone conformations

Rosenberg *et al.* [249] conducted a highly relevant work exploring the relationship between the identity of synonymous codons and the distribution of backbone dihedral angles in translated amino acids. They stated that the nature of the codon has a significant influence on the local protein conformation. However, the statistical methodology implemented in [249] presents important incorrectness that prevents the correct interpretation of the results. More precisely, the implemented procedure to define p-values is unsuitable to correctly perform statistical tests. Following this observation, we repeated the analyses using the data provided by Rosenberg *et al.*, but applying the statistical hypothesis testing methods presented in Chapter 3. Our analyses show that synonymous codons have a non-negligible effect on (ϕ, ψ) distributions of translated amino acid residues, as suggested by Rosenberg *et al.*, but differ regarding the strength of significance of the differences between distributions depending on the secondary structure type. They also indicate that synonymous codon effects are stronger when considered in the context of the local sequence.

This work has been submitted and is available at *bioRxiv 2022.11.29.518303*, Statistical tests to detect differences between codon-specific Ramachandran plots, with Pablo Mier, Pau Bernadó, Pierre Neuvial and Juan Cortés. It is presented here with minor changes for the sake of coherence in the manuscript.

Contents

4.1 Introduction

In their recent work, Rosenberg *et al.* [249] studied the dependence between the identity of synonymous codons and the distribution of the backbone dihedral angles of the translated amino acids. It has been shown that the use of synonymous codons is highly relevant in multiple biological processes including, among others, mRNA splicing, translational rates and protein folding [220, 44]. While the correlation between synonymous codons and secondary structure in translated proteins has been widely studied [218, 258], Rosenberg et al. evaluated the effect of codon identity on a finer scale, analyzing whether the distribution of (ϕ, ψ) dihedral angles within secondary structure elements is significantly altered when synonymous codons are used. However, their statistical methodology is formally incorrect, casting doubt on the obtained results. The origin of the incorrectness is described in Section 4.2. Then, using the methodology introduced in Chapter 3, we reanalyzed the data presented in [249]. Our results, presented in Section 4.3, confirm the influence of the codon on the distribution of the dihedral angles, but differ from those of Rosenberg *et al.* in the strength of significance of the differences depending on the secondary structure type. Finally, we assessed whether these findings may be affected by the structural classification or the local sequence context. These additional analyses show that codon-specific effects have similar significance in different areas of Ramachandran space, although the effect may be stronger for a particular type of secondary structure, such as β-strands compared to α-helices. They also indicate that synonymous codon effects are stronger when considered in the context of the local sequence.

4.2 Incorrectness of the methodology of Rosenberg *et al.*

The goal of Rosenberg *et al.* was to assess the effect of synonymous codons on the distribution of (ϕ, ψ) dihedral angles by comparing codon-specific Ramachandran plots. Keeping the notation of [249], if (c, c') denotes a pair of synonymous codons and $\mathcal{X}$ a type of secondary structure, they aimed at testing the null hypothesis $H_{0,(c,c')|\mathcal{X}}$ that both codon-specific distributions are the same. To do so, the authors introduced a metric to quantify differences between the distributions corresponding to different codons. Then, to assess the significance of such differences, Rosenberg *et al.* proposed to draw $B = 25$ pairs of bootstrapped samples, and to compare them with their synonymous codon counterparts using a permutation test procedure, with $K = 200$ permutations. For each bootstrap sample $b \in \{1, \ldots, B\}$, if n_b denotes the number of permutations where the permuted

metric is larger than the base metric (obtained from non-permuted data), they proposed the quantity

$$p_{(c,c'),\mathcal{X}} = \frac{1 + \sum_{b=1}^{B} n_b}{1 + BK} \tag{4.1}$$

as a p-value for $H_{0,(c,c')|\mathcal{X}}$. We can reformulate (4.1) in order to gain insight into its statistical behavior. First, let us define

$$p_b = \frac{1 + n_b}{1 + K}\,, \tag{4.2}$$

which is a well-defined p-value for the b-th permutation test. Letting

$$\overline{p}_B = \frac{1}{B} \sum_{b=1}^{B} p_b\,, \tag{4.3}$$

we can show that, for sufficiently large K, $p_{(c,c'),\mathcal{X}}$ is approximately the empirical mean of the B p-values associated to individual permutation tests. The following result is proved in Appendix B.

Proposition 4.2.1. *Let $p_{(c,c'),\mathcal{X}}$ be the p-value defined in* (4.1) *for a given null hypothesis $H_{0,(c,c')|\mathcal{X}}$, and p_b be the p-value for the b-th permutation test, defined in* (4.2)*, for $b = 1, \ldots, B$. Let $\overline{p}_B$ be the empirical mean of $(p_b)_{1 \leq b \leq B}$. Then for any $K > 0$, it holds that:*

$$0 \leq \overline{p}_B - p_{(c,c'),\mathcal{X}} \leq \frac{1}{K}\,. \tag{4.4}$$

However, $\overline{p}_B$ is not a valid p-value. Let us recall that a p-value p is statistically valid if and only its distribution under the null hypothesis is Super-Uniform [172, Section 3.3]. A random variable is said to be Super-Uniform if is stochastically greater than a uniform random variable or, in other words, if its cumulative distribution function (CDF) F is upper bounded by that of the Uniform distribution (denoted by $\mathrm{U}[0,1]$ below), that is:

$$F(x) \leq x \text{ for all } x \text{ in } [0,1]. \tag{4.5}$$

Moreover, the closer the p-value distribution under the null hypothesis is to $\mathrm{U}[0,1]$, the more powerful the corresponding test is. Condition (4.5) is satisfied for classical permutation p-values such as p_b (with the CDF getting closer to the $\mathrm{U}[0,1]$ distribution as K increases), but not for averages of p-values like $\overline{p}_B$, as we show below. Instead, all the p_b could be correctly aggregated by taking their minimum and correcting the result for multiple testing (Bonferroni aggregation).

Proposition 4.2.2. *Let $U_1, \ldots, U_n$ be n real-valued random variables uniformly distributed on $[0,1]$. For all $n \geq 2$, their empirical mean $\overline{U_n} = \frac{1}{n}\sum_{i=1}^{n} U_i$ is not super-uniform.*

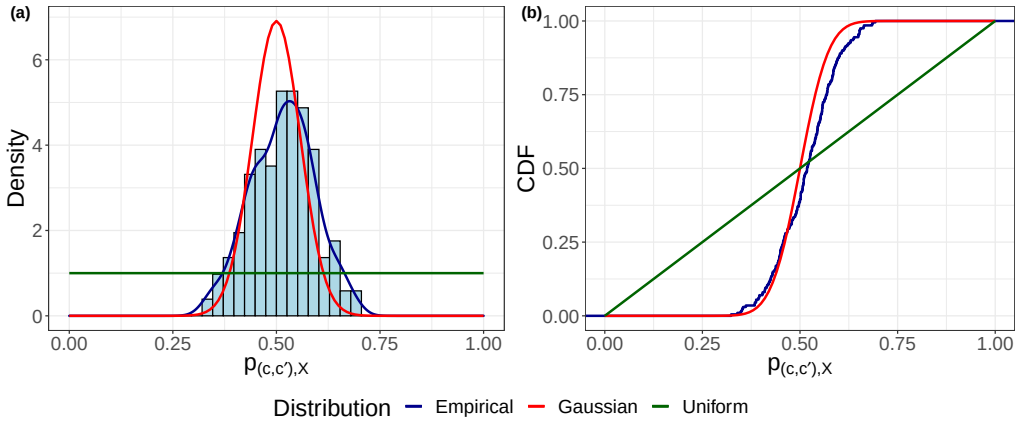


Figure 4.1: Simulation of the null distribution of $p_{(c,c'),\mathcal{X}}$ for $K = 200$ and $B = 25$, chosen in [249]. Left panel (a): histogram and kernel density estimate. Right panel (b): empirical Cumulative Distribution Function (CDF). Red lines: asymptotic Gaussian distribution $\mathcal{N}(1/2, 1/\sqrt{12B})$; green lines: uniform distribution on $[0, 1]$.

The proof of Proposition 4.2.2 is presented in Appendix B. If the p_b were independent, then, by the Central Limit Theorem (e.g. [29, Theorem 27.1]), the distribution of $\overline{p}_B$ would be asymptotically Gaussian $\mathcal{N}(1/2, 1/\sqrt{12B})$ as B tends to infinity. This distribution does not verify (4.5), and therefore tests based on such a distribution are mathematically invalid. In the setting of [249], the p_b are not independent since they have been computed by bootstrapping from one initial sample. However, for small values of B (including the choice $B = 25$ in [249]), the null distribution of (4.1) deviates only slightly from the asymptotic independence setting. This is illustrated in Figure 4.1, where the null distribution of (4.1) is simulated using the parameters chosen in [249]. Details on the simulation and further analyses of the effect of K and B are included in Appendix B.

The empirical distribution of $p_{(c,c'),\mathcal{X}}$ presented in Figure 4.1 does not satisfy Condition (4.5). Moreover, it is extremely conservative for large values of the statistic realization that is, low p-values, yielding an important number of false negatives and thus ignoring substantial differences appearing between the compared samples.

Finally, since the scores $p_{(c,c'),\mathcal{X}}$ are not valid p-values, they cannot be incorporated in a multiple testing procedure [247]. In particular, the Benjamini-Hochberg procedure [18] used in [249] needs the p-values to be Super-Uniform under the null hypothesis to control the False Discovery Rate (FDR). Consequently, using and adjusting (4.1) for multiplicity will yield misleading analyses of the overall behaviour of all the null hypotheses and therefore, inaccurate results when the specificities of individual amino acids are studied *a posteriori*.

Beyond the above-mentioned methodological issues, the approach proposed in [249] presents several practical limitations. It needs, on the one hand, a prior parametric es-

timation of the underlying densities, whose parameters would need to be optimized. On the other hand, it requires a substantial reduction of sample sizes, which may imply an important loss of information in some cases and thus a substantial power reduction. Indeed, the maximum sample size in [249] is set to $N_{\max} = 200$, whereas, for instance, the median sample size for α-helical conformations is 1414 and only 1.16% of the samples have sizes below $N_{\max}$. The goodness-of-fit tests presented in Chapter 3 are non-parametric and use the information provided by entire datasets, as the test statistic is based on the 2-Wasserstein distance. Here, we implemented the testing procedure (N_g-geod) introduced in Section 3.3.1, to detect differences between the codon-specific Ramachandran plots provided in [249].

4.3 Results

In Section 4.3.1, we present the results of implementing (N_g-geod) to detect differences between codon-specific (ϕ, ψ) distributions stratified by DSSP classification. Then, in Section 4.3.2, we repeat the same analysis with a less restrictive classification based only on conformational regions of the Ramachandran space. Finally, we consider the case where (ϕ, ψ) distributions are defined for triplets of amino acids, following the conclusions of Chapter 2.

4.3.1 Structural classification based on DSSP

For each amino acid, we tested all the pairwise differences between all the (ϕ, ψ) distributions of synonymous codons. To facilitate the comparison with the results in [249], we kept only pairs of samples with sizes $n, m \geq 30$ and we divided all conformations according to their secondary structure according to DSSP [151]: Extended strand (E) and α-helix (H). We also performed the analysis for all the conformations not belonging to any of these classes, which we named Others. The same multiplicity correction as in [18] was performed to the computed p-values. The results are presented in Figure 4.2.

The p-value distributions presented in Figure 4.2 indicate that significant differences between codon-specific Ramachandran plots are found for a substantial number of tested hypotheses: 78% for H, 87% for E and and 92% for Others. The results for α-helical structures strongly differ from those presented in [249], where no significant difference was retrieved (see Figure 4 in the original study). In addition, the proportion of significant differences for E is also considerably higher than in [249], where only 39% of the synonymous pairs were identified as structurally distinct. We believe that the observed discrepancies originate from the above-discussed methodological incorrectness of the methods applied in the original study, and in particular the substantial lack of power of the chosen statistic.

Results presented in Figure 4.2 clearly show how the effect of codon on the (ϕ, ψ) distribution is stronger for less rigid structural elements, as suggested in [249]. Indeed, we observe that the null hypothesis is more strongly rejected for extended strand structures

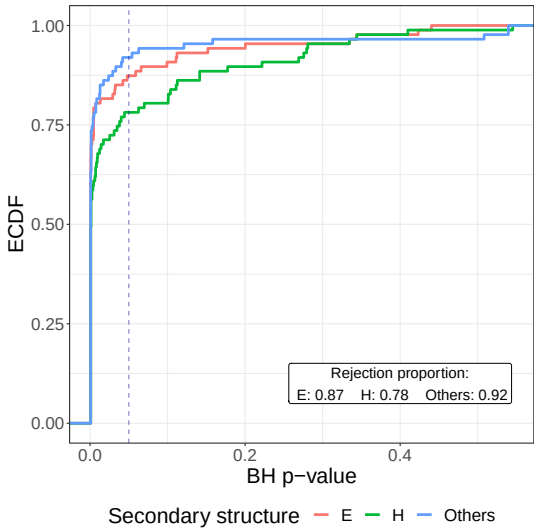


Figure 4.2: Empirical cumulative distribution function (ECDF) of corrected p-values corresponding to testing the equality of (ϕ, ψ) distribution pairs corresponding to different synonymous codons, for conformations in extended strand (E, red), α-helix (H, green) and other (Others, blue) secondary structures. The dashed blue line corresponds to a target FDR set to 0.05, determining the proportion of rejections among each set of tested hypotheses.

than for α-helical ones, but even more strongly in regions that do not belong to any of these categories (blue curve in Figure 4.2). Outside H and E structures, (ϕ, ψ) angles are less constrained, making them potentially more sensitive to the translated codon. These differences in dihedral angle restrictions can be illustrated by measuring the dispersion of (ϕ, ψ) samples belonging to each secondary structure. We defined an estimator D measuring the concentration of one sample around its torus barycenter, which confirmed the previous statements. See, for instance, the average values for the three secondary structures: $\bar{D}_{\text{Others}} = 0.06 > \bar{D}_{\text{E}} = 0.01 > \bar{D}_{\text{H}} = 0.002$. Details and further analyses are provided in Appendix B.

4.3.2 Structural classification as non-overlapping regions of the Ramachandran space

The results presented above, as those of Rosenberg *et al.*, were based on the structural classification provided by DSSP. We performed the same analyses using a less restrictive classification, only considering conformational regions on the Ramachandran space based on non-overlapping angular intervals and disregarding the formation of hydrogen bonds:

$$\mathcal{A} = (-180°, 0°] \times (-120°, 50°], \quad \mathcal{B} = (-180°, 0°] \times (-50°, 240°]. \tag{4.6}$$

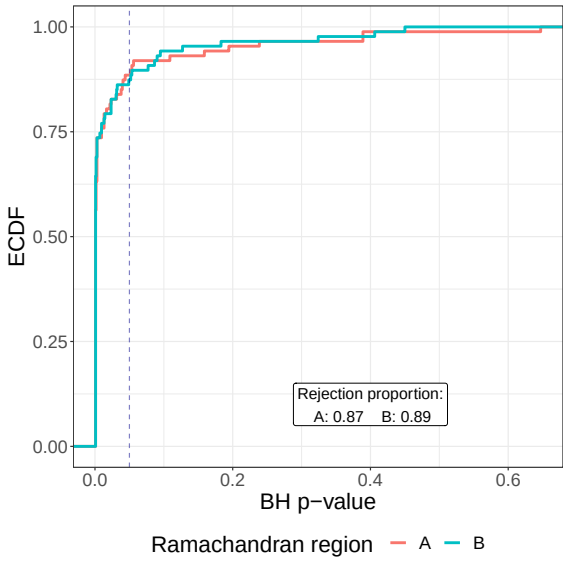


Figure 4.3: Empirical cumulative distribution function (ECDF) of corrected p-values corresponding to testing the equality of (ϕ, ψ) distribution pairs corresponding to different synonymous codons, for conformations in $\mathcal{A}$ and $\mathcal{B}$ classes, defined by the angular intervals (4.6). The dashed blue line indicates an arbitrary level of significance of $\alpha = 0.05$, determining the proportion of rejections among each set of tested hypotheses.

Note that classes $\mathcal{A}$ and $\mathcal{B}$ are not limited to α-helices and extended strands. For instance, poly-l-proline type II (PPII) structures are included in $\mathcal{B}$. Moreover, a substantial number of conformations that were not classified as α-helical (H) or extended stand (E) by DSSP (named 'Others' in Figure 4.2) belong now to the $\mathcal{A}$ or $\mathcal{B}$ classes. More precisely, 37.69% and 44.41% of 'Others' conformations are now contained in $\mathcal{A}$ or $\mathcal{B}$, respectively.

The corresponding results are presented in Figure 4.3. They show that the differences on the rejection power between extended and helical conformations disappear in this case. This indicates that codon-specific effects have similar significance in different areas of Ramachandran space, although the effect may be stronger for a particular type of secondary structure, such as β-strands compared to α-helices.

4.3.3 Tripeptide-specific (ϕ, ψ) distribution analysis

Rosenberg *et al.* considered codon-specific Ramachandran plots corresponding to amino acids with arbitrary neighbors. However, the invalidity of Flory's Isolated Pair Hypothesis [99] and the interdependence of neighbor effects have been demonstrated in -besides several studies [216, 260, 280]- Section 3.4.3 and Chapter 2 respectively.

The consideration of neighboring residues is particularly relevant here, because the dataset in [249] exhibits important discrepancies in the proportion of left and right neigh-

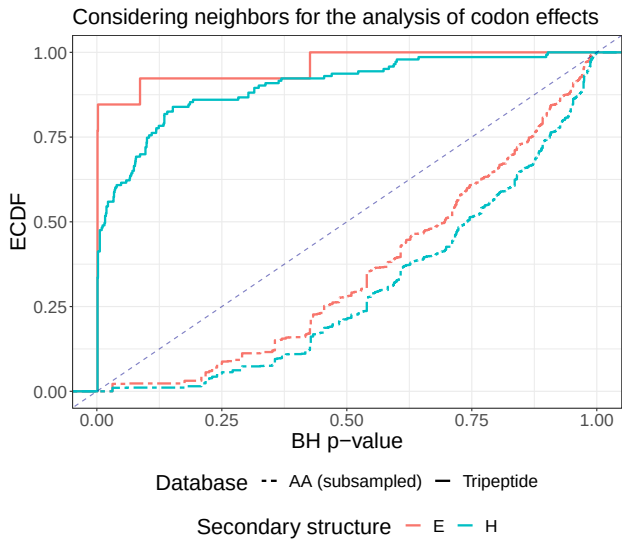


Figure 4.4: Empirical cumulative distribution function (ECDF) of p-values, corrected for multiplicity, corresponding to testing the equality of every synonymous codon (ϕ, ψ) distribution pairs, for conformations in extended strand (E) and α-helix (H) secondary structures. Each line type corresponds to the same analysis performed in one different database. The dashed blue line indicates the cumulative distribution function of a uniform distribution.

boring amino acid types among synonymous codons (see Figure B.3 in Section B.4). When repeating the analyses by considering codon-specific Ramachandran plots for triplets of amino acids, the overall conclusions do not change. However, subtle differences appear if we analyze results more in detail. Here, we illustrate how the codon effect was found to be stronger when neighbors were considered. The quantitative comparison of p-values in both cases (with and without fixing neighbors) is possible only if sample sizes are similar. Therefore, we repeated the analysis for the single-amino-acid Ramachandran plots but reducing sample sizes to $n = 50$, which is the average sample size in the datasets considering triplets. More precisely, for each pair of synonymous codons, we extracted and compared 20 pairs of subsamples of size $n = 50$. Then, the distribution of p-values for the subsampled single-amino-acid datasets can be compared to the ones obtained from the tripeptide datasets. Such comparison is presented in Figure 4.4.

The distributions of p-values presented in Figure 4.4 show that, for comparable sample sizes, the deviations encountered between the distributions for synonymous codons are larger when tripeptides are considered. Indeed, p-values for the analysis using triplets are substantially closer to zero than the ones considering single amino acids. This means that the effect of codon is stronger when neighbors are taken into account or, equivalently, that ignoring neighbor identities -as in [249]- underestimates the codon effect on (ϕ, ψ) distributions.

4.4 Discussion and conclusions

Although quantitatively different, our results, based on an appropriate statistical methodology, confirm those presented in the original study by Rosenberg *et al.* [249], indicating that the nature of the codon has some influence on the fine details of the local conformation in proteins. While the correlation between synonymous codons and secondary structure in the translated proteins is a well known phenomenon, differences at the (ϕ, ψ) level for the most populated conformational states remain an intriguing and somehow counterintuitive observation. In fact, we cannot exclude artifacts related to the procedures. In particular, the nature of the dataset used could explain some of the subtle differences that we have observed. This dataset was derived from a limited set of *Escherichia coli* proteins for which the structure has been experimentally determined, and it was assumed that the gene used for the production of the protein was the same as in the original organism, which is a reasonable assumption in this case, but probably not in general. Moreover, high-resolution crystallographic structures are elucidated in a highly-packed context and at low temperature, severely reducing their inherent conformational fluctuations.

We believe that the detailed understanding of the codon effect on the fine structural features of proteins will only be achieved when extensive structural databases including the corresponding gene sequence are available. With the availability of extensive and accurate datasets, the comparative analysis of codon-specific Ramachandran plots at the amino acid and/or triplet level will be possible using the statistical methods presented here, thus enabling an unambiguous assessment on the influence of the gene sequence on polypeptide structure.

Software availability

The code that reproduces the analyses presented in this chapter is available at https://github.com/gonzalez-delgado/synco. The two testing procedures defined in Chapter 3 for assessing differences between (ϕ, ψ) distributions are implemented in the R package `torustest`, available at https://github.com/gonzalez-delgado/torustest.

Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) through grant ANR-19-P3IA-0004, the LabEx CIMI (ANR-11-LABX-0040) and EpiGenMed (ANR-10-LABX-12-01) within the French State Programme "Investissements d'Avenir", and by the European Research Council under the European Union's H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030] awarded to Pau Bernadó.

Appendix A

Appendix of Chapter 3

Contents

A.1 Geodesics on $\mathbb{T}^2$: practical considerations

This Section is devoted to address some practical questions that arise when defining the test proposed in Section 3.3.1. In Appendix A.1.1, we propose a sampling method to prevent the practitioner from explicitly choosing the N_g geodesics, letting them be chosen randomly with respect to a given distribution. In Section A.1.2, we propose an algorithm to project a pair of samples on $\mathbb{T}^2$ to a given closed geodesic.

A.1.1 Sampling closed geodesics

As the closed geodesics on $\mathbb{T}^2$ are given by the canonical projections of straight lines on $\mathbb{R}^2$ with rational slope, sampling from the set of all closed geodesics is equivalent to sampling from $\mathbb{Q}$, which is a countable set. This prevents the sampling to be uniform, in the sense that geodesics can not be equiprobable. Indeed, if $\mathbb{P}(q) = c$ for all $q \in \mathbb{Q}$, by countable additivity $\mathbb{P}(\mathbb{Q}) = \sum_{q\in\mathbb{Q}} c$, which is zero if $c = 0$ and ∞ otherwise. In consequence, as we have to assign different weights to rational slopes, we will opt for *simpler* geodesics to be more probable, in order to ease computational implementations. To achieve so, we can consider the random variable $Q = A/B$, studied in detail in [227], where B follows a

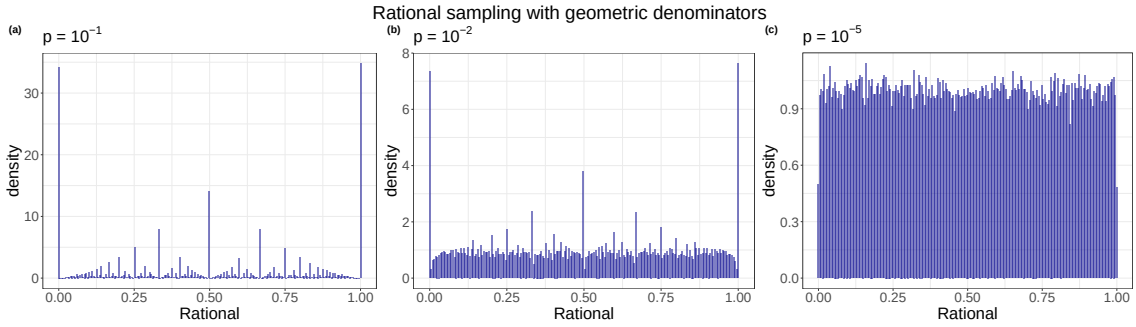


Figure A.1: Histograms representing the distribution of the random variable Q, for different values of the parameter p. For $p = 0.1$ (a), rationals with small values of A and B have more weight and, therefore, simpler geodesics are prioritized.

geometric distribution of parameter p and, for a given denominator $B = b$, A is uniform on $\{0, 1, \ldots, b\}$. Note that Q maps into $\mathbb{Q} \cap [0, 1]$. As p increases, A and B take smaller values and the corresponding geodesics revolt less over the torus. Conversely, when $p \to 0$, $\mathbb{P}(Q = q) \to 0$ for all $q \in \mathbb{Q} \cap [0, 1]$, and Q is *asymptotically equiprobable* [227]. However, small values of p yield extremely high values of A and B and, consequently, unmanageable geodesics with a too-big number of revolutions. The distribution of Q for different values of the parameter p is illustrated in Figure A.1. Here, we will ask $p \geq 0.1$ for computational simplicity.

Note that rationals in $\mathbb{Q} \cap [0, 1]$ yield to straight lines in $\mathbb{R}^2$ whose director vector (B, A) lies in the first (eq. fifth) octant. To cover all the set of closed geodesics, we uniformly assign an octant to each realization of Q and transform its coordinates appropriately. As we would like all the N_g p-values to be independent, we must only accept samples with N_g different geodesics. This may be a problem if N_g is too big, and might require decreasing the value of p. Nevertheless, for a small number ($\lesssim 30$) of geodesics we can keep $p \sim 0.1$ and easily get samples with no repetitions. If one needs to perform the test for large values of N_g, we recommend to explicitly choose geodesics a priori to avoid this problem, leaving the sampling method for controlled values of N_g. The complete sampling procedure is described in Algorithm 1, which takes N_g and p as arguments and retrieves N_g director vectors. In Algorithm 1, $\mathcal{M}_{N_g \times 2}(\mathbb{Z})$ denotes the set of $(N_g \times 2)$-matrices with entire entries and we define $\mathring{\text{gcd}}$ as

$$\mathring{\text{gcd}}(b, a) = \begin{cases} \text{gcd}(b, a) & \text{if} \quad a \neq 0, \\ b & \text{otherwise,} \end{cases}$$

for $a, b \in \mathbb{Z}$ with $b \neq 0$.

Algorithm 1 Geodesics sampling

Require: $N_g \in \mathbb{N}$, $p = 0.1$
Ensure: $G \in \mathcal{M}_{N_g \times 2}(\mathbb{Z})$
  $G \leftarrow 0 \in \mathcal{M}_{N_g \times 2}(\mathbb{Z})$
  while $|\{i = 1, \dots, N_g \,:\, G_{ik} = G_{jk}\ \forall k \in \{1,2\} \quad \text{for any } j \in \{1, \dots, N_g\} \backslash \{i\}\}| > 0$ **do**
    for $i \leftarrow 1$ to N_g **do**
      $b \leftarrow \mathcal{G}(p)$
      $a \leftarrow \mathcal{U}(\{0, 1, \dots, b\})$
      $u \leftarrow (b, a)/\mathring{gcd}(b, a)$ ▷ Director vector in $\mathbb{R}^2$.
      $o \leftarrow \mathcal{U}(\{1, 2, 3, 4\})$ ▷ Octant of the upper semi-circle.
      if $o = 2$ **then**
        $u \leftarrow (a, b)/\mathring{gcd}(b, a)$
      else if $o = 3$ **then**
        $u \leftarrow (-b, a)/\mathring{gcd}(b, a)$
      else if $o = 4$ **then**
        $u \leftarrow (-a, b)/\mathring{gcd}(b, a)$
      end if
      $G_{i\cdot} \leftarrow u$
    end for
  end while

A.1.2 Projection to a closed geodesic

Let $a, b \in \mathbb{Z}$, with $b \neq 0$, and $u = (b, a)$ the director vector of a straight line r_u^0 containing the origin $(0, 0)$. Let I_b be the real interval $(\min(b, 0), \max(b, 0))$, being I_a analogously defined. We aim to project a pair of samples into the geodesic given by the canonical projection of r_u^0. To do so, we first consider the finite set $\mathcal{P}_u$ of the points in $I_b \times I_a$ where r_u^0 cuts the lines $x = z_b$, $y = z_a$ for $z_b \in I_b \cap \mathbb{Z}$ and $z_a \in I_a \cap \mathbb{Z}$:

$$\mathcal{P}_u = \{(x, z) \,:\, x \in I_b, z \in \mathbb{Z}\} \cap \{(z, y) \,:\, y \in I_a, z \in \mathbb{Z}\} \cap r_u^0.$$

An example is presented in Figure A.2a. Then, we consider the set $\mathcal{L}_u$ of straight lines of director vector u and containing the points of $\mathcal{P}_u \cup \{(0, 1), (1, 0), (0, 0)\}$ transferred to $[0, 1] \times [0, 1]$ by subtracting the integer part of its coordinates. If we denote r_v^p the straight line containing $p = (p_x, p_y) \in \mathbb{R}^2$ and having v as director vector, we can write $\mathcal{L}_u$ as follows

$$\mathcal{L}_u = \{r_u^q \,:\, q = (p_x - [p_x], p_y - [p_y]),\ p \in \mathcal{P}_u \cup \{(0, 1), (1, 0), (0, 0)\}\}.$$

This is illustrated in Figure A.2b. In a first step, each point in $[0, 1] \times [0, 1]$ will be projected to the closest straight line in $\mathcal{L}_u$. Then, projections (x_u, y_u) outside $[0, 1] \times [0, 1]$ will be replaced by the elements $(x'_u, y'_u) \in [0, 1] \times [0, 1]$ such that $(x_u, y_u)\mathcal{R}(x'_u, y'_u)$, where $\mathcal{R}$ is

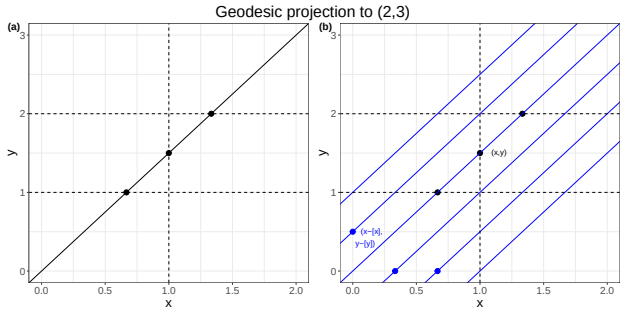


Figure A.2: First steps of the projection algorithm for the closed geodesic corresponding to the straight line of director vector $u = (2, 3)$. The three points in black constitute the ensemble $\mathcal{P}_u$. In (b), points of $\mathcal{P}_u$ are transferred to $[0, 1] \times [0, 1]$ by subtracting to their coordinates their integer parts. The blue lines are the elements of $\mathcal{L}_u$.

the one defined in the beginning of Section 3.2. These two steps are depicted in Figure A.3.

The last step is to relocate all the projections on $\mathbb{R}/\mathbb{Z}$. To do so, we put the segments $\mathcal{L}_u \cap ([0, 1] \times [0, 1])$ in order, following the spiral path. This corresponds to transfer back the points to the straight line r_u^0 of Figure A.2a. Let $(x_u, y_u) \in r_u^p \in \mathcal{L}_u$. The element $t_u \in \mathbb{R}/\mathbb{Z}$ will be parameterized as

$$t_u = \frac{\|\tilde{p}\| + \|(x_u, y_u)\|}{\|u\|} \in [0, 1),$$

where $\tilde{p} \in \mathcal{P}_u \cup \{(0, 0)\}$ is the one such that $p_x = \tilde{p}_x - [\tilde{p}_x]$ and $p_y = \tilde{p}_y - [\tilde{p}_y]$.

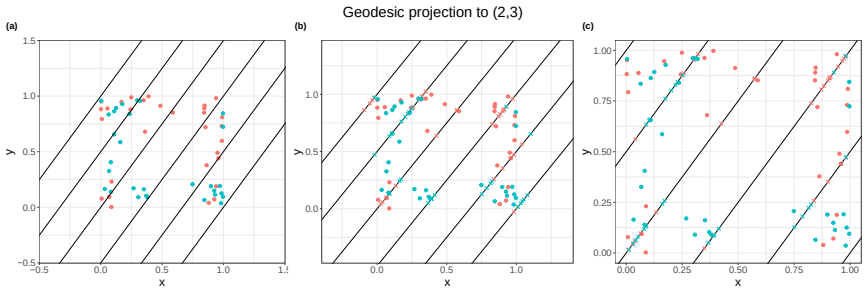


Figure A.3: Projection to the closed geodesic given by the director vector $u = (2, 3)$ of a pair of samples of size $n = m = 30$ drawn from a uniform distribution on $\mathbb{T}^2$. Black lines are the elements of $\mathcal{L}_u$. In (a), the given samples distinguished by colors. In (b), their projections to the closest line in $\mathcal{L}_u$ are represented by colored crosses. In (c), projections outside $[0, 1] \times [0, 1]$ are relocated in $[0, 1] \times [0, 1]$ according to the equivalence relation $\mathcal{R}$.

A.2 Proofs

A.2.1 Proofs of Section 3.2

Proof of Theorem 3.2.1. Recall that we denote the interior of the support of a measure μ (over $\mathbb{T}^d$ or $\mathbb{R}^d$) as $\mathcal{X}_\mu$. Since $\mathbb{T}^d$ is a Polish space, Theorem 4.1 in [299] implies that there exists a solution π^* of (3.2). Additionally, Theorem 5.10 in [299] establishes that $\text{supp}(\pi^*)$ is d^p-cyclically monotone. More precisely, by Theorem 5.10 in [299], this support lies on the graph of the d^p-differential

$$\partial^{d^p} f(\boldsymbol{x}) = \{\boldsymbol{y} : \ f(\boldsymbol{z}) \leq f(\boldsymbol{x}) + d^p(\boldsymbol{z}, \boldsymbol{y}) - d^p(\boldsymbol{x}, \boldsymbol{y}), \text{ for all } \boldsymbol{z} \in \mathbb{T}^d\}$$

of a function f solving (3.3). Its graph is denoted by $\partial^{d^p} f = \{(\boldsymbol{x}, \boldsymbol{y}) : \ \boldsymbol{y} \in \partial^{d^p} f(\boldsymbol{x})\}$. These definitions of d^p-differential and d^p-concave functions apply verbatim to $\|\cdot\|^p$-differential and $\|\cdot\|^p$-concave functions with the obvious notation. Let Γ be the set defined in (3.6), $\{(\mathbf{x}_k + \mathbf{p}_k, \mathbf{y}_k + \mathbf{p}_k)\}_{k=1}^n \subset \Gamma$ be a sequence and $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ be a bijection. Then, the definition of Γ implies that

$$\begin{aligned}
\sum_{k=1}^n \|\mathbf{x}_k - \mathbf{y}_k\|^p &= \sum_{k=1}^n d^p(\boldsymbol{x}_k, \boldsymbol{y}_k) \\
&\leq \sum_{k=1}^n d^p(\boldsymbol{x}_k, \boldsymbol{y}_{\sigma(k)}) \\
&\leq \sum_{k=1}^n \|\mathbf{x}_k + \mathbf{p}_k - \mathbf{y}_{\sigma(k)} - \mathbf{p}_{\sigma(k)}\|^p,
\end{aligned}$$

which means that Γ is $\|\cdot\|^p-$cyclically monotone. Therefore, $\Gamma \subset \partial^{\|\cdot\|^p} \varphi_p$, for some $\|\cdot\|^p$-concave function φ_p. Now, recall from Theorem 3.3 and Proposition 3.4 in [102], that

1. The set of differentiablity

$$\text{dom}(\nabla \varphi_p) = \left\{ \mathbf{x} \in \mathbb{R}^d : \ \partial^{\|\cdot\|^p} \varphi_p = \left\{ \mathbf{x} - \left(\frac{1}{p} \|\nabla \varphi_p(\mathbf{x})\| \right)^{\frac{2-p}{p-1}} \nabla \varphi_p(\mathbf{x}) \right\} \right\}$$

   has full Lebesgue measure in $\text{dom}(\varphi_p) = \{\mathbf{x} \in \mathbb{R}^d : \ \varphi_p(\mathbf{x}) \in \mathbb{R}\} \supset \mathcal{X}_{\mu_P}$,

2. The relation $\mathbf{S}_p(\mathbf{x}) = \mathbf{x} - \left(\frac{1}{p} \|\nabla \varphi_p(\mathbf{x})\| \right)^{\frac{2-p}{p-1}} \nabla \varphi_p(\mathbf{x})$ defines a Borel function in $\text{dom}(\nabla \varphi_p)$, and

3. The equality $\{\mathbf{S}_p(\mathbf{x})\} = \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in \Gamma\}$ holds for all $\mathbf{x} \in \text{dom}(\nabla \varphi_p)$.

Since $\Gamma \subset \partial^{\|\cdot\|^p} \varphi_p$, this means that, for all $\mathbf{x} \in \text{dom}(\nabla \varphi_p)$, there exists an unique $\mathbf{y}_\mathbf{x} = \mathbf{S}_p(\mathbf{x})$ such that $(\mathbf{x}, \mathbf{y}_\mathbf{x}) \in \Gamma$. We observe that, due to the fact that $\mu_P \ll \ell_d$, the measure $\gamma^* = (\mathbf{Id} \times \mathbf{S}_p) \# \mu_P$ on $\mathbb{R}^d \times \mathbb{R}^d$ is well defined, its support is $\|\cdot\|^p$-cyclically monotone

and its first marginal is μ_P. We claim that the second marginal is μ_Q. Let $(\boldsymbol{x}, \boldsymbol{y}) \in \pi^*$ be such that $\mathbf{x} + \mathbf{p} \in \mathrm{dom}(\nabla\varphi_p)$, for all $\mathbf{p} \in \mathbb{Z}^d$. Then, for any representative pair, call it $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$, there exist $\mathbf{p}, \mathbf{p}' \in \mathbb{Z}^d$ such that $(\mathbf{x} + \mathbf{p}, \mathbf{y} + \mathbf{p}') \in \Gamma$. Since

$$\begin{aligned}\{\mathbf{S}_p(\mathbf{x}) + \mathbf{p}\} &= \{\mathbf{y} + \mathbf{p} : \ (\mathbf{x}, \mathbf{y}) \in \Gamma\} \\ &= \{\mathbf{y} : \ (\mathbf{x} + \mathbf{p}, \mathbf{y}) \in \Gamma\} \\ &= \{\mathbf{S}_p(\mathbf{x} + \mathbf{p})\} = \{\mathbf{y} + \mathbf{p}'\},\end{aligned}$$

the relation $\boldsymbol{y} = \overline{\mathbf{S}_p(\boldsymbol{x})}$ holds. Since $\mathbf{x} + \mathbf{p} \in \mathrm{dom}(\nabla\varphi_p)$, for all $\mathbf{p} \in \mathbb{Z}^d$, which is the intersection of sets of full μ_P-measure, the relation $\boldsymbol{y} = \overline{\mathbf{S}_p(\boldsymbol{x})}$ happens μ_P-a.e. This means that $\pi^* = (\mathbf{Id} \times \overline{\mathbf{S}_p(\cdot)})\#P$, which proves automatically the claim. Consequently, the existence is proven.

The uniqueness follows from the proof of Corollary 2.4. in [102]. Indeed, we can define the set

$$S = \bigcup_{\pi^* \text{ solving (3.2)}} \Gamma(\pi^*),$$

where $\Gamma(\pi^*)$ is defined as in (A.1) for each π^* solving (3.2). Therefore, taking any finite sequence $\{(\mathbf{x}_k + \mathbf{p}_k, \mathbf{y}_k + \mathbf{p}_k)\}_{k=1}^n \subset S$, there exists at most n different probability measures π_k, for $k = 1, \dots, n$, such that $(\mathbf{x}_k + \mathbf{p}_k, \mathbf{y}_k + \mathbf{p}_k) \in \Gamma(\pi_k)$. As all of them are solutions of (3.2) we have, due to the linearity of the optimization in (3.2) and the convexity of the set $\Gamma(P, Q)$, that the mean $\pi_0 = \frac{1}{n}\sum_{k=1}^n \pi_k$ is also a solution. Then, its support is contained in a d^p-cyclically monotone set, and $\Gamma(\pi_0)$ is $\|\cdot\|^p$-cyclically monotone, since it contains the sequence $\{(\mathbf{x}_k + \mathbf{p}_k, \mathbf{y}_k + \mathbf{p}_k)\}_{k=1}^n \subset S$. Consequently, S is $\|\cdot\|^p$-cyclically monotone.

To conclude, repeating the previous arguments, there exists a $\|\cdot\|^p$-concave function f^S such that $S \subset \partial^d f^S$. Moreover, for any other φ_p, defined as before, it holds that $\partial^d\varphi_p \subset \partial^d f^S$. Then, the equality

$$\mathbf{x} - \left(\frac{1}{p}\|\nabla\varphi_p(\mathbf{x})\|\right)^{\frac{2-p}{p-1}} \nabla\varphi_p(\mathbf{x}) = \mathbf{x} - \left(\frac{1}{p}\|\nabla f^S(\mathbf{x})\|\right)^{\frac{2-p}{p-1}} \nabla f^S(\mathbf{x})$$

holds μ_P-a.e. This proves the uniqueness of $\mathbf{S}_p$ and, consequently, the one of $\mathbf{T}_p$

□

Proof of Theorem 3.2.2. We set $(\mathbf{x}, \mathbf{y}) \in \Gamma$ and observe that $d(\boldsymbol{x}, \boldsymbol{y}) = \|\mathbf{x} - \mathbf{y}\|$. Since $(\boldsymbol{x}, \boldsymbol{y}) \in \mathrm{supp}(\pi^*)$, Theorem 5.10 in [299] establishes that if (f, g) solves (3.3), it holds

$$f(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in \mathbb{T}^d} \{d(\boldsymbol{x}, \boldsymbol{y})^p - g(\boldsymbol{y})\}.$$

Since, for each $(\mathbf{x}, \mathbf{y})$, there exists $\mathbf{p} \in \mathbb{Z}^d$ such that $d(\boldsymbol{x}, \boldsymbol{y}) = \|\mathbf{x} - \mathbf{y} - \mathbf{p}\|$, we can directly replace $\mathbf{y}$ by $\mathbf{y} + \mathbf{p}$ in the infimum without altering any of the terms. This yields the equality

$$f(\boldsymbol{y}) = \inf_{z, \mathbf{y} \in \in \mathbb{R}^d \ \boldsymbol{z} = \boldsymbol{x}} \{\|\mathbf{z} - \mathbf{y}\|^p - g(\boldsymbol{y})\},$$

and allows to define the following periodic $\|\cdot\|^d$-concave function in $\mathbb{R}^d$:

$$\hat{\varphi}_p(\mathbf{x}) = \inf_{\mathbf{y}\in\in\mathbb{R}^d}\{\|\mathbf{x}-\mathbf{y}\|^p - g(\boldsymbol{y})\} = f(\boldsymbol{x}).$$

We claim that $\nabla\hat{\varphi}_p = \nabla\varphi_p$ for μ_P-a.e., which implies the equality of both $\hat{\varphi}_p$ and φ_p, in each connected component of $\mathrm{supp}(\mu_P)$. By assumption, $\mathrm{supp}(\mu_P) = \bigcup_{p\in\mathbb{Z}^d} p + A$ is a union of connected sets. By periodicity we can restrict our study to the connected set A, where the claim yields $\nabla\hat{\varphi}_p = \nabla\varphi_p$ for ℓ_d-a.e. We can apply Theorem 2.6 in [74] to conclude that $\varphi_p = \hat{\varphi}_p + C$ in A, thus in $\mathrm{supp}(\mu_P)$. We prove now the claim. Let π^* be a measure solving (3.2), we know (from Theorem 5.10 in [299]) that its support lies in the graph of $\partial^{d^p} f$. Therefore, we can define the following $\|\cdot\|^p$-cyclically monotone set (note that this is true by repeating the same arguments as for Γ):

$$\begin{aligned}\Gamma(\partial^{d^p} f) = \{&(\mathbf{x}+p, \mathbf{y}+p) : \\ &(\boldsymbol{x},\boldsymbol{y}) \in \partial^{d^p} f,\ \mathbf{x}\in[0,1]^d,\ d(\boldsymbol{x},\boldsymbol{y}) = \|\mathbf{x}-\mathbf{y}\| \text{ and } p\in\mathbb{Z}^d\},\end{aligned} \tag{A.1}$$

which satisfies the relation $\Gamma(\pi^*) \subset \Gamma(\partial^{d^p} f)$, with the notation of the proof of Theorem 3.2.1. Recall that the relation $\Gamma(\pi^*) \subset \partial^{\|\cdot\|^p}\varphi_p$ also holds. Moreover, by definition we have $\Gamma(\partial^{d^p} f) \subset \partial^{\|\cdot\|^p}\hat{\varphi}_p$. Since μ_P-a.e. the sets $\partial^{\|\cdot\|^p}\hat{\varphi}_p(\mathbf{x})$ and $\partial^{\|\cdot\|^p}\varphi_p(\mathbf{x})$ are singletons, and, for μ_P-a.e. $\mathbf{x}$, there exists at least one $\mathbf{y}\in\mathbb{R}^d$ such that $(\mathbf{x},\mathbf{y}) \in \Gamma(\pi^*)$, then $\partial^{\|\cdot\|^p}\hat{\varphi}_p(\mathbf{x}) = \partial^{\|\cdot\|^p}\varphi_p(\mathbf{x})$. This implies that the functions $\mathbf{S}_p(\mathbf{x}) = \mathbf{x} - \left(\frac{1}{p}\|\nabla\varphi_p(\mathbf{x})\|\right)^{\frac{2-p}{p-1}}\nabla\varphi_p(\mathbf{x})$ and $\hat{\mathbf{S}}_p(\mathbf{x}) = \mathbf{x} - \left(\frac{1}{p}\|\nabla\hat{\varphi}_p(\mathbf{x})\|\right)^{\frac{2-p}{p-1}}\hat{\varphi}_p(\mathbf{x})$ are equal μ_P-a.e., which proves the claim. Note that, under continuity of the optimal transport potential, their uniqueness only need to be fulfilled μ_P-a.e. □

Proof of Lemma 3.2.4. Set $\boldsymbol{x},\boldsymbol{z}\in\mathrm{dom}(f)$. Then, by definition

$$\begin{aligned}|f(\boldsymbol{x}) - f(\boldsymbol{z})| &= |\inf_{\boldsymbol{y}\in\mathbb{T}^d}\{d^p(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{y})\} - \inf_{\boldsymbol{y}\in\mathbb{T}^d}\{d^p(\boldsymbol{z},\boldsymbol{y}) - g(\boldsymbol{y})\}| \\ &= |\inf_{\boldsymbol{y}\in\mathbb{T}^d}\{d^p(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{y})\} + \sup_{\boldsymbol{y}\in\mathbb{T}^d}\{-d^p(\boldsymbol{z},\boldsymbol{y}) + g(\boldsymbol{y})\}| \\ &\le \sup_{\boldsymbol{y}\in\mathbb{T}^d}|d^p(\boldsymbol{x},\boldsymbol{y}) - d^p(\boldsymbol{z},\boldsymbol{y})\}|.\end{aligned}$$

The mean value theorem yields the inequality $a^p - b^p \le p|a-b|(a^{p-1}+b^{p-1})$, which holds for any $a,b\ge 0$. Then, the triangle inequality for d leads to

$$\begin{aligned}|f(\boldsymbol{z}) - f(\boldsymbol{x})| \le p\, d(\boldsymbol{z},\boldsymbol{x}) \sup_{\boldsymbol{y}\in\mathbb{T}^d}&|d^{p-1}(\boldsymbol{x},\boldsymbol{y}) + d^{p-1}(\boldsymbol{z},\boldsymbol{y})\} \\ &\le 2\,p\,d(\boldsymbol{z},\boldsymbol{x}) \sup_{\boldsymbol{z},\boldsymbol{x}\in\mathbb{T}^d}\left(d^{p-1}(\boldsymbol{z},\boldsymbol{x})\right) \le 2\,p\,d^{\frac{p-1}{2}}\,d(\boldsymbol{z},\boldsymbol{x})\end{aligned}$$

where the $d^{\frac{p-1}{2}}$ term comes from the trivial bound of the diameter of $\mathbb{T}^d$. This concludes the proof. □

Proof of Theorem 3.2.5. Set $\bar{p} \in \mathcal{X}_P$ and assume that $f_p(\bar{p}) = 0$. Set $\epsilon_m \to 0$ and consider the sequence of balls $\mathbb{B}_{\epsilon_m}(\bar{p}) \subset \text{supp}(P)$, centered at $\bar{p}$ with radius ϵ_n. Since the ball is a continuity set of P, after Portmanteau's Theorem, $P_n \xrightarrow{w} P$ implies that for each m there exists a n_m such that P_n gives mass to $\mathbb{B}_{\epsilon_m}(\bar{p})$ for all $n \geq m_n$. Then, we can extract a sequence $\bar{p}_n \to \bar{p}$ such that $\bar{p}_n \in \mathcal{X}_{P_n}$. As a consequence, we have that $f_n(\bar{p}_n) \in \mathbb{R}$, and we can set $a_n = -f_n(\bar{p}_n)$ and define $h_n = f_n + a_n$. Recall from Lemma 3.2.4 that all such functions are L-Lipschitz in their respective domains. Kirszbraun's Theorem (Theorem B in [166]) implies that, without loss of generality, we can consider that h_n (resp. f_p) are $2p$-Lipschitz functions defined in the whole $\mathbb{T}^d$. The previous reasoning implies that $\{h_n\}_{n\in\mathbb{N}}$ is point-wise bounded for the compact sequence $\{\bar{p}_n\}_{n\in\mathbb{N}}$. Since all such functions are $2p$-Lipschitz, then Arzelá-Ascoli's Theorem concludes that every subsequence $\{h_{n_k}\}_{k\in\mathbb{N}}$ admits a convergent subsequence $\{h_{n_{k_j}}\}_{j\in\mathbb{N}}$. Let h be one of those limits. Note that the d^p-conjugation is continuous in the sense that

$$\begin{aligned}|h_n^{d^p}(\boldsymbol{x}) - h^{d^p}(\boldsymbol{x})| = |\inf_{\boldsymbol{y}}\{d^p(\boldsymbol{y},\boldsymbol{x}) - h_n(\boldsymbol{x})\} - \inf_{\boldsymbol{y}}\{d^p(\boldsymbol{y},\boldsymbol{x}) - h(\boldsymbol{x})\}| \\ \leq \sup_{\boldsymbol{y}}\{h_n(\boldsymbol{x}) - h(\boldsymbol{x})\} = \|h_n - h\|_\infty,\end{aligned}$$

for all $\boldsymbol{y} \in \mathbb{T}^d$. By assumption, we have

$$A_n = \int h_n d\alpha_n + \int h_n^{d^p} d\beta_n - \int f_p d\alpha - \int f_p^{d^p} d\beta \to 0,$$

and

$$\begin{aligned}\int h d\alpha + \int h^{d^p} d\beta = \\ \int h d(\alpha - \alpha_n) + \int h^{d^p} d(\beta - \beta_n) + \int (h_n - h) d\alpha_n + \int (h_n^{d^p} - h^{d^p}) d\beta_n.\end{aligned}$$

Then, the inequality $|\int (h_n - h) d\alpha_n| \leq \|h_n - h\|_\infty$ gives $\int h d\alpha + \int h^{d^p} d\beta = 0$. The function h is thus an optimal transport potential. The uniqueness described in Theorem 3.2.2 and the fact that $\bar{p}_n \to \bar{p}$ and $h_n(\bar{p}_n) = f_p(\bar{p}) = 0$ conclude that f_p is the unique possible limit of such subsequences in $\text{dom}(f_p)$.

□

Proof of Theorem 3.2.6. Note that as Theorem 3.2.5 holds, since probability measures are supported in a compact set, the torus, then the reasoning of [77] can be imitated. Here the main steps of the proof for the one-sample case are given. For further details about the proof we refer to the original text.

Efron-Stein inequality, see Chapter 3.1 in [37], states that if $(X_1', \ldots, X_n')$ is an independent copy of $(X_1, \ldots, X_n)$, then we have the bound

$$\begin{aligned}\text{Var}(f(X_1, \ldots, X_n)) \leq \sum_{i=1}^n \mathbb{E}(f(X_1, \ldots, X_n) \\ - f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).)_+^2.\end{aligned}$$

Moreover, if $X_1, \dots, X_n$ are i.i.d, such inequality can be written as

$$\mathrm{Var}(f(X_1, \dots, X_n)) \leq n\mathbb{E}(f(X_1, \dots, X_n) - f(X_1', \dots, X_n))_+^2.$$

Set the empirical measures $P_n = \frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ and $P_n' = \frac{1}{n}(\delta_{X_1'} + \sum_{k=2}^n \delta_{X_k})$, and the values $R_n = \mathcal{T}_p(P_n, Q) - \int f_p dP_n$ and $R_n' = \mathcal{T}_p(P_n', Q) - \int f_p dP_n'$. Let f_n and f_n' be solutions of the dual problem (3.3) of $\mathcal{T}_p(P_n, Q)$ and $\mathcal{T}_p(P_n', Q)$ respectively. Then from (3.3) we derive that

$$\begin{aligned}(R_n - R_n')_+ \leq & \frac{1}{n}|f_n(X_1) - f_p(X_1) - f_n(X_1') + f_p(X_1')| \\ & + |f_n'(X_1) - f_p(X_1) - f_n'(X_1') + f_p(X_1')|,\end{aligned}$$

which together with Theorem 3.2.5 yields

$$n(R_n - R_n')_+ \xrightarrow{a.s.} 0.$$

Since the probability measures are supported in the torus, which is compact, then $n^2\mathbb{E}(R_n - R_n')_+^2 \to 0$. Finally, we conclude by the so-called Efron-Stein's inequality.

□

A.2.2 Proofs of Section 3.3

Proof of Lemma 3.3.1. If G is the distribution function of the uniform distribution on $\mathbb{R}/\mathbb{Z}$, we have that

$$(G - \alpha)^{-1}(t) = \inf\{s \,:\, s > t + \alpha\} = t + \alpha. \tag{A.2}$$

Plugging (A.2) in (3.18), we have

$$\mathcal{T}_2(P^c, U) = \inf_{\alpha \in \mathbb{R}} \int_0^1 \left(F^{-1}(t) - t - \alpha\right)^2 dt, \tag{A.3}$$

where the optimal value for α can be found by analytically minimizing the function

$$\begin{aligned}H(\alpha) = \int_0^1 \left(F^{-1}(t) - t - \alpha\right)^2 dt = \int_0^1 \left(F^{-1}(t) - t\right)^2 dt + \\ \alpha^2 - 2\alpha \int_0^1 \left(F^{-1}(t) - t\right) dt,\end{aligned}$$

which satisfies $H'(\alpha) = 0 \Leftrightarrow \alpha = \int_0^1 \left(F^{-1}(t) - t\right) dt$. □

Proof of Lemma 3.3.2. Let $\mathcal{D}([0,1])$ denote the Banach space of right-continuous functions on $[0,1]$ with left limits. Donsker's Theorem [30, Theorem 14.3], states the weak convergence in $\mathcal{D}([0,1])$ of the empirical process $\sqrt{n}(F_n - F)$ for $n \to \infty$ to the standard Brownian bridge $\mathbb{B}(F(t))$. As the operator $h : \mathcal{D}([0,1]) \longrightarrow \mathbb{R}$ defined as

$$h(f) = \int_0^1 \left(f(t) - \int_0^1 f(s)\, ds\right)^2 dt = \int_0^1 f(t)^2\, dt - \left(\int_0^1 f(s)\, ds\right)^2 \tag{A.4}$$

is continuous, the continuous mapping Theorem [291, Theorem 1.3.6] yields that

$$n\mathcal{T}_2(P_n^c, U) \xrightarrow[n]{w} \int_0^1 \mathbb{B}(t)^2\, dt - \left(\int_0^1 \mathbb{B}(t)\, dt\right)^2, \tag{A.5}$$

when $P^c = U$, which concludes the proof. □

Proof of Proposition 3.3.3. Keeping the notation of the proof of Lemma 3.3.2, after Theorem 1 in [239] we have that, when $P^c = Q^c$,

$$\sqrt{\frac{nm}{n+m}}\left(G_m^{-1}(F_n) - \mathbb{I}\right) \xrightarrow[n,m]{w} \mathbb{B}(t), \tag{A.6}$$

in $\mathcal{D}([0,1])$, where $\mathbb{I}$ denotes the identity function. Finally, using the same arguments as in the proof of Lemma 3.3.2, the result is proved. □

Proof of Proposition 3.3.4. Note that

$$\mathbb{P}(\pi_{nm}^c = 1) = \mathbb{P}\left(T_{nm}^c \geq c_{nm}^c(\alpha)\right) = \mathbb{P}\left(\mathcal{T}_2(G_m\#P_n, U) \geq \frac{n+m}{nm} c_{nm}^c(\alpha)\right).$$

On one hand, we have that

$$c_{nm}^c(\alpha) = \inf\{t > 0 : F_{nm}^c(t) \geq 1-\alpha\} = \inf\{t > 0 : \mathbb{P}_{H_0}(T_{nm}^c > t) \leq \alpha\} = \frac{nm}{n+m} \inf\{t > 0 : \mathbb{P}_{H_0}(\mathcal{T}_2(G_m\#P_n, U) > t) \leq \alpha\}.$$

Under the null hypothesis, $G_m\#P_n \xrightarrow{w} U$. Thus, $\mathcal{T}_2(G_m\#P_n, U) \to 0$ in probability (recall Section 3.2.2). In consequence, for every $\varepsilon > 0$ and every $\alpha > 0$, we have

$$\frac{n+m}{nm}\, c_{nm}^c(\alpha) \leq \varepsilon \tag{A.7}$$

for sufficiently large n, m. On the other hand, when $P^c \neq Q^c$, $\mathcal{T}_2(G_m\#P_n, U) \to \mathcal{T}_2(G\#P, U) > 0$ in probability which, together with (A.7), proves the result. □

Size of (N_g-geod). Let us prove that (N_g-geod) controls the type I error at any significance level $\alpha > 0$. Indeed, if H_0 denotes the null hypothesis (3.13), we have

$$\mathbb{P}_{H_0}\left(\pi_{nm,N_g}^g = 0\right) = \mathbb{P}_{H_0}\left(\min_{i=1}^{N_g} p_i \leq \frac{\alpha}{N_g}\right) = \mathbb{P}_{H_0}\left(\bigcup_{i=1}^{N_g}\left\{p_i \leq \frac{\alpha}{N_g}\right\}\right) \leq \sum_{i=1}^{N_g} \mathbb{P}_{H_0}\left(p_i \leq \frac{\alpha}{N_g}\right) = N_g \frac{\alpha}{N_g} = \alpha, \tag{A.8}$$

where the first equality in (A.8) is ensured as every p_i follows a uniform distribution under the null. □

Proof of Proposition 3.3.5. Suppose that $P_j^c \neq Q_j^c$ w.l.o.g. for some $j \in \{1, \ldots, N_g\}$. After (N_g-geod), we have

$$\mathbb{P}(\pi^g_{nm,N_g} = 1) = \mathbb{P}\left(\min_{i=1}^{N_g} p_i \leq \frac{\alpha}{N_g}\right) \geq \mathbb{P}\left(p_j \leq \frac{\alpha}{N_g}\right).$$

Then, as the right side of the previous inequality tends to 1 after Proposition 3.3.4, so does the left side, which ends the proof. □

Proof of Remark 3.3.6. Suppose that $\mu_P \ll \ell_2$ and project with respect a given direction $\mathbf{e}_1$. As an immediate consequence of the monotone convergence theorem, $\ell_2(A \times \mathbb{R}) = 0$, for any Lebesgue null set $A \subset \mathbb{R}$. Consequently, by hypothesis $0 = \mu_P(A \times \mathbb{R}) = \mu_P^1(A)$. Here, μ_P^1 is the projected measure of μ_P to the direction $\mathbf{e}_1$. Then, for any null set B in $\mathbb{R}/\mathbb{Z}$ the leveraged set $\tilde{B} = \bigcup_{s\in\mathbb{Z}}(s + B)$ is a Lebesgue null set, so that $\mu_P^1(\tilde{B}) = 0$ and $P^c(B) = 0$. □

Proof of Theorem 3.3.7. Note that $\mathcal{T}_2(P_n, Q_m) = \mathcal{T}(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is a function of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. For each $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{T}^d$ and $\mathbf{x}' \in \mathbb{T}^d$ let π and π' be both joint measures such that

$$\begin{aligned}\mathcal{T} := \sum_{i,j} d(\mathbf{x}_i - \mathbf{y}_j)^2 \pi_{i,j} &= \mathcal{T}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_m) \\ s.t. \quad \sum_{i,j} \pi_{i,j} &= \frac{1}{n}, \quad j = 1, \ldots, m, \\ \sum_{i,j} \pi_{i,j} &= \frac{1}{m}, \quad i = 1, \ldots, n,\end{aligned}$$

and

$$\begin{aligned}\mathcal{T}' := \sum_{j} d(\mathbf{x}'_1 - \mathbf{y}_j)^2 \pi'_{1,j} + \sum_{i>1,j} d(\mathbf{x}_i - \mathbf{y}_j)^2 \pi'_{i,j} &= \mathcal{T}(\mathbf{x}'_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_m) \\ s.t. \quad \sum_{i,j} \pi'_{i,j} &= \frac{1}{n}, \quad j = 1, \ldots, m, \\ \sum_{i,j} \pi'_{i,j} &= \frac{1}{m}, \quad i = 1, \ldots, n.\end{aligned}$$

Then we have that

$$\mathcal{T}' \leq \sum_{j} d(\mathbf{x}'_1 - \mathbf{y}_j)^2 \pi_{1,j} + \sum_{i,j} d(\mathbf{x}_i - \mathbf{y}_j)^2 \pi_{i,j},$$

which implies

$$\begin{aligned}\mathcal{T}' - \mathcal{T} &\leq \sum_j \left(d(\mathbf{x}_i - \mathbf{y}_j)^2 - d(\mathbf{x}_1 - \mathbf{y}_j)^2 \right) \pi_{1,j} \\ &\leq \sum_j \frac{1}{2} \pi_{1,j} = \frac{1}{2n},\end{aligned}$$

where the second inequality comes from the fact that $d^2(\mathbf{x}, \mathbf{y}) \leq 1/2$ in $\mathbb{T}^d$. By symmetry we also obtain the reverse inequality. Doing the same with $\mathbf{y}_1'$ and $\mathbf{y}_1$ we obtain the bound $\frac{1}{2m}$. By using McDiarmid's inequality, see [195], we derive that

$$\mathbb{P}\left(\mathcal{T}_2(P_n, Q_m) - \mathbb{E}\mathcal{T}_2(P_n, Q_m) > t\right) \leq \exp\left(-\frac{nm}{n+m} 8t^2\right).$$

□

Proof of Proposition 3.3.9. Let $P = Q$. After Lemma 3.3.8, for every $\varepsilon > 0$ these exists $N_\varepsilon \in \mathbb{N}$ such that for all $n, m \geq N_\varepsilon$, $\mathbb{E}\mathcal{T}_2(P_n, Q_m) \leq \varepsilon$. Using explicitly the convergence speed, we can find the relationship between ε and N_ε:

$$\frac{\log N_\varepsilon}{N_\varepsilon} = \frac{\varepsilon}{C}, \tag{A.9}$$

where $C > 0$ is an unspecified constant. Then, directly from Theorem 3.3.7, we can bound (3.23) as

$$\begin{aligned}\mathbb{P}_{H_0}\left(\mathcal{T}_2(P_n, Q_m) > t\right) \leq \exp\left(-\frac{nm}{n+m} 8(t - \mathbb{E}\mathcal{T}_2(P_n, Q_m))^2\right) \leq \\ \exp\left(-\frac{nm}{n+m} 8(t-\varepsilon)^2\right),\end{aligned}$$

for all $n, m \geq N_\varepsilon$. □

Proof of Proposition 3.3.10. Let us first prove that (UB) is asymptotically of level α. Let $\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}$ such that for all $n, m \geq N_\varepsilon$, the test (UB) controls type I error. As we are taking the limit $n, m \to \infty$, we can choose n, m large enough such that they surpass N_ε. Then, consistency is ensured by Proposition 3.3.9.

To conclude, we prove the consistency under fixed alternatives such that $\mathcal{T}_2(P, Q) > \varepsilon$. First, note that

$$\begin{aligned}\mathbb{P}(\pi^{ub}_{nm,\varepsilon} = 1) &= \mathbb{P}\left(\mathcal{T}_2(P_n, Q_m) \geq \varepsilon + \sqrt{-\frac{n+m}{8nm}\log\alpha}\right) \\ &= \mathbb{P}\left(\sqrt{\frac{mn}{n+m}}(\mathcal{T}_2(P_n, Q_m) - \mathcal{T}_2(P, Q)) \geq \sqrt{\frac{mn}{n+m}}(\varepsilon - \mathcal{T}_2(P, Q)) + \sqrt{-\frac{1}{8}\log\alpha}\right).\end{aligned}$$

Now, (3.12) implies, under the alternative, the stochastically boundedness of the left hand side. However, the right hand side is clearly unbounded if $\mathcal{T}_2(P,Q) > \epsilon$. Consequently,

$$\lim_{n,m\to\infty} \mathbb{P}(\pi^{ub}_{nm,\varepsilon} = 1) = 1,$$

which concludes the proof. □

A.3 Supplementary figures

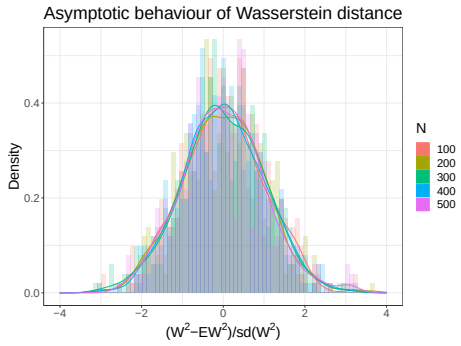


Figure A.4: Normalized asymptotic deviations from the mean of squared Wasserstein distance between two bivariate von Mises distributions of same means $(\mu,\nu) = (0,0)$ and different concentration parameters $(\kappa_1,\kappa_2,\kappa_3) = (0,0,0)$ and $(\kappa_1,\kappa_2,\kappa_3) = (2,2,0)$. The figures show the corresponding histograms and the associated kernel density estimates, for different sample sizes.

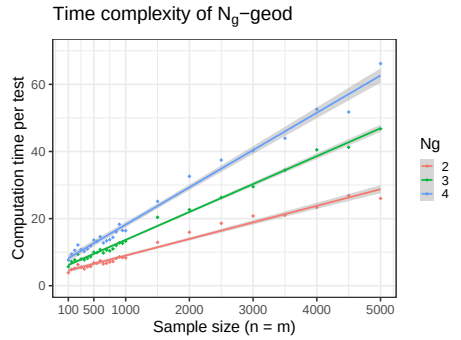


Figure A.5: Empirical time complexity of (N_g-geod) for $N_g = 2,3,4$. Each point corresponds to the average computation time per test among 200 repetitions of (N_g-geod) for two equally sized samples drawn from a bivariate von Mises distributions of equal means $(\mu,\nu) = (0,0)$ and different concentration parameters $(\kappa_1,\kappa_2,\kappa_3) = (0,0,0)$ and $(\kappa_1,\kappa_2,\kappa_3) = (1,1,0)$. The lines correspond to a linear regression performed for each value of N_g.

Appendix B

Appendix of Chapter 4

Contents

B.1 Proofs of Section 4.2

Proof of Proposition 4.2.1. Recalling that the p-value for the b-th permutation test is $p_b = \frac{n_b+1}{K+1}$, we have

$$p_{(c,c'),\mathcal{X}} = \frac{1+\sum_{b=1}^{B} n_b}{1+BK} = \frac{1+\sum_{b=1}^{B}(p_b(K+1)-1)}{1+BK} = \frac{(K+1)\sum_{b=1}^{B} p_b - (B-1)}{1+BK} \tag{B.1}$$

$$= \frac{B(K+1)\overline{p}_B - (B-1)}{1+BK}\,. \tag{B.2}$$

Therefore, we obtain

$$p_{(c,c'),\mathcal{X}} - \overline{p}_B = \frac{\overline{p}_B(B(K+1)-(1+BK))-(B-1)}{1+BK} = \frac{(B-1)(\overline{p}_B-1)}{1+BK}\,. \tag{B.3}$$

Since $0 \leq p_b \leq 1$ for all b, we have $0 \leq \overline{p}_B \leq 1$ as well, so that

$$0 \leq \overline{p}_B - p_{(c,c'),\mathcal{X}} \leq \frac{(B-1)}{1+BK} \leq \frac{1}{K}\,, \tag{B.4}$$

where the last inequality holds for any B and K since $(B-1)K \leq 1+BK$. □

Proof of Proposition 4.2.2. Let U be a random variable uniformly distributed in $[0,1]$. As a consequence of Theorem 1 in [235], a random variable X taking values in $[0,1]$ is super-uniform if and only if

$$\mathbb{E}(u(U)) \leq \mathbb{E}(u(X)) \quad \text{for all non-decreasing function u.} \tag{B.5}$$

Therefore, it suffices to find a non-decreasing function u such that $\mathbb{E}(u(U)) > \mathbb{E}(u(\overline{U_n}))$ for all $n \geq 2$. Let $u : [0,1] \longrightarrow [0,1]$ be such that $u(t) = t^2$ for all $t \in [0,1]$. Then, as $\mathbb{E}(\overline{U_n}) = \mathbb{E}(U)$ and $\mathbb{E}(X^2) = \mathrm{Var}(X) + \mathbb{E}(X)^2$ for any real-valued random variable X, it suffices to prove that

$$\mathrm{Var}\left(\overline{U_n}\right) < \mathrm{Var}(U) = \frac{1}{12} \quad \forall\, n \geq 2\,. \tag{B.6}$$

First, we have

$$\mathrm{Var}\left(\overline{U_n}\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} U_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{Var}(U_i) + 2\sum_{i<j}\mathrm{Cov}(U_i,U_j)\right] = \tag{B.7}$$

$$\frac{1}{12n} + \frac{2}{n^2}\sum_{i<j}\mathrm{Cov}(U_i,U_j) = \frac{1}{12n} + \frac{1}{n^2}\sum_{i<j}\left(\mathbb{E}(U_iU_j) - \mathbb{E}(U_i)\mathbb{E}(U_j)\right) = \tag{B.8}$$

$$\frac{1}{12n} + \frac{2}{n^2}\sum_{i<j}\left(\mathbb{E}(U_iU_j) - \frac{1}{4}\right) = \frac{1}{12n} - \frac{1}{2n^2}\binom{n}{2} + \frac{2}{n^2}\sum_{i<j}\mathbb{E}(U_iU_j)\,. \tag{B.9}$$

As the expectation of the product of two random variables defines an inner product on the set of random variables equally supported, we can apply Cauchy–Schwarz inequality and upper bound the last expectation in (B.9) as

$$\mathbb{E}(U_iU_j) \leq \sqrt{\mathbb{E}(U_i^2)\mathbb{E}(U_j^2)} = \frac{1}{3}\,. \tag{B.10}$$

However, the maximum $\frac{1}{3}$ is achieved if and only if both random variables are equal. Indeed, an equality in (B.10) holds if and only if the two variables are linearly dependent [11]. If, what's more, they are identically distributed, linear dependence is equivalent to equality. Consequently, at least one of the pairs $i < j$ must satisfy $\mathbb{E}(U_iU_j) < \frac{1}{3}$ or, on the contrary, we would have $U_1 = \cdots = U_n$, contradicting the hypothesis $n \geq 2$. Therefore, we can upper bound (B.9) as

$$\mathrm{Var}(\overline{U_n}) < \frac{1}{12n} - \frac{1}{2n^2} + \frac{2}{3n^2}\binom{n}{2} = \frac{1}{12n} + \frac{1}{6n^2}\binom{n}{2} = \frac{1}{12} \quad \forall\, n \geq 2\,, \tag{B.11}$$

which concludes the proof. □

B.2 Numerical study of p-value null distribution

In this section, we illustrate the behaviour of the non-uniform p-values $p_{(c,c'),\mathcal{X}}$ under the null hypothesis. As explained in Section 4.2, the B individual p-values p_b, $b = 1, \ldots, B$, are not independent as they are computed by bootstrapping from one initial sample. If, on the contrary, the p_b were computed from independent samples, the empirical mean $\overline{p}_B$ would converge in distribution to a Gaussian (Theorem 27.1 in [29]):

$$\sqrt{12B}\left(\overline{p}_B - \frac{1}{2}\right) \underset{B\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}(0,1)\,. \tag{indep}$$

We aim at analysing how the dependence induced by bootstrapping alters the asymptotic distribution of (indep), as well as the effect of the number B of bootstrap iterations and the number K of permutations for each individual test. We simulated the distribution of $p_{(c,c'),\mathcal{X}}$ under the null hypothesis following the algorithm detailed in [249] (see *Full procedure* in Methods Section, end of p. 7). The original samples were drawn from a uniform distribution and had fixed sizes $N = 2000$. Bootstrapped samples were extracted with size $N_{\max} = 200$. For the independence scenario, we replaced the bootstrapped samples by new equally sized samples drawn from a uniform distribution. As the explicit form of the test statistic is not provided in [249], we used the Wilcoxon statistic to illustrate the behaviour of $p_{(c,c'),\mathcal{X}}$. For each pair of values of K, B, the null distribution of $p_{(c,c'),\mathcal{X}}$ was simulated with 200 Monte Carlo iterations. Results are presented in Figure B.1, where the empirical distribution is compared to the asymptotic independence scenario (indep).

The first row in Figure B.1 shows the null distribution of $p_{(c,c'),\mathcal{X}}$ if samples are not bootstrapped but drawn independently at each iteration $b = 1, \ldots, B$. The encountered empirical distribution matches the Gaussian (indep) more faithfully as K increases, which was expected as the difference $\|p_{(c,c'),\mathcal{X}} - \overline{p}_B\|$ is upper bounded by $1/K$. In the same way, we should expect that the simulated $p_{(c,c'),\mathcal{X}}$ distributions are closer to the real (and unknown for this dependency scenario) null distribution of $\overline{p}_B$ when moving from the left to the right column in Figure B.1. When samples are bootstrapped as in [249], dependency between the p_b appears and as B increases (from the second to the last row in Figure B.1) values deviate from the independence scenario (indep). When B remains small (as for $B = 25$, the value chosen in [249]), the deviation from (indep) is slight. This can be explained as $N_{\max} \ll N$, and bootstrapping few times samples with small size compared to the one of the original sample is close to drawn samples independently from the entire population. As B increases, so does the dependency between the individual p-values. This dilates the empirical distribution of $p_{(c,c'),\mathcal{X}}$ and extends the difference to (indep). A similar phenomenon was observed in [85] when studying the effect of unobserved covariates on the null distribution of p-values.

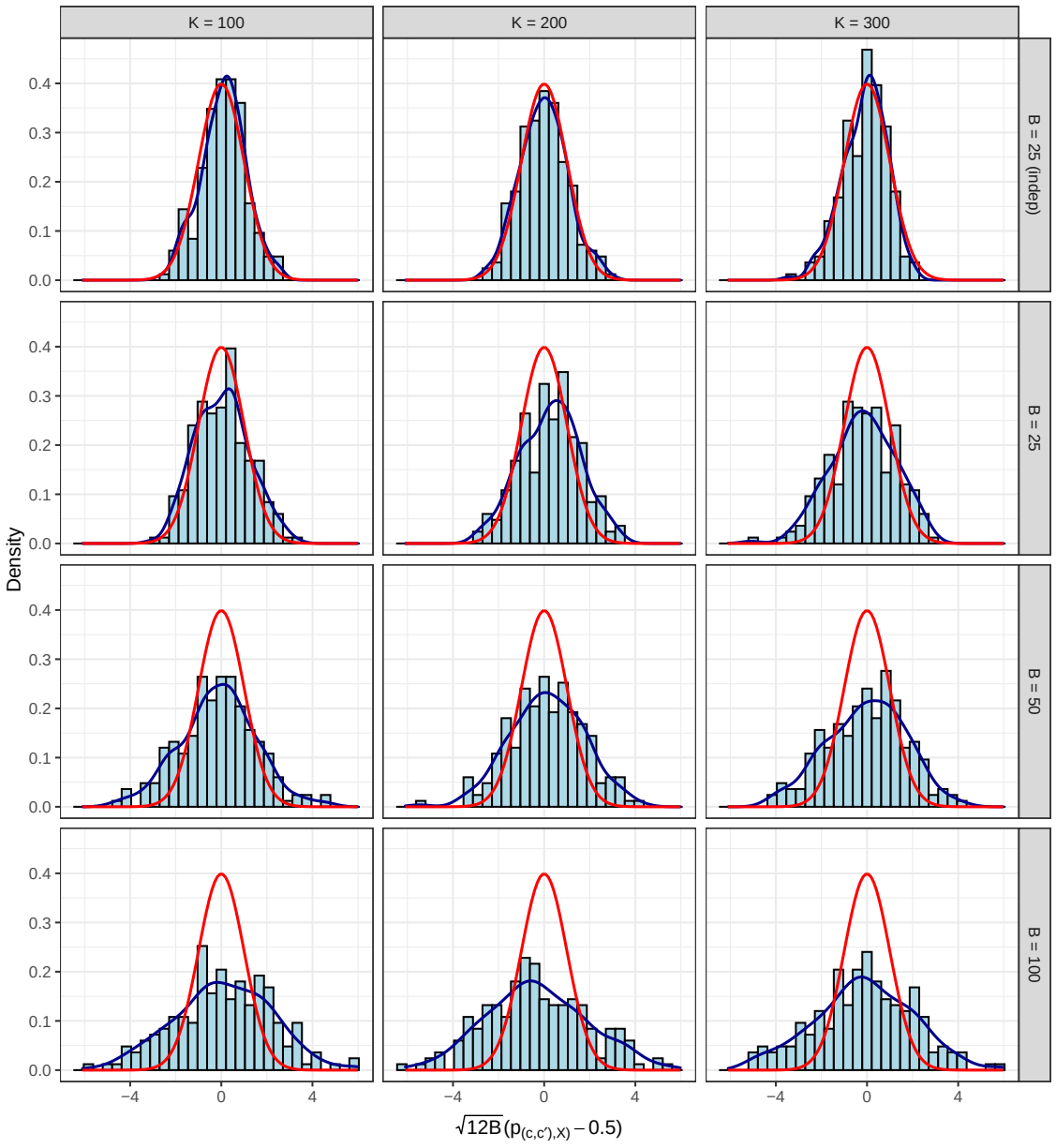


Figure B.1: Simulation of the null distribution of $p_{(c,c'),\mathcal{X}}$ for different values of parameters K (in columns) and B (in rows). The first row corresponds to the independence scenario (indep), whose asymptotic standard Gaussian density is depicted in red in all cases. The blue line corresponds to the non-parametric kernel density estimate of the encountered empirical distribution. Note that, for the sake of comparison to (indep), the presented p-values have been re-scaled as $\sqrt{12B}(p_{(c,c'),\mathcal{X}} - 0.5)$.

B.3 Dispersion of (ϕ, ψ) samples for each secondary structure type

Let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) real-valued random variables. The sample variance, defined as

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2\,, \tag{B.12}$$

is a measure of dispersion for $X_1, \ldots, X_n$ or, conversely, of its concentration around its empirical mean $\overline{X_n}$. To define an analogous estimator of (B.12) for n i.i.d. random variables $\{(\phi_i, \psi_i)\}_{i=1}^n$ taking values on the two-dimensional flat torus $\mathbb{T}^2$, we may consider

$$S^2_{(\phi,\psi)} = \frac{1}{n}\sum_{i=1}^{n} d_{\mathbb{T}^2}((\phi_i, \psi_i), (\phi_{n,\mathcal{F}}, \psi_{n,\mathcal{F}}))^2\,, \tag{B.13}$$

where $d^2_{\mathbb{T}}$ denotes the geodesic distance on the torus [107] and $(\phi_{n,\mathcal{F}}, \psi_{n,\mathcal{F}})$ denotes the sample barycenter (or Fréchet mean). However, the computation of Fréchet mean on the torus is not a trivial task. As our aim here is not theoretical, we will replace the barycenter $(\phi_{n,\mathcal{F}}, \psi_{n,\mathcal{F}})$ by the *extrinsic* barycenter on $\mathbb{T}^2$ [148], which is defined through a transformation to the Euclidean space $\mathbb{R}^4$ as

$$(\phi_{n,E}, \psi_{n,E}) = (\text{atan2}(s_\phi, c_\phi), \text{atan2}(s_\psi, c_\psi))\ ,$$

where $\text{atan2}(y, x)$ is the $\theta \in [-\pi, \pi)$ such that $\cos\theta = x$ and $\sin\theta = y$, and

$$(c_\phi, s_\phi, c_\psi, s_\psi) = \frac{1}{n}\sum_{i=1}^{n}(\sin\phi_i, \cos\phi_i, \sin\psi_i, \cos\psi_i)$$

is the Euclidean mean of the transformed sample. In conclusion, our dispersion estimator is defined as

$$D = \frac{1}{n}\sum_{i=1}^{n} d_{\mathbb{T}^2}((\phi_i, \psi_i), (\phi_{n,E}, \psi_{n,E}))^2\,, \tag{B.14}$$

which can be easily implemented. We computed (B.14) for every codon-specific Ramachandran plot with more than 30 points (the same criteria as to perform the statistical test). The results, classified by secondary structure as in Figure 4.2, are presented in Figure B.2. The empirical distributions of (B.14) clearly illustrate how $\alpha-$helical conformations (H) are highly restricted, the corresponding dihedrals being strongly concentrated around its barycenter. The dispersion of (ϕ, ψ) considerably increases for extended strand (E), and even more for the remaining DSSP structure classes merged together. These differences may be summarized by the average dispersion $\overline{D}$ for each secondary structure: $\bar{D}_{\text{Others}} = 0.06 > \bar{D}_{\text{E}} = 0.01 > \bar{D}_{\text{H}} = 0.002$, as stated in Section 4.3.1.

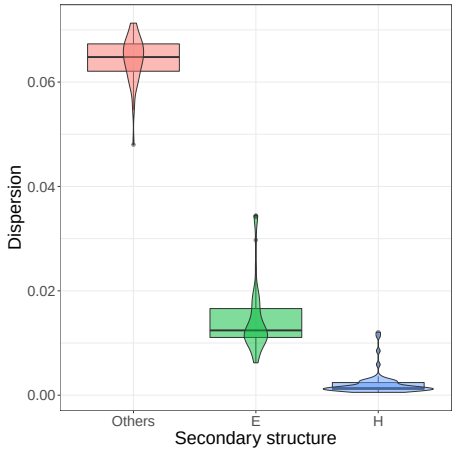


Figure B.2: Empirical distribution (boxplots, violin plots) of (B.14), for conformations in extended strand (E, green), α-helix (H, blue) and other (Others, red) secondary structures. Values higher than the 0.95 quantile for each group were excluded.

B.4 Supplementary figures

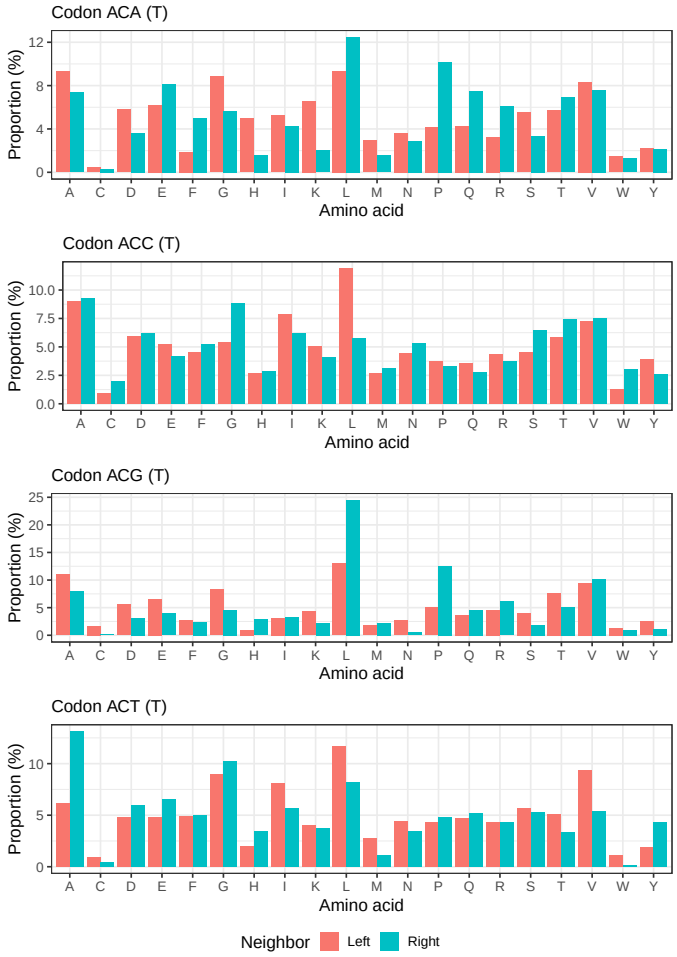


Figure B.3: Proportions (in percentage) of left and right neighboring amino acid types for the four synonymous codons of threonine in the database provided in [249].

Part II

Global structural analysis of highly flexible proteins

Chapter 5

WASCO: A Wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins

The structural investigation of intrinsically disordered proteins (IDPs) requires ensemble models describing the diversity of the conformational states of the molecule. Due to their probabilistic nature, there is a need for new paradigms that understand and treat IDPs from a purely statistical point of view, considering their conformational ensembles as well-defined probability distributions. In this chapter, we define a conformational ensemble as an ordered set of probability distributions and provide a suitable metric to detect differences between two given ensembles at the residue level, both locally and globally. The underlying geometry of the conformational space is properly integrated, one ensemble being characterized by a set of probability distributions supported on the three-dimensional Euclidean space (for global-scale comparisons) and on the two-dimensional flat torus (for local-scale comparisons). The inherent uncertainty of the data is also taken into account to provide finer estimations of the differences between ensembles. Additionally, an overall distance between ensembles is defined from the differences at the residue level. We illustrate the interest of the approach with several examples of applications for the comparison of conformational ensembles: (*i*) produced from molecular dynamics (MD) simulations using different force fields, and (*ii*) before and after refinement with experimental data. We also show the usefulness of the method to assess the convergence of MD simulations, and discuss other potential applications such as in machine-learning-based approaches. The numerical tool has been implemented in Python through easy-to-use Jupyter Notebooks available at https://gitlab.laas.fr/moma/WASCO.

This work has been published in *Journal of Molecular Biology, 168053, 2023*, with Amin Sagar, Christophe Zanon, Kresten Lindorff-Larsen, Pau Bernadó, Pierre Neuvial and Juan Cortés. It is presented here with minor changes for the sake of coherence in the manuscript.

Contents

5.1 Introduction

The comparison of protein structures is a crucial problem in structural biology. In the early works [241, 185], the use of root-mean-square deviation (RMSD) was introduced and discussed as a metric between conformations of folded proteins, and later extended to its ensemble version [43]. More recently, Lindorff-Larsen and Ferkinghoff-Borg [176] defined three metrics that allow overall comparison between ensembles of ordered/structured systems, with stronger mathematical guarantees, but using RMSD as a distance between individual conformations, which complicates its extension to disordered structures. Cazals *et al.* [47] used a graph-based representation of the conformational space based on a set of low-energy conformations (i.e. local minima of the potential energy landscape) and compared them with the more suitable Wasserstein distance. To do so, they used the least-RMSD as ground metric between conformations. The methods presented in [176] and [47] are well suited to examine conformational ensembles of molecules that present a well-characterized energy landscape. However, their application to molecules with energy landscapes where low-energy conformations are difficult to identify, as it is the case of IDPs, is inappropriate.

A few recent works have dealt with the comparison of conformational ensembles of IDPs. Huihui and Ghosh [131] focused on averaged conformational properties over ensembles as informative descriptors of their function. They proposed a sequence-decoration

metric that classifies IDPs using only their primary structure together with their charge configuration. The same idea of comparing average descriptors was applied by Lazar *et al.* [167], who proposed an ensemble comparison tool based on differences between average pairwise distances. Due to the huge conformational variability of IDPs, it is, however, important to take into account both the average properties as well as the distribution around those averages. Describing IDP conformations as being drawn from probability distributions determining their structure may yield to an important loss of information (or even misleading results) if the whole distribution is reduced to its mean. Even when comparing two (possibly multivariate) Gaussian distributions, the difference between the two depends both on the means and variances [163, 318]; thus, methods for comparing ensembles should ideally include also higher order moments of the probability distributions. This is why a statistical approach that integrates the entire probability law defining an ensemble is crucial to correctly capture the existing differences between disordered ensembles.

The probability distributions describing the ensembles need to be compared using a suitable metric, well-adapted to the geometric features of the underlying spaces. The Wasserstein distance [299], sometimes called "earth mover's distance", integrates the geometry of the space where the distributions are supported and provides strong mathematical guarantees. Moreover, it has a physical interpretation, as it is defined as the minimum transportation cost needed to reconfigure the mass of one probability distribution to recover the other. All this makes Wasserstein distance substantially preferable to other metrics currently used in the literature (e.g. Kullback-Leibler divergence, Helliger distance), as discussed in Section 5.2.

In this chapter, we define a set of probability distributions that characterize at local and global level the highly variable conformations in an ensemble of disordered proteins, and to which we can have access in practice. These probability laws can then be compared using the Wasserstein distance, allowing the identification of residue-specific and overall discrepancies. We also propose an approach to integrate the intrinsic uncertainty of the data within the metric, which enables a more clear identification of the relevant differences between the ensembles. The method has been implemented inside a purely non-parametric framework, avoiding model assumptions, dimensionality reduction or further simplifications that may yield significant loss of information.

In the following sections, we provide an overall description of the proposed methodology, which is further detailed in Appendix C, together with several cases of applications that illustrate how our method identifies residue-specific and overall discrepancies between conformational ensembles of IDPs or flexible peptides generated for example by molecular dynamics simulations or stochastic sampling techniques. Finally, we discuss current limitations and possible extensions of WASCO, as well as the great potential interest of this type of metric for its integration in machine-learning-based (ML-based) methods applied to generate or to refine conformational ensembles of IDPs.

5.2 Methods

Due to the intrinsic probabilistic nature of IDPs, descriptors of their conformational ensembles should be conceived from a purely statistical point of view. To do so, we seek to locally and globally describe conformational ensembles using well-defined probability distributions and to develop statistical tools allowing their comparison. The main questions to answer are therefore: (1) which is the best way to define those probability distributions? and (2) how these distributions have to be compared to provide quantitative information about similarities and differences between ensembles?

5.2.1 Defining conformational ensembles as a set of probability distributions

IDP ensembles can be described at both local and global scales, providing complementary information. We aim at defining an ordered set of probability distributions that account for the highly variable structure of the ensemble and, above all, that can be estimated in practice from a set of sampled conformations.

The most important aspects of the local structure can be described by the dihedral angles (φ, ψ) for each amino acid residue along the sequence. Therefore, for each residue, the ensemble is locally characterized by a two-dimensional random variable (φ, ψ) or, in other words, by a probability distribution supported on the two-dimensional flat torus $\mathbb{T}^2$ [189, 35]. If we denote such distribution as P_i^l, for the residue at the i-th position, we define the local structural descriptor of an ensemble as the L-tuple

$$(P_1^l, \ldots, P_L^l), \quad P_i^l \in \mathcal{P}(\mathbb{T}^2) \quad \text{for all } i = 1, \ldots, L, \tag{5.1}$$

where L is the sequence length and $\mathcal{P}(\mathbb{T}^2)$ denotes the space of probability distributions supported on $\mathbb{T}^2$.

Describing the global structure is a less trivial task. The use the absolute positions of the atoms and an absolute reference frame for the entire ensemble is not an appropriate description as it is sensitive to rigid-body motions. Therefore, our approach uses the relative positions of all pairs of residues along the sequence, which are invariant under rigid-body motion. More precisely, we define the position of a given residue as the position of its C_β atom when it exists and of its C_α atom otherwise. If $i, j \in \{1, \ldots, L\}$, $i \neq j$, denote two different sequence positions, let $\overrightarrow{R_{i,j}}$ be the three-dimensional random variable determining the relative position of j-th residue with respect to the i-th one. If we denote $P_{i,j}^g$ the probability distribution associated to $\overrightarrow{R_{i,j}}$, we define the global structural descriptor of an ensemble as the $(L(L-1)/2)$-tuple

$$(P_{1,2}^g, P_{1,3}^g, \ldots, P_{L-1,L}^g), \quad P_{i,j}^g \in \mathcal{P}(\mathbb{R}^3) \quad \text{for all } i = 1, \ldots, L-1,\ j = i+1, \ldots, L, \tag{5.2}$$

where L is the sequence length and $\mathcal{P}(\mathbb{R}^3)$ denotes the space of probability distributions supported on the three-dimensional Euclidean space.

5.2.2 Accessing empirical probability distributions from sampled conformations

Estimating the local structural descriptor (5.1) is immediate as we have direct access to dihedral angles (φ, ψ) from the sample of conformations. Therefore, the local structural descriptor will be estimated by its *empirical* counterpart

$$(P^l_{1;n}, \ldots, P^l_{L;n}), \tag{5.3}$$

where each $P^l_{i,n}$, $i = 1, \ldots, L$, is the empirical probability distribution of P^l_i, and n is the number of conformations constituting the sample. Such empirical probability distributions are commonly represented through Ramachandran maps [237].

Obtaining a sample of $\overrightarrow{R_{i,j}}$ from the set of conformations is less direct. To compute a set of comparable $\overrightarrow{R_{i,j}}$ vectors from all conformations, their coordinates must be expressed on the same reference system. To do so, we first define a reference frame at the i-th residue, using only the positions of the i-th C', C_α and N^H atoms. This frame, whose construction is detailed in the Supplementary Information (SI), is a meaningful representation of the spatial pose of each residue.

The reference frame associated to each residue $i \in \{1, \ldots, L\}$ allows to express the relative positions of all residues $j \neq i$ with respect to i. Moreover, the definition of a reference system allows the *superposition* of all the conformations in the ensemble. This is illustrated in Figure 5.1, for three conformations. Consequently, for every $j \neq i$, we will have access to n realizations of the random variable $\overrightarrow{R_{i,j}}$ or, in other words, to a point cloud in the three-dimensional Euclidean space, representing a sample drawn from the distribution of $P^g_{i,j}$. Therefore, the global structural descriptor of the ensemble (5.2) will be estimated by its *empirical* counterpart

$$(P^g_{1,2;n}, P^g_{1,3;n}, \ldots, P^g_{L-1,L;n}), \tag{5.4}$$

where $P^g_{i,j,n}$ is the empirical probability distribution of $P^g_{i,j}$, for all $i = 1, \ldots, L-1$, $j = i+1, \ldots, L$. An example of a pair of samples of $\overrightarrow{R_{i,j}}$ is presented in Figure C.8.

5.2.3 Distances between local and global structural descriptors

After defining the local and global structural descriptors of an ensemble as an ordered set of probability distributions, the choice of a suitable metric allowing inter-ensemble comparisons becomes the subsequent question to deal with. The basic properties that such a metric should have are:

1. Satisfying the mathematical properties that define a distance (i.e. being 0 if and only if the two compared distributions are identical, symmetry and triangle inequality),

2. Integrating the geometry of the underlying space.

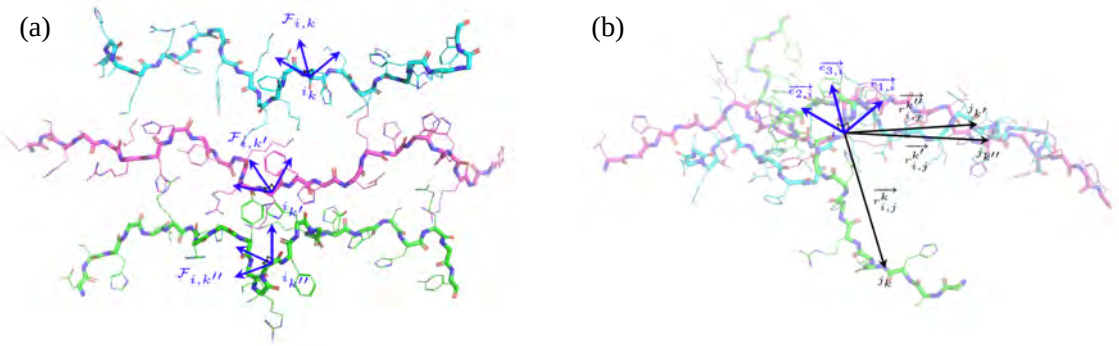


Figure 5.1: Illustration of how samples of global structural descriptors are obtained, for a pair of positions i, j along the sequence. In (a), the reference frame is built for every conformation at residue i. In (b), all the frames are superimposed using this reference frame. Then, for any $j \neq i$, the vectors $\overrightarrow{r_{i,j}}$ constitute a sample of $\overrightarrow{R_{i,j}}$.

The use of metrics between probability distributions is not new in structural biology. For instance, Ting *et al.* [280] used Hellinger distance to detect differences between (φ, ψ) distributions. However, this metric does not take into account the geometry of the underlying space (in particular here, its periodicity). A symmetrized Kullback-Leibler (KL) or the Jensen-Shannon (JS) divergence was used in [176, 279] to compare ensembles of ordered systems. This metric has a firm interpretation, based on information theory (in particular the JS divergence is the square of a metric). However, it still misses the geometrical reliability and does not satisfy triangle inequality, which makes comparisons between multiple ensembles difficult to interpret.

Besides satisfying conditions 1 and 2, the Wasserstein distance, derived from the theory of Optimal Transport (OT), provides both strong theoretical guarantees [299] and attractive empirical performance [228]. Informally, it represents the minimum transportation cost needed to reconfigure the mass of one probability distribution to recover the other. We refer to [228] for an in-depth introduction to OT. Most of the applications of OT are related to the very active field of machine learning (ML), notably in the framework of generative networks [9], robustness [262] or fairness [76], among others. With some notable exemptions [47, 19, 248, 67, 107], Wasserstein distance has not been widely used in structural biology. More related to our work, in [47], Cazals *et al.* used Wasserstein distance to compare energy landscapes sampled from conformational ensembles. We might also refer to our work [107] presented in Chapter 3 introducing statistical tests to assess differences between (ϕ, ψ) distributions. The incorporation of the underlying geometry to its definition makes it a well-adapted metric to measure distances between local and global structural descriptors of the ensembles. Details and important considerations regarding its practical computation in this context are given in Appendix C.

5.2.4 The comparison tool

Consider two ensembles A, B, associated to two protein sequences of equal length L, and let n_A, n_B be their number of conformations, respectively. We define the differences between local structural descriptors of A and B as the L-tuple of Wasserstein distances

$$(\mathcal{W}_1^{l,A,B}, \dots, \mathcal{W}_L^{l,A,B}) = \left(\mathcal{W}(P_{1;n_A}^{l,A}, P_{1;n_B}^{l,B}), \dots, \mathcal{W}(P_{L;n_A}^{l,A}, P_{L;n_B}^{l,B})\right), \tag{5.5}$$

where $P_{i;n_A}^{l,A}$ (resp. $P_{i;n_B}^{l,B}$) denotes the i-th distribution of the empirical local structural descriptor (5.3) of ensemble A (resp. B). Statistical tests to assess whether any $\mathcal{W}_i^{l,A,B}$ is significantly different from zero have been defined in Chapter 3. The second of the introduced techniques (UB) is better adapted to our problem, as it only detects the more important discrepancies and accepts slight differences that may arise from experimental or computational procedures. This is discussed in detail in Section 3.5. Consequently, together with the L-tuple (5.5) of distances comparing local structural descriptors, we are able to supply a L-tuple of (UB) p-values (corrected for multiplicity [125]) accounting for the statistical significance of the corresponding distances:

$$(p_1^{A,B}, \dots, p_L^{A,B}). \tag{5.6}$$

Recall that a small p-value $p_i^{A,B}$ indicates strong evidence that the *true* distance that $\mathcal{W}_i^{l,A,B}$ estimates is different from zero. In other words, small p-values show significant differences between the corresponding local structural descriptors. Therefore, the vector (5.6) enables the identification of those residues where the differences are more important, and those residues for which differences can be assigned as non-significant.

Analogously, the difference between global structural descriptors of A and B is defined as the $(L(L-1)/2)$-tuple

$$(\mathcal{W}_{1,2}^{g,A,B}, \dots, \mathcal{W}_{L-1,L}^{g,A,B}) = \left(\mathcal{W}(P_{1,2;n_A}^{g,A}, P_{1,2;n_B}^{g,B}), \dots, \mathcal{W}(P_{L-1,L;n_A}^{g,A}, P_{L-1,L;n_B}^{g,B})\right), \tag{5.7}$$

where $P_{i,j;n_A}^{g,A}$ (resp. $P_{i,j;n_B}^{g,B}$) denotes the i,j distribution of the empirical global structural descriptor (5.4) of ensemble A (resp. B). In this case, we are not able to provide a vector of p-values assessing the significance of the global differences. This is due to the intrinsic limitations of the Optimal Transport theory when the ground space has dimension $d \geq 3$. Note that (5.7) can be more naturally represented as a triangular $(L-1)\times(L-1)$ matrix $W^{g,A,B}$, whose elements are given by $(W^{g,A,B})_{ij} = \mathcal{W}_{i,j}^{g,A,B}$. Graphically, the matrix $W^{g,A,B}$ is represented using a color scale to fill the coefficients according to distance values. As the diagonal will remain empty, it will be filled with the local distances (5.5). This will also allow to assess whether changes on local structural descriptors are related with changes in global structural descriptors and to compare both scales within the same representation.

Accounting for uncertainty

The variability in experimental and simulated structures causes uncertainties and statistical noise that may substantially bias the distance estimation. For example, when running a MD simulation, independent replicas of the same simulation setup may results in non-negligible differences that distort the analysis of the comparison matrices. The same may occur when comparing two uniformly chosen subsets of conformations from an ensemble generated by stochastic sampling techniques [219, 88]. In order to soften the effect of uncertainty and to obtain *net* estimates of the differences between a pair of ensembles, we will use (if available) independent replicas of the same ensemble. These replicas may also be produced by uniform subsampling of the set of conformations. However, special care must be taken when subsampling MD trajectories as the convergence of the simulation must be ensured for the subsamples to be representative of the entire ensemble.

Let $A_1, \ldots, A_{n_I}$ (resp. $B_1, \ldots, B_{n_I}$) be n_I independent replicas of ensemble A (resp. B). The corrected difference between local structural descriptors of A and B is defined as the L-tuple

$$(\widetilde{\mathcal{W}}_1^{l,A,B}, \ldots, \widetilde{\mathcal{W}}_L^{l,A,B}), \tag{5.8}$$

where each corrected distance is defined as

$$\widetilde{\mathcal{W}}_i^{l,A,B} = \left(\frac{1}{n_I} \sum_{s=1}^{n_I} \mathcal{W}_i^{l,A_s,B_s} - \frac{1}{2(n_I-1)} \sum_{s=2}^{n_I} \left(\mathcal{W}_i^{l,A_1,A_s} + \mathcal{W}_i^{l,B_1,B_s} \right) \right)_+, \quad \text{for all } i = 1, \ldots, L, \tag{5.9}$$

where, for any real number x, $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise. The first term in (5.9) is an average of n_I Wasserstein distances between n_I paired independent replicas of A and B. As it was shown in [269], an average of Wasserstein distances between sub-samples of the same population is a pertinent estimate of the Wasserstein distance between the two entire populations that, in addition, conserves the properties that mathematically define a distance. Therefore, this first term estimates the Wasserstein distance between the entire populations of A and B (conceived as the union of all independent replicas), softening the variability. To this *brutto* inter-ensemble difference, we subtract an average of the Wasserstein distances between independent replicas of the same population (intra-ensemble). Note that, for the sake of computational simplicity, we just compared the first replica of each ensemble with the subsequent ones. This alignment is arbitrary and can be set otherwise. Of course, distances between all pairs of replicas can be added to this term. The more combinations are added to (5.9), the finer will be the estimate of the (unknown) true Wasserstein distance between the ensembles but, as replicas are independent, different alignments for a given number of combinations should not yield substantial discrepancies on the quality of this estimate. The same applies if n_I is different for A and B; both terms in (5.9) can be accordingly adapted. As it is illustrated in Section 5.3, the use of corrected distances (5.9) contribute to reduce the noise coming from structural uncertainty and help to emphasize residue-specific differences in the matrix representation. For the distances between global structural descriptors, the correction is performed analogously.

Setting an interpretable scale

When defining an absolute distance or score between conformational ensembles, providing the clues to ease its interpretation is crucial. The problem of interpreting unbounded metrics with no intrinsic reference values has been widely discussed since the introduction of RMSD for the comparison of pairs of conformations [241, 185]. Here, we do not seek to define any cutoff to binarize the resulting matrices, but to provide a more informative continuous scale. To do so, we aim at quantifying the magnitude of the inter-ensemble distances compared to the intra-ensemble ones, using the uncertainty estimate as a reference. If we denote as $\mathcal{W}_{\text{inter}}^{l,A,B}$ (resp. $\mathcal{W}_{\text{intra}}^{l,A,B}$) the first (resp. second) term in (5.9), the score

$$\frac{\widetilde{\mathcal{W}}_i^{l,A,B}}{\mathcal{W}_{\text{intra}}^{l,A,B}} = \frac{\left(\mathcal{W}_{\text{inter}}^{l,A,B} - \mathcal{W}_{\text{intra}}^{l,A,B}\right)_+}{\mathcal{W}_{\text{intra}}^{l,A,B}}, \tag{5.10}$$

corresponds to the relative difference between the inter-ensemble and intra-ensemble differences. Once again, this score is analogously defined for differences between global structural descriptors.

An overall distance between ensembles

In some situations, it may be of interest to perform overall comparisons between multiple ensembles. To do so, moving from a residue-specific analysis to a comparison at the whole structure level might be preferable. The definition of a score for the overall ensemble has been addressed for ordered systems [176]. Here, we propose to define such a score by aggregating all the residue-specific distances computed using the above-described methods. We recall that if $d_1, \ldots, d_L$ are L distances defined on L metric spaces $\mathcal{X}_1, \ldots, \mathcal{X}_L$, the function $\sqrt{d_1^2 + \cdots + d_L^2}$ is a distance on the product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$. Consequently,

$$\mathcal{OW}^{l,A,B} = \left(\sum_{i=1}^{L} \left(\mathcal{W}_i^{l,A,B}\right)^2\right)^{1/2} \tag{5.11}$$

is a distance on the product space of all dihedral angles along the sequence and, therefore, serves to quantify the *overall local discrepancy* between a pair of ensembles. Analogously,

$$\mathcal{OW}^{g,A,B} = \left(\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \left(w_{ij}\mathcal{W}_{i,j}^{g,A,B}\right)^2\right)^{1/2}, \quad \text{with } w_{ij} > 0 \text{ for all } i, j \in \{1, \ldots, L\}, \tag{5.12}$$

is a distance on the product space of all pairwise relative positions of the residues in both ensembles, and serves to quantify the *overall global discrepancy*. Note that we have assigned a positive weight w_{ij} to each global distance in (5.12). This allows to consider distances between specific residue pairs as more relevant than the others when computing the overall discrepancy [124]. For instance, we can highlight differences between global

structural descriptors that appear for residue pairs that are far from each other in the sequence, i.e. large $|i-j|$, and neglect distances between neighboring residue pairs, i.e small $|i-j|$. This can be done by choosing w_{ij} as an appropriate increasing function of $|i-j|$, as

$$w_{ij} = w(i,j) = \frac{1}{\tanh 1} \tanh\left(\left(\frac{|i-j|}{L-1}\right)^{\frac{1}{2}}\right), \tag{5.13}$$

which satisfies $w_{i,i} = 0$ for all i and $w_{1,L} = w_{L,1} = 1$.

The drawback of this definition of the overall distance is that it does not take into account the previously mentionned uncertainty. To solve this problem, the same strategy to define a global score can be performed by replacing each $\mathcal{W}_i^{l,A,B}$ (resp. $\mathcal{W}_{i,j}^{g,A,B}$) by its corresponding corrected distance $\widetilde{\mathcal{W}}_i^{l,A,B}$ (resp. $\widetilde{\mathcal{W}}_{i,j}^{g,A,B}$) in (5.11) (resp. (5.12)). However, this strategy makes the triangle inequality for the overall metric no longer satisfied. Both scores can be implemented by the practitioner and used depending on the specific comparison context.

5.2.5 The Jupyter notebook

The WASCO comparison tool has been implemented through an easy-to-use Jupyter Notebook. It is available at https://gitlab.laas.fr/moma/WASCO, together with its installation guidelines and detailed implementation instructions. The notebook takes a pair of ensembles as input and returns the comparison results through the matrix defined in Section 5.2.4, containing global and local differences. Users can choose to correct the computed distances by uncertainty (5.9). When independent replicas are not provided as input, subsampling is used to emulate them. If this correction is performed, results are displayed in the interpretable scale (5.10). The overall scores (5.11), (5.12), aggregating the corrected distances, are also returned by the tool.

Ensembles can be provided as input in several of the most common data formats. WASCO accepts one .xtc file per replica, together with a .pdb file including the topology information of the molecule, one multiframe .pdb file per replica or a folder per replica containing one .pdb file per conformation. The user can also choose to compare ensembles for sequence segments (of equal length) instead of the entire sequence. Details are provided in the notebook documentation.

Due to the large number of Wasserstein distances to be computed ($L(L-1)/2+L$ per pair of replicas), the computation time might be considerably high. The number of conformations constituting the ensemble also has a significant impact, due to computational limitations of the existing OT algorithms when sample sizes and dimension increase. In order to return results within a reasonable amount of time, WASCO computes Wasserstein distances in parallel. The required CPU time depends on the number of conformations, replicas and sequence length of the ensembles. For small proteins of $L \sim 30$ and ensembles of reasonable size $n_A, n_B \sim 10^4$, the CPU time using 20 threads is less than 15 minutes using a standard computing server. However, for larger proteins of $L \sim 150$ and large

ensembles with $n_A, n_B \sim 10^5$, the CPU time using 20 threads goes up to some hours. Additionally, comparing large ensembles of substantially longer sequences ($L \gg 150$) might cause memory problems, as all pairwise relative positions for every conformation need to be stocked. Therefore, the suitability of the sizes of the ensembles must be considered before launching WASCO. Adapting WASCO to longer sequences with large conformational ensembles remains an objective for future work.

The output of WASCO is given through a matrix, whose entries are the values of the score (5.10) computed for local and global distances, when independent replicas are provided. Otherwise, the matrix depicts the values of the non-corrected inter-ensemble distances (5.5), (5.7). The values for the discrepancies between the global structural descriptors (5.4) are provided in the lower triangle. The differences between the local structural descriptors (5.3) are displayed along the diagonal. Details on the interpretation of the matrix are given in Section C.1.3 and illustrated in Figure C.4. These guidelines are also presented in the software documentation.

5.3 Results

In this section, we present several applications to illustrate the different possibilities enabled by WASCO. In all cases, the distances between local and global structural descriptors were corrected for uncertainty using (5.9), as independent replicas were available. The results are depicted through the score (5.10), representing the relative difference between the inter-ensemble distances and the uncertainty. Both overall local and global discrepancies between pairs of ensembles were computed plugging the corrected distances in (5.11) and (5.12). The weight function (5.13) was used to highlight differences between residue pairs far from each other in the sequence and reduce differences between neighboring amino acids. Note that this weighting is considered only to compute the overall distance (5.12), and not to depict distance values in the matrix representation, which correspond to the interpretable scale (5.10). An additional analysis illustrating the application of WASCO to assess the convergence of MD simulations is included in Section C.2.2.

5.3.1 Comparison of ensembles produced by MD simulations using different force-fields

We applied WASCO to compare the results of MD simulations using different force-fields presented in [140] for two flexible peptides showing a significant propensity to form poly-l-proline type II (PPII) structures. Four different force-fields, having demonstrated relatively good performances to simulate IDPs were applied: AMBER ff99SB-disp, AMBER ff99SB-ILDN, CHARMM36IDPFF, and CHARMM36m (details and references to these force-fields can be found in [140]). For simplicity, we will refer to these force-fields as disp, ildn, c36idp and c36m, respectively. As independent replicas for each simulation were available, we could perform the correction for uncertainty (5.9).

Figure 5.2 presents the output of WASCO for several pairwise comparisons of conformational ensembles of Histatin-5 (Hst5) obtained with the different force-fields. The matrices and the overall dissimilarities suggest that the generated structures are closer (in Wasserstein distance) when they are simulated using c36idp and c36m (which we can define as group-I), or disp and ildn (group-II). This is not surprising as group-I are versions of CHARMM and group-II are versions of AMBER. Indeed, matrices (a) and (b), comparing force fields inside group-I and inside group-II respectively, present overall global differences which are small compared to those of panels (c) and (d), which compare force-fields of different groups. The same conclusion can be reached by comparing the magnitude of the scales of both pairs of matrices. The two remaining comparisons (ildn vs. ildn and c36m vs. disp) are not included in Figure 5.2 as the corresponding matrices are qualitatively equivalent to (c) and (d). Similar observations have been made when comparing ensembles of folded proteins generated using related force-fields [192, 279].

Matrices returned by WASCO also allow a residue-specific analysis of the distances. In Figure 5.2, panels (c) and (d) show that the most relevant global differences appear in regions close to the diagonal (i.e. between residue pairs close in the sequence), where the inter-ensemble corrected distances rise up to 6-7 times the intra-ensemble ones. This is not the case when comparing force-fields inside the same group, as the largest differences appear in more internal matrix regions (i.e. between residue pairs more distant in the sequence). However, these corrected differences represent less than the half of the intra-ensemble distances. The information displayed on the diagonal allows the detection of the residues where the local conformation change more abruptly between force-fields. These local changes are restricted to smaller regions, contrary to the observed behaviour of global differences, which appeared for more extent regions inside the lower triangle and not for isolated pairs of amino acids. In some cases, substantial local distances appear in residues where global structure also changes (see, for example, residues next to the N-terminus in (a,c)). However, this correspondence is not observed in all matrices. We repeated the same analysis for MD simulations of PEP3 with the same force-fields. Results are presented in Section C.2.1.

5.3.2 Structural impact of SAXS ensemble refinement

Using Hst5 as an example, we applied WASCO to evaluate the structural impact of SAXS refinement with the Ensemble Optimization Method (EOM)[22] on the resulting ensemble. We first compared the Hst5 ensemble simulated with Flexible-Meccano [219, 21] with the refined one using previously reported SAXS data[253]. The results are presented in Figure 6.4. Note that a previous EOM analysis of these data suggested that Hst5 in solution is slightly more extended that the random coil model generated with Flexible-Meccano [253]. Small but non-negligible differences were observed at the central part of the peptide (from residues 6 to 13). Most probably, the SAXS-based refinement selected

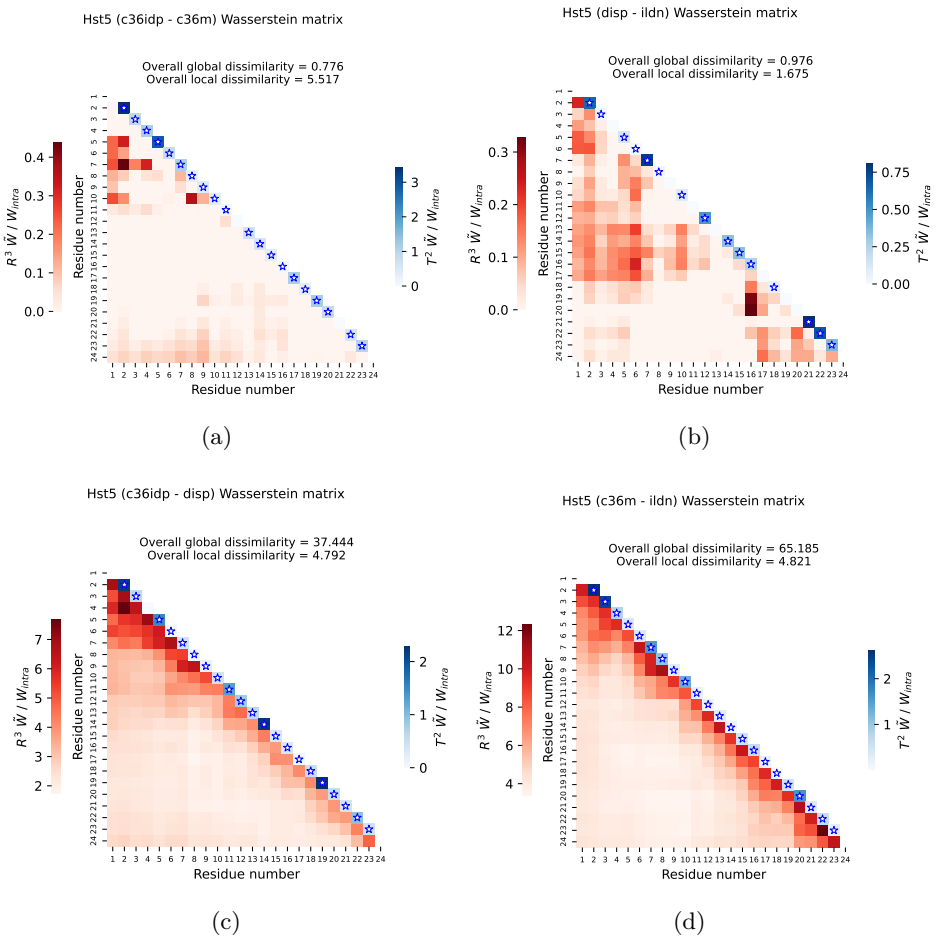


Figure 5.2: Comparison of Molecular Dynamics simulations of Hst5 ensemble using different force-fields. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (5.10), representing the relative difference between the inter-ensemble distances and the uncertainty.
The coefficients in the lower-triangle (in red) correspond to the global differences. The coefficients along the diagonal (in blue) correspond to the local differences. Blue stars indicate that the corresponding local corrected distance is significantly different from zero (the associated p-value (5.6) is smaller than $\alpha = 0.05$). Note the different scales used in the different plots.

conformations with an extended central region to account for the overall expansion of peptide in the solution [22]. Moreover, we observed highly significant local distances that propagate locally towards the interior of the matrix. In other words, these residues with large local distances conformationally influence their closest neighbours. Intriguingly, this propagation seems to only occur locally towards the C-terminus.

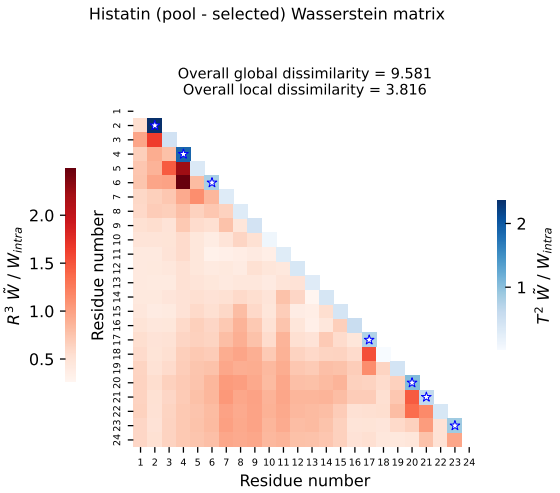


Figure 5.3: Comparison of Hst5 ensemble before and after filtering with experimental SAXS data. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (5.10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower triangle (in red) correspond to the global differences. The coefficients along the diagonal (in blue) correspond to the local differences. Blue stars indicate that the corresponding local corrected distance is significantly different from zero (the associated p-value (5.6) is smaller than $\alpha = 0.05$).

We next assessed whether the direction in which conformations are built have a structural effect and change the refined ensemble. To do so, we generated two Hst5 ensembles using a stochastic sampling method similar to Flexible-Meccano but using a different strategy [88], where the chains were built either from N-to-C or from C-to-N. When using these two ensembles to fit the experimental curve, the resulting distance matrices displayed very similar features for local and global distances (Figure C.9(a,b)), suggesting that the chain-building direction does not have a relevant effect. In both cases, a systematic increase in the distances is observed for the central residues, as observed in the previous analysis (Figure 6.4).

In a recent study, ENCORE was used to show that refined ensembles were closer to each other than different input ensembles [279]. This can also be illustrated using WASCO, by comparing the Hst5 ensembles generated in both directions before and after the filtering with SAXS data (Figure C.9(c,d)). These comparisons clearly showed that both global and local differences were smaller for the refined ensembles than for the input ones, as observed when comparing the maximum values of the corresponding color scales. As we were comparing very similar ensembles, we expected the distances to be small. Nevertheless, we observe one significant local difference on the diagonal in Figure C.9c that disappeared after refinement.

5.4 Discussion

We have presented a novel method to compare conformational ensemble models of highly flexible proteins. WASCO is based on a non-parametric framework: local and global structural descriptors of the conformational space are defined as distributions and do not rely on probabilistic or statistic models. This allows capturing the entire variability of the ensemble without information loss. The distributions are compared using the Wasserstein distance, which has strong mathematical guarantees and respects the geometry of the underlying space. To this metric, we incorporated the structural uncertainty presented in experimental and simulated ensembles. Using this strategy, WASCO highlights the relevant differences between ensembles. We have illustrated several possible applications of WASCO as an additional tool for the investigation of IDPs and flexible peptides. It provides complementary information with respect to other tools to analyze and compare conformational ensembles based on global descriptors, such as the radius of gyration [286] or secondary structure propensities [140]. Besides, the presented approach is advantageous with respect to simpler comparison techniques based on average descriptors, such as the difference of median distance matrices introduced in [167]. This is illustrated with an example in Section C.2.3. Thanks to its accuracy to identify differences between ensembles, WASCO has great potential interest for integration into ML-based methods for generating or refining conformational ensembles of IDPs [178, 138, 316]. More precisely, metrics based on WASCO can be used to evaluate the performance of these methods, or as a loss function when training neural network models.

WASCO has been implemented in an open-source Jupyter Notebook, which enables an easy use of the methods as well as their adaptation or extension to particular needs. The main drawback of the current implementation is its limitation to deal with considerably large ensembles of long IDPs. Adapting WASCO to larger chains remains for future work. Other interesting directions for future work will be the extension of WASCO to compare ensembles of multi-domain proteins, and to operate with coarse-grained models. The extension of WASCO to compare ensembles for chains of different length is also an interesting but challenging work. Note however that the Jupyter Notebook enables the user to select sequence fragments of equal length for the comparison.

Software availability

WASCO has been implemented as an easy-to-use Jupyter Notebook, available at https://gitlab.laas.fr/moma/WASCO.

Acknowledgements

We are grateful to Francesco Pesce, Sthitadhi Maiti and Matthias Heyden for providing useful data. We thank Gabriella Gerlach, Frederik Emil Thomasen and Philipp Schake

for their helpful discussions and valuable feedback on WASCO implementation.

This work has been partially supported by the French National Research Agency (ANR) through grant ANR-19-P3IA-0004, the LabEx CIMI (ANR-11-LABX-0040) and EpiGenMed (ANR-10-LABX-12–01) within the French State Programme "Investissements d'Avenir", by the European Research Council under the European Union's H2020 Framework Programme (2014–2020)/ ERC Grant agreement n° [648030] awarded to PB and by the Lundbeck Foundation BRAINSTRUC initiative (R155-2015-2666). The CBS is a member of France-BioImaging (FBI) and the French Infrastructure for Integrated Structural Biology (FRISBI), 2 national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively).

Chapter 6

Post-clustering inference under dependence

The recent work by Gao et al. [104] laid the foundation for post-clustering inference. For the first time, the authors established a theoretical framework allowing to test for differences between means of estimated clusters. Additionally, they studied the estimation of unknown parameters while controlling the selective type I error. However, their theory was developed for independent observations identically distributed as p-dimensional Gaussian variables with spherical covariance matrix. Here, we aim at extending this framework to a more convenient scenario for practical applications, where arbitrary dependence structures between observations and features are allowed. We show that a p-value for post-clustering inference under general dependency can be defined, and assess the theoretical conditions allowing the compatible estimation of one covariance matrix. The theory is developed for hierarchical agglomerative clustering algorithms with several types of linkages, and for k-means algorithm. We illustrate our method with synthetic data and real data of protein structures.

Contents

6.1 Introduction

Post-selection inference has gained substantial attention in recent years due to its potential to address practical problems coming from various scientific disciplines. The problem of using data to answer a question that has been chosen based on the same data was formalized in [97], where the basis of selective hypothesis testing was rigorously set with the definition of the selective type I error. This paved the way to perform selective testing when null hypotheses are chosen through clustering algorithms, bypassing the naive data splitting that reveals unsuitable in this context. However, their proposed approach, referred to as *data carving*, as well as more recent approaches like *data fission* [173] are difficult to implement in practice because they require knowledge of the covariance structure between variables. Moreover, they often involve the non-trivial calibration of a tuning parameter that controls the proportion of information allocated for model selection and for inference. The seminal work [104] established for the first time a theoretical framework allowing selective testing after clustering, when observations are independent and identically distributed as p-dimensional Gaussian random variables with spherical covariance matrix. This corresponds to the following matrix normal model [127]:

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \sigma^2\mathbf{I}_p), \tag{6.1}$$

where $\boldsymbol{\mu} \in \mathcal{M}_{n\times p}(\mathbb{R})$ and $\sigma > 0$. Under (6.1), the authors in [104] defined a p-value that controls the selective type I error when testing for a difference in means between a pair of estimated clusters. This p-value can be efficiently computed for hierarchical clustering algorithms with common linkage functions. Moreover, the authors in [104] made another remarkable contribution by addressing the estimation of σ while controlling the selective type I error, which had been overlooked in previous works [173, 243] and that is essential in real problems. They showed that if σ is asymptotically over-estimated the p-value is asymptotically super-uniform, and provided an estimator $\hat{\sigma}$ that can be used in practice.

Despite the notable contribution of [104], the model (6.1) is somewhat limited in more complex applications. In real problems, features describing observations are unlikely to be independent and have identical variance, but rather present more general covariance structures $\boldsymbol{\Sigma}$. In the same way, observations might present non-negligible dependence structures when, for instance, they are drawn from time series models or simulated with physical models involving time evolution. The practical motivation of the present work is to perform inference after clustering protein conformations. Protein structures are non-static and their conformational variability is essential to understand the relationship between sequence, structure and function [157]. Due to the high complexity of the conformational space, clustering techniques have emerged as powerful tools to characterize the structural

variability of proteins, by extracting families of representative states [59, 8, 263, 224]. Usually, Euclidean distances between pairs of amino acids are considered as p-dimensional descriptors of protein conformations [59, 167, 45]. These distances are highly correlated and hardly match the model (6.1). Moreover, protein data is often simulated with Molecular Dynamics approaches that recreate the time-evolution of the protein according to physical models [3]. In that case, independence between observations is not admissible.

Accordingly, our aim is to go one step further and extend the framework introduced in [104] to a more general setting where arbitrary dependence structures between observations and features are admitted. We present a straightforward adaptation of [104] where the model (6.1) is extended to

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}), \tag{6.2}$$

where $\boldsymbol{\mu} \in \mathcal{M}_{n\times p}(\mathbb{R})$, $\mathbf{U} \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\boldsymbol{\Sigma} \in \mathcal{M}_{p\times p}(\mathbb{R})$. Our techniques follow the same reasoning steps as the ones in [104] and show that a p-value for testing differences between estimated cluster means can be defined under (6.2). The Chapter is organized as follows:

- In Section 6.2, we present the definition of a p-value for post-clustering inference under the general model (6.2), and show that its efficient computation is straightforward if it is achievable under (6.1).
- In Section 6.3, we explore the scenarios that allow the asymptotic over-estimation of either $\mathbf{U}$ or $\boldsymbol{\Sigma}$ while respecting the asymptotic control of the selective type I error. We provide an estimator that can be used in several common practical scenarios.
- In Section 6.4 we revisit the framework presented in Section 6.2 when, for technical reasons, additional information is imposed to the conditioning event that defines the p-value. This allows in particular to perform selective inference after k-means clustering, following the work in [51].
- In Section 6.5, we illustrate all the results through numerical experiments on synthetic data. Finally, in Section 6.6, we show how this theory can be applied to perform inference after clustering protein structures.

6.2 Selective inference for clustering under general dependency

In [104], the authors consider the problem of selective inference after hierarchical clustering in the case of independent observations and features. Here, we aim to extend the method to admit general dependence structures. We consider n observations of p features drawn from the matrix normal distribution (6.2), where $\mathbf{U}$ and $\boldsymbol{\Sigma}$ are required to be positive definite. Each row of $\mathbf{X}$ is a vector of features in $\mathbb{R}^p$. The dependence between such features is given by $\boldsymbol{\Sigma}$, and $\mathbf{U}$ encodes the dependency between observations. If observations are

independent with unit variance, we have $\mathbf{U} = \mathbf{I}_n$, and if features are independent with equal variance we can write $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ for a given $\sigma > 0$. These two assumptions define the model in [104]. Here, we show that this model can be generalized to arbitrary $\mathbf{U}$, $\mathbf{\Sigma}$, defining a p-value that controls the selective type I error rate for clustering.

Let us first recall the setting introduced in [104]. We will denote by X_i (resp. μ_i) the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) and, for a group of observations $\mathcal{G} \subseteq \{1, \dots, n\}$, $X_{\mathcal{G}}$ will denote the submatrix of $\mathbf{X}$ with rows X_i for $i \in \mathcal{G}$. We also consider the mean of $\mathcal{G}$ in $\mathbf{X}$, denoted by

$$\mu_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i, \tag{6.3}$$

and its empirical counterpart

$$\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i. \tag{6.4}$$

From now on, we use the notation $\mathbf{M} = (M_{ij})_{ij}$ to denote real matrices. Let $\mathcal{C}$ be a clustering algorithm, $\mathbf{x}$ a realization of the random variable $\mathbf{X}$ and $\hat{C}_1, \hat{C}_2$ an arbitrary pair of clusters in $\mathcal{C}(\mathbf{x})$. The goal of post-clustering inference is to assess the null hypothesis

$$H_0^{\{\hat{C}_1, \hat{C}_2\}} : \mu_{\hat{C}_1} = \mu_{\hat{C}_2} \tag{6.5}$$

by controlling the *selective type I error for clustering* at level α, i.e. by ensuring that

$$\mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left(\text{reject } H_0^{\{\hat{C}_1, \hat{C}_2\}} \text{ based on } \mathbf{X} \text{ at level } \alpha \;\middle|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \right) \leq \alpha \quad \forall\, \alpha \in [0, 1]. \tag{6.6}$$

The ideal scenario to define a p-value for (6.5) satisfying (6.6) would be to only condition on the event $\{\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})\}$, which is the broader conditioning set that allows selective type I error control. However, making the p-value analytically tractable often needs the refinement of the conditioning set by adding more technical events (see Section 6.4). In [104], the authors consider a test statistic of the form $\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_2$ and introduce the quantity

$$\begin{aligned} p(\mathbf{x}; \left\{\hat{C}_1, \hat{C}_2\right\}) = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \Big(& \|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_2 \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2 \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}), \\ & \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{x}\,,\, \mathrm{dir}\left(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\right) = \mathrm{dir}\left(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\right) \Big), \end{aligned} \tag{6.7}$$

where $\pi^{\perp}_{\nu} = \mathbf{I}_n - \nu\nu^T / \|\nu\|_2^2$, $\mathrm{dir}(u) = u / \|u\|_2 \mathbb{1}\{u \neq 0\}$ and the components of $\nu(\hat{C}_1, \hat{C}_2)$ are defined as

$$\nu(\hat{C}_1, \hat{C}_2)_i = \mathbb{1}\left\{i \in \hat{C}_1\right\} / |\hat{C}_1| - \mathbb{1}\left\{i \in \hat{C}_2\right\} / |\hat{C}_2|. \tag{6.8}$$

Theorem 1 in [104] proves that (6.7) is a p-value for (6.5). Moreover, if $\mathcal{C}$ is a hierarchical clustering algorithm the p-value (6.7) can be explicitly characterized and efficiently computed for several types of linkages. Otherwise, it can be approximated with a Monte Carlo procedure.

Here, we aim at extending (6.7) for the general model (6.2). The main idea is to replace the norm $\|\cdot\|_2$ in the test statistic by the more general norm

$$\|x\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} = \sqrt{x^T \mathbf{V}^{-1}_{\hat{C}_1,\hat{C}_2} x}, \quad \forall\, x \in \mathbb{R}^p, \tag{6.9}$$

where $\mathbf{V}_{\hat{C}_1,\hat{C}_2} \in \mathcal{M}_{p\times p}(\mathbb{R})$ integrates the information about the scale matrices in (6.2). Let us first introduce some notation. For a pair of non-overlapping groups of observations $\mathcal{G}_1, \mathcal{G}_2 \subseteq \{1,\dots,n\}$, we define the $p(|\mathcal{G}_1| + |\mathcal{G}_2|)$ column vector

$$X_{\mathcal{G}_1,\mathcal{G}_2} = (\mathrm{vec}(X^T_{\mathcal{G}_1}), \mathrm{vec}(X^T_{\mathcal{G}_2})), \tag{6.10}$$

which concatenates the column vectors of observations in $\mathcal{G}_1$ with the ones in $\mathcal{G}_2$. Similarly, we denote as $\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2}$ the principal submatrix of $\mathbf{U}$ containing the rows and columns in $\mathcal{G}_1 \cup \mathcal{G}_2$. Finally, we consider $\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} \in \mathcal{M}_{p\times p(|\mathcal{G}_1|+|\mathcal{G}_2|)}$ the linear operator verifying $\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} X_{\mathcal{G}_1,\mathcal{G}_2} = \bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}$, that we can write explicitly as the block matrix

$$\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} = \left(\underbrace{\tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p \ \cdots \ \tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p}_{|\mathcal{G}_1|} \ \underbrace{-\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p \ \cdots \ -\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p}_{|\mathcal{G}_2|} \right). \tag{6.11}$$

We can now define the matrix $\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}$ in (6.9) as

$$\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2} = \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}(\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \mathbf{\Sigma})\mathbf{D}^T_{\mathcal{G}_1,\mathcal{G}_2}, \tag{6.12}$$

where $\otimes$ denotes the Kronecker product of matrices. Note that (6.9) is a well-defined norm if and only if $\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}$ is a positive definite matrix, which here is guaranteed as $\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}$ has full rank and $\mathbf{U}$ and $\mathbf{\Sigma}$ are positive definite [127]. The following result extends Theorem 1 in [104] by proving that the quantity

$$\begin{aligned} p_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \left\{\hat{C}_1, \hat{C}_2\right\}) = \mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}} \Big(\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}), \\ \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x}\,,\, \mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\left(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\right) = \mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\left(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\right) \Big), \end{aligned} \tag{6.13}$$

where $\mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(u) = u/\|u\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} \mathbb{1}\{u \neq 0\}$, is a computationally tractable p-value for (6.5) that controls the selective type I error rate for arbitrary dependence structures $\mathbf{U}, \mathbf{\Sigma}$.

Theorem 6.2.1. *Let* $\mathbf{x}$ *be a realization of* $\mathbf{X}$ *and* $\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{P}(\{1,\dots,n\})$ *with* $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$. *Then,* $p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})$ *is a p-value for the test* $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$*:* $\mu_{\mathcal{G}_1} = \mu_{\mathcal{G}_2}$ *that controls the selective type I error for clustering* (6.6) *at level* α. *Furthermore, it satisfies*

$$p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}, \mathcal{S}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) \right), \tag{6.14}$$

where $\mathbb{F}_p(t, \mathcal{S})$ *is the cumulative distribution function of a* χ_p *random variable truncated to the set* $\mathcal{S}$ *and*

$$\begin{aligned} \mathcal{S}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \Big\{ \phi \geq 0 \,:\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x} + \\ \left(\frac{\phi}{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}} \right) \nu(\mathcal{G}_1, \mathcal{G}_2) \mathrm{dir}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}) \Big) \Big\}. \end{aligned} \tag{6.15}$$

Theorem 6.2.1 is proved in Appendix D.1. One can easily verify that replacing $\mathbf{U} = \mathbf{I}_n$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ in Theorem 6.2.1 yields exactly Theorem 1 in [104]. The only difference is that, here, the information about the variance has been extracted from the statistic null distribution, which now remains the same independently of $\mathbf{U}, \mathbf{\Sigma}$, and moved it *into* the test statistic itself by making it dependent on the scale matrices. Note that this formulation replaces the Euclidean distance considered in [104] by the *Mahalanobis distance* [184]. Recall that, if $x, y \in \mathbb{R}^p$ and P is a probability distribution supported on $\mathbb{R}^p$ with covariance matrix C, the Mahalanobis distance between x and y with respect to P is given by $\|x-y\|_C$, where $\|\cdot\|_C$ is defined as (6.9). Consequently, the formulation in Theorem 6.2.1 corresponds to consider as statistic the Mahalanobis distance between the empirical means $\bar{X}_{\hat{C}_1}$ and $\bar{X}_{\hat{C}_2}$ with respect to the null distribution of their difference $\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}$, which is a centered multivariate normal of covariance matrix $\mathbf{V}_{\hat{C}_1,\hat{C}_2}$ (see proof of Theorem 6.2.1). That distance generalizes to multiple dimensions the idea of quantifying how many standard deviations away a point is from the mean of its distribution, and therefore integrates the dependence structure between columns and rows in $\mathbf{X}$.

Analogously, computing the p-value (6.13) depends only on the characterization of the one-dimensional set

$$\hat{\mathcal{S}}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} = \mathcal{S}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \left\{\hat{C}_1, \hat{C}_2\right\}) = \left\{\phi \geq 0 \,:\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}\left(\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)\right)\right\}, \tag{6.16}$$

where $\mathcal{S}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}, \cdot)$ is defined in (6.15) and where

$$\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi) = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x} + \left(\frac{\phi}{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}\right)\nu(\hat{C}_1, \hat{C}_2)\mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}). \tag{6.17}$$

The set (6.17) is the analogous set of $\mathbf{x}'(\phi)$ in [104, Equation (13)] for the norm (6.9), and its interpretation is equivalent. Indeed, we can rewrite both $\mathbf{x}'(\phi)$ and (6.17) as

$$\mathbf{x}'(\phi) = \mathbf{x} + \frac{\nu(\hat{C}_1, \hat{C}_2)}{\|\nu(\hat{C}_1, \hat{C}_2)\|_2^2}\left(\phi - \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2\right)\mathrm{dir}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}), \tag{6.18}$$

$$\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi) = \mathbf{x} + \frac{\nu(\hat{C}_1, \hat{C}_2)}{\|\nu(\hat{C}_1, \hat{C}_2)\|_2^2}\left(\phi - \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\right)\mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}). \tag{6.19}$$

Consequently, we can interpret (6.17) as a perturbed version $\mathbf{x}'(\phi)$ of $\mathbf{x}$, but where the perturbation is based on the norm (6.9) instead of $\|\cdot\|_2$, i.e. the points are pulled apart or pushed together using $\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}$ as a reference instead of $\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2$. Thus, the set (6.16) is analogously defined as the set of non-negative ϕ for which applying the clustering algorithm $\mathcal{C}$ to the perturbed data set $\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)$ yields $\hat{C}_1$ and $\hat{C}_2$. As shown in [104], the set

$$\hat{\mathcal{S}} = \left\{\phi \geq 0 \,:\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\}, \tag{6.20}$$

can be explicitly characterized for hierarchical clustering. Fortunately, we don't need to re-adapt the work in [104] to the set (6.16), as its points are given by a scale transformation of the points in $\hat{\mathcal{S}}$.

Lemma 6.2.2. *Let* $\mathbf{x}$ *be a realization of* $\mathbf{X}$ *and* $\hat{C}_1, \hat{C}_2$ *an arbitrary pair of clusters in* $\mathcal{C}(\mathbf{x})$. *Let* $\hat{\mathcal{S}}$ *denote the set* (6.20) *defined in [104, Equation (12)]. Then,*

$$\hat{\mathcal{S}}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} = \frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2} \hat{\mathcal{S}}, \tag{6.21}$$

where $\hat{\mathcal{S}}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}$ *is defined in* (6.16).

Consequently, the work in [104, Section 3] can be applied here to characterize the set (6.16) and, therefore, to compute the p-value defined in (6.13). An explicit characterization of (6.16) is possible when $\mathcal{C}$ is a hierarchical clustering algorithm with squared Euclidean distance, along with either single linkage or a linkage satisfying a linear Lance-Williams update [104, Equation 20], e.g. average, weighted, Ward, centroid or median linkage functions. Otherwise, the p-value (6.13) can be approximated with a Monte Carlo procedure, adapting the importance sampling approach presented in [104, Section 4.1]. Following the same notation, we sample

$$\omega_1, \ldots, \omega_N \overset{i.i.d.}{\sim} \mathcal{N}\left(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}, 1\right)$$

and approximate (6.13) as

$$p_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \left\{\hat{C}_1, \hat{C}_2\right\}) \approx \frac{\sum_{i=1}^{N} \pi_i \mathbb{1}\left\{\omega_i \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{x}'(\omega_i))\right\}}{\sum_{i=1}^{N} \pi_i \mathbb{1}\left\{\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{x}'(\omega_i))\right\}}, \tag{6.22}$$

for $\pi_i = f_1(\omega_i)/f_2(\omega_i)$, where f_1 is the density of a χ_p random variable, and f_2 is the density of a $\mathcal{N}(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}, 1)$ random variable.

6.3 Unknown dependence structures

The selective inference framework introduced for model (6.2) in Section 6.2 assumes that both scale matrices $\mathbf{U}$ and $\mathbf{\Sigma}$ are known, which is a quite unrealistic scenario. Under the independence assumption made in [104], where $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ and $\mathbf{U} = \mathbf{I}_n$, the authors showed in Theorem 4 that over-estimating σ yields asymptotic control of the selective type I error, and provided such an estimator $\hat{\sigma}$ that can be used in practice. Under the general model (6.2), the scale matrices $\mathbf{U}$ and $\mathbf{\Sigma}$ are non-identifiable:

$$\mathbf{X} \sim \mathcal{MN}_{np}(\boldsymbol{\mu}, \mathbf{U}, \mathbf{\Sigma}) \Leftrightarrow \text{vec}(\mathbf{X}) \sim \mathcal{N}_{np}(\text{vec}(\boldsymbol{\mu}), \mathbf{U} \otimes \mathbf{\Sigma}). \tag{6.23}$$

This makes their simultaneous estimation an arduous task in practice. Non-unique Maximum Likelihood Estimates (MLE) exist for $\mathbf{U}$ and $\mathbf{\Sigma}$ [84], which depend on each other and can be computed through an iterative algorithm. However, even in the unlikely scenario where we had access to enough realizations of $\mathbf{X}$, the interdependence of the computed

MLEs would still prevent us from assessing the control of selective type I error after estimation. In this Section, we investigate the situation where only one of the scale matrices is known, and assess theoretical conditions that allow asymptotic control of the selective type I error when estimating the other one. We also provide an estimator that satisfies these conditions for some common dependence models.

Let us recall that, for the model (6.2), we have

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}) \Leftrightarrow \mathbf{X}^T \sim \mathcal{MN}_{p\times n}(\boldsymbol{\mu}^T, \boldsymbol{\Sigma}, \mathbf{U}). \tag{6.24}$$

Therefore, the methods presented in this Section can be equally applied to estimate $\mathbf{U}$ or $\boldsymbol{\Sigma}$ when the other is known, by transposing $\mathbf{X}$ if needed. From now on, we assume that the dependence structure between observations $\mathbf{U}$ is known, and study under which conditions we can suitably estimate $\boldsymbol{\Sigma}$. In other words, if $\hat{\boldsymbol{\Sigma}}(\mathbf{x})$ is an estimate of $\boldsymbol{\Sigma}$ for a given realization $\mathbf{x}$ of $\mathbf{X}$, we study under which conditions the p-value

$$p_{\hat{\mathbf{V}}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\hat{\mathbf{V}}_{\mathcal{G}_1,\mathcal{G}_2}}; \mathcal{S}_{\hat{\mathbf{V}}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})\right), \tag{6.25}$$

where $\hat{\mathbf{V}}_{\mathcal{G}_1,\mathcal{G}_2} = \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}(\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \hat{\boldsymbol{\Sigma}}(\mathbf{x}))\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}^T$, asymptotically controls the selective type I error. Theorem 6.3.1 generalizes Theorem 4 in [104] for the estimation of $\boldsymbol{\Sigma}$ under model (6.2) by relying on the Loewner partial order, defined below. The proof is included in Appendix D.2.

Definition 6.3.1 (Definition 7.7.1 in [127])**.** *Let A, B be two square matrices of equal size. The binary relation $A \succeq B$ if and only if A, B are Hermitian and $A - B$ is positive semidefinite is called the Loewner partial order between square matrices.*

Theorem 6.3.1. *For $n \in \mathbb{N}$, let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \boldsymbol{\Sigma})$. Let $\mathbf{x}^{(n)}$ be a realization of $\mathbf{X}^{(n)}$ and $\hat{C}_1^{(n)}, \hat{C}_2^{(n)}$ a pair of clusters estimated from $\mathbf{x}^{(n)}$. Let $\succeq$ denote the Loewner partial order between square matrices [127], i.e. $A \succeq B$ if and only if A, B are Hermitian and $A - B$ is positive semidefinite. If $\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ is a positive definite estimator of $\boldsymbol{\Sigma}$ such that*

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \boldsymbol{\Sigma} \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) = 1, \tag{6.26}$$

then, for any $\alpha \in [0, 1]$, we have

$$\limsup_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(p_{\hat{\mathbf{V}}_{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}}}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right) \leq \alpha \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) \leq \alpha, \tag{6.27}$$

where $p_{\hat{\mathbf{V}}_{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}}}$ is defined in (6.25)*.*

Note that the Loewner partial order is a natural extension to Hermitian matrices of the usual order in $\mathbb{R}$. If we replace $\boldsymbol{\Sigma}$ by $\sigma^2\mathbf{I}_p$ in Theorem 6.3.1, the condition $\hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma}$ becomes $\hat{\sigma} \geq \sigma$, as in [104, Theorem 4]. We aim now at providing an estimator of $\boldsymbol{\Sigma}$

satisfying the conditions in Theorem 6.3.1. The asymptotic properties of such an estimator strongly depend on the asymptotic dependence structure between observations, given by the sequence of matrices $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ of Theorem 6.3.1. Let us first consider

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}\left(\mathbf{X}\right) = \frac{1}{n-1}\left(\mathbf{X}-\bar{\mathbf{X}}\right)^{T}\mathbf{U}^{-1}\left(\mathbf{X}-\bar{\mathbf{X}}\right), \tag{6.28}$$

where $\bar{\mathbf{X}}$ is a $n \times p$ matrix having as rows the mean across rows of $\mathbf{X}$, i.e.

$$\bar{\mathbf{X}} = \mathbf{1}_n \otimes \frac{1}{n}\sum_{k=1}^{n} X_k, \tag{6.29}$$

where $\mathbf{1}_n$ is a column n-vector of ones. We can also write (6.28) element-wise:

$$\hat{\Sigma}_{ij} = \frac{1}{n-1}\sum_{l,s=1}^{n}\left(X_{li}-\bar{X}_i\right)\left(U^{-1}\right)_{ls}\left(X_{sj}-\bar{X}_j\right), \quad \forall\, i,j \in \{1,\ldots,p\}, \tag{6.30}$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ki}$. Note that the estimator $\hat{\mathbf{\Sigma}}$ is a positive definite matrix if the matrix $\mathbf{X}-\bar{\mathbf{X}}$ has full rank. In that case, (6.28) satisfies the conditions of Theorem 6.3.1 if we make some assumptions about how the matrices $\boldsymbol{\mu}^{(n)}$ and $\mathbf{U}^{(n)}$ in Theorem 6.3.1 grow up as n increases. We first adopt the assumptions about $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$ made in [104] to prove the equivalent of Theorem 6.3.1 for the independence scenario.

Assumption 2 (Assumptions 1 and 2 in [104])**.** *For all $n \in \mathbb{N}$, there are exactly K^* distinct mean vectors among the first n observations, i.e.*

$$\left\{\mu_i^{(n)}\right\}_{i=1,\ldots,n} = \{\theta_1,\ldots,\theta_{K^*}\}. \tag{6.31}$$

Besides, the proportion of the first n observations that have mean vector θ_k converges to $\pi_k > 0$, i.e.

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\mu_i^{(n)} = \theta_k\right\} = \pi_k, \tag{6.32}$$

for all $k \in \{1,\ldots,K^*\}$, where $\sum_{k=1}^{K^*}\pi_k = 1$.

If observations are independent and we set $\mathbf{U}^{(n)} = \mathbf{I}_n$, Assumption 2 is the only requirement for (6.28) to asymptotically over-estimate $\mathbf{\Sigma}$ in the sense of Theorem 6.3.1. However, for general $\mathbf{U}^{(n)}$, the quantities

$$\frac{1}{n}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mathbb{1}\{\mu_l^{(n)} = \theta_k\}\,\mathbb{1}\{\mu_s^{(n)} = \theta_{k'}\} \tag{6.33}$$

are also required to converge as n tends to infinity. Furthermore, we need to know its limit explicitly to assess whether $\hat{\mathbf{\Sigma}} \succeq \mathbf{\Sigma}$ asymptotically. This requires relatively strong conditions on the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ that can be difficult to verify for a given model of dependence, as well as an additional mild condition on the sequence $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$, needed for non-diagonal $\mathbf{U}^{(n)}$. Let's begin by stating the latter.

Assumption 3. *If $\mathbf{U}^{(n)}$ is non-diagonal for all $n \in \mathbb{N}$, for any $k, k' \in \{1, \dots, K^*\}$, the proportion of the first n observations at distance $r \geq 1$ in $\mathbf{X}^{(n)}$ having means θ_k and $\theta_{k'}$ converges, and its limit converges to $\pi_k \pi_{k'}$ when the lag r tends to infinity. More precisely,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-r} \mathbb{1}\{\mu_i = \theta_k\}\, \mathbb{1}\{\mu_{i+r} = \theta_{k'}\} = \pi^r_{kk'} \underset{r\to\infty}{\longrightarrow} \pi_k\, \pi_{k'}. \tag{6.34}$$

Note that we are asking the proportion of pairs of observations having a given a pair of means to approach the product of individual proportions (6.32) when both observations are far away in $\mathbf{X}^{(n)}$. Stronger conditions need to be imposed to the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$. Together with Assumptions 2 and 3, the following Assumption is a sufficient condition for (6.33) to converge with tractable limit.

Assumption 4. *Let $\left\{\mathbf{U}^{(n)}\right\}_{n\in\mathbb{N}}$ be a sequence of real positive definite matrices, and let $(U^{(n)})^{-1}_{ij}$ denote the i, j entry of $(\mathbf{U}^{(n)})^{-1}$ for any $n \in \mathbb{N}$. Then, every superdiagonal of $(\mathbf{U}^{(n)})^{-1}$ defines asymptotically a convergent sequence, whose limits sum up to a real value. More precisely, for any $i \in \mathbb{N}$ and any $r \geq 0$,*

$$\lim_{n\to\infty} \left(U^{(n)}\right)^{-1}_{i\,i+r} = \Lambda_{i\,i+r}, \quad \text{where} \quad \lim_{i\to\infty} \Lambda_{i\,i+r} = \lambda_r \quad \text{and} \quad \sum_{r=0}^{\infty} \lambda_r = \lambda \in \mathbb{R}. \tag{6.35}$$

Moreover, for each $r \geq 0$, the sequence $\left\{(U^{(n)})^{-1}_{i\,i+r}\right\}_{n\in\mathbb{N}}$ satisfies any of the following conditions:

- *(i) It is dominated by a summable sequence i.e. $\left|\left(U^{(n)}\right)^{-1}_{i\,i+r} - \Lambda_{i\,i+r}\right| \leq \alpha_i \;\forall\, n \in \mathbb{N}$, with $\{\alpha_i\}_{i=1}^{\infty} \in \ell_1$,*
- *(ii) For each $i \in \mathbb{N}$, it is non-decreasing or non-increasing.*

If Assumptions 2, 3 and 4 hold for a given pair of sequences $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$, $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$, the following result ensures that (6.28) asymptotically over-estimates (in the sense of the Loewner partial order) the dependence structure $\boldsymbol{\Sigma}$ between features.

Proposition 6.3.2. *Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^{(n)}$ and $\mathbf{U}^{(n)}$ satisfy Assumptions 2, 3 and 4 for some $K^* > 1$. Let $\hat{\boldsymbol{\Sigma}}$ be the estimator defined in* (6.28)*. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \boldsymbol{\Sigma}\right) = 1. \tag{6.36}$$

The proof of Proposition 6.3.2 makes use of the following Lemma, that makes explicit the need of Assumptions 2, 3 and 4. Both results are proved in Appendix D.2.

Lemma 6.3.3. *Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^{(n)}$ and $\mathbf{U}^{(n)}$ satisfy Assumptions 2, 3 and 4 for some $K^* > 1$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mathbb{1}\{\mu^{(n)}_l = \theta_k\}\, \mathbb{1}\{\mu^{(n)}_s = \theta_{k'}\} = 2(\lambda - \lambda_0)\pi_k\pi_{k'} + \lambda_0 \pi_k \delta_{kk'}, \tag{6.37}$$

for any $k, k' \in \{1, \dots, K'\}$, and where $\pi_k, \pi_{k'}$ and λ_0, λ are defined in Assumptions 2 and 4 respectively.

Finally, it suffices to estimate $\mathbf{\Sigma}$ using an independent and identically distributed copy of $\mathbf{X}^{(n)}$ to have (6.26) provided (6.36). Our final result comes as an immediate consequence of the previous statement and Proposition 6.3.2.

Proposition 6.3.4. *Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \mathbf{\Sigma})$, where $\boldsymbol{\mu}^{(n)}$ and $\mathbf{U}^{(n)}$ satisfy Assumptions 2, 3 and 4 for some $K^* > 1$. Let $\mathbf{x}^{(n)}$ be a realization of $\mathbf{X}^{(n)}$ and $\hat{C}_1^{(n)}$, $\hat{C}_2^{(n)}$ a pair of clusters estimated from $\mathbf{x}^{(n)}$. Let $\mathbf{Y}^{(n)}$ an independent and identically distributed copy of $\mathbf{X}^{(n)}$. Then, the estimator $\hat{\mathbf{\Sigma}}\left(\mathbf{Y}^{(n)}\right)$ defined in* (6.28) *satisfies the conditions of Theorem 6.3.1, i.e.*

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}} \left(\hat{\mathbf{\Sigma}}\left(\mathbf{Y}^{(n)}\right) \succeq \mathbf{\Sigma} \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right) \right) = 1. \tag{6.38}$$

Assessing whether a model of dependence satisfies the hypotheses of Proposition 6.3.4 (more precisely, Assumption 4) is not trivial as it requires full knowledge of how the inverse matrices $\left(\mathbf{U}^{(n)}\right)^{-1}$ grow up when dimension increases. However, we are able to show that Assumption 4 is satisfied for some simple dependence models and, consequently, that selective type I error can be controlled when $\mathbf{\Sigma}$ is estimated in such cases. The following remarks are proved in Appendix D.2.

Remark 6.3.5 (Diagonal)**.** *Let $\mathbf{U}^{(n)} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. If the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ is convergent, then the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 4.*

Remark 6.3.5 trivially covers the case of independent observations. Besides, if the matrix $\mathbf{X}$ is transposed, any general dependence structure between observations $\mathbf{U}$ can be estimated if independent features with known variances are provided. Another simple model that satisfies Assumption 4 is the one defined by constant variances and covariances (also known as compound symmetry). In that case, $\mathbf{U}^{(n)}$ is the sum of a constant and a diagonal matrix.

Remark 6.3.6 (Compound symmetry)**.** *Let $a, b \in \mathbb{R}$ with $b \neq a \geq 0$. If $\mathbf{U}^{(n)} = b\mathbf{1}_{n\times n} + (a-b)\,\mathbf{I}_n$, where $\mathbf{1}_{n\times n}$ is a $n \times n$ matrix of ones, then $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 4.*

We can extend the complexity of $\mathbf{U}^{(n)}$ to auto-regressive covariance structures of any lag. This is mainly thanks to the fact that the inverses of such matrices are tractable and banded, i.e. their non-zero entries are confined to a centered diagonal band. Under model (6.2), assuming that $\mathbf{U}^{(n)}$ is the covariance matrix of an auto-regressive process of order P means that

$$\frac{1}{\sqrt{\Sigma_{jj}}} X_{ij}^{(n)} = \frac{1}{\sqrt{\Sigma_{jj}}} \sum_{s=1}^{P} \beta_s X_{i-s\,j}^{(n)} + \varepsilon_i, \quad \forall\, j \in \{1, \dots, p\}, \tag{6.39}$$

where $\{\varepsilon_i\}_{i=1,\dots,n}$ are i.i.d univariate centered normal variables and $\{\beta_s\}_{s=1,\dots,P} \subset \mathbb{R}$ are the model coefficients. Then, for any $j \in \{1,\dots,p\}$, the entries of $\mathbf{U}^{(n)}$ would be given by

$$U_{ii'} = \mathrm{Cov}\left(\frac{X_{ij}}{\sqrt{\Sigma_{jj}}}, \frac{X_{i'j}}{\sqrt{\Sigma_{jj}}}\right), \quad \forall i, i' \in \{1,\dots,n\},\ \forall\, j \in \{1,\dots,p\}. \tag{6.40}$$

If the model (6.39) is assumed, the covariance matrix $\mathbf{U}^{(n)}$ and its inverse have a tractable structure. For example, for the simplest auto-regressive process where $P = 1$, and the i-th observation depends linearly only on the $(i-1)$-th one, the entries of $\mathbf{U}^{(n)}$ have the form $U^{(n)}_{ij} = \sigma^2\rho^{|i-j|}$, for $\sigma > 0$. To ensure the the positive definiteness of $\mathbf{U}^{(n)}$, we need $|\rho| < 1$ (see the form of eigenvalues in [284]). This is equivalent to ask the the process to be stationary. Then, the inverse of $\mathbf{U}^{(n)}$ is a tridiagonal matrix of the form

$$\left(\mathbf{U}^{(n)}\right)^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} 1 & -\rho & & & & \\ -\rho & 1+\rho^2 & -\rho & & & \\ & -\rho & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & 1+\rho^2 & -\rho \\ & & & & -\rho & 1 \end{pmatrix}. \tag{6.41}$$

The super and sub-diagonals trivially satisfy condition (i) in Assumption 4 with $\lambda_{\pm 1} = -\rho/(1-\rho^2)$. Then, the entries of the main diagonal define the sequences

$$\sigma^2(1-\rho^2)\left\{\left(U^{(n)}\right)^{-1}_{ii}\right\}_{n\in\mathbb{N}} = \begin{cases} \{1, 1, \dots\} & \text{if } i = 1, \\ \{\xi_1, \dots, \xi_{i-1}, 1, 1+\rho^2, 1+\rho^2, \dots\} & \text{if } i > 1, \end{cases}$$

for every $i \in \mathbb{N}$, where the entries $\sigma^2(1-\rho^2)\,(U^{(n)})^{-1}_{ii} = \xi_n$ for $i > n$ can be chosen as needed. Note that these sequences do not satisfy condition (i) in Assumption 4, but they are non-decreasing (choosing appropriately the ξ_k). Consequently, Assumption 4 holds and we have $\Lambda_{11} = 1/(\sigma^2((1-\rho^2))$, $\Lambda_{ii} = \lambda_0 = (1+\rho^2)/(\sigma^2((1-\rho^2))$ for all $i > 1$ and, finally, $\lambda = (1-\rho)^2/(\sigma^2((1-\rho^2))$. For any $P \geq 1$, the inverse matrices are banded with $2P+1$ non-zero diagonals and we can follow the same reasoning. However, for $P > 2$, we need to require the coefficients $\beta_1, \dots, \beta_P$ to have the same sign.

Remark 6.3.7 (Auto-regressive)**.** *Let $\mathbf{U}^{(n)}$ be the covariance matrix of an auto-regressive process of order $P \geq 1$ such that, if $P > 2$, $\beta_k\beta_{k'} \geq 0$ for all $k, k' \in \{1,\dots,P\}$. Then, the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 4.*

6.4 Non-maximal conditioning sets

The methodology presented in Section 6.2 sets up the framework to perform selective inference after hierarchical clustering. Exploring its adaptation to further clustering algorithms

involves, as shown in [51], the redefinition of p-values by constraining the conditional event that define (6.7) and (6.13). In this Section, we revisit the procedure of post-clustering inference introduced in Section 6.2 and rewrite it in a more general form that allows its straightforward adaptation to the scenario where more conditioning is imposed.

When defining a p-value for (6.5) that controls the selective type I error (6.6), one may think on conditioning only on having selected the pair of clusters that define the null hypothesis, i.e. on the event

$$\hat{M}_{12}(\mathbf{X}) = M_{12}(\mathbf{X}; \{\hat{C}_1, \hat{C}_2\}) = \{\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})\}. \tag{6.42}$$

However, this is not enough to ensure its analytical tractability. When considering a matrix normal distribution for the p-dimensional observations, two further conditions are imposed as shown in [104]. Following Section 6.2, this yields conditioning on the event

$$\hat{M}_{12}(\mathbf{X}) \cap \left\{ \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x}\,,\, \mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\left(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\right) = \mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\left(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\right) \right\}, \tag{6.43}$$

which is the maximal event for which any analytically tractable p-value has been shown to control (6.6) under the general model (6.2). If we denote by $\hat{T}_{12}(\mathbf{X}) = T_{12}(\mathbf{X}; \{\hat{C}_1, \hat{C}_2\})$ the second set in (6.43), we can rewrite (6.13) as

$$p_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}\left(\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} \;\middle|\; \hat{M}_{12}(\mathbf{X}) \cap \hat{T}_{12}(\mathbf{X}) \right). \tag{6.44}$$

Then, from Theorem 6.2.1 and its proof we can rewrite the truncation set in (6.14) as

$$\mathcal{S}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = \left\{ \phi \in \mathbb{R} \,:\, \hat{M}_{12}\left(\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)\right) \right\}, \tag{6.45}$$

where $\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)$ is defined in (6.17). Consequently, (6.13) is analytically tractable as

$$p_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}, \left\{ \phi \geq 0 \,:\, \hat{M}_{12}\left(\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)\right) \right\} \right), \tag{6.46}$$

where $\mathbb{F}_p$ is defined in Theorem 6.2.1. Uncoupling $\hat{M}_{12}(\mathbf{X})$ and $\hat{T}_{12}(\mathbf{X})$ in (6.44) allows us to characterize the null distribution of the p-value in terms of the conditioning event (6.42). This is useful to study the scenarios where, for technical reasons, subsets of (6.42) are chosen to define the p-value for (6.5). That was the case in [51], where the framework of [104] under model (6.1) was adapted to perform selective inference after k-means clustering. To allow the efficient computation of their truncation set, they conditioned -on $\hat{T}_{12}(\mathbf{X})$ and- on all the intermediate clustering assignments for the n observations [51, Equation (9)], which is a subset of (6.42). In accordance with (6.45) and (6.46), this more restrictive conditioning yielded the same p-value (6.7) as in [104] except from a different truncation set, based on the finer conditioning event. The following result characterizes this framework under our general model (6.2) and for any non-maximal conditioning event. Thus, it is a generalization of Theorem 6.2.1.

Theorem 6.4.1. *Let* $\mathbf{x}$ *be a realization of* $\mathbf{X}$ *and* $\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{P}(\{1, \ldots, n\})$ *with* $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$. *Let* $\emptyset \neq E_{12}(\mathbf{X}) \subset M_{12}(\mathbf{X}) = M_{12}(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\})$, $T_{12}(\mathbf{X}) = T_{12}(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\})$ *and*

$$p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \;\middle|\; E_{12}(\mathbf{X}) \cap T_{12}(\mathbf{X}) \right). \quad (6.47)$$

Then, $p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12})$ *is a p-value for the test* $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$*:* $\mu_{\mathcal{G}_1} = \mu_{\mathcal{G}_2}$ *that controls the selective type I error for clustering* (6.6) *at level* α. *Furthermore, it satisfies*

$$p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}, \left\{ \phi \geq 0 \,:\, E_{12}\left(\mathbf{x}'_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\phi) \right) \right\} \right), \quad (6.48)$$

where $\mathbb{F}_p(t, \mathcal{S})$ *is the cumulative distribution function of a* χ_p *random variable truncated to the set* $\mathcal{S}$ *and* $\mathbf{x}'_{V_{\mathcal{G}_1,\mathcal{G}_2}}(\phi)$ *is defined in* (6.17).

Note that, following (6.46), replacing $E_{12}(\mathbf{X})$ by $M_{12}(\mathbf{X})$ yields exactly Theorem 6.2.1. We omit the proof of (6.48) as it is identical to the one of (6.14) in Theorem 6.2.1. The control of the selective type I error is proved in Appendix A.2.2.

Once again, the efficient computation of (6.48) depends on the efficient computation of the truncation set $E_{12}(\mathbf{x}'_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\phi))$. As we showed for the maximal conditioning event in Lemma 6.2.2, it suffices to characterize the truncation set when the perturbed dataset $\mathbf{x}'$ is defined with respect to any norm.

Lemma 6.4.2. *Let* $\mathbf{x}$ *be a realization of* $\mathbf{X}$ *and* $\hat{C}_1, \hat{C}_2$ *an arbitrary pair of clusters in* $\mathcal{C}(\mathbf{x})$. *Let* $\mathbf{x}'$ *denote the set* (6.19) *defined in [104, Equation (12)]. Then,*

$$E_{12}\left(\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi) \right) = \frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2} E_{12}\left(\mathbf{x}'(\phi) \right). \quad (6.49)$$

We omit the proof of Lemma 6.4.2 as it is identical to the one of Lemma 6.2.2. In [51], the authors characterized $E_{12}(\mathbf{x}'(\phi))$ when E_{12} are all the intermediate clustering assignments of a k-means algorithm. That allows us once again to benefit from their efficient computation procedure and compute the truncation set under model (6.2) using Lemma 6.4.2. Consequently, we are able to perform selective inference after k-means clustering when observations and features have arbitrary dependence structures. The estimation procedure presented in Section 6.3 remains identical for this case.

6.5 Numerical experiments

This section is devoted to assess the performance of the test for the difference of cluster means in different scenarios simulated with synthetic data. We consider the following three settings for the scale matrices $\mathbf{U}$ and $\mathbf{\Sigma}$:

(*a*) $\mathbf{U} = \mathbf{I}_n$ and $\mathbf{\Sigma}$ is the covariance matrix of an AR(1) model, i.e. $U_{ij} = \sigma^2\rho^{|i-j|}$, with $\sigma = 1$ and $\rho = 0.5$.

(*b*) $\mathbf{U}$ is a compound symmetry covariance matrix, i.e. $\mathbf{U} = b + (a-b)\mathbf{I}_n$, with $a = 0.5$ and $b = 1$. $\mathbf{\Sigma}$ is a Toeplitz matrix, i.e. $\Sigma_{ij} = t(|i-j|)$, with $t(s) = 1 + 1/(1+s)$ for $s \in \mathbb{N}$.

(*c*) $\mathbf{U}$ is the covariance matrix of an AR(1) model with $\sigma = 1$ and $\rho = 0.1$. Σ is a diagonal matrix with diagonal entries given by $\Sigma_{ii} = 1 + 1/i$.

We simulate matrix normal data in settings (*a*), (*b*) and (*c*) and perform k-means and hierarchical agglomerative clustering (HAC) with average, centroid, single and complete linkages. In Section 6.5.1 we illustrate the uniformity of the p-values (6.13) under a global null hypothesis when both scale matrices are known. In Section 6.5.2, we consider the case where the dependence between observations is known and the covariance matrix between features $\mathbf{\Sigma}$ is estimated. We show, as proved in Section 6.3, that p-values are super-uniform for large enough sample sizes. Finally, in Section 6.5.3 we assess the relative efficiency of the four linkages in terms of power, for the three dependence scenarios.

6.5.1 Uniform p-values under a global null hypothesis

To illustrate the null distribution of p-values, we followed the same steps as in [104, Section 5.1]. For $n = 100$ and $p \in \{5, 20, 50\}$, we simulated $M = 2000$ samples drawn from model (6.2) in settings (*a*), (*b*) and (*c*) with $\boldsymbol{\mu} = \mathbf{0}_{n\times p}$ a zero matrix, so the null hypothesis (6.5) holds for any pair of clusters in $\mathcal{C}(\mathbf{X})$. For each simulated sample, we used k-means and HAC to estimate three clusters and tested (6.5) for two randomly selected clusters. Results for HAC with average linkage are displayed in Figure 6.1, where the empirical cumulative distribution functions (ECDF) of the simulated p-values are shown. The results for k-means and HAC with centroid, single and complete linkage are analogous to those for average linkage and we present them in Appendix D.4.1. The p-values for HAC with complete linkage are computed as their Monte Carlo approximation (6.22) with $N = 2000$ iterations. In all cases, the p-values follow a uniform distribution when the null hypothesis (6.5) holds, excluding a slight deviation from uniformity found for HAC with complete linkage under (*c*). The reasons explaining this deviation rely on the difficulty of simulating independent realizations of auto-regressive processes (see Appendix D.4.1).

6.5.2 Super-uniform p-values for unknown $\mathbf{\Sigma}$

In this section, we illustrate that p-values (6.25) are asymptotically super-uniform when Σ is asymptotically over-estimated in the sens of Loewner partial order, as proved in Theorem 6.3.1. We used the estimator (6.28) that asymptotically over-estimates $\mathbf{\Sigma}$ if Assumptions 2, 3 and 4 hold, which for the three dependence scenarios (*a*), (*b*) and (*c*) is guaranteed following Remarks 6.3.5, 6.3.6 and 6.3.7 respectively. The estimate was

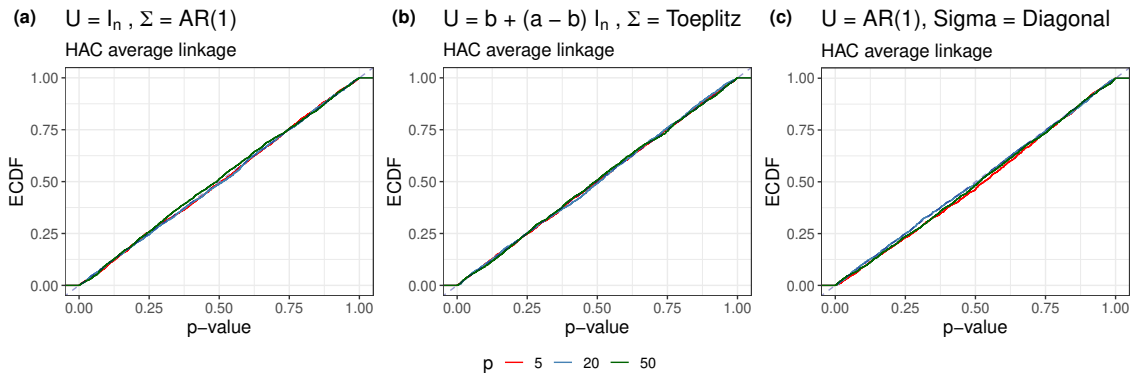


Figure 6.1: Empirical cumulative distribution functions (ECDF) of p-values (6.13) with $\mathcal{C}$ being a hierarchical clustering algorithm with average linkage. The ECDF were computed from $M = 2000$ realizations of (6.2) under the three dependence settings (a), (b) and (c) with $\boldsymbol{\mu} = \mathbf{0}_{n\times p}$, $n = 100$ and $p \in \{5, 20, 50\}$.

computed using an independent and identically distributed copy of the sample where the clustering was performed, following Proposition 6.3.4.

We followed the same steps as in [104, Section D.1]. For $n = 500$ and $p = 10$, we simulated $M = 5000$ samples drawn from (6.2) in settings (a), (b) and (c) with $\boldsymbol{\mu}$ being divided into two clusters:

$$\mu_{ij} = \begin{cases} \frac{\delta}{j} & \text{if } i \leq \frac{n}{2}, \\ -\frac{\delta}{j} & \text{otherwise,} \end{cases} \qquad \forall\, i \in \{1, \ldots, n\},\ \forall\, j \in \{1, \ldots, p\}, \tag{6.50}$$

with $\delta \in \{4, 6\}$. For k-means and HAC with average, centroid, single and complete linkage we set $\mathcal{C}$ to chose three clusters. Then, we kept the samples for which (6.5) held when comparing two randomly selected clusters. Results for HAC with average linkage are presented in Figure 6.2. The results for k-means and HAC with centroid, single and complete linkage are analogous and we present them in Appendix D.4.2. All simulations illustrate the asymptotic super-uniformity of p-values (6.13) under the null hypothesis, when $\boldsymbol{\Sigma}$ is asymptotically over-estimated using (6.28). Moreover, as the distance between clusters δ decreases, the over-estimation is less severe and the null distribution of p-values approaches the one of a uniform random variable.

It is important to remark that Figure 6.2 serves only to illustrate the validity of Theorem 6.3.1, but in no way to interpret the conservativeness of p-values when $\boldsymbol{\Sigma}$ is over-estimated. The deviation from uniformity of the null distribution of (6.25) or, equivalently, the power of the corresponding test, depends on the measure of the conditioning set, which in Figure 6.2 is determined by the frequency of iterations satisfying (6.5).

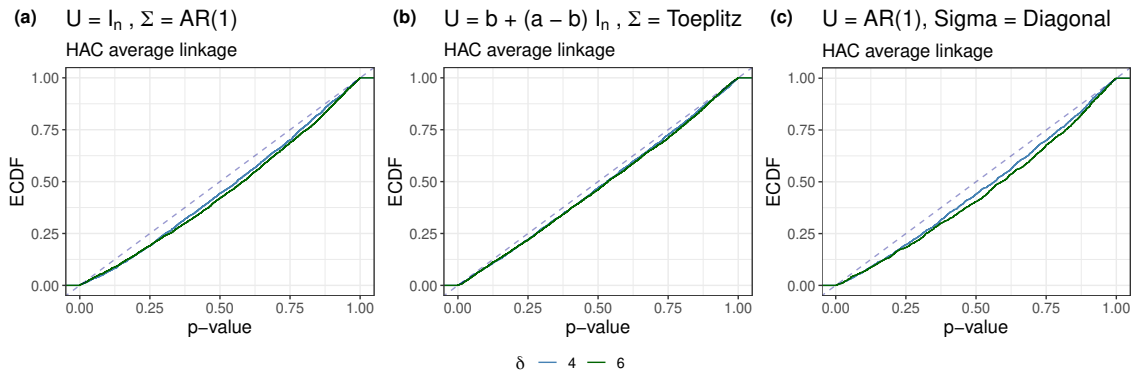


Figure 6.2: Empirical cumulative distribution functions (ECDF) of p-values (6.13) with $\mathcal{C}$ being a hierarchical clustering algorithm with average linkage. The ECDF were computed from $M = 5000$ realizations of (6.2) under the three dependence settings (a), (b) and (c) with $n = 500$, $p = 10$ and $\boldsymbol{\mu}$ given by (6.50) with $\delta \in \{4, 6\}$. Only samples for which the null hypothesis held were kept, as described in Section 6.5.2.

6.5.3 Power analysis

We conclude the numerical simulations on synthetic data by assessing the relative efficiency of the five clustering algorithms considered in terms of power. As in [104, Section 5.2], we consider the *conditional* power of the p-value (6.13), which is the probability of rejecting the null (6.5) for a randomly selected pair of clusters when it holds. To estimate the conditional power, we simulated $M = 5000$ samples drawn from (6.2) under the three settings (a), (b) and (c) with $\boldsymbol{\mu}$ dividing the $n = 50$ observations into three true clusters:

$$\mu_{ij} = \begin{cases} -\frac{\delta}{2} & \text{if } i \leq \lfloor \frac{n}{3} \rfloor, \\ \frac{\sqrt{3}\delta}{2} & \text{if } \lfloor \frac{n}{3} \rfloor < i \leq \lfloor \frac{2n}{3} \rfloor, \\ \frac{\delta}{2} & \text{otherwise,} \end{cases} \quad \forall\, i \in \{1, \dots, n\}, \forall\, j \in \{1, \dots, p\}, \tag{6.51}$$

for $p = 10$ and 14 evenly-spaced values of $\delta \in [4, 10.5]$. Then, we estimated the conditional power as the proportion of rejections at level $\alpha = 0.05$ among the samples for which the null hypothesis (6.5) did not hold (which were above the 90% of n in all settings). Figure 6.3 depicts the conditional power as a function of δ for the three scenarios (a), (b) and (c) and the five considered clustering algorithms. The p-values for HAC with complete linkage were estimated using the approximation (6.22) with $N = 2000$ iterations.

Figure 6.3 shows that, in all cases, conditional power increases with the distance between true clusters. Regarding HAC, we observe that average linkage presents the best relative efficiency among the four considered linkages in all the dependence settings, followed closely by complete linkage, which seems to weaken in (b). This might suggest that conditional power depends on the scale matrices and some scenarios might strongly differ

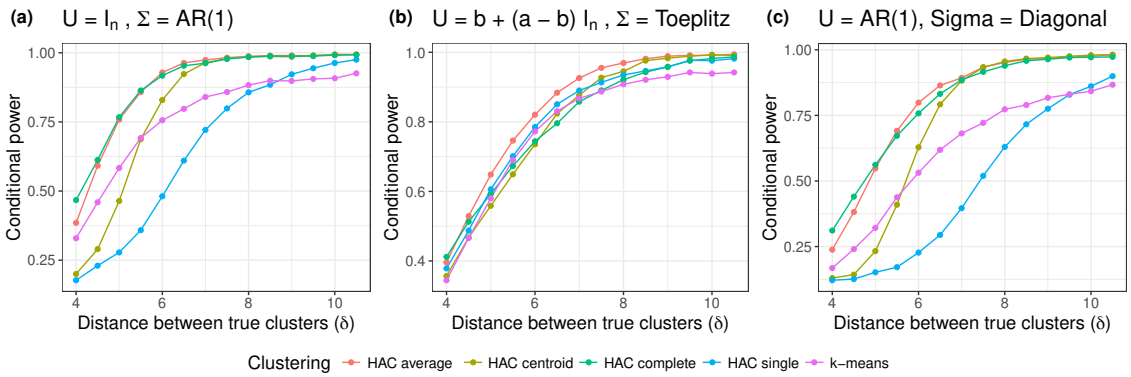


Figure 6.3: Conditional power for the test proposed in Section 6.2 under model (6.2) with the three dependence settings (a), (b) and (c) and the mean matrix defined in (6.51). The conditional power is estimated as the proportion of rejection at level $\alpha = 0.05$ among the subset of the $M = 5000$ realizations of (6.2) for which the null hypothesis (6.5) holds.

from the overall observed behavior. Indeed, the qualitative difference between average or complete linkage and centroid or single linkage that is observed in (a) and (c) considerably lessens in (b). In (a) and (c), the performance of single linkage is undoubtedly the lowest, and large differences between clusters are required to attain satisfactory levels of conditional power. However, single linkage shows the second best performance in (b).

The relative efficiency of the k-means algorithm in terms of conditional power is one of the worst among all the considered algorithms. This behavior was already pointed out by the authors in [51], that referred to the fact that conditioning on too much information entails a loss of power [141, 181, 50, 97]. Recall that the truncation set for k-means post-clustering inference defined in [50] is non-maximal to allow its efficient computation (see Section 6.4 and [51, Equation (9)]). This approach, although respecting the selective type I error as shown in Theorem 6.4.1, sacrifices the efficiency in terms of power of the corresponding test as illustrated in Figure 6.3.

6.6 Application to clustering of protein structures

Proteins are dynamic molecules essential in all living organisms. Their numerous functions are closely related to their non-static structure, which exhibits high variability within numerous protein families [175, 265, 88]. The characterization of such intrinsic structural complexities represents a highly active area of research in the field of Structural Biology. In this pursuit, clustering methods of protein conformations have provided valuable insights into this challenging problem [59, 8]. One of the main descriptors that are considered to characterize a conformation is the set of pairwise Euclidean distances between every pair

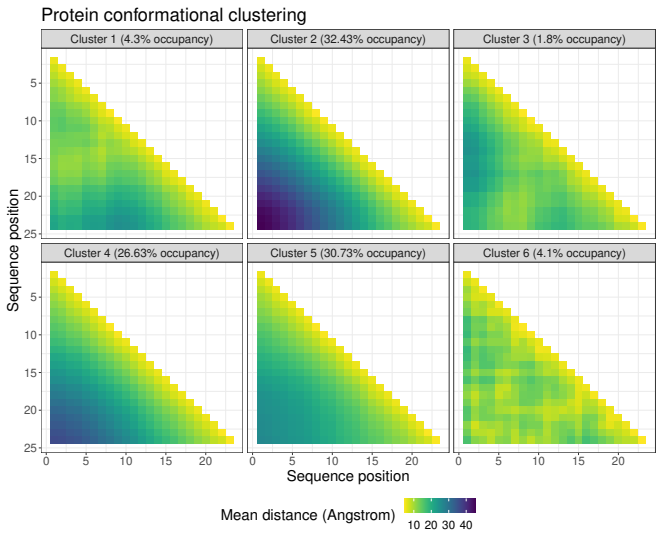


Figure 6.4: Average pairwise distances between every pair of amino acids across the conformations of each cluster. The clusters were found after performing hierarchical clustering with average linkage on the protein data presented in Section 6.6.

of amino acids along the protein sequence [229, 213, 167], usually referred to as distance maps. As these distances are strongly correlated, assuming a constant diagonal covariance matrix as in [104] seems very unrealistic. Instead, we opt for the more convenient model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \boldsymbol{\Sigma}), \tag{6.52}$$

where $\boldsymbol{\Sigma}$ can be estimated using (6.28). Each row of $\mathbf{X}$ corresponds to a protein conformation, featured by a vector of Euclidean distances between every pair of amino acids, which constitute the columns of $\mathbf{X}$. We performed hierarchical agglomerative clustering with average linkage (as it showed the best relative efficiency in Section 6.5.3) to estimate six clusters among $n = 2000$ conformations of the protein ensemble Hst5. The corresponding sequence is 24 amino acids long, so $p = 23 \cdot 24/2 = 276$. Data were simulated with Flexible-Meccano [219, 21] and refined using previously reported SAXS data [253]. Note that Flexible-Meccano is a sampling algorithm that simulates independent conformations, contrary to Molecular Dynamics simulation techniques that present temporal dependence between observations. This justifies our choice of $\mathbf{U} = \mathbf{I}_n$. Moreover, we had access to an independent replica of the simulated ensemble that we used to estimate $\boldsymbol{\Sigma}$, as it is usual for generated protein ensembles. Figure 6.4 shows the average distance map across all conformations in a given cluster or, in other words, the empirical cluster means $\bar{X}_{\hat{C}_1}, \ldots, \bar{X}_{\hat{C}_6}$ as defined in (6.4). Table 6.1 presents the p-values corresponding to every pair of clusters, corrected for multiple testing using Holm adjustment [125].

Cluster	1	2	3	4	5
2	$2.187589{\cdot}10^{-4}$				
3	$3.039844{\cdot}10^{-11}$	$1.41{\cdot}10^{-3}$			
4	$1.070993{\cdot}10^{-10}$	0.300540	$2.98464{\cdot}10^{-4}$		
5	$3.038979{\cdot}10^{-16}$	0.093018	$6.015797{\cdot}10^{-5}$	0.105446	
6	$1.729616{\cdot}10^{-6}$	0.010612	$9.290826{\cdot}10^{-9}$	$2.105{\cdot}10^{-3}$	$5.624624{\cdot}10^{-5}$

Table 6.1: p-values (6.13) computed under model (6.52) retrieved after testing (6.5) on the protein data presented in Section 6.6. The hierarchical clustering algorithm was set to find six clusters using average linkage. In blue, p-values that do not reject the null at level $\alpha = 0.05$.

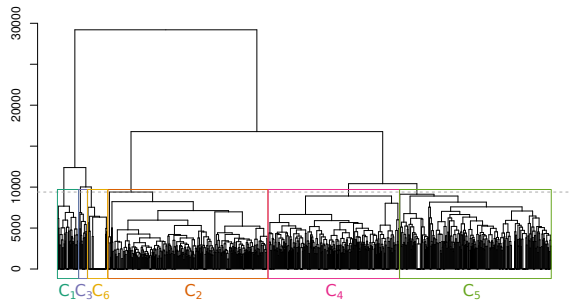


Figure 6.5: HAC dendogram for the Hst5 protein ensemble data, with the six estimated clusters marked with colored rectangles.

The p-values presented in Table 6.1 show significant differences between the most part of the average distance maps depicted in Figure 6.4. It is interesting to look at the non-rejecting pair of clusters at level $\alpha = 0.05$, marked in blue in Table 6.1, that might suggest that clusters 2, 4 and 5 could be merged into a single group. Indeed, when looking at the corresponding empirical means $\bar{X}_{\hat{C}_2}$, $\bar{X}_{\hat{C}_4}$ in Figure 6.4, we appreciate that these two clusters are characterized by high distances between pair of amino acids far away from each other in the sequence, which indicate a lack of interactions between the sequence termini and a more extended structure of the corresponding conformations. This feature appears as an exclusive and prominent characteristic of clusters 2, 4 and 5, which might explain the non-rejection of the corresponding nulls. For the rest of rejecting pairs of clusters, clear differences in distance patterns are retrieved in Figure 6.4, accounting for significant changes on Hst5 structure between the corresponding groups. Results in Table 6.1 are coherent with the HAC dendogram, presented in Figure 6.5, showing that clusters 2, 4, and 5 form a subgroup that is promptly separated from the rest.

6.7 Discussion

The seminal work [104] laid the foundation for selective inference after clustering by introducing the theoretical framework allowing to test differences between cluster means, conditioning on having estimated those clusters. Furthermore, the authors tackled the problem of estimating unknown parameters while controlling the selective type I error, which had been overlooked in previous works [173, 243] but is crucial for the practical application of this theory. Their contribution paved the way into extending post-clustering inference to more general frameworks that arise in complex real applications, where observations or features present non-negligible dependence structures. In this Chapter, we generalize the model considered in [104] to non-independent observations and features, as well as the adequate estimation of the dependence structure, from the uni-dimensional case in [104] to the matrix framework presented here. These extensions, presented in Sections 6.2 and 6.3 respectively, and numerically illustrated in Sections 6.5 and 6.6, represent the main contributions of this Chapter.

The theoretical framework presented in Section 6.2 covers any known dependence structure for observations and features. The main idea is to replace the Euclidean norm in [104] by the Mahalanobis distance with respect to the null distribution of the difference of means (6.9) to define the test statistic (Theorem 6.2.1). This removes the information about the variance from the statistic null distribution, which is now independent of $\mathbf{U}$ and $\mathbf{\Sigma}$. Although the simultaneous estimation of both scale matrices $\mathbf{\Sigma}$ and $\mathbf{U}$ is difficult to manage under (6.2), we have to set the framework allowing the estimation of one of them when the other is known. The key idea is to redefine the asymptotic over-estimation in terms of the Loewner partial order, which maintains the asymptotic control of the selective type I error (Theorem 6.3.1). Following Proposition 6.3.4, an i.i.d. copy of $\mathbf{X}$ is required to estimate $\mathbf{\Sigma}$. Resorting to data splitting here is unfeasible if $\mathbf{U}$ is not block diagonal with identical blocks. Nevertheless, in numerous practical applications several copies of $\mathbf{X}$ are naturally available, as it is the case in the analysis of simulated protein ensembles presented in Section 6.6. To allow post-clustering inference in real scenarios, we provided an estimator of $\mathbf{\Sigma}$ that asymptotically over-estimates $\mathbf{\Sigma}$ when $\mathbf{U}$ satisfies Assumptions 2, 3 and 4. Future work would consist on showing that these assumptions are satisfied for new models of dependence between observations, besides the one presented in Remarks 6.3.5, 6.3.6 and 6.3.7.

A model that generalizes the auto-regressive structure is the Toeplitz covariance matrix, whose entries depend only in the distance to the main diagonal $U_{ij} = t(|i-j|)$, where t is any real-valued function. Assessing whether the inverse of $\mathbf{U}^{(n)}$ satisfies the conditions in Assumption 4 is challenging and we were not able to state so in general. Extensive work has been done on the asymptotic behavior of continuous functions of Toeplitz matrices [108]. However, it mainly concerns their average behavior rather than their element-wise one. Further results more adapted to our problem appear if we impose $\mathbf{U}^{(n)}$ to be banded. In that case, the entry-wise convergence of the elements $\left(\mathbf{U}^{(n)}\right)^{-1}_{i\,i+r}$ has been assessed in

[62] for the tridiagonal case. This, together with the exponential decay of the entries of banded matrices [79], is enough to prove the first part of Assumption 4 for tridiagonal Toeplitz matrices. Unfortunately, the existing results do not ensure that any of the conditions (i) or (ii) in Assumption 4 holds. Assessing that remaining step is mathematically very challenging and it is out of the scope of the present Chapter.

Clustering is a multidimensional method that incorporates information from p descriptors to classify n observations. However, the encountered groups are often distinguished by a subset of variables, whose determination is essential in various fields of application [215, 295]. The framework presented in [104] was adapted to feature-level post-clustering inference in [121], testing for the difference of the g-th coordinate of cluster means, for a fixed $g \in \{1, \dots, p\}$. In that case, clustering is performed on the complete dataset $\mathbf{X}$ but the inference is carried out on the g-th column, modelled by a n-dimensional Gaussian of covariance matrix $\sigma_g^2 \mathbf{I}_n$, for a $\sigma_g > 0$. Note that the possible dependence structure between features is not taken into account for inference, but only the covariance between observations. Following a similar reasoning as in [104], the authors in [121] define a p-value that controls the selective type I error, but whose efficient analytic computation is not proposed, resorting to Monte Carlo approximation. Following the strategy presented here, adapting the framework of [121] to arbitrary dependence between observations is straightforward, but it would entail the same limitations regarding the efficient computation of the p-value. The analytical determination of the truncation set would be a highly valuable contribution. Additionally, the non-trivial extension of the over-estimation strategy presented in Section 6.3 to this framework would be essential to allow the practical implementation of the feature-level selective test.

Another potential avenue for exploration would involve adapting the efficient computation of the truncation set, as presented in [104, 51], to other clustering algorithms. The combination of dimensionality reduction algorithms, such as t-SNE [294] and UMAP [199], with clustering techniques has gained immense popularity in various fields of Biology due to its remarkable empirical efficiency [59, 82, 8, 83, 15, 2]. A notable contribution would be to develop methods that avoid computationally expensive Monte Carlo approximations and efficiently compute the truncation set in scenarios where, for example, $\mathcal{C}$ represents the composition of a dimensionality reduction algorithm with hierarchical or k-means clustering.

As discussed in Section 6.4, performing analytically tractable post-clustering inference needs the addition of technical events to the conditioning set, paying a price in power. Investigating whether these conditions might be relaxed is an interesting path for future research. The problem of power loss due to extra conditioning is not exclusive to this method. Techniques like data fission [173] need to calibrate the conditioning information and consequences in terms of power are analogous. However, it is still unknown whether power loss is more drastic in one method or the other. A substantial contribution would be to establish a framework allowing for a proper comparison of this effect when performing post-clustering inference using data fission and the approach proposed in [104]. Never-

theless, extending this comparison to practical applications would be unfeasible as long as the estimation of the covariance structure with statistical guarantees cannot be carried out in both methods.

Software availability

The methods introduced in the present Chapter are implemented in the R package `PCIdep`, available at https://github.com/gonzalez-delgado/PCIdep. The package makes use of the R package `clusterpval`, providing the approaches of [104], and the R package `KmeansInference`, providing the approaches of [51].

Acknowledgements

We thank Amin Sagar and Pau Bernadó for providing useful protein data.

This work was supported by the French National Research Agency (ANR) under grant ANR-11-LABX-0040 (LabEx CIMI) within the French State Programme "Investissements d'Avenir" and under grant ANR-22-CE45-0003 (CORNFLEX project).

Chapter 7

WARIO: Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins

The high conformational variability of flexible proteins makes their structural characterization a non-trivial task. Commonly used approaches to describe proteins based on a single structure are unsuitable in this context. Although some extensions have been proposed based on average-based techniques, their practical applicability remains limited to proteins that fluctuate around a stable conformation, but they are not capable of disentangling the underlying the variability of the highly-flexible system such as intrinsically disordered proteins. This Chapter proposes to extend the classical contact maps to the ensemble framework by incorporating the intrinsic probabilistic nature of disordered proteins. In that regard, an ensemble is characterized through a weighted family of contact maps. To do so, conformations are first described with a refined definition of contact that suitably accounts for the geometry of the inter-residue interactions and the sequence context. Then, a clustering algorithm is performed to retrieve representative groups. The performance of the method is illustrated through its application to characterize conformational ensembles of highly flexible proteins and compare it to other existing approaches. Its implementation as an easy-to-use Jupyter notebook is available at https://gitlab.laas.fr/moma/WARIO.

Contents

7.1 Introduction

Proteins are dynamic molecules essential in all living organisms. In many cases, their functions are intricately linked to their non-static structure, which can exhibit significant variability within large protein families, such as intrinsically disordered proteins (IDPs) [175, 265, 88]. Therefore, understanding the relationship between protein sequence and functional spectrum requires a suitable characterization of their structural behavior. Contact and distance maps have served as one of the main tools for characterizing the structure of rigid proteins [229, 213, 275], demonstrating their suitability to detect structural domains [250, 164, 255, 137]. More recently, contact maps have become the key tool of numerous Machine Learning models for structure prediction [116, 317, 4, 145, 304, 302, 233, 91, 52]. Indeed, contact maps provide a simplified and reliable representation of the protein structure and, consequently, their accurate prediction can assist *de novo* protein modeling [111]. However, the highly variable structural features that disordered systems exhibit are very unlikely to be captured by classical contact maps. Their naive extension to conformational ensembles, consisting on estimating contact probabilities by averaging binary contacts across every conformation, has been used to describe interaction propensities in ordered systems [201, 312, 112, 57]. However, in the presence of structural disorder, this approach looses its suitability. More precisely, it ignores the contact patterns

outside the diagonal that appear for sets of conformations with low occupancy. This can be easily illustrated with an example. Figure 7.1 shows the average contact map for all the conformations in an ensemble of CHCHD4, one of the proteins used as an example in this work (see Section 7.3 for more explanations on this protein). Contact was defined as the binary indicator of the pairwise Euclidean distance between C_β atoms (C_α for glycines) being smaller than $8\mathring{A}$. Following Figure 7.1, the frequency of a long-range contact among the states of CHCHD4 is negligible. However, as it will be shown throughout this work, this contradicts the real structural behavior of the system. In short, the methods commonly used to describe the structure of folded proteins are unsuitable in this context. Consequently, the characterization of disordered ensembles represents a non-trivial task that requires novel approaches integrating the statistical behavior of these systems.

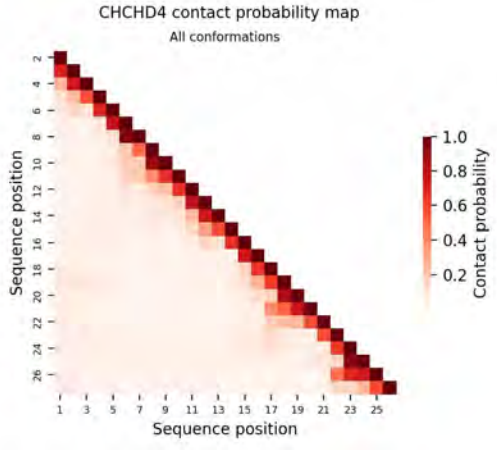


Figure 7.1: Contact probability map for CHCHD4 ensemble. Each contact probability is estimated as the proportion of contacts at threshold $8\mathring{A}$ using the Euclidean distance between the C_β (C_α for glycine) atoms.

The methodological contributions for the characterization of conformational ensembles may be classified into two major families: average-based approaches and clustering-based approaches. The first family includes methods that adapt descriptors commonly applied to folded structures by averaging them across the set of conformations. In [167], the authors propose to use the median C_α-C_α distances and their standard deviations to construct a matrix-based characterization of the ensemble. Another interesting approach is RING [191], where structured proteins are represented as a graph with edges accounting for the interactions between amino acids. An important feature of this work is the incorporation of the relative orientation between interacting residues. RING was recently extended to its ensemble version [57] by averaging the descriptors used for the structured case. However, reducing the high conformational variability of flexible proteins to averaged descriptors yields important loss of information and hides relevant structural features.

This phenomenon was also discussed in Chapter 5 regarding the comparison of protein ensembles.

Clustering-based techniques constitute another important family of works in the literature, often incorporating the projection of structural descriptors into low-dimensional spaces. In recent years, the use of advanced statistical methods has demonstrated its potential in dealing with complex probabilistic systems that appear in various areas of Biology [126]. One of these techniques are non-linear dimensionality reduction algorithms, such as t-SNE [294] or UMAP [199], which have shown efficient empirical performances when identifying underlying structures in complex data [80, 81, 15, 174, 232, 83]. This motivates their combination with clustering algorithms that detect such structures and classify them into well-defined groups. Indeed, this strategy is becoming a standard technique supported by its successful empirical efficiency [82, 2, 110, 15, 83]. Conformational spaces can be thought as high-dimensional manifolds with non-Euclidean geometries. Consequently, these techniques emerge as very attractive tools for unraveling structural features within conformational ensembles of disordered proteins. Two recent works [8, 59] combine t-SNE and UMAP with clustering algorithms to describe the structural variability of ensembles of highly flexible proteins. The main idea is to perform clustering on the low-dimensional space to provide representative families of conformations accounting for the structural distribution of the ensemble. Conformations are featured by commonly-used descriptors as all-atom coordinates and compared using RMSD [242, 186], whose suitability to compare non-folded conformations is dubious. The same strategy of clustering C_α coordinates based on RMSD dissimilarity was already performed in [177] to describe and compare ensembles of globular proteins. Backbone torsion angles and Euclidean distances between all residue pairs are also employed in [59] to feature conformations. The use of pairwise distances as structural descriptors has been widely incorporated to characterize protein ensembles [167, 45, 254, 130, 131]. Although being suitable to describe their overall structural patterns, clustering pairwise distances tends to generate clusters that match conformations with good alignments (see e.g. [8, Figure 5]). The same occurs when comparing conformations using overall metrics such as RMSD. As it is shown in this Chapter, these approaches loose the information of contacts that appear between pairs of residues in infrequent states, and misses finer patterns that take part into the structural variability of the ensemble. Together with C_α coordinates, inter-residue Lennard-Jones contact energies have been also proposed in the recent work [8] to feature conformations, also compared using RMSD.

We believe that, rather than aligning highly flexible states, a faithful characterization of disordered ensembles should classify conformations based on how residue-residue interactions manifest themselves in protein dynamics. In that regard, we opt for a strategy that exploits the potential of contact maps but that is wisely adapted to the statistical behavior of the system. This is done by first performing a well-adapted clustering algorithm that unravels the underlying conformational variability of the system and then characterizing such distribution through its representative contact motifs. We propose to characterize a

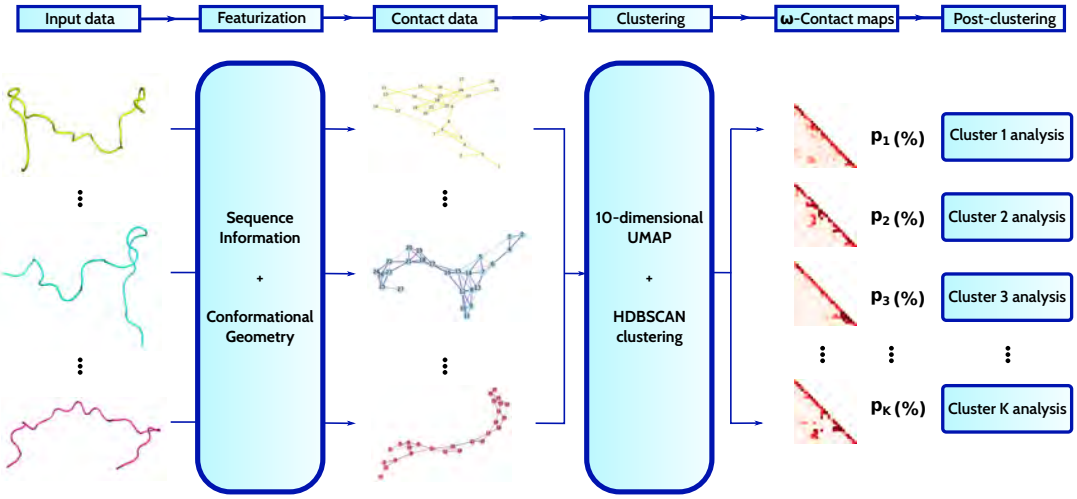


Figure 7.2: Overview of the pipeline implementing WARIO.

conformational ensemble by a *weighted family of contact maps*, representing its structural diversity through a set of contact patterns that appear with a given frequency along the dynamics of the protein. To do so, HDBSCAN [46] clustering is performed on the conformational space featured with contact-based information, passing through an UMAP low-dimensional space. To avoid the arbitrariness of contact thresholds, contact are redefined as a continuous weight function in $[0, 1]$ that acts as proxy for the interaction between residue pairs. This weight will be a function of their Euclidean distance, the distance between their relative orientation and a set of empirically-determined ideal orientations, their distance on the sequence (range) and their identities. The incorporation of relative orientation is shown crucial for an accurate detection of local structural motifs, which is essential for the structural analysis of intrinsically disordered proteins (IDPs). In addition to the contact pattern, several descriptors associated to each cluster are provided, like secondary structure propensities or average radius of gyration. The pipeline describing the method, that we named WARIO, is illustrated in Figure 7.2.

This Chapter is organized as follows:

- Section 7.2 we detail the methodology defining the complete clustering pipeline.
  - In Section 7.2.1, we start by relaxing the threshold-based contact definition for Euclidean distances through the introduction of sequence dependent contact intervals.
  - We continue by addressing the role of relative orientation in short-range contacts in Section 7.2.2 and how it can be combined with the Euclidean distance to define a metric accounting for residue-residue interactions. The precise form

of this combination is determined through empirical analysis of the interactions between amino acids. This is presented in Section 7.2.3.

  - Then, in Section 7.2.4, we define the contact function for amino acid pairs as a decreasing function of their interaction distance, whose form is once again empirically calibrated.
  - In Section 7.2.5 we detail how clustering is performed on the previously featured data. That allows us to define the ensemble characterization as a weighted family of contact maps. Then, each cluster can be futher analyzed through several proposed descriptors.
  - Section 7.2.6 presents the implementation in Python of the complete pipeline as an easy-to-use Jupyter notebook.

- Section 7.3 is devoted to show the performance of the method on conformational ensembles of four highly flexible proteins. In Section 7.4, we provide a meta-analysis of the methodology defining WARIO. First, in Section 7.4.1, we discuss how WARIO complements other existing distance-based approaches as EnGens [59]. We also demonstrate that refining the contact definition by removing arbitrary thresholds and incorporating relative orientation significantly improves the performance of the method. This is presented in Section 7.4.2.

- In Section 7.5, we discuss the suitability and possible extensions of the method, as well as its great potential for its integration in machine-learning-based (ML-based) methods applied to generate or to refine conformational ensembles of IDPs.

7.2 Methods

This section details the methodology that defines WARIO. To calibrate the functional form of functions that describe contact and interaction between amino acids, we made use of a set of 15177 experimentally-determined high-resolution structures of protein domains extracted from the SCOPe 2.07 release [49]. Throughout this section, this set will be referred to as the *structural database*.

7.2.1 Contact intervals for the Euclidean distance

Contact between amino acids is usually defined by setting universal thresholds to the Euclidean distance between their positions [212]. By universal, we mean that these thresholds are fixed independently of the amino acids identities or their distance along the sequence. However, when looking at how contact distances distribute in nature, we directly observe that residue-residue interactions concentrate around distance values that change according to these parameters. To account for this, we computed the Euclidean distance between every pair of C_α atoms (C_β for glycines) for every structure in the structural database, and represented their empirical distribution stratifying residue identities and range (the

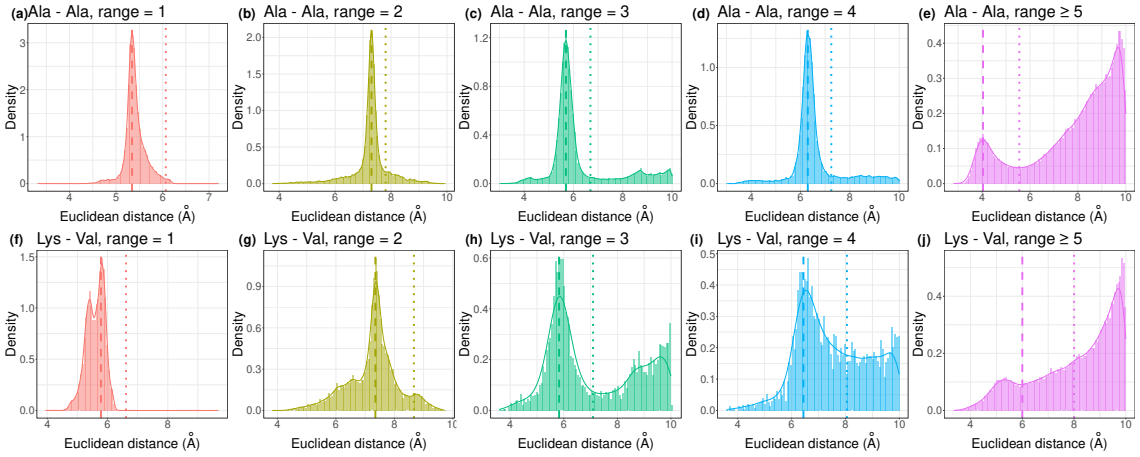


Figure 7.3: Empirical distribution of the Euclidean distance between (a-e) Ala-Ala and (f-j) Lys-Val residues in the empirical database, stratified by range groups. Distributions are depicted through a histogram and a kernel density estimate. Vertical dashed and dotted lines indicate the lower and upper limits of the contact intervals for the Euclidean distance respectively.

distance between both residues along the sequence in number of amino acids). Figure 7.3 presents the encountered distributions truncated to the interval $[0\mathring{A}, 10\mathring{A}]$ for two pairs of residues at ranges in $\{1, 2, 3, 4\}$ and $[5, \infty)$.

Figure 7.3 illustrates how the residue-residue Euclidean distance truncated to $[0\mathring{A}, 10\mathring{A}]$ is not identically distributed across amino acid identities and ranges. Distance values concentrate around sequence dependent maxima with sequence dependent variance. Therefore, contact descriptors computed from Euclidean distance must take this information into account and avoid universal thresholds that contradict the empirical behavior. The sequence-specific distance distributions presented in Figure 7.3 allow us to relax the threshold-based definition of contact for Euclidean distances. Let A_i, A_j denote a pair of amino acid identities and $S_{ij} = 1, 2, \ldots$ denote a sequence range. Let $f_{ij}^{\mathbb{R}^3}$ denote the density function of the Euclidean distance distribution for A_i-A_j pairs at range S_{ij} estimated from the empirical database and truncated to the interval $[0\mathring{A}, 10\mathring{A}]$. The *Euclidean contact interval* for A_i-A_j pairs at range S_{ij} is defined as the real interval

$$C_{ij}^{\mathbb{R}^3}(A_i, A_j, S_{ij}) = C_{ij}^{\mathbb{R}^3} = [\Delta_{a;i,j}^{\mathbb{R}^3}, \Delta_{b;i,j}^{\mathbb{R}^3}], \tag{7.1}$$

where $\Delta_{a;i,j}^{\mathbb{R}^3}$ is the abscissa smaller than $8\mathring{A}$ presenting the highest maximum of $f_{ij}^{\mathbb{R}^3}$ and $\Delta_{b;i,j}^{\mathbb{R}^3}$ is the closest abscissa from the right to $\Delta_{a;i,j}^{\mathbb{R}^3}$ presenting a minimum of $f_{ij}^{\mathbb{R}^3}$. Both limits are depicted in Figure 7.3 with dashed and dotted lines respectively. For low maxi-

mum prominences[6] (as in Figure 7.3(j)) the Euclidean contact interval is set to $[6\mathring{A}, 8\mathring{A}]$ by default. Note that, as distance distributions are not significantly different when varying $S_{ij} \geq 5$, we are setting $C_{ij}^{\mathbb{R}^3}(A_i, A_j, S) = C_{ij}^{\mathbb{R}^3}(A_i, A_j, S')$ for every $S, S' \geq 5$. The complete list of contact intervals and the counterparts of Figure 7.3 for every amino acid pair and range class are available at https://gitlab.laas.fr/moma/WARIO.

The intervals (7.1) allow a continuous description of residue-residue interactions by removing binary contact classifications. The upper limit of (7.1) represents the distance value at which the interaction probability starts to be significant, and continuously increases until reaching the lower limit of (7.1), beyond which interaction occurs with high probability. Replacing thresholds by intervals is the key idea to define continuous functions accounting for contact forcefulness that increase smoothly as the interaction probability starts to be significant. Their explicit applicability in this work is detailed in the following sections.

7.2.2 Distance to ideal orientations

Relative orientation plays a determinant role in residue-residue interactions [183, 313, 57, 146]. This idea was already incorporated in RING [191, 57], where contact thresholds were defined by integrating the values of backbone angles mediating multiple types of interactions. Here, we propose to capture this effect through a meaningful representation of the spatial pose of each amino acid. This can be achieved by defining a residue-specific reference frame at each C_β atom (C_α for glycines) as it was done in Chapter 5. The detailed construction of the reference system is included in C.1.1. An outline of its definition is presented here. To encode the angular configuration of the backbone at the residue level, we first define a virtual atom $\widetilde{C_\beta}$, which exists also for glycines. The position of $\widetilde{C_\beta}$ is an estimate of the position of the true C_β when it exists, but it is defined for every residue using only the coordinates of the C_α, C and N atoms. We denote as $\overrightarrow{C}$ and $\overrightarrow{N}$ the vectors going from C_α to C and N atoms, respectively, and we define $\overrightarrow{CN} = \overrightarrow{N} - \overrightarrow{C}$. The residue-specific reference frame is defined as follows:

$$\begin{cases} \overrightarrow{e_1} = \overrightarrow{\widetilde{C}_\beta}/\|\overrightarrow{\widetilde{C}_\beta}\| \\ \overrightarrow{e_2} = \overrightarrow{CN}/\|\overrightarrow{CN}\| \times \overrightarrow{e_1} \\ \overrightarrow{e_3} = \overrightarrow{e_1} \times \overrightarrow{e_2}. \end{cases} \tag{7.2}$$

An illustration of (7.2) is presented in Figure 7.4. Note that the third basis vector $\overrightarrow{e_3}$ is parallel to $\overrightarrow{CN}$ under the hypothesis that the atoms C, N, C_α and $\widetilde{C_\beta}$ form a perfect tetrahedron. Let L denote the sequence length and $i \in \{1, \ldots, L\}$ the position of the i-th residue. Denoting as $\mathcal{F}_i = \{\overrightarrow{e_{1,i}}, \overrightarrow{e_{2,i}}, \overrightarrow{e_{3,i}}\}$ the reference system (7.2) built on the i-th

[6] The difference between the maximum value and the one of its nearest minimum. Here, maxima with prominences lower than 0.05 are neglected.

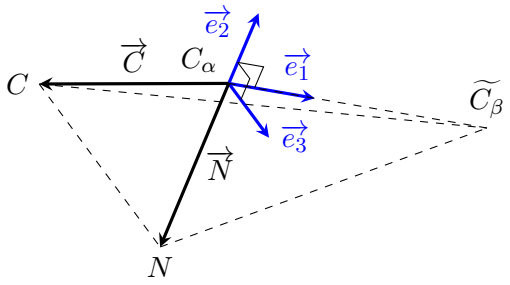


Figure 7.4: The three vectors $\{\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}\}$ defining the reference frame, built from the virtual atom $\widetilde{C_\beta}$ and vectors $\overrightarrow{C}$ and $\overrightarrow{N}$.

residue, its relative orientation with respect to another residue at position $j \neq i$ will be measured by considering the angles between the first and third basis vectors:

$$\theta_{1;i,j} = \arccos\langle \overrightarrow{e_{1,i}}, \overrightarrow{e_{1,j}} \rangle, \qquad \theta_{3;i,j} = \arccos\langle \overrightarrow{e_{3,i}}, \overrightarrow{e_{3,j}} \rangle, \tag{7.3}$$

where $\langle \cdot\,, \cdot \rangle$ denotes the inner product in $\mathbb{R}^3$. The reason why the angles (7.3) were chosen to capture the role of orientation in residue-residue interactions is that they present preferred configurations in nature. This was observed in the structural database for short range contacts i.e. for $S_{ij} = |i - j| < 5$. Figure 7.5 depicts an example of the empirical distribution of (7.3) when the Euclidean distance between both amino acids has crossed the upper limit of the contact interval (7.1) i.e. it is smaller than $\Delta^{\mathbb{R}^3}_{b;i,j}$. Indeed, residues beyond the upper limit present preferred relative orientations that are specific to their identities and range. These preferred orientations might not be unique, as in Figure 7.5(c-d) and represent the contact poses with highest probability in nature. For each pair of amino acid identities and range class, we took up to three maxima from the density estimates of the empirical distributions of (7.3). The maximum with the highest density value was always kept, and the subsequent maxima were kept if their prominence with respect to the first maximum was not negligible. We refer to these maxima as the *ideal orientations* for A_i-A_j pairs at range S_{ij}, and we denote them as

$$\theta^*_{1;i,j} = \theta^*_{1;i,j}(A_i, A_j, S_{ij}) \quad \text{and} \quad \theta^*_{3;i,j} = \theta^*_{3;i,j}(A_i, A_j, S_{ij}) \tag{7.4}$$

for the angles between the first and third basis vectors respectively. Note that (7.4) are non-empty subsets of $[0°, 180°]$ containing up to three values. The complete list of $\theta^*_{1;i,j}$ and $\theta^*_{3;i,j}$ sets and their corresponding counterparts of Figure 7.5 are available at https://gitlab.laas.fr/moma/WARIO.

Following the fact that the angles (7.3) concentrate around a set of sequence-specific ideal orientations when both amino acids interact, it is possible to define how close to the ideal contact setting is the relative orientation of a pair of residues. For two amino acids at positions $i \neq j$ in the sequence, this is done by considering the *distance between the*

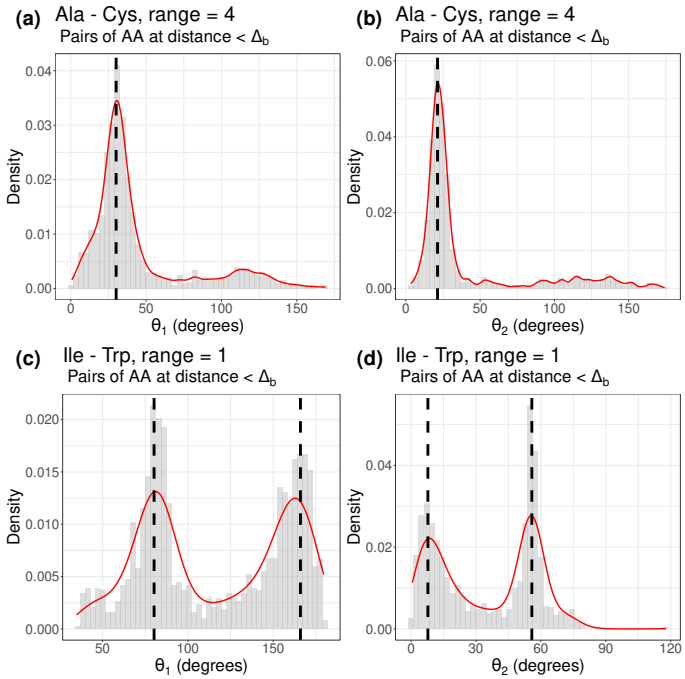


Figure 7.5: Empirical distribution of angles (7.3) computed from the empirical database for pairs of residues at Euclidean distance smaller than $\Delta_b^{\mathbb{R}^3}(A_i, A_j, S_{ij})$, stratified by amino acid identities and range. In (a-b), distributions for Ala-Cys pairs at range 4 and, in (c-d), distributions for Ile-Trp pairs at range 1. The significant maxima of the kernel density estimates (red curves) are marked with a dashed black line.

pair $\{\mathcal{F}_i, \mathcal{F}_j\}$ *and its ideal orientation*:

$$d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = \frac{1}{4}\, h\left(\min_{\theta \in \theta^*_{1;i,j}} |\theta_{1;i,j} - \theta|\right)^2 + \frac{1}{4}\, h\left(\min_{\theta \in \theta^*_{3;i,j}} |\theta_{3;i,j} - \theta|\right)^2, \tag{7.5}$$

where $h(x) = \sin(x)$ if $x \leq 90°$ and $h(x) = 1 - \cos(x)$ otherwise. This choice makes h a monotonic function on $[0°, 180°]$. Note that the quantity $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\})$ in (7.5) takes values in $[0, 1]$, with $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 0$ being a perfect match to the ideal orientation and $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 1$ the strongest disagreement with such setting. Remark also that we have omitted the explicit dependence of (7.5) on A_i, A_j and S_{ij} to lighten notation. As we mentioned before, preferred orientations were only found when $S_{ij} = |i - j| < 5$. We refer to this setting as short-range and to the case $S_{ij} \geq 5$ as long-range. Consequently, the relative orientation of the residue pair will only be considered for short-range interactions. In that case, we need to find a suitable strategy to combine distance and orientation information to correctly account for contact. This is addressed in the following section.

7.2.3 Interaction distance

The aim of this Section is to define a suitable equilibrium between Euclidean and orientation distances that correctly acts as proxy for the interaction between residue pairs. Let $i \neq j$ denote two sequence positions and $\mathcal{F}_i$, $\mathcal{F}_j$ the i-th and j-th reference frame defined in (7.2). We denote by $d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$ the Euclidean distance between the positions of both residues. We propose to combine $d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$ with (7.5) as

$$(1 - \omega_{\theta^*})^2 \, d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) + \omega^2_{\theta^*} \, d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}), \tag{7.6}$$

where the weight $\omega_{\theta^*} \in [0, 1]$ governs the distance-orientation balance. Of course, the main problem here is the suitable choice of ω_{θ^*}. This should be done by considering the following guidelines:

(i) Relative orientation must only be considered for short-range interactions,

(ii) Relative orientation must only be considered when both residues are close in Euclidean distance, i.e. closer than the upper limit of their Euclidean contact interval (7.1),

(iii) Relative orientation must significantly enhance the contact forcefulness if it is close to the ideal setting, and remain ineffective otherwise.

The first conclusion that can be extracted is that for ω_{θ^*} to satisfy ($i - iii$) it must be a function of the pair of frames, the amino acid identities and the sequence range $\omega_{\theta^*} = \omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j, A_i, A_j, S_{ij})$. To lighten notation, we will omit the explicit dependence on range and residue identities and write only $\omega_{\theta^*} = \omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j)$. The first point ($i$) can be easily guaranteed by asking $\omega_{\theta^*} = 0$ if $S_{ij} \geq 5$. For long-range interactions, contact will be exclusively encoded by the Euclidean distance between both residues. Ensuring (ii) remains to ask ω_{θ^*} to be a decreasing function of $d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$, whose smooth decay concentrates in the Euclidean contact interval (7.1). Finally, satisfying (iii) demands that ω_{θ^*} is also decreasing with $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\})$. Note that the word *significantly* has been added in (iii). In other words, ω_{θ^*} needs to calibrate distance and orientation in a way that they are comparable beyond the Euclidean contact interval when orientation plays a non-negligible role. This can be ensured if the following relation holds

$$(1 - \omega_{\theta^*})^2 \, d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) \sim \omega^2_{\theta^*} \, d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) \quad \text{for all } \; d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) \leq \Delta^{\mathbb{R}^3}_{a;i,j}, \tag{7.7}$$

where $\Delta^{\mathbb{R}^3}_{a;i,j}$ is the lower limit of the Euclidean contact interval for A_i-A_j pairs at range S_{ij}, defined in (7.1). All these conditions are verified if the following functional form is chosen to define ω_{θ^*}.

$$\omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j) = \begin{cases} 1 - \tanh\left[4 \left(d^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) + g_{ij}\left(d^2_{\mathbb{R}^3}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)\right)\right)^2\right] & \text{if} \quad S_{ij} < 5, \\ 0 & \text{otherwise,} \end{cases} \tag{7.8}$$

where

$$g_{ij}(x) = \frac{1}{2}\left(\frac{x}{\Delta^{\mathbb{R}^3}_{b;i,j}}\right)^{\frac{d^{\mathbb{R}^3}_{ij}}{2}} \quad \text{for all } x \geq 0 \quad \text{and} \quad d^{\mathbb{R}^3}_{ij} = \frac{\log\left(\operatorname{argtanh}\left(1/\Delta^{\mathbb{R}^3}_{a;i,j}\right)\right)}{\log\left(\operatorname{argtanh}\left(\Delta^{\mathbb{R}^3}_{a;i,j}/\Delta^{\mathbb{R}^3}_{b;i,j}\right)\right)}, \quad (7.9)$$

where $\Delta^{\mathbb{R}^3}_{a;i,j}$ (resp. $\Delta^{\mathbb{R}^3}_{b;i,j}$) is the lower (resp. upper) limit of the Euclidean contact interval for A_i-A_j pairs at range S_{ij}, defined in (7.1). With this, it is possible to define the *interaction distance* between the pair of residues A_i-A_j with frames $\mathcal{F}_i$, $\mathcal{F}_j$ at range S_{ij} in the sequence as the function

$$d_{\text{int}}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) = \left(1 - \omega_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)\right)^2 d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) + \omega^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) d^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right), \quad (7.10)$$

where we have omitted the dependence on amino acid identities and range for simplicity and $\omega^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)$ is defined in (7.8). A clear visualization of the orientation weight (7.8) and the interaction distance (7.10) is presented in Figure 7.6 for Ala-Ala pairs at range 3. The curves in Figure 7.6 show the definition (7.8) satisfies conditions $(i - iii)$. Note first that for Euclidean distances greater than the upper limit of the Euclidean contact interval, orientation is not considered to describe interaction. Its contribution smoothly increases when crossing the Euclidean contact interval from right to left, becoming comparable to the one of the Euclidean distance after crossing the lower limit. The rise of ω_{θ^*} is stronger when the relative orientation of $\{\mathcal{F}_i,\ \mathcal{F}_j\}$ gets closer to its ideal setting, and weaker otherwise. Indeed, orientation has no effect in the worst scenario $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 0$. In other words, the role of orientation is to enhance the contact forcefulness defined by the Euclidean distance when it is close to the ideal setting. It is important to remark that dependence of ω_{θ^*} on Euclidean distance and orientation occurs smoothly in all directions. This is possible thanks to the definition of Euclidean contact intervals (7.1), that allows to concentrate the smooth variation of ω_{θ^*} within a sequence dependent range of values extracted in accordance to the observed empirical behavior.

7.2.4 Contact function definition

This section is devoted to define contact between amino acids as a continuous function taking values in $[0, 1]$ and correctly acting as an indicator of their interaction strength. In other words, contact will be defined as a decreasing function of the interaction distance:

$$\omega^C_{ij}(\{\mathcal{F}_i, \mathcal{F}_j\}) = t_{ij}\left(d_{\text{int}}(\{\mathcal{F}_i, \mathcal{F}_j\})\right) \quad \text{with } t_{ij} : [0, \infty) \longrightarrow [0, 1] \quad \text{decreasing.} \quad (7.11)$$

The contact function ω^C_{ij} will take values close to 1 (resp. 0) when the interaction distance between residues at positions $i \neq j$ is close to (resp. far from) 0. Once again, we will ask the contact function to decrease smoothly with d_{int} and to concentrate its decay inside an empirically determined interval. To calibrate its functional form, we proceed analogously to the previous sections and start by computing the empirical distribution

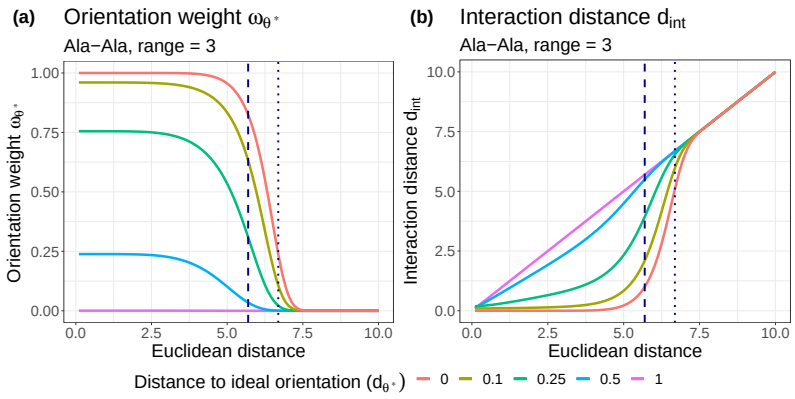


Figure 7.6: For Ala-Ala pairs at range 3, (a) the weight function (7.8) and (b) the interaction distance (7.10). Both quantities are depicted as a function of the Euclidean distance between residue positions and stratified by distance to the ideal orientation. Dashed and dotted vertical lines indicate respectively the lower and upper limit of the Euclidean contact interval.

of the interaction distance (7.10) for every pair of amino acids at ranges in $\{1, 2, 3, 4\}$ extracted from the structural database. The results for Ala-Ala and Lys-Val pairs are presented in Figure 7.7.

Figure 7.7 illustrates the effect of incorporating (7.5) to the Euclidean distance when accounting for short-range residue-residue interactions. If we compare panels in Figure 7.7 with their counterparts of Figure 7.3, we see how the interaction distance (7.10) enhances -by translating their Euclidean distance value to the left- those residue pairs whose relative orientation is close to the ideal one. This translation is very clear for pairs at range 1, for which the uni-modal distributions of Figure 7.3 become bi-modal in Figure 7.7, but it also appreciable for ranges higher than one, where the probability mass moves to smaller distance values thanks to the residue pairs with low values of (7.5). Note that the shift is more visible for contacts at range 1 due to the high concentration of the distance distribution around its mean. Indeed, distances and orientations between consecutive residues are very physically restricted. For longer ranges, the shift is equally present but less appreciable through Figure 7.7 due to the higher variance of the distance distributions. To conclude, defining (7.10) allows us to filter residue-residue interactions that, besides corresponding to amino acids close in Euclidean distance, present ideal relative orientations. We introduce now the analogous contact interval of (7.1) for the interaction distance (7.10). Let A_i, A_j denote a pair of amino acid identities and $S_{ij} = 1, 2, \ldots$ denote a sequence range. Let f_{ij}^{int} denote the density function of the interaction distance distribution for A_i-A_j pairs at range S_{ij} estimated from the structural database and truncated to the interval $[0\mathring{A}, 10\mathring{A}]$.

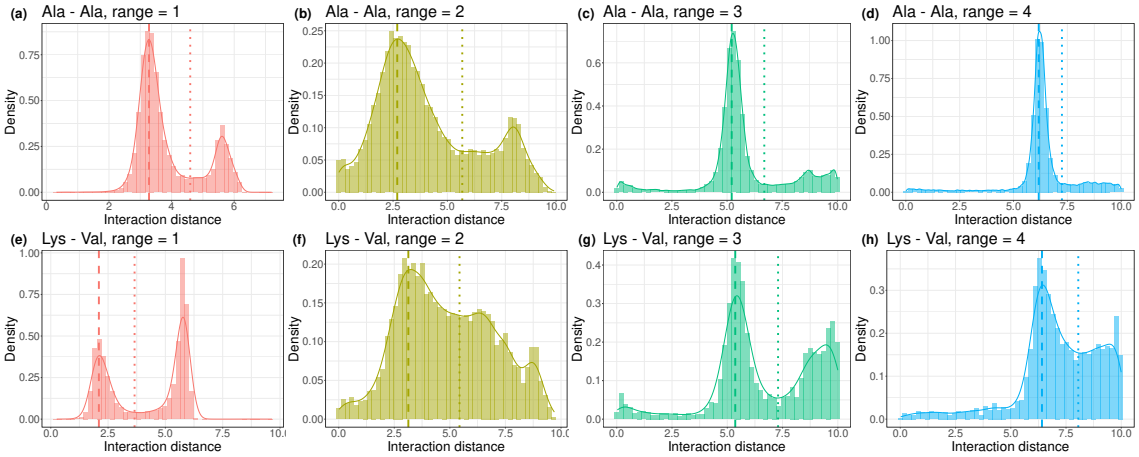


Figure 7.7: Empirical distribution of the interaction distance (7.10) between (a-e) Ala-Ala and (f-j) Lys-Val residues in the empirical database, stratified by range groups. Distributions are depicted through a histogram and a kernel density estimate. Dashed and dotted vertical lines indicate respectively the lower and upper limit of the contact interval.

The *contact interval* for A_i-A_j pairs at range S_{ij} is defined as the real interval

$$C_{ij}^{\text{int}}(A_i, A_j, S_{ij}) = C_{ij}^{\text{int}} = [\Delta_{a;i,j}^{\text{int}}, \Delta_{b;i,j}^{\text{int}}], \tag{7.12}$$

where $\Delta_{a;i,j}^{\text{int}}$ is the smaller abscissa presenting a maximum of f_{ij}^{int} and $\Delta_{b;i,j}$ is the closest abscissa from the right to $\Delta_{a;i,j}^{\text{int}}$ presenting a minimum of f_{ij}^{int}. As the interaction distance (7.10) corresponds to the Euclidean one for $S_{ij} \geq 5$, we have

$$\Delta_{a;i,j}^{\text{int}} = \Delta_{a;i,j}^{\mathbb{R}^3}, \quad \Delta_{b;i,j}^{\text{int}} = \Delta_{b;i,j}^{\mathbb{R}^3} \quad \text{for all } S_{ij} \geq 5.$$

Then, we choose the decreasing function t_{ij} in (7.11) to concentrate its smooth decay in (7.12). This can be done by choosing

$$t_{ij}(x) = 1 - \tanh\left[\left(\frac{x}{\Delta_{b;i,j}^{\text{int}}}\right)^{d_{ij}^{\text{int}}}\right] \quad \text{for all } x \geq 0 \quad \text{and} \quad d_{ij}^{\text{int}} = \frac{\log\left(\operatorname{argtanh}\left(1/\Delta_{a;i,j}^{\text{int}}\right)\right)}{\log\left(\operatorname{argtanh}\left(\Delta_{a;i,j}^{\text{int}}/\Delta_{b;i,j}^{\text{int}}\right)\right)}. \tag{7.13}$$

The curve of t_{ij} is illustrated in Figure 7.8 for Ala-Ala pairs at range 3. It shows how the contact function (7.11) represents a relaxation of the classical step function based on a universal threshold. Here, contact is described by a continuous function whose transition from low to high values is smooth and concentrated inside an empirically determined sequence-specific interval.

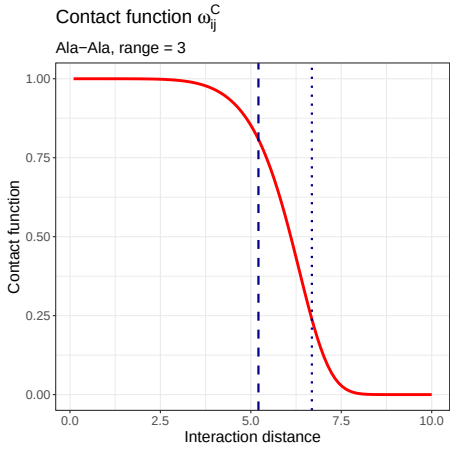


Figure 7.8: Contact function (7.11) as a function of interaction distance (7.10) for Ala-Ala pairs at range 3. Vertical dashed and dotted lines indicate respectively the lower and upper limit of the contact interval.

7.2.5 Clustering pipeline and ensemble characterization

This section details the clustering algorithm for the ensemble characterization. First, data will be featured by the contact function values for every pair of amino acid residues along the sequence. Consequently, an ensemble corresponding to a protein of length L and having n conformations will be described by the $n \times L(L-1)/2$ matrix

$$\mathbf{W}_C = \begin{pmatrix} \omega^C_{11;1} & \omega^C_{12;1} & \cdots & \omega^C_{ij;1} & \cdots & \omega^C_{L(L-1);1} \\ \omega^C_{11;2} & \omega^C_{12;2} & \cdots & \omega^C_{ij;2} & \cdots & \omega^C_{L(L-1);2} \\ \vdots & \vdots & & & & \vdots \\ \omega^C_{11;n} & \omega^C_{12;n} & \cdots & \omega^C_{ij;n} & \cdots & \omega^C_{L(L-1);n} \end{pmatrix}, \tag{7.14}$$

where $\omega^C_{ij;k}$ denotes the value of ω^C_{ij} defined in (7.11) for the k-th conformation, for $k \in \{1, \ldots, n\}$. Note that this formulation is equivalent to consider each conformation as a graph , as it was explicitely done in RING [191, 57]. Here, the set of nodes is given by the set of residues and every pair of residues at positions i,j is linked by an edge weighted by $\omega^C_{ij;k}$. This is the idea depicted in Figure 7.2. Then, clustering the rows of (7.14) comes down to clustering the set of graphs. Note that the graph representation is just an alternative visualization of the data, but the presented methodology does not resort to elements of graph theory.

The clustering algorithm performed on (7.14) is based on the combination of a dimensionality reduction technique with an efficient clustering method, similarly to state-

of-the-art approaches [59, 8]. Here, we choose UMAP [199] to first embed the data (7.14) into a 10-dimensional space. This choice is motivated by its ability to preserve the high-dimensional topology of data and efficiently reveal population structure [81, 83]. Then, we apply the clustering algorithm HDBSCAN [46] to the embedding, which we believe to be one of the most sophisticated density based techniques. One of its practical advantages in this context is that it takes as input parameter the minimum cluster occupancy and selects automatically the retrieved number of classes. This seems more natural in our setting as the practitioner might have more intuition in the desired "resolution" of the characterization through the setting of a minimum number of conformations rather than through the direct choice of a number of classes.

Once the clustering is performed, each class is characterized through a cluster-specific contact map. Let K be the number of retrieved classes and $\mathcal{C}_k \subset \{1, \dots, n\}$ be the subset of conformations constituting the k-the cluster, for $k \in \{1, \dots, K\}$. Of course $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$ for all $k \neq k'$. Keeping with the notation of (7.14), we define the k-th *cluster-specific ω-contact map* as the $(L-1) \times (L-1)$ matrix

$$\overline{W}_{\mathcal{C}_k} = \left(\frac{1}{|\mathcal{C}_k|} \sum_{l \in \mathcal{C}_k} \omega^C_{ij;l} \right)_{ij} \qquad \text{for } i < j \in \{1, \dots, L\}. \tag{7.15}$$

where $|\mathcal{C}_k|$ denotes the cardinal of $\mathcal{C}_k$. The matrix (7.15) is the average of all the rows in (7.14) that belong to the k-th cluster, represented in matrix form. Its entries are the cluster averages of the contact function (7.11) values for every pair of residues along the sequence, and it accounts for the contact patterns that dominate the cluster. To the matrix (7.15) we can assign a weight p_k given by the cluster occupancy proportion

$$p_k = \frac{|\mathcal{C}_k|}{n}. \tag{7.16}$$

This allows us to define the *ensemble characterization* as the K-tuple of weighted cluster-specific ω-contact maps:

$$\mathcal{E} = \left(\left(\overline{W}_{\mathcal{C}_1}, p_1 \right), \dots, \left(\overline{W}_{\mathcal{C}_K}, p_K \right) \right). \tag{7.17}$$

The representation (7.17) provides a compact characterization of how residue-residue interactions distribute within the ensemble. Instead of averaging contacts across conformations as in Figure 7.1, WARIO first disentangles their underlying distribution through the clustering algorithm and represent it as a weighted family of representative motifs. Thanks to the proportions (7.16), it is easy to extract a representative family of conformations by sampling from the distribution

$$P_{\text{rep}}(\mathcal{E}) = p_1 \mathcal{U}(\mathcal{C}_1) + \cdots + p_K \mathcal{U}(\mathcal{C}_K), \tag{7.18}$$

where $\mathcal{U}(\mathcal{C}_k)$ denotes the discrete uniform distribution on $\mathcal{C}_k$, for $k \in \{1, \dots, K\}$. Note that the HDBSCAN algorithm might not classify every conformation. In that case, before sampling from (7.18) the proportions (7.16) must be normalized to one.

Every cluster of conformations can be analyzed a posteriori through any suitable descriptor. Here, we propose to evaluate the secondary structure propensities based on the structural classification provided by DSSP [152] and to compute the cluster average radius of gyration. These descriptors provide a preliminary picture of the class family that characterizes the ensemble, beyond the contact patterns that define each group. More specific descriptors that align with the practitioner's needs can be easily added to the post-clustering analysis, using methods implemented in tools like SOURSOP [165].

7.2.6 The Jupyter Notebook

WARIO has been implemented through an easy-to-use Jupyter Notebook. It is available at https://gitlab.laas.fr/moma/WARIO, together with its installation guidelines and detailed implementation instructions. The notebook takes a confrontational ensemble as input and returns the ensemble characterization (7.17). The data featurization is performed at a first stage, allowing the user to adjust the resolution of the clustering algorithm afterwards. Then, clustering is performed and results are saved and depicted. Cluster-specific secondary structure propensities and average radius of gyration are also provided.

Ensembles can be given as input in several of the most common data formats. WARIO accepts one .xtc file together with a topology file in any format admitted by MDTraj [196], one multiframe .pdb file or a folder containing one .pdb file per conformation. The user can also choose to characterize sequence segments instead of the ensemble entire sequence. Details are provided in the notebook documentation. It should be noted that the current implemention of WARIO requires an all-atom representation of the protein backbone. Its extension to coarse-grained models represents a future work.

The main output of WARIO is given through a weighted set of ω-contact maps depicting the interaction patterns that characterize each cluster. Together, a plot with cluster-specific DSSP propensities and average radius of gyration is provided. The notebook also allows to create new files -in the same format as the one provided by the user- to divide conformations by clusters. These files can be used for further analysis of the retrieved contact patterns and the computation of any other descriptor that aligns with the practitioner's needs.

7.3 Results

This section is devoted to illustrate the ability of WARIO to characterize ensembles of highly-flexible proteins. We implement the pipeline described in Section 7.2 to four ensembles drawn from Molecular Dynamics (MD) simulations and presenting high structural variability. The corresponding results are presented in the following four sub-sections. We do not present detailed explanations about the MD techniques implemented to generate the ensembles, as the aim of this Section is to evaluate the performance of the presented methodology.

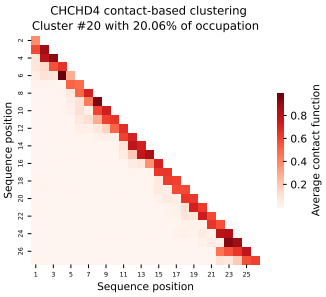
7.3.1 Characterization of CHCHD4

We first applied WARIO to characterize the intrinsically disordered domain of the protein ensemble CHCHD4 (Coiled-Coil-Helix-Coiled-Coil-Helix Domain Containing 4). This protein plays a crucial role in the import of intermembrane space-targeted proteins [123, 96]. An example is the case of the CHCHD4-AIF complex, which is responsible for regulating the import and correct folding of cysteine-containing proteins in the mitochondrial intermembrane space [115, 245]. As it was recently shown, this complex is also involved in lung cancer development [240]. Only the folded domain of CHCHD4 (residues 45-109) has been experimentally resolved [12]. However, the interaction with AIF involves exclusively the intrinsically disordered N-terminal of CHCHD4 (27 residues) [115], that we characterize here and refer to as simply "CHCHD4" from now on.

We made use of a MD trajectory drawn from the concatenation of 50 independent simulations of 200 ns each (Ha-Duong *et al.*, unpublished). The ensemble contained $n = 100050$ conformations and the sequence had $L = 27$ amino acids. We set the minimum cluster size to the 1% of the total number of conformations. WARIO identified 23 clusters with different levels of occupancy. The two most populated clusters contain approximately the 20% and 16% of conformations, and the remaining 21 clusters contain around 1-3% of conformations each. The overall cluster distribution can be visualized through the projection to a 2-dimensional UMAP space. This type of representation gives us a general outlook of how weights (7.16) distribute across the ensemble characterization (7.17), but its analysis cannot be extended further due to the non-trivial interpretation of the coordinates in UMAP space. It is provided in Appendix E.2.

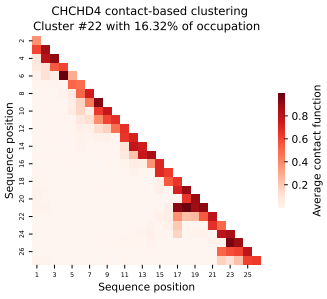
When looking at the ensemble characterization (7.17) for CHCHD4, the first overall conclusion that we can extract is that the two more occupied clusters do not present long-range contacts. Nevertheless, all the remaining low occupied groups of conformations presented specific contact motifs far from the diagonal of (7.15). That emphasizes the need of fine clustering algorithms implemented of suitable data accounting for contact information. Broader characterizations might miss low populated clusters whose contact patterns may be determinant for the practitioner's interest. Let us first take a look at the two most populated clusters of CHCHD4. Their corresponding ω-contact maps are presented in Figure 7.9. We also depict 10 conformations randomly selected from the cluster and aligned at the residues presenting prominent contact patterns.

Indeed, conformations in clusters 20 and 22 (Figure 7.9) present only short-range residue residue interactions. Contact specificities appear at the C-terminal (cluster 20) and at residues 17-21 (cluster 22). The absence of long-range contacts in the most occupied classes equally manifest when looking at the cluster average radius of gyration, which is substantially larger for clusters 20 and 22 (15.36$\mathring{A}$, 13.98$\mathring{A}$ respectively) than for the rest of low-occupied groups, presenting values around 10-12$\mathring{A}$. These remaining groups contain the $1-3\%$ of conformations each and are characterized by long-range contact patterns. Four examples are presented in Figure 7.10. Indeed, contacts between amino acids far

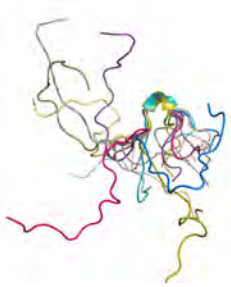


(a) Cluster-specific ω-contact map.

(b) Alignment at res. 24-26.



(c) Cluster-specific ω-contact map.



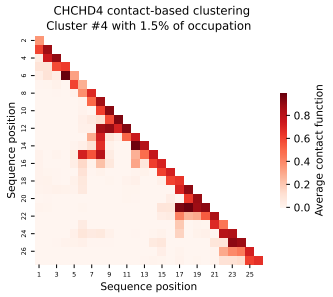
(d) Alignment at res. 17-21.

Figure 7.9: (a,c) Cluster-specific ω-contact maps (7.15) for the two most populated clusters of CHCHD4. (b,d) 10 conformations randomly selected from each cluster (a,c respectively) and aligned at residues exhibiting off-diagonal contact patterns.

away from each other in the sequence stand out in the ω-contact maps. In some cases, clusters are characterized by the presence of structural motifs at some given residues, as for cluster 8 (Figure 7.10(c,d)). The complete family of ω-contact maps for CHCHD4 as well as the secondary structure propensities and average radius of gyration for every cluster are presented in Appendix E.2.

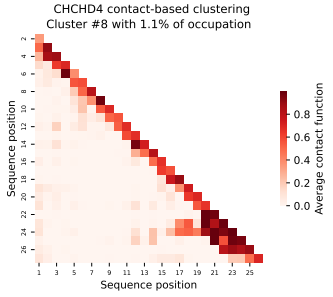
7.3.2 Characterization of Huntingtin

The N-terminal region of huntingtin, the so-called exon-1, is the causative agent of Huntington's disease, a deadly neurodegenerative pathology [257]. This fragment, which we will call huntingtin from now on, contains a poly-glutamine tract, poly-Q, that is flanked by 17 amino acids (N17) at N- and and a proline rich region at C-. Importantly, when the number of glutamines in the poly-Q exceeds a pathological threshold of 35, the protein forms large aggregates in neurons that cause the pathology. The structural changes occurring in the protein above the threshold have been the object of an intense research [287, 87].



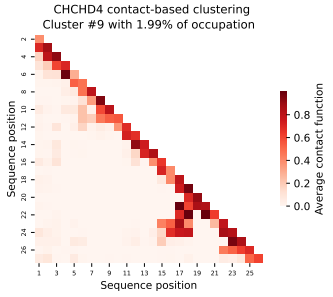
(a) Cluster-specific ω-contact map.

(b) Alignment at res. 5-15.



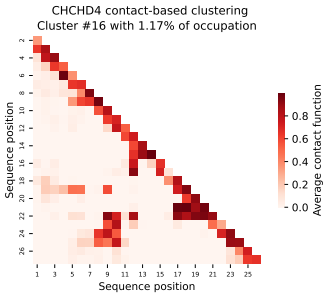
(c) Cluster-specific ω-contact map.

(d) Alignment at res. 21-25.



(e) Cluster-specific ω-contact map.

(f) Alignment at res. 16-24.



(g) Cluster-specific ω-contact map.

(h) Alignment at res. 8-12, 22-26.

Figure 7.10: Left column: CHCHD4 cluster-specific ω-contact maps (7.15). Right column: 10 CHCHD4 conformations randomly selected from the cluster in the same row, left column, and aligned at residues exhibiting off-diagonal contacts.

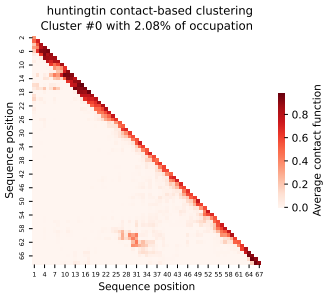
Here, we aim at characterizing a 20 microsecond MD trajectory of a pathogenic construct encompassing the N17, a poly-Q tract of 46 glutamines and 5 prolines. Details of this simulation can be found in the original publication [87]. The trajectory analysed contains $n = 96000$ conformations.

Here, the sequence had $L = 68$ amino acids and WARIO found $K = 43$ clusters (almost twice as for CHCHD4 with 27 residues). The main overall difference with respect to the previous analysis is the lack of prominent clusters in terms of occupancy. For Huntingtin, all the 43 clusters presented comparable sizes of around 1-3% the number of conformations when setting to the 1% of n the minimum cluster size. Once again, the overall distribution of the weights (7.16) can be illustrated by projecting (7.14) to a two-dimensional UMAP space (see Appendix E.2).

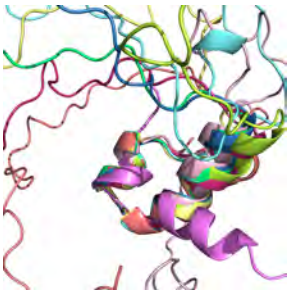
The ensemble characterization (7.17) for Huntingtin gathers the family of structural configurations yielded by the helix displacement across the protein dynamic. Each cluster-specific ω-contact maps account for a different helical pattern at a given sequence subset, as illustrated in Figure 7.11 for four examples (see Appendix E.2 for the complete characterization). Among the 43 clusters, very few long-range contacts appear and the structural dynamic is mainly governed by the short-range helical motifs appearing close to the diagonal. This is also appreciated when looking at how secondary structure DSSP propensities evolve across clusters (see Appendix E.2).

This analysis serves also to assess the effect of the clustering resolution, calibrated by the choice of the minimum cluster size. In some cases, the detection of very few frequent structural motifs or contact patterns might be essential, and those might be unnoticed if the resolution is not low enough. Look, for example, at the cluster-specific ω-contact map for the 11-th cluster, presented in Figure 7.12(a) and corresponding to the 2.56% of conformations. Besides the long-range contacts appearing between $\sim$50-55 and $\sim$60-65 residues, a contact pattern that characterizes β-sheet structures leans out around residue 35. This is barely appreciated as its corresponding (7.11) average value remains around 0.1-0.2. Nevertheless, that means that between 10% and 20% of cluster 11 conformations might present this extended motif at that sequence segment. This can be confirmed by refining the clustering resolution and looking at how conformations belonging to the 11-th cluster spread out among the new partition.

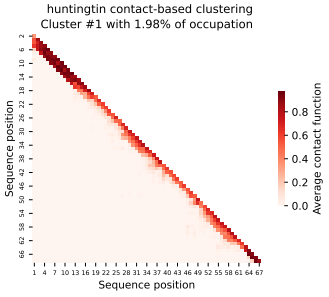
We repeated the clustering algorithm by setting the minimum cluster size to the 0.1% of the total number of conformations. WARIO retrieved now 440 clusters, which is probably a too fine representation of the conformational variability. However, re-calibrating the clustering resolution in that way is useful to extract determinant contact patterns that might be hidden inside the broader classification. The new cluster 144 is a subset of the previous cluster 11 and it contains the 0.2% of the conformations. Its ω-contact map is depicted in Figure 7.12(b), where the β-sheet contact pattern has clearly stood out. Indeed, that structural motif appears when looking at random conformations extracted from the cluster, as shown in Figure 7.12(c). Among the 440 cluster-specific ω-contact maps, the one in Figure 7.12(b) was the only presenting β-sheet motifs at any sequence



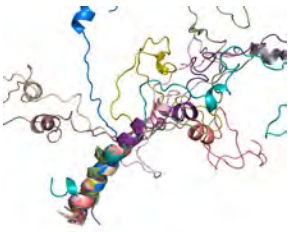
(a) Cluster-specific ω-contact map.



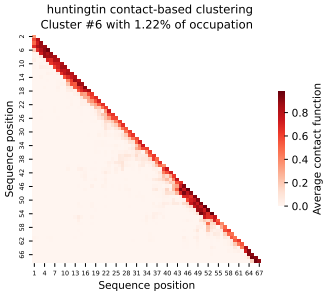
(b) Alignment at res. 1-25.



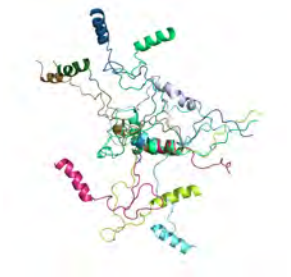
(c) Cluster-specific ω-contact map.



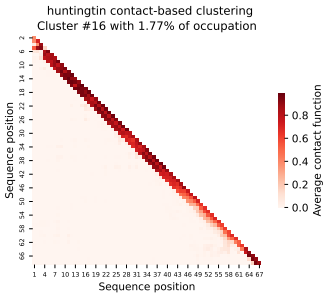
(d) Alignment at res. 1-25.



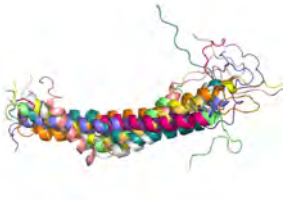
(e) Cluster-specific ω-contact map.



(f) Alignment at res. 45-55.

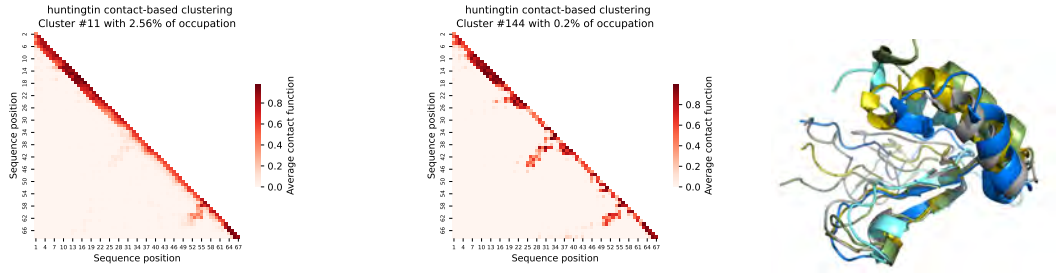


(g) Cluster-specific ω-contact map.



(h) Alignment at all res.

Figure 7.11: Left column: Huntingtin cluster-specific ω-contact maps (7.15). Right column: 10 Huntingtin conformations randomly selected from the cluster in the same row, left column, and aligned at residues exhibiting off-diagonal contacts.



(a) (b) (c)

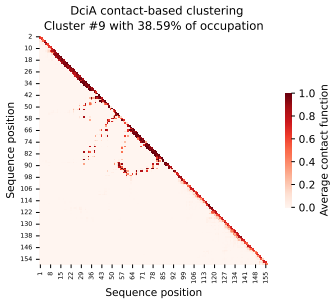
Figure 7.12: (a, resp. b) Cluster-specific ω-contact map for the 11-th (resp. 144-th) cluster of Huntingtin characterization after setting to the 1% (resp. 0.1%) of n the minimum cluster size. (c) Ten conformations randomly selected from the 144-th cluster and aligned at the extended sheet structure.

segment. That shows how WARIO is capable of detecting that 192 conformations among the 96000 that constitute the ensemble present a particular contact pattern that is specific and exclusive to them.

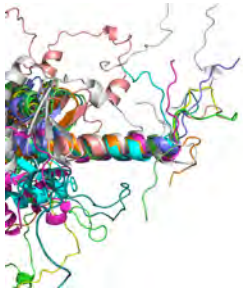
7.3.3 Characterization of DciA

Here we illustrate an example of a protein ensemble with a longer sequence. We consider the DciA protein having $L = 157$ amino acids. DciA is a gene that exhibits widespread prevalence among diverse bacterial species, encompassing a significant number of pathogenic strains like Vibrio cholerae, Yersinia pestis, Mycobacterium tuberculosis, and Pseudomonas aeruginosa. Notably, in the case of Pseudomonas aeruginosa, in-depth investigations have provided evidence of a direct and specific interaction between the DciA gene and DnaB protein. Remarkably, experimental knockout of the DciA gene has been found to induce a consequential impediment in the initiation of replication [42, 190]. Structurally, DciA presents a folded N-terminal domain at residues 1-111, and a disordered C-terminal region at residues 112-157 [48]. We implemented WARIO to extract a family of weighted contact motifs that elucidates the structural variability of DciA disordered domain. Data were drawn from a MD simulation, whose details can be found in [48, Section 2.6]. The retrieved conformational ensemble was refined by experimental SAXS data reported in [48], leading to a sub-ensemble containing $n = 1034$ conformations.

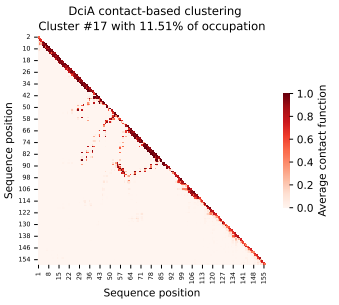
After setting to the 2% of n the minimum cluster size, WARIO retrieved a characterization (7.17) containing 18 cluster-specific ω-contact maps (see Appendix E.2 for the two-dimensional UMAP projection). A prominent cluster in terms of occupancy was found, containing the 38.59% of conformations, followed by the second most occupied class with the 11.51% of conformations. Their cluster-specific ω-contact maps are presented in Figure 7.13, together with an illustration of then randomly selected conformations. As we found in Section 7.3.1 for CHCHD4, the most occupied clusters do not present long-range



(a) Cluster-specific ω-contact map.



(b) Alignment at res. 10-40.



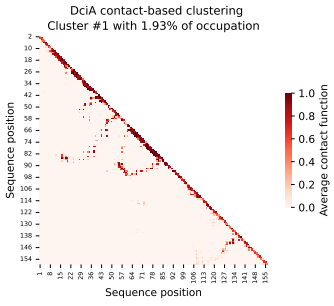
(c) Cluster-specific ω-contact map.

(d) Alignment at res. 5-40.

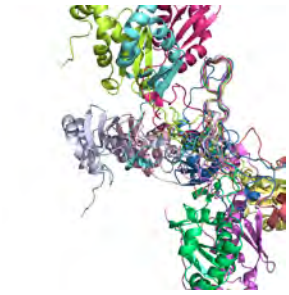
Figure 7.13: (a,c) Cluster-specific ω-contact maps (7.15) for the two most populated clusters of DciA. (b,d) 10 conformations randomly selected from each cluster (a,c respectively) and aligned at residues presenting helical motifs near the N-terminal.

contacts in the disordered domain. At residues 112-157, the only structural motif is an helix between residues 95 and 105 for cluster 17, which does not appear for cluster 9. Both clusters also differ at the starting residue for the N-terminal helix (see Figure 7.13(b,d)).

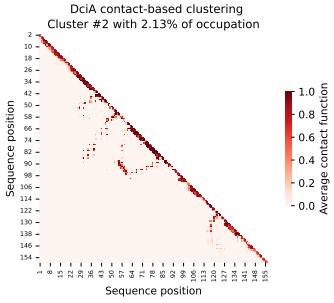
The remaining 16 clusters captured the off-diagonal contact patterns that appear along the structural evolution of DciA, as well as the formation of helical motifs in different segments of the disordered domain. Their occupation oscillates between the 1% and 3% of the total number of conformations. Four examples are depicted in Figure 7.14. Cluster 1 in panels (a,b) is mainly characterized by contacts between 127-134 and 138-146 residues. A similar motif appear in cluster 2 between residues 120-127 and 122-130, as shown in panels (c,d). From cluster 7 we may highlight the off-diagonal contact appearing between residues 100-105 and 128-130 (panels (e,f)). Finally, cluster 8 presents a disordered domain with helical residues between position 130 and 140, as illustrated in panels (g,h). The complete characterization (7.17) for DciA is included in Appendix E.2, together with the cluster-specific secondary structure propensities and average radii of gyration.



(a) Cluster-specific ω-contact map.

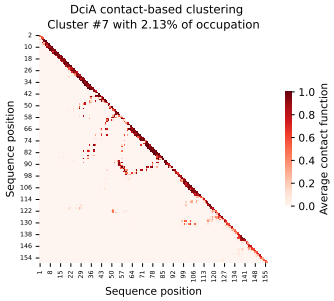


(b) Alignment at res. 127-146.

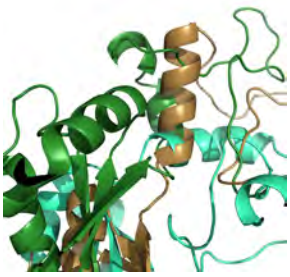


(c) Cluster-specific ω-contact map.

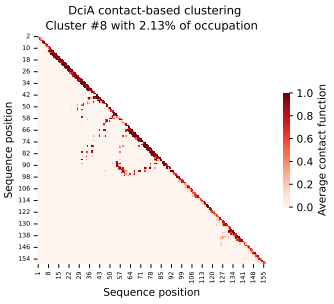
(d) Alignment at res. 120-130.



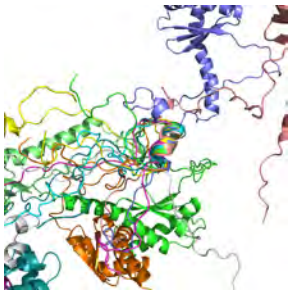
(e) Cluster-specific ω-contact map.



(f) Alignment at res. 100-105, 130-135.



(g) Cluster-specific ω-contact map.



(h) Alignment at res. 135-140.

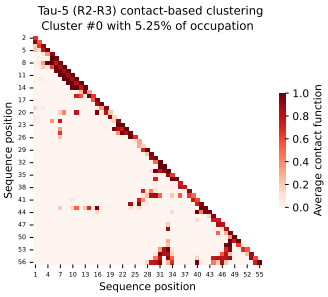
Figure 7.14: Left column: DciA cluster-specific ω-contact maps (7.15). Right column: 10 (3 for panel (f)) DciA conformations randomly selected from the cluster in the same row, left column, and aligned at residues exhibiting off-diagonal contacts.

7.3.4 Characterization of the Tau-5 domain of AR-NTD

For a last example, WARIO was applied to characterize the R2 and R3 partially helical regions in the transactivation unit 5 (Tau-5) domain (residues 350-448) of the disordered AR-NTD (Androgen Receptor N-terminal Transactivation Domain) [320]. We made use of a MD trajectory of the R2 and R3 regions of Tau-5 in its apo form, from now on denoted by Tau-$5_{\mathrm{R2-R3}}$, extracted from [319] and containing $n = 57144$ conformations. The sequence length was $L = 56$ residues. After setting the minimum cluster size to the 0.5% of the total number of conformations, WARIO characterized the ensemble with 63 cluster-specific weighted ω-contact maps (see Appendix E.2 for the two-dimensional UMAP projection). Three of them had significantly higher occupancies than the rest, containing the 8.44%, 5.25% and 3.9% of conformations (clusters 62, 0 and 61 respectively). Let us first take a look at clusters 0 and 61, whose specific contact patterns are presented in Figure 7.15. In contrast to the behavior that we observed for widely populated clusters in Sections 7.3.1 and 7.3.3, the second most occupied group of conformations presents a very specific pattern of long-range contacts (panel (a) in Figure 7.15). Moreover, the C-terminal region at residues 25-56 presents a very precise alignment across the cluster conformations (see Figure 7.15(b)). Cluster 61 is characterized by the lack of long-range contacts and the presence of helical structure near the N-terminal (see panels (c,d) in Figure 7.15), as well as in some other sequence segments.

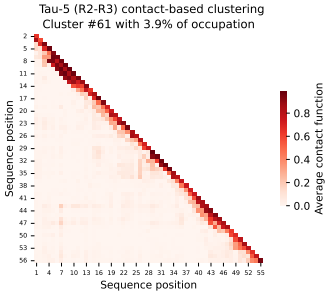
When looking at the most occupied cluster, we encounter a similar behaviour as the one for the 11-th cluster of Huntingtin (Figure 7.12). Indeed, the 62-th cluster-specific ω-contact map (panel (a) in Figure 7.16) does not present long-range contacts but short helical structures at some sequence segments. However, we can make out the contact pattern of a β-sheet near the C-terminal, with low contact function values. By repeating the same strategy as in Section 7.3.2, and re-implementing the clustering algorithm setting to the 0.1% of n the minimum cluster size, a subset of the 62-th cluster is retrieved whose main characteristic is the presence of a β-sheet structure at the end of the sequence. This is illustrated in Figure 7.16(b,c). This shows again the pertinence of re-adjusting the calibration parameter when looking for very specific and low frequent structures is of interest.

Some examples of the remaining ω-contact maps for Tau-$5_{\mathrm{R2-R3}}$ are presented in Figure 7.17. The first overall conclusion that we can extract is that the ensemble characterization (7.17) for Tau-$5_{\mathrm{R2-R3}}$ is mostly determined by the contact patterns in the C-terminal region at residues 25-56. Indeed, long-range residue-residue interactions present very specific motifs across clusters that yield adequate alignments for residues at such domain (see panels (b,d,f) in Figure 7.17). Note that, specially for clusters 6, 22 and 45, the protein tends to fold at the C-terminal, presenting a β-sheet alike contact pattern. Note first that the residues at which that motif appears are not the same as the ones in Figure 7.15(b,c), where the pattern is slightly relocated some residues up along the diagonal. once again illustrates the idea that the β-sheets in Figure 7.17(b,c) are exclusive to the corresponding group of conformations and thus that WARIO distinguishes slightly different motifs with

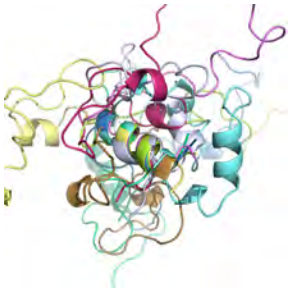


(a) Cluster-specific ω-contact map.

(b) Alignment at res. 25-56.

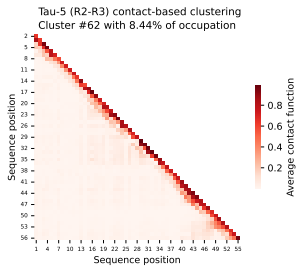


(c) Cluster-specific ω-contact map.

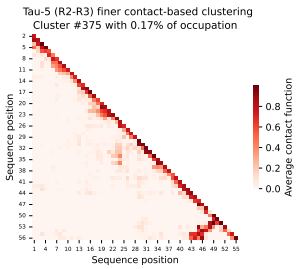


(d) Alignment at res. 5-15.

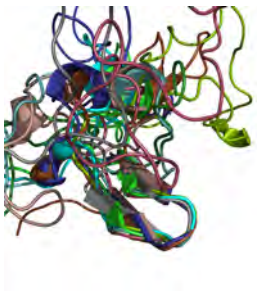
Figure 7.15: (a,c) Cluster-specific ω-contact maps (7.15) for the second and third most populated clusters of Tau-5_{R2-R3}. (b,d) 10 conformations randomly selected from each cluster (a,c respectively) and aligned at residues presenting off-diagonal contacts.



(a)



(b)



(c)

Figure 7.16: (a, resp. b) Cluster-specific ω-contact map for the 62-th (resp. 375-th) cluster of Tau-5_{R2-R3} characterization after setting to the 0.5% (resp. 0.1%) of n the minimum cluster size. (c) Ten conformations randomly selected from the 375-th cluster and aligned at the extended sheet structure.

high precision. Besides, there is a difference in terms of secondary structure when comparing the C-terminal patterns in Figures 7.15(c) and 7.17(a,c,g). See, for instance, the DSSP secondary structure average propensities for the 375-th cluster (Figure 7.16(b,c), for the finer minimum cluster size) and the 45-th cluster in Figure 7.17(g,h). They are presented in Figure 7.18. Note that the DSSP classification stands out extended sheet structure (E) at residues 46-50, 53,56 in cluster 375 (panel (a)) that are not present in cluster 45 (panel (b)). Consequently, despite the visual similarity of the C-terminal region in the ω-contact maps of Figures 7.16(b) and 7.17(g), the partition made by WARIO distinguishes contacts that entail the formation of secondary structure from those who do not. The complete characterization (7.17) for Tau-$5_{\mathrm{R2-R3}}$ is presented in Appendix E.2.

7.4 Methodological meta-analysis of WARIO

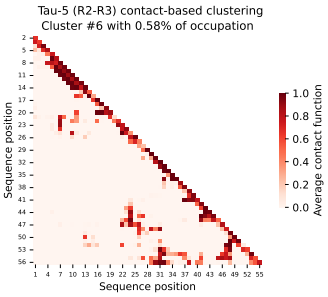
This section analyzes the performance of WARIO from a methodological perspective, evaluating the relevance of the choices made in Section 7.2 and comparing the suitability of the method with respect to other existing approaches in the literature. First, in Section 7.4.1, we compare WARIO with the approach presented in [59], where data is featured by all the Euclidean distances between residue pairs. Then, in Section 7.4.2, we illustrate how relaxing the definition of contact and adapting it to the sequence context of the interacting amino acids has a significant impact on the ensemble characterization.

7.4.1 Comparison with distance-based methods

The classification strategy that combines a dimensionality reduction technique with a clustering algorithm has already been used in previous works to partition ensembles of flexible proteins [8, 59]. One of the most commonly used descriptors to feature conformations is the set of all residue-residue Euclidean distances. For instance, it has been used in [59] to find representative families of conformations. Here, we implement the UMAP+HDBSCAN pipeline on the data featured with pairwise Euclidean distances between all C_β atoms (C_α for glycines) to characterize CHCHD4 and compare the output with the results presented in Section 7.3.1.

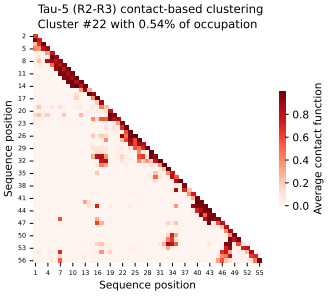
The clustering pipeline on the distance dataset retrieved 10 clusters, among which one contained the 67% of conformations. Recall that WARIO retrieved 23 clusters for CHCHD4 where the two most occupied clusters contained the 20% and 16% of states. In Figure 7.19 we present the average distance maps for the four most occupied distance clusters, together with 30 random conformations drawn from each one and aligned at all residues.

Even if the classification technique in [59] is basically identical to the one presented here, it is clear that the partition of the set of conformations will strongly depend on how states are described. If the algorithm is provided with all the pairwise distances, clusters



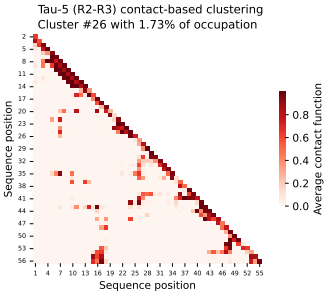
(a) Cluster-specific ω-contact map.

(b) Alignment at res. 25-56.



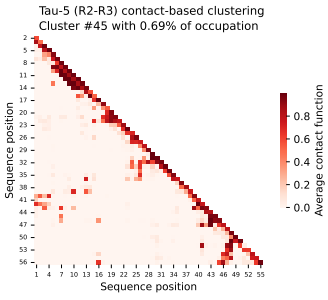
(c) Cluster-specific ω-contact map.

(d) Alignment at res. 25-56.

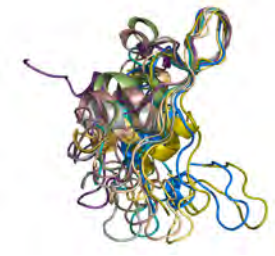


(e) Cluster-specific ω-contact map.

(f) Alignment at res. 25-56.

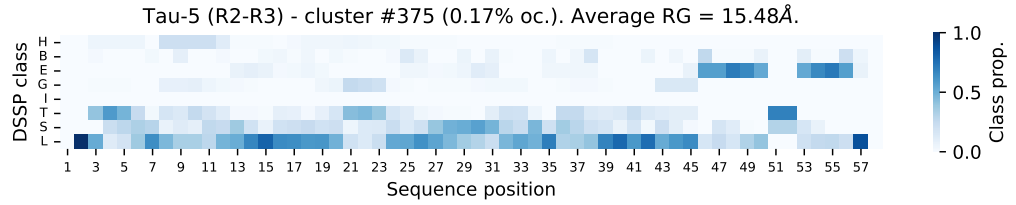


(g) Cluster-specific ω-contact map.

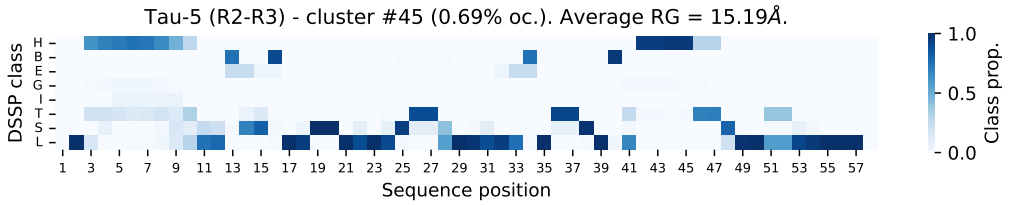


(h) Alignment at res. 25-35.

Figure 7.17: Left column: Tau-$5_{\mathrm{R2-R3}}$ cluster-specific ω-contact maps (7.15). Right column: 10 Tau-$5_{\mathrm{R2-R3}}$ conformations randomly selected from the cluster in the same row, left column, and aligned at residues exhibiting off-diagonal contacts.



(a)



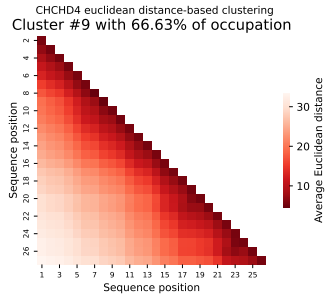
(b)

Figure 7.18: Average DSSP secondary structure propensities and average radius of gyration across cluster conformations. In (a), cluster 375 after setting to the 0.1% of n the minimum cluster size. In (b), cluster 45 after setting to the 0.5% of n the minimum cluster size.

will tend to group conformations having similar global structures or, in other words, to match conformations with good alignments. We believe that this strategy is ill-suited to the context of highly-flexible proteins, as the complex variability of the conformations might intrinsically contradict the fact that structures can be globally classified into well-defined groups with aligned conformations. This is reflected in Figure 7.19, where we can see how groups are rather constructed to put together conformations that align well, but the classification remains broad when compared to the structural diversity found in Section 7.3.1, where the target was not the global structural alignment but the detection of common contact patterns manifesting locally. Consequently, we believe that the use of pairwise distances as conformational features might be suitable to classify states when the objective is to find well aligned groups of structures, which is not really relevant for highly-flexible proteins. However, contact information needs to be taken into account when trying to disentangle the whole structural variability of the protein through the detection of all local interaction patterns that appear with low frequencies, and that are missed when distances are considered.

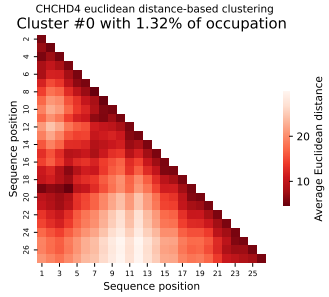
7.4.2 The importance of refining contact definition

We assessed whether the effort made in Section 7.2 to define contact as a continuous function that integrates sequence and geometrical information is worth it to characterize ensembles. To do so, we kept the same strategy of characterizing an ensemble by a



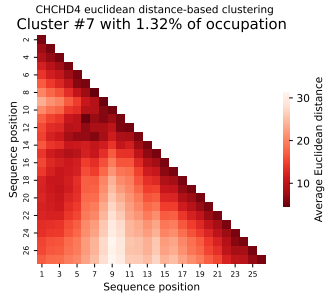
(a) Average distance map

(b) Alignment at all res.



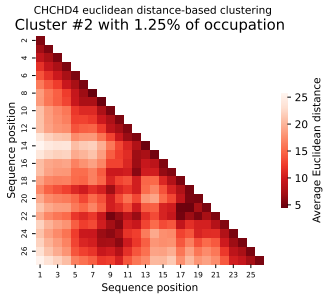
(c) Average distance map

(d) Alignment at all res.



(e) Average distance map

(f) Alignment at all res.



(g) Average distance map

(h) Alignment at all res.

Figure 7.19: Left column: CHCHD4 cluster-specific average distance maps after implementing the UMAP+clustering pipeline to the set of all Euclidean residue-residue distances. Right column: 30 CHCHD4 conformations randomly selected from the cluster in the same row, left column, and aligned at all residues.

weighted family of contact maps, but starting from the classical contact definition i.e. by considering the matrix

$$\mathbf{C} = \begin{pmatrix} c_{11;1} & c_{12;1} & \cdots & c_{ij;1} & \cdots & c_{L(L-1);1} \\ c_{11;2} & c_{12;2} & \cdots & c_{ij;2} & \cdots & c_{L(L-1);2} \\ \vdots & \vdots & & & & \vdots \\ c_{11;n} & c_{12;n} & \cdots & c_{ij;n} & \cdots & c_{L(L-1);n} \end{pmatrix}, \tag{7.19}$$

where $c_{ij,k} = \mathbb{1}\{d_{\mathbb{R}^3}(\mathcal{F}_i^k, \mathcal{F}_j^k) \leq 8\mathring{A}\}$ and $\mathcal{F}_i^k$ denotes the i-th reference frame (7.2) for the k-th conformation. As the entries of (7.19) are binary, we chose the Jaccard distance to project the data into the 10-dimensional UMAP space. Then, the clustering was performed using the Euclidean distance between points in the low-dimensional space. Of course, using the classical contact definition based on thresholds imposes the need of metrics that are well-defined for this type of data. The choice of such metric is not straightforward and neither is its suitability in the low-dimensional projection. Whether we can correctly compare points in the UMAP space with the Euclidean distance when the high-dimensional space is $\{0,1\}^p$ is not a trivial question to address. Moving to the continuous scenario removes these issues and ensures a less intrincate implementation of the entire clustering pipeline using exclusively the Euclidean distance between points.

We observed a significant disagreement between methods when looking at the number of classified conformations. When using (7.14), WARIO classified the 78% of conformations. This proportion decreased to 65% when using (7.19). The number of retrieved clusters was almost the same as in Section 7.3.1, where WARIO found 23 classes versus the 22 retrieved here, using the same value for the minimum cluster size: the 1% of n. When looking at the cluster-specific contact maps, we find similar contact trends when comparing both approaches. We can identify groups of conformations similarly classified with both approaches by detecting visually matching contact maps. Three examples are presented in Figure 7.20. This was expected as the definition proposed in Section 7.2 is a refinement of the classical one, and no extreme disagreements should appear.

However, remarkable differences appear when diving into short-range contacts, for which relative orientation played a role in the interaction distance (7.10). To illustrate this, we focus on the last row of Figure 7.20. Both contact maps seem to indicate the presence of helical motifs near the C-terminal. We already showed it in Figure 7.10(c,d) for the continuous contact definition. Conformations belonging to the corresponding cluster exhibit α-helix structure at residues 21-24, which is confirmed by the DSSP propensities presented in Figure 7.21(a). However, despite the visual similarity of panels (e) and (f) in Figure 7.20, we can appreciate that values for the continuous contact function (panel (e)) are slightly higher at the C-terminal than the ones for the binary definition (panel (f)). This means that residues 21-24 are *closer in interaction distance* (7.10) *than in Euclidean*

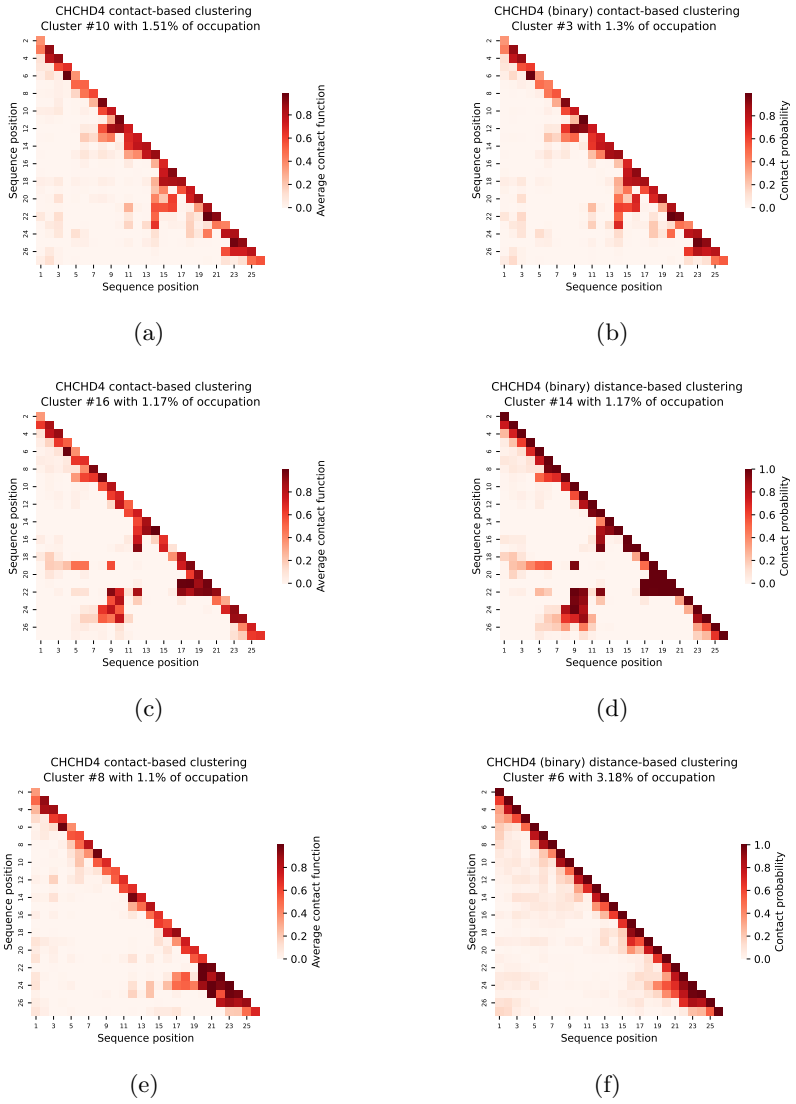
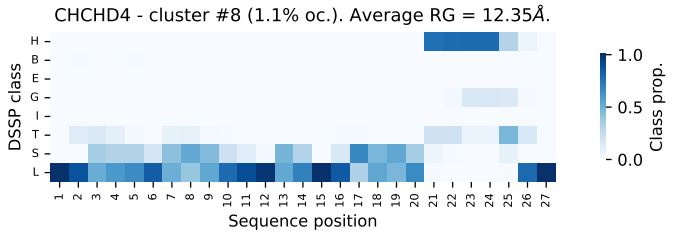


Figure 7.20: Left (resp. right) column: cluster-specific ω-contact maps (resp. average contact maps) for CHCHD4 after performing the UMAP+clustering pipeline on (7.14) (resp. (7.19)). Maps in the same row are those who visually match each other among both classification techniques.

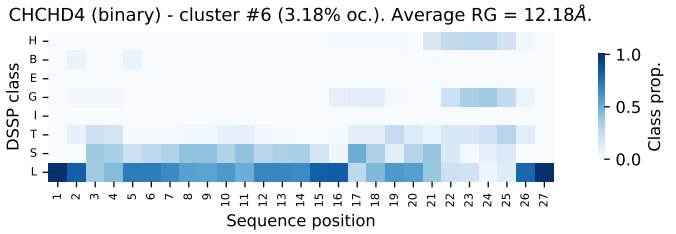
distance. In other words, taking relative orientation into account enhances contact identification when it is close to the preferred behavior observed in nature.

Indeed, the proportion of α-helix structures at 21-24 in cluster 6 for the binary contact clustering (see Figure 7.21(b)) is considerably smaller. This can be alternatively illustrated by looking at the conformations from such cluster, shown in Figure 7.22, which differ from the structured behavior depicted in Figure 7.10(d). Consequently, redefining contact as

a continuous function (7.11) that integrates sequence information and relative orientation is crucial to make the classification coherent in terms of secondary structure.



(a) UMAP+HDBSCAN on (7.14).



(b) UMAP+HDBSCAN on (7.19).

Figure 7.21: Average DSSP secondary structure propensities across cluster conformations after performing the UMAP+HDBSCAN pipeline on (7.14), for cluster 8 (a) and on (7.19), for cluster 6 (b), for the CHCHD4 ensemble.

Figure 7.22: 10 conformations randomly selected from cluster 6 after performing the UMAP+HDBSCAN pipeline on (7.19) for CHCHD4, aligned at residues 21-24.

7.5 Discussion

The method presented here provides a compact and meaningful characterization of an ensemble through a weighted family of contact maps. The idea of using a graph-based characterization built from contact information to characterize ensembles was already presented in [57]. However, due to the high structural variability of highly flexible proteins, there is a need to previously cluster the contact distribution to correctly account for the complexity of the conformational space. Rather than implementing average-based approaches, WARIO first unravels the most determinant interaction patterns that characterize the ensemble and then represent them as easily interpretable cluster-specific contact maps, to which we can add a weight accounting for their frequency in the protein dynamics. This is done after refining the contact definition and getting rid of the threshold-based setting that has been commonly used. We have illustrated the importance of adapting these thresholds to the sequence context and make them dependent on the relative orientation of the pair of residues. This last point is essential to correctly account for local structural motifs that appear at the short-range level. We believe that the usefulness of incorporating orientation to capture long-range contacts is less evident, due to their questionable interest regarding the structural analysis of IDP. Besides, we are not able to empirically determine whether preferred orientation settings exist in that context. The suitability of WARIO to detect how residue-residue interactions distribute across conformations has been illustrated for four ensembles of disordered proteins. This is possible thanks to the use of contact information to feature conformations, while avoiding global descriptors as pairwise distances. WARIO has been implemented through an easy-to-use Jupyter notebook, which has been made available to the community.

The proposed ensemble characterization (7.17) is clearly defined and easily interpretable. Nevertheless, it strongly depends on the minimum cluster size that is given as input to HDBSCAN. The output dependence on hyper-parameters is an intrinsic characteristic of every clustering algorithm and cannot be avoided in any case. The best thing we can do is to provide interpretability to these parameters and ensure that their variation has a clear meaning within the biological context. In our case, the minimum cluster size is easily interpretable as the resolution of the ensemble characterization (7.17). The smaller the size, the finer the classification will be and less frequent contact patterns will be detected, as illustrated in Figures 7.12 and 7.16. However, too low resolutions will result in redundant group classifications that are more difficult to interpret. The choice of the clustering resolution should be made based on the practitioner's needs, and its potential readjustment can be studied in each case. It is important to emphasize that there is no “true number of clusters" in the ensemble, and every classification algorithm aims at easily representing the diversity of the system states rather than revealing an inherent population grouping that does not actually exist.

An effective solution to deal with the dependence on the minimum cluster size would be to incorporate statistical techniques that provide evidence of differences between the

encountered clusters. In other words, procuring evidence of whether the resolution is too high and several clusters can be merged together into a larger one, or vice versa. This problem is an actively growing field of study in selective inference [97] and is referred to as post-clustering inference. However, these methods are highly dependent on the type of algorithm that is used and on the interdependence of the observations and descriptors employed. Despite remarkable recent advances [104, 51], their application to evaluate the output of WARIO remains a distant prospect for now.

The applicability of WARIO can be directly extended to the study of protein complexes and protein-protein interactions between systems with varying levels of disorder. Note that, however, WARIO operates in all-model representation of the protein backbone. This is essential for the definition of the residue-specific reference systems and, therefore, for the integration of relative orientation to the contact functions. A potential avenue for future work would be the adaptation of WARIO to coarse-grain models and the non-trivial assessment of relative orientations in that context.

Software availability

WARIO has been implemented as an easy-to-use Jupyter Notebook, available at https://gitlab.laas.fr/moma/WARIO.

Acknowledgements

We are grateful to Tâp Ha-Duong for providing useful data and for his helpful discussions and valuable feedback.

This work was supported by the French National Research Agency (ANR) under grant ANR-11-LABX-0040 (LabEx CIMI) within the French State Programme “Investissements d’Avenir” and under grant ANR-22-CE45-0003 (CORNFLEX project).

Appendix C

Appendix of Chapter 5

Contents

C.1 Methodology details

C.1.1 Building a residue-specific reference frame

Reference frame definition

We seek to define a reference frame that determines the global pose (position and orientation) of a given residue and that allows to describe the relative pose of other residues along the sequence. As we want this reference system to be universally defined (independently of the residue identity), we first define a virtual atom $\widetilde{C_\beta}$, which exists also for glycines. The position of $\widetilde{C_\beta}$ is an estimate of the position of the true C_β when it exists, but it is defined for every residue using only the atoms that are always present. Its definition allows the construction of a universal frame that locally represents the geometry of the backbone.

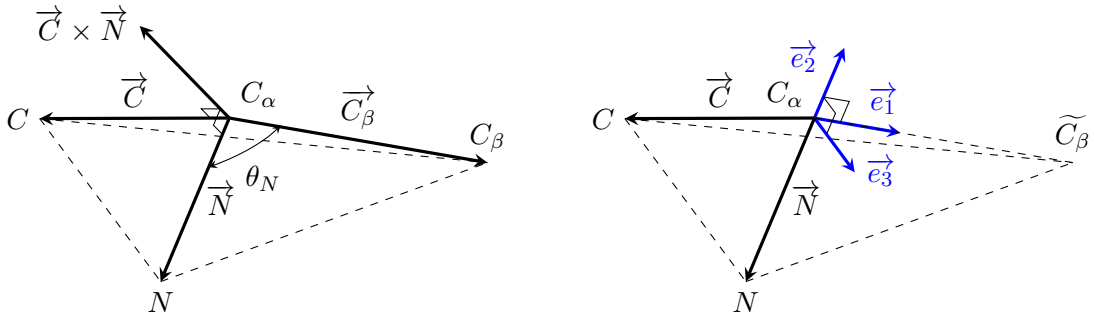


Figure C.1: (a) Illustration of vectors and angles involved in the construction of the residue-specific reference frame. The vector $\overrightarrow{C_\beta}$ can be determined from vectors $\overrightarrow{C}$ and $\overrightarrow{N}$ together with the angles θ_N (the only depicted for simplicity), θ_C and θ_{CN}. (b) The three vectors $\{\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}\}$ defining the reference frame, built from the virtual atom $\widetilde{C_\beta}$ and vectors $\overrightarrow{C}$ and $\overrightarrow{N}$.

Let $\overrightarrow{C}$ and $\overrightarrow{N}$ be the vectors going from C_α to C and N atoms, respectively. If a C_β atom is present, let $\overrightarrow{C_\beta}$ denote the vector going from C_α to C_β. In such case, $\overrightarrow{C_\beta}$ can be determined using the vectors $\overrightarrow{C}, \overrightarrow{N}$ and $\overrightarrow{C} \times \overrightarrow{N}$ together with their angles with respect to $\overrightarrow{C_\beta}$, denoted θ_C, θ_N and θ_{CN} respectively. See Figure C.1a for an illustration. This can be done by solving the following linear system, whose unknown variables are the three coordinates of C_β.

$$\begin{cases} \|\overrightarrow{N}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_N = \overrightarrow{N} \cdot \overrightarrow{C_\beta} \\ \|\overrightarrow{C}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_C = \overrightarrow{C} \cdot \overrightarrow{C_\beta} \\ \|\overrightarrow{C} \times \overrightarrow{N}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_{CN} = (\overrightarrow{C} \times \overrightarrow{N}) \cdot \overrightarrow{C_\beta}. \end{cases} \tag{C.1}$$

To define a *universal* C_β, denoted $\widetilde{C_\beta}$, we will estimate fixed values for θ_N, θ_C and θ_{CN} from all non-glycine residues of a set of protein structures and *define* the $\widetilde{C_\beta}$ coordinates as the solution of (C.1), independently of the residue identity. Details on angles estimation are given in the following section. Consequently, for a given residue, the virtual atom $\widetilde{C_\beta}$ is determined from the coordinates of its C_α, N and C atoms. This allow us to define a reference system at each sequence position through the following three vectors, where $\overrightarrow{CN} = \overrightarrow{N} - \overrightarrow{C}$.

$$\begin{cases} \overrightarrow{e_1} = \overrightarrow{\widetilde{C}_\beta} / \|\overrightarrow{\widetilde{C}_\beta}\| \\ \overrightarrow{e_2} = \overrightarrow{CN} / \|\overrightarrow{CN}\| \times \overrightarrow{e_1} \\ \overrightarrow{e_3} = \overrightarrow{e_1} \times \overrightarrow{e_2}. \end{cases} \tag{C.2}$$

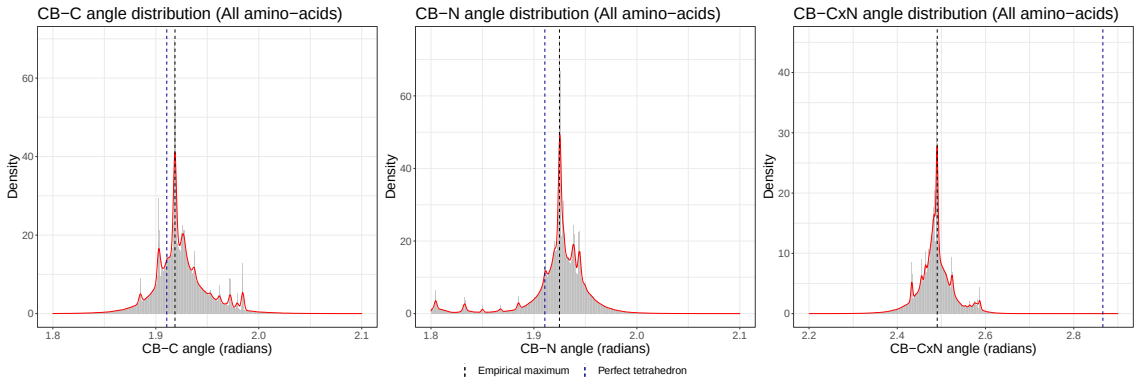


Figure C.2: From left to right: empirical distributions of θ_C, θ_N and θ_{CN} respectively, extracted from a set of 15177 protein structures, considering all non-glycine residues. The red line corresponds to a kernel density estimate, whose maximum (vertical black dashed line) was used as angle estimate. The blue dashed line depicts the theoretical value of each angle under the hypothesis that the four atoms bound to the C_α form a regular tetrahedron.

Once the reference system of the i-th residue, denoted $\mathcal{F}_i = \{\overrightarrow{e_{1,i}}, \overrightarrow{e_{2,i}}, \overrightarrow{e_{3,i}}\}$, has been built, its origin will be placed at the C_β atom when it exists, or at the C_α otherwise. This allows the computation of relative positions and distances with respect to C_β atoms for all non-glycine residues.

Estimation of θ_C, θ_N and θ_{CN}

We estimated three fixed values for θ_C, θ_N and θ_{CN}, to be replaced in the linear system (C.1). After that, the vector $\overrightarrow{C_\beta}$ is determined for each residue along the sequence by solving (C.1) after plugging in the corresponding coordinates of C_α, C and N atoms. As mentioned in the previous section, this allows the definition of a residue-specific reference frame, built independently of the residue identity.

To estimate the three angles, we used a set of 15177 experimentally-determined high-resolution structures of protein domains extracted from the SCOPe 2.07 release [49]. For each structure, θ_C, θ_N and θ_{CN} were computed and stored for every non-glycine residue. The three corresponding histograms, together with a kernel density estimate, are presented in Figure C.2, for all residue types. The residue-specific counterparts of Figure C.2 did not show important fluctuations from the overall densities. Therefore, for simplicity, we did not estimate three angles per residue type, but three universal values.

The three distributions of Figure C.2 show that all the angle distributions are strongly concentrated around their kernel density maximum. Consequently, these values were chosen as an estimate of θ_C, θ_N and θ_{CN}. Due to the symmetry of the empirical distributions,

choosing the mean would provide similar estimates. Figure C.2 depicts the theoretical angle values under the hypothesis that C, N, C_β and H (when present) are the vertices of a regular tetrahedron, with C_α as its centroid. One could think of using these values as estimates, but the deviation from the experimental value of θ_{CN} is too high, showing how the fluctuations from the regular polyhedron are not homogeneous along its faces.

C.1.2 Wasserstein distance: practical implementation

The Wasserstein distance can be easily computed from a pair of samples drawn from the corresponding probability distributions. However, a major drawback of the algorithms that compute the Wasserstein distance is their inability to handle large datasets ($\gtrsim 10^3$ points). The current implementations in Python [98] or R [259] only admit datasets with $\lesssim 5 \cdot 10^3$ points, which is usually not enough for conformational ensembles of IDPs. To the best of our knowledge, there are no existing algorithms that solve an OT problem for large sample sizes and that are easily implementable, considerably fast (which, in our case, is essential due to the large number of Wasserstein distances to compute), and that accept non-euclidean ground distances (like the distance in the torus).

Here, we propose an approximation method to "simplify" the input empirical distributions and compute the Wasserstein distance from a pair of smaller samples sizes. The efficiency of this approach in terms of error is illustrated via simulations on real protein data, but we provide no theoretical bounds. The proposed algorithm consists in clustering the original distribution and defining its clustered version as a discrete probability distribution supported on the set of clusters whose mass is given by the proportion of points assigned to each cluster. Then, the Wasserstein distance is computed between the pair of clustered distributions, whose samples have admissible sizes. The method is implemented for both local and global structural descriptors, which are empirical probability distributions supported on $\mathbb{T}^2$ and $\mathbb{R}^3$ respectively.

The accuracy in terms of relative and mean-square error is presented in Figure C.3. Note that the approximation algorithm has a considerably better performance when implemented for local structural descriptors, which was expected due to the boundedness of the corresponding ground space. Accuracy in $\mathbb{R}^3$ is slightly worse, as cloud points representing the relative position of residues are in general more disperse, and therefore the clustered distribution needs a larger number of centroids to better capture its variability. Nevertheless, we observe that, in both cases, the error estimates for a proportion of $\sim 10\%$ of clusters with respect to the entire dataset size (the proportion we will be using in practice) are acceptable for our practical purposes. To enrich the interpretation, we performed the same accuracy analysis but by computing the Wasserstein distance between subsamples drawn uniformly from the corresponding datasets. As shown in Figure C.3, the effect of clustering significantly improves the quality of the approximation.

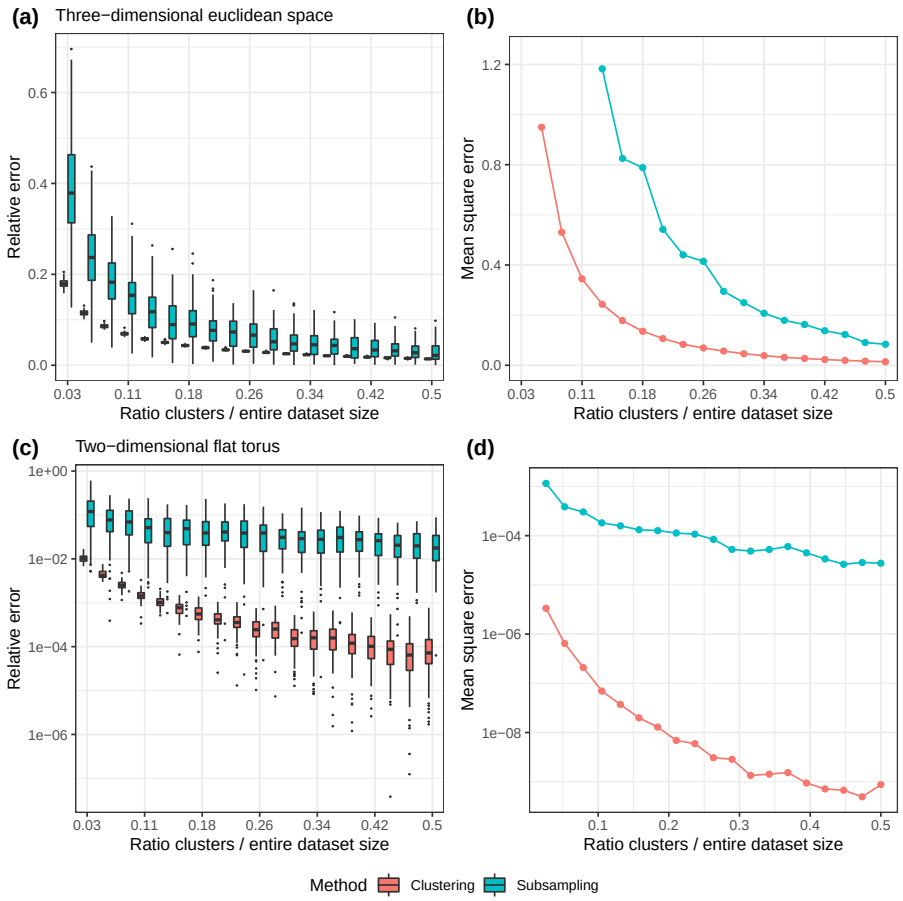


Figure C.3: From left to right (columns): relative and mean square error estimates of the Wasserstein distance between the clustered distribution as an estimate of the Wasserstein distance between the original datasets. In abscissas, the proportion of the number of clusters with respect to the entire dataset size. The first row (a,b) corresponds to samples drawn from local structural descriptors (dihedral angles) and the second (c,d) to samples drawn from global structural descriptors (pairwise relative positions of residues).

C.1.3 The matrix representation

The result of the comparison analysis is represented through a matrix, $\mathbf{W}$. We will denote by $\mathcal{W}_{ij}$ the entries of $\mathbf{W}$, where $i, j \in \{1, \ldots, L\}$. The matrix will be lower triangular (i.e. $\mathcal{W}_{ij} = 0$ if $j > i$). Figure C.4 illustrates the main elements of the matrix representation, which are described below.

1. The matrix is headed by a title describing the comparison, introduced by the user.

2,3. Below the title, the overall local and global discrepancies are depicted (equations (5.11) and (5.12) respectively). By default, they are computed by aggregating and

weighting the corrected distances as described in Section 5.2.4. These features can be modified by the user.

4,5. The matrix entries are represented using two independent color scales, for local and global differences. Both scales correspond to the score (5.10), which can be computed when several independent replicas of each ensemble are available. Otherwise, distances cannot be corrected by uncertainty and the scale will correspond to the (non-corrected) inter-ensemble local and global distances (equations (5.5) and (5.7) respectively).

6. The entries $\mathcal{W}_{ij}$ for $i < j$ correspond to the scores (5.10) computed for the i, j-th global structural descriptors, i.e. the score comparing the relative position distribution of the i-th and j-th residues in the two ensembles. If no independent replicas are available, the entry corresponds to the i, j-th global distance in (5.7).

7. The entries $\mathcal{W}_{ii}$ correspond to the scores (5.10) computed for the i-th local structural descriptors, i.e. the score comparing the (ϕ, ψ) distribution of the i-th residue in the two ensembles. If no independent replicas are available, the entry corresponds to the i-th local distance in (5.5).

8. The entries $\mathcal{W}_{ii}$ are marked with a star if their associated p-value (5.6) is less than the significance level $\alpha = 0.05$.

9. The axes labels correspond to the residue position, counting from the N-terminal, relative to the sequence segment that is being compared (and not to the absolute position in the entire sequence).

C.2 Additional results

C.2.1 Comparison of PEP3 ensembles produced by MD simulations using different force-fields

We replicated the analysis described in Section 5.3.1 for MD simulations of PEP3 using the same force-fields. Results are presented in Figure C.5. Here, the discrimination between the two force-field families is not observed. Nonetheless, we still observe that structures simulated with disp and ildn are very close in Wasserstein distance (Figure C.5b). Indeed, the overall global dissimilarity is substantially smaller than these of the remaining comparisons. Only inter-ensemble corrected differences representing about 20% of the intra-ensemble ones appear for residues at the C-terminus. The distances between c36idp and c36m are now higher than for Hst5, and corrected differences of the same magnitude than the intra-ensemble ones appear in the interior of the matrix. The same behavior is observed when comparing force-fields of different groups for PEP3. See, for instance, that

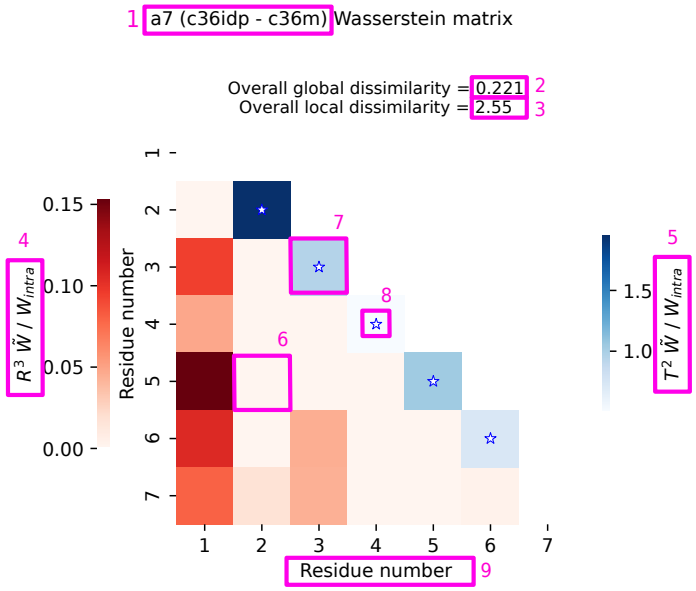


Figure C.4: Schematic representation of the output of WASCO. All the elements marked with numbers are described in Section C.1.3.

substantial differences arise between relative positions of residues at opposite terminus in panel (d), which are highly weighted when computing the overall global discrepancy. One intriguing observation is that while there are substantial differences between disp and ildn (and between c36idp and c36m), simulations with c36idp and c36m used the same water model (the CHARMM-modified TIP3P water model) and the disp and ildn simulations also used very similar water models (TIP4P-D and a slightly variant of this) [140]. Overall, these results are complementary to those presented in [140], which mainly focused on secondary structure differences among ensembles, and they show the ability of WASCO to identify differences at both local and global scales.

C.2.2 Assessment of the convergence of MD simulations

Ensemble comparisons have previously been used to assess convergence in MD simulations of folded proteins [120, 279, 192]. We here propose to use the overall ensemble distances (5.11), (5.12) to examine the convergence of an MD simulation of a disordered protein. Moreover, this can be done on-the-fly to assess whether the simulation can be stopped. Let T denote the current simulation time and let $0 < t_1 < t_2 < \cdots < t_k = T$ be k time points. If we denote by A_t the conformational ensemble simulated up to time t, we can

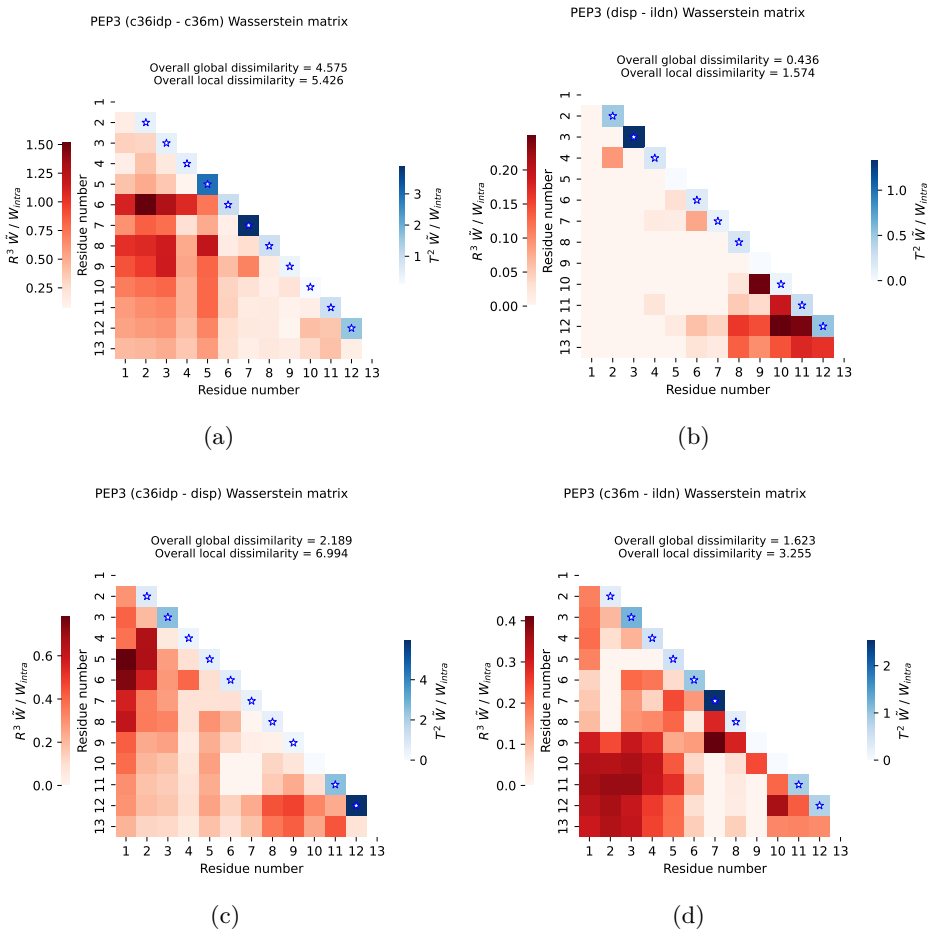


Figure C.5: Comparison of Molecular Dynamics simulations of PEP3 ensemble using different force fields. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (5.10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower-triangle (in red) correspond to the global differences. The coefficients along the diagonal (in blue) correspond to the local differences. Blue stars indicate that the corresponding local corrected distance is significantly different from zero (the associated p-value (5.6) is smaller than $\alpha = 0.05$).

compute the *online* overall distances

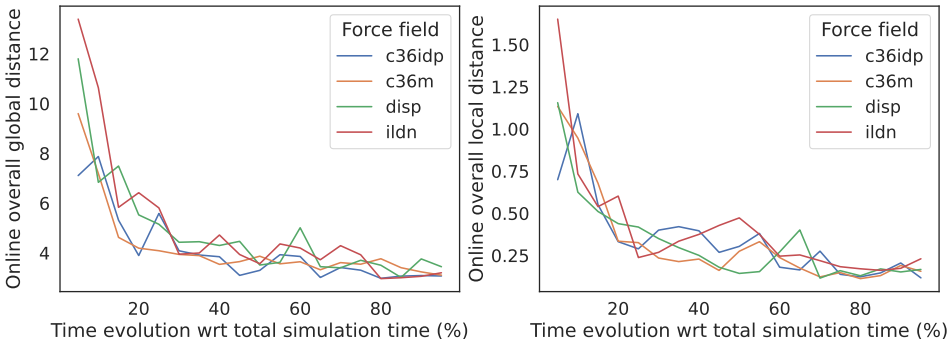
$$\mathcal{OW}_i^l = \mathcal{OW}^{l,A_{t_{i-1}},A_{t_i}}, \tag{C.3}$$

defined in (11) of the main text, for all $i = 2, \ldots, k$. Analogously, we compute the *online* overall global distances

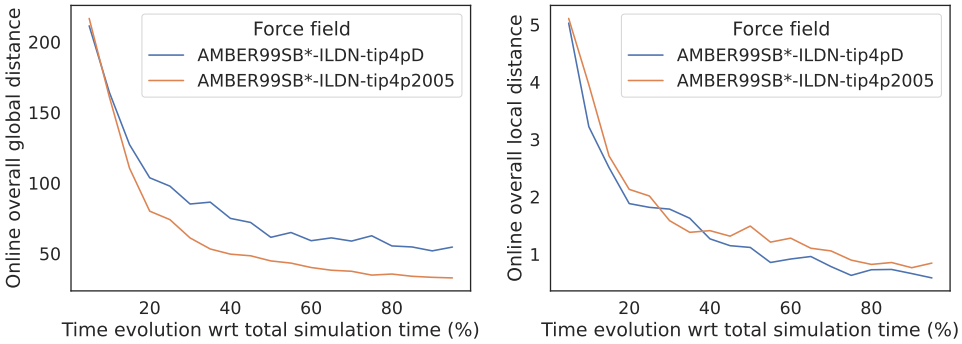
$$\mathcal{OW}_i^g = \mathcal{OW}^{g,A_{t_{i-1}},A_{t_i}}, \tag{C.4}$$

as defined in (12) of the main text.

For each i, $\mathcal{OW}_i^l$ (resp. $\mathcal{OW}_i^g$) corresponds to the overall local (resp. global) distance between the ensemble from $t = 0$ to $t = t_i$ and the ensemble from $t = 0$ to $t = t_{i-1}$. In other words, (C.3) (resp. (C.4)) is the distance between the ensembles simulated up to time t_{i-i} and up to time t_i. Consequently, it quantifies whether the new simulated trajectories between t_{i-1} and t_i yielded a non-negligible contribution to the ensemble structure (if (C.3) is not small) or, otherwise, whether proceeding the simulation up to t_i does not yield any substantial contribution (if (C.3) is close to zero). Then, the representation of



(a) PEP3



(b) K-18 domain of Tau

Figure C.6: (a) Online convergence analysis for PEP3 ensemble simulated with force-fields c36idp, c36m, disp and ildn. (b) Online convergence analysis for K-18 domain of Tau ensemble simulated with AMBER ff99SB*-ILDN and TIP4P-D water model. In abscissas, the percentage of simulation time, divided in 20 equally spaced time intervals. In ordinates, the overall distances between the ensembles simulated at the extremes of the time intervals. The left (resp. right) column presents the evolution of $\mathcal{OW}_i^g$ (resp. $\mathcal{OW}_i^l$) with respect to time.

$\mathcal{OW}_i^l$, $\mathcal{OW}_i^g$ with respect to the t_i indicates whether the simulation has converged or not. Note that the distances $\mathcal{OW}_i^l$, $\mathcal{OW}_i^g$ can never be equal to zero, as they are empirical distances which *converge* to zero when the sample size tends to infinity. Therefore, the profiles will approach a non-zero plateau under convergence, whose ordinate will decrease when sample size increases. The criteria to assume convergence will be therefore the reach of such a plateau at a *reasonable* ordinate, meaning that it must be small enough if sample sizes are considerably large. Nevertheless, this criteria provides a stronger evidence of non-convergence, as the achievement of an asymptote for (C.4), even if necessary, may not be sufficient to guarantee convergence. If we resolve that the simulation must keep going until time $T' > T$, it suffices to add $\mathcal{OW}^{l,A_T,A_{T'}}$ and $\mathcal{OW}^{g,A_T,A_{T'}}$ to each curve and recheck.

Figure C.6a presents the evolution of the online overall distances for PEP3 simulated with the four force-fields introduced in Section 5.3.1. We observe that all the curves exhibit an asymptote at a value close to zero after 80% of simulation time, which is compatible with convergence in all cases. This is not the case for the simulation in Figure C.6b, corresponding to a 1,000 ns simulation of the K-18 domain of Tau using the AMBER ff99SB*-ILDN force-field and the TIP4P-D water model (Sthitadhi Maiti and Matthias Heyden, unpublished). Here, we clearly observe that curves do not reach an asymptote and present a decreasing behavior during all the time evolution. This result was expected due to the length of the protein (129 amino acids) and the reduced simulated time.

C.2.3 Comparison of ensembles using distance matrices

As it is discussed in Section 5.1, the use of average descriptors to compare IDP ensembles may yield a substantial loss of information when the underlying distributions describing their structure exhibit a high and complex variability. The work presented in [167] computes the median C_α-C_α distance for every pair of residues $i < j$, denoted $\bar{d}_{ij}$, as well as its corresponding standard deviation, denoted σ_{ij}. If $\bar{d}_{ij}^A$, σ_{ij}^A (resp. $\bar{d}_{ij}^B$, σ_{ij}^B) denote the previously defined descriptors for ensemble A (resp. B), the difference between both ensembles is given by a matrix with entries M_{ij}, where

$$M_{ij} = \begin{cases} \Delta\bar{d}_{ij} = |\bar{d}_{ij}^A - \bar{d}_{ij}^B| & \text{if} \quad i < j, \\ \Delta\sigma_{ij} = |\sigma_{ij}^A - \sigma_{ij}^B| & \text{if} \quad j > i, \\ 0 & \text{otherwise.} \end{cases} \tag{C.5}$$

In [167], the entries M_{ij} are neglected if they are not significantly different from zero (according to a Mann-Whitney-Wilcoxon test for the distance distributions). Here, we skipped this step for simplicity. We computed the matrix with entries M_{ij} for the comparison analysis presented in Section 5.3.1, using one replica per ensemble. The counterpart of Figure 5.2 is depicted in Figure C.7. As could be anticipated, the conclusions stated in Section 5.3.1 are difficult to extract from the matrices in Figure C.7. First, the overall behaviour between force-fields suggested by Figure 5.2 is not observed in the distance matrices, as the corresponding color scales do not present significant discrepancies in the distance magnitudes between comparisons (see, on the contrary, the differences between

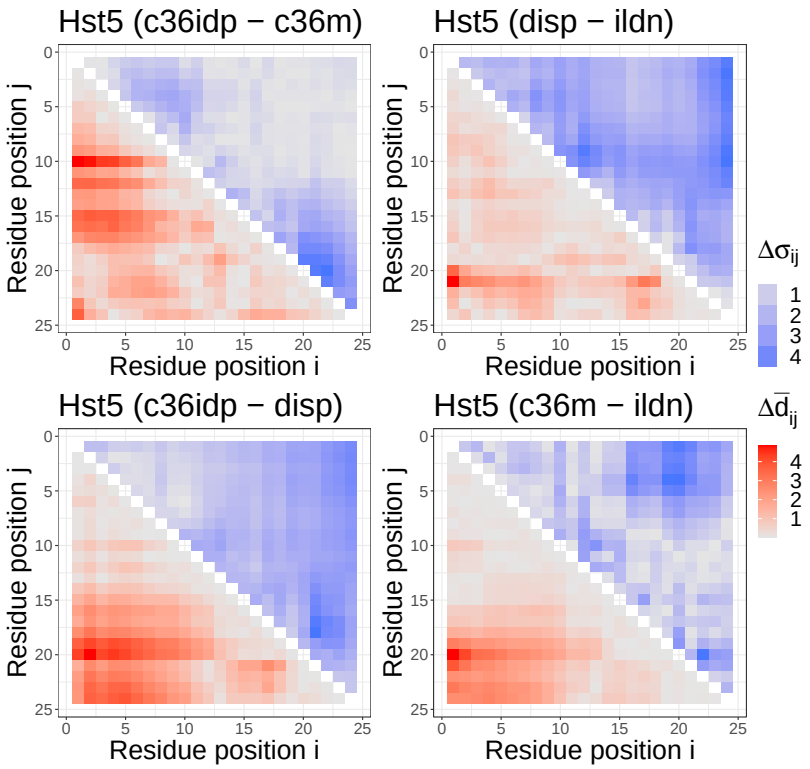


Figure C.7: Comparison of Molecular Dynamics simulations of Hst5 ensemble using different force-fields, using the methodology described in [167]. The matrix entries correspond to the absolute differences defined in (C.5).

rows in Figure 5.2). When looking at the differences located in the interior of the matrices, some similarities might arise between Figures 5.2 and C.7 for the top left comparison (c36idp vs. c36m), where the more important discrepancies appear between residues close to the N-terminus. However, the remaining comparisons exhibit contradictory behaviors between both methods, as the regions where the more relevant discrepancies appear differ. See, notably, comparisons on the bottom row. In Figure 5.2, only residues close to each other present big changes on their relative position, and no discrepancies are found in the interior region of the matrix. The opposite behavior is found in Figure C.7. The fact that the distance matrix (C.5) ignores the uncertainty (intra-ensemble distances) might partially explain the encountered discrepancies between methods.

C.3 Supplementary figures

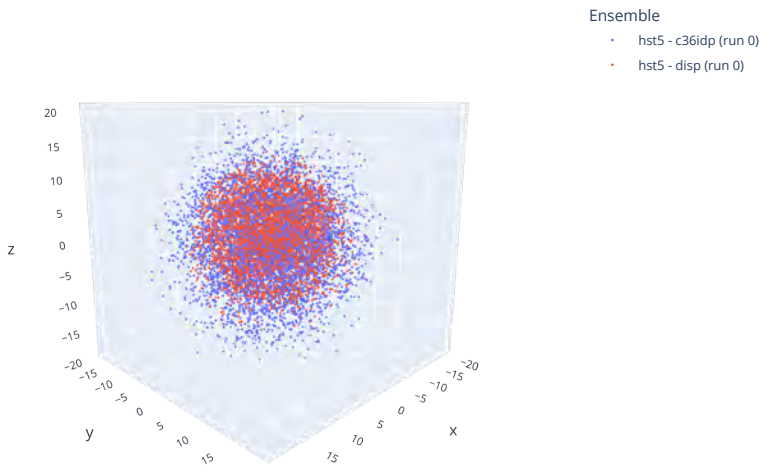
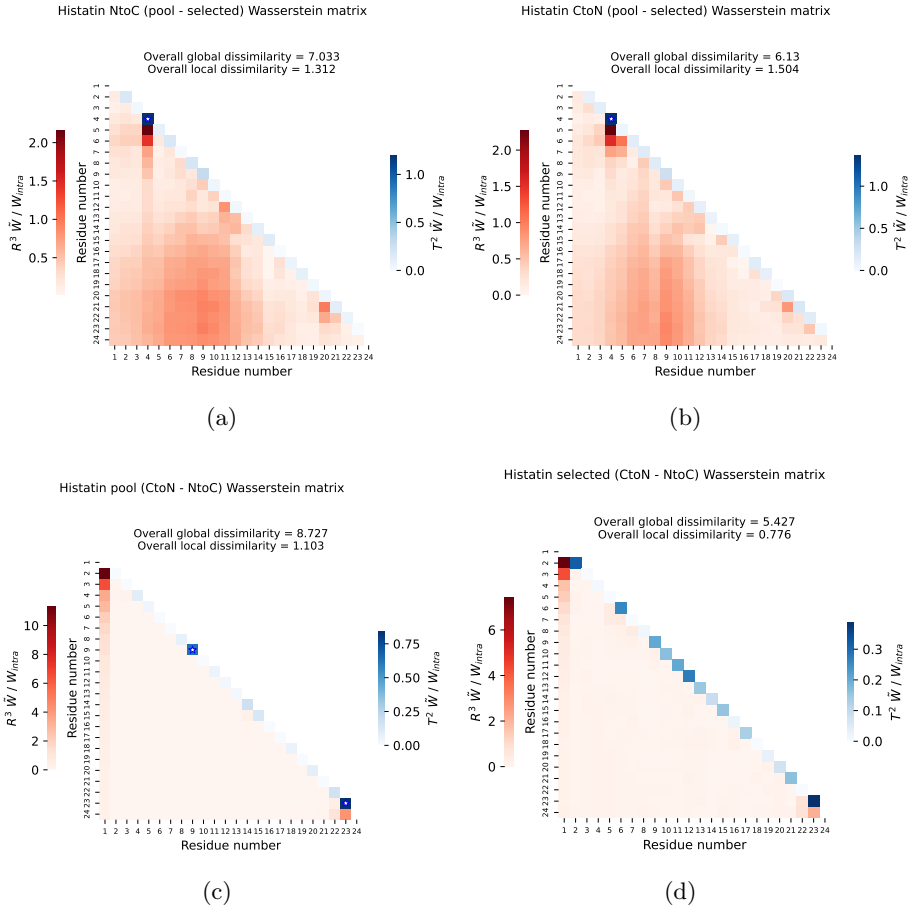


Figure C.8: Two samples of $\overrightarrow{R}_{3,10}$ corresponding to a pair of ensembles of Hst5 simulated with force-fields CHARMM36IDPSFF (c36idp) and AMBER ff99SB-disp (disp). Each sample is represented by a point cloud in the three-dimensional euclidean space.



(a) (b)

(c) (d)

Figure C.9: Comparison of Hst5 ensembles before and after filtering with experimental SAXS data. The ensemble was simulated from (a) N-to-C or from (b) C-to-N. (c) Comparison of Hst5 ensembles generated from N-to-C and C-to-N. (d) comparison of the N-to-C and C-to-N SAXS refined. In all matrices, The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (5.10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower triangle (in red) corresponds to the global differences. Coefficients along the diagonal (in blue) correspond to the local differences. Blue stars indicate that the corresponding local corrected distance is significantly different from zero (the associated p-value (5.6) is smaller than $\alpha = 0.05$).

Appendix D

Appendix of Chapter 6

Contents

D.1 Proofs of Section 6.2

Proof of Theorem 6.2.1. We use the same steps as in the proof of Theorem 1 in [104]. We begin by deriving the null distribution of the test statistic $\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}$ under the null $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$. First, we have

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \mathbf{\Sigma}) \Leftrightarrow \mathbf{X}^T \sim \mathcal{MN}_{p\times n}(\boldsymbol{\mu}^T, \mathbf{\Sigma}, \mathbf{U}) \Leftrightarrow \mathrm{vec}(\mathbf{X}^T) \sim \mathcal{N}_{np}(\mathrm{vec}(\boldsymbol{\mu}^T), \mathbf{U} \otimes \mathbf{\Sigma}), \tag{D.1}$$

where $\mathrm{vec}(\mathbf{X}^T)$ is a column vector concatenating the n vectors of p-dimensional observations that constitute $\mathbf{X}$. If we restrict $\mathrm{vec}(\mathbf{X}^T)$ to the observations (6.10) in $\mathcal{G}_1 \cup \mathcal{G}_2$, we have

$$X_{\mathcal{G}_1,\mathcal{G}_2} \sim \mathcal{N}_{p(|\mathcal{G}_1|+|\mathcal{G}_2|)}(\mu_{\mathcal{G}_1,\mathcal{G}_2}, \mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \mathbf{\Sigma}) \tag{D.2}$$

where $\mu_{\mathcal{G}_1,\mathcal{G}_2} = (\mathrm{vec}(\mu_{\mathcal{G}_1}^T), \mathrm{vec}(\mu_{\mathcal{G}_2}^T))$. Then, we can apply the linear transformation $\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}$ (6.11) to obtain the difference of means and get

$$\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} = \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} X_{\mathcal{G}_1,\mathcal{G}_2} \sim \mathcal{N}_p\left(\bar{\mu}_{\mathcal{G}_1} - \bar{\mu}_{\mathcal{G}_2}, \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}(\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \mathbf{\Sigma})\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}^T\right), \tag{D.3}$$

that, under $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$, gives

$$\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} \overset{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}{\sim} \mathcal{N}_p\left(0, \mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}\right), \tag{D.4}$$

where we replaced $\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}$ by its definition (6.12). $\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2}$ is positive definite as it is a principal submatrix of $\mathbf{U}$. The Kronecker product of two positive definite matrices is also positive definite and, as $\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}$ is a full rank linear operator, $\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}$ is positive definite [127, Observation 7.1.8]. Consequently, $\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}$ is invertible and defines the norm (6.9) in $\mathbb{R}^p$. This, together with (D.4), yields

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|^2_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \overset{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}{\sim} \chi^2_p. \tag{D.5}$$

Let us now build the p-value for $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$, by slightly adapting the reasoning in [104]. On one hand, for any $\nu \in \mathbb{R}^n$, we have

$$\mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{X} + (\mathbf{I}_n - \boldsymbol{\pi}_\nu^\perp \mathbf{X}) = \boldsymbol{\pi}_\nu^\perp \mathbf{X} + \left(\frac{\|\mathbf{X}^T\nu\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}}{\|\nu\|_2^2}\right) \nu \, \mathrm{dir}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{X}^T\nu)^T. \tag{D.6}$$

Following (6.8), $\mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2) = \bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}$ and $\|\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2^2 = 1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|$. Thus, we can write

$$\mathbf{X} = \boldsymbol{\pi}_\nu(\mathcal{G}_1,\mathcal{G}_2)^\perp \mathbf{X} + \left(\frac{\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}}{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}}\right) \nu(\mathcal{G}_1,\mathcal{G}_2) \, \mathrm{dir}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2})^T. \tag{D.7}$$

On the other hand, from the proof in [104] we have $\boldsymbol{\pi}_\nu(\mathcal{G}_1,\mathcal{G}_2)^\perp \mathbf{X} \perp\!\!\!\perp \mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2)$, which implies

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \perp\!\!\!\perp \boldsymbol{\pi}_\nu(\mathcal{G}_1,\mathcal{G}_2)^\perp \mathbf{X} \tag{D.8}$$

and, from the independence of the length and direction (in any norm) of a centered multivariate normal vector (D.4), we have

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \perp\!\!\!\perp \mathrm{dir}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}). \tag{D.9}$$

We can now plug (D.7) in the definition of our p-value (6.13) and, applying (D.8) and (D.9), we can derive

$$p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \;\Big|\; \|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \in \mathcal{S}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\})\Big), \tag{D.10}$$

where the set $\mathcal{S}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\})$ is defined in (6.15). Consequently, if we denote by $\mathbb{F}_p(t, \mathcal{S})$ the cumulative distribution function of a χ_p random variable truncated to the set $\mathcal{S}$, from (D.10) and (D.5) we have

$$p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}, \mathcal{S}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\})\right), \tag{D.11}$$

which proves the first statement (6.14). The control of selective type I error is proved identically to the reasoning in the proof of [104, Theorem 1]. □

Proof of Lemma 6.2.2. Let us first show that the perturbed data sets $\mathbf{x}'(\phi)$, defined in [104, Equation (13)] and $\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)$, defined in (6.17) are the same up to a scale transformation, i.e. that

$$\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi) = \mathbf{x}'\left(\frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}\,\phi\right) \quad \forall\, \phi \geq 0. \tag{D.12}$$

Note first that we can write

$$\left(\frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}\,\phi - \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2\right)\mathrm{dir}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}) = \left(\phi - \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}\right)\mathrm{dir}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}), \tag{D.13}$$

where $\mathrm{dir}(u) = u/\|u\|_2 \mathbb{1}\{u \neq 0\}$. Replacing (D.13) in (6.19), we have (D.12). Finally, it suffices to remark that

$$\begin{aligned}\hat{\mathcal{S}}_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}} &= \left\{\phi \geq 0 \,:\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}\left(\mathbf{x}'_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}(\phi)\right)\right\} = \\ &\left\{\phi \geq 0 \,:\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}\left(\mathbf{x}'\left(\frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}\,\phi\right)\right)\right\} \\ = \left\{\frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2}\,\phi \,:\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\} &= \frac{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{V}_{\hat{C}_1,\hat{C}_2}}}{\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2}\,\hat{\mathcal{S}},\end{aligned}$$

which concludes the proof. □

D.2 Proofs of Section 6.3

Proof of Theorem 6.3.1. The proof follows the same steps as the one of [104, Theorem 4]. In the same way, we will simplify notation by using $\hat{p}_n$ to denote $p_{\hat{\mathbf{V}}_{\hat{C}_1^{(n)},\hat{C}_2^{(n)}}}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right)$, p_n to denote $p_{\mathbf{V}_{\hat{C}_1^{(n)},\hat{C}_2^{(n)}}}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right)$, $\hat{\mathbf{V}}_n$ to denote $\hat{\mathbf{V}}_{\hat{C}_1^{(n)},\hat{C}_2^{(n)}}$ and $\mathbf{V}_n$ to denote $\mathbf{V}_{\hat{C}_1^{(n)},\hat{C}_2^{(n)}}$. If we show that

$$\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma} \Rightarrow \hat{p}_n \geq p_n, \tag{D.14}$$

then the result follows using the same reasoning as in the proof of [104, Theorem 4], replacing the usual order $\geq$ in $\mathbb{R}$ by the Loewner partial order $\succeq$ between matrices. Consequently, we only need to prove (D.14). First note that, as the Kronecker product is distributive, we have

$$\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma} \Rightarrow \hat{\mathbf{V}}_n \succeq \mathbf{V}_n. \tag{D.15}$$

Next, by Corollary 7.7.4(a) and Theorem 7.7.2(a) in [127], we can write

$$\begin{aligned}
&\hat{\mathbf{V}}_n \succeq \mathbf{V}_n \;\Leftrightarrow\; \mathbf{V}_n^{-1} \succeq \hat{\mathbf{V}}_n^{-1} \\
\Rightarrow &\left(\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\right)^T \mathbf{V}_n^{-1} \left(\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\right) \\
\geq &\left(\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\right)^T \hat{\mathbf{V}}_n^{-1} \left(\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\right) \\
\Leftrightarrow &\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\mathbf{V}_n} \geq \|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\hat{\mathbf{V}}_n}.
\end{aligned} \tag{D.16}$$

Let us then state that, if $\mathbb{F}_p(t, c, \mathcal{S})$ denotes the cumulative distribution function of a $c \cdot \chi_p$ distribution truncated to the set $\mathcal{S}$, for $c > 0$, it follows that

$$\mathbb{F}_p(t, c, a\,\mathcal{S}) = \mathbb{F}_p\left(\frac{t}{a}, \frac{c}{a}, \mathcal{S}\right), \tag{D.17}$$

for any $a > 0$. With a slight abuse of notation we write $\mathbb{F}_p(t, 1, \mathcal{S}) = \mathbb{F}_p(t, \mathcal{S})$ where $\mathbb{F}_p(t, \mathcal{S})$ is the CDF involved in (6.14). Consequently, taking

$$a = \frac{\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\hat{\mathbf{V}}_n}}{\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\mathbf{V}_n}} \leq 1, \tag{D.18}$$

we have

$$\begin{aligned}
1 - \hat{p}_n &= \mathbb{F}_p\left(\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\hat{\mathbf{V}}_n}\,,\, \mathcal{S}_{\hat{\mathbf{V}}_n}\right) = \mathbb{F}_p\left(\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\hat{\mathbf{V}}_n}\,,\, a\,\mathcal{S}_{\mathbf{V}_n}\right) \\
&= \mathbb{F}_p\left(\frac{1}{a}\,\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\hat{\mathbf{V}}_n}\,,\, \frac{1}{a}\,,\, \mathcal{S}_{\mathbf{V}_n}\right) = \mathbb{F}_p\left(\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\mathbf{V}_n}\,,\, \frac{1}{a}\,,\, \mathcal{S}_{\mathbf{V}_n}\right) \\
&\leq \mathbb{F}_p\left(\|\overline{X^{(n)}}_{\hat{C}_1^{(n)}} - \overline{X^{(n)}}_{\hat{C}_2^{(n)}}\|_{\mathbf{V}_n}\,,\, 1\,,\, \mathcal{S}_{\mathbf{V}_n}\right) = 1 - p_n,
\end{aligned} \tag{D.19}$$

where the last inequality follows from Lemma A.3 in [104]. This shows (D.14). We conclude by proving the statement (D.17). First, if we denote by $f(t, c, \mathcal{S})$ the probability density function of a $c \cdot \chi_p$ distribution truncated to the set $\mathcal{S}$, we have

$$f(t, c, a\,\mathcal{S}) = \frac{1}{a}\, f\left(\frac{t}{a}, \frac{c}{a}, \mathcal{S}\right). \tag{D.20}$$

Indeed, following the first lines of the proof of [104, Lemma A.3], we can rewrite $f(t, c, a\,\mathcal{S})$ as

$$f(t, c, a\,\mathcal{S}) = \frac{t^{p-1}\,\mathbb{1}\{t \in a\,\mathcal{S}\}}{\int u^{p-1} \exp(-\frac{u^2}{2c^2}), \mathbb{1}\{t \in a\,\mathcal{S}\}\, du} \exp\left(-\frac{t^2}{2c^2}\right), \tag{D.21}$$

that we can easily express in terms of t/a as

$$f(t, c, a\,\mathcal{S}) = \frac{\left(\frac{t}{a}\right)^{p-1}\,\mathbb{1}\{\frac{t}{a} \in \mathcal{S}\}}{\int \left(\frac{u}{a}\right)^{p-1} \exp(-\frac{(u/a)^2}{2(c/a)^2}), \mathbb{1}\{\frac{t}{a} \in \mathcal{S}\}\, du} \exp\left(-\frac{(t/a)^2}{2(c/a)^2}\right) = \frac{1}{a}\, f\left(\frac{t}{a}, \frac{c}{a}, \mathcal{S}\right), \tag{D.22}$$

where the last equality follows from taking the variable change $y = u/a$ in the integral. Finally, we have

$$\mathbb{F}_p(t, c, a\,\mathcal{S}) = \int_0^t f(x, c, a\,\mathcal{S})\,dx = \frac{1}{a}\int_0^t f\left(\frac{x}{a}, \frac{c}{a}, \mathcal{S}\right)\,dx = \int_0^{\frac{t}{a}} f\left(u, \frac{c}{a}, \mathcal{S}\right)\,du = \mathbb{F}_p\left(\frac{t}{a}, \frac{c}{a}, \mathcal{S}\right),$$

which proves (D.17). □

Proof of Remark 6.3.5. The case of diagonal matrices is straightforward as both $\mathbf{U}^{(n)}$ and $(\mathbf{U}^{(n)})^{-1}$ are defined by a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$. Every diagonal entry of the inverse satisfies $(U^{(n)})^{-1}_{ii} = \frac{1}{\lambda_i}$ for all $n \in \mathbb{N}$ and, as we asked the λ_i to converge to λ, which is strictly positive due to the positive definiteness of $\mathbf{U}^{(n)}$, Assumption 4 is satisfied. □

Proof of Remark 6.3.6. Let $\mathbf{U}^{(n)} = b\mathbf{1}_{n\times n} + (a-b)\,\mathbf{I}_n$. Note that, as $\mathbf{U}^{(n)}$ is positive definite, the coefficients a, b verify $a > b$. This follows the fact that $\max_{i,j}|A_{ij}| \leq \max_{ii} A_{ii}$ for any positive definite matrix A. Following the Sherman–Morrison formula [14], we can derive an explicit expression for the sequence of inverse matrices:

$$\left(\mathbf{U}^{(n)}\right)^{-1} = \frac{1}{a-b}\,\mathbf{I}_n + \frac{-b}{(a-b)(nb+a-b)}, \quad \forall\, n \in \mathbb{N}. \tag{D.23}$$

Consequently, for every $r \geq 0$ and every $i \in \mathbb{N}$, we have

$$\left(\mathbf{U}^{(n)}\right)^{-1}_{i\,i+r} = \begin{cases} \frac{1}{a-b} + \frac{-b}{(a-b)(nb+a-b)} & \text{if } r = 0, \\ \frac{-b}{(a-b)(nb+a-b)} & \text{if } r > 0, \end{cases}$$

which are monotone, so condition (ii) in Assumption 4 is satisfied. Then, we have

$$\Lambda_{i\,i+r} = \begin{cases} \frac{1}{a-b} & \text{if } r = 0, \\ 0 & \text{if } r > 0, \end{cases}$$

for all $i \in \mathbb{N}$, $\lambda_0 = 1/(a-b)$ and $\lambda_r = 0$ for $r > 0$. Consequently, Assumption 4 holds. □

Proof of Remark 6.3.7. The inverse of an auto-regressive covariance matrix of lag $P \geq 1$ is banded with $2P-1$ non-zero diagonals. Its explicit form is derived in [297] for a stationary process of any lag, and the cases $P \leq 3$ are discussed in detail in [308]. From these results we can derive the behavior of the sequences $\{(U^{(n)})^{-1}_{i\,i+r}\}$ as n increases. The diagonal elements define the sequences

$$\sigma^2\left\{\left(U^{(n)}\right)^{-1}_{ii}\right\}_{n\in\mathbb{N}} =$$

$$\begin{cases} \{1 + \sum_{k=1}^{i-1}\beta_k^2, 1 + \sum_{k=1}^{i-1}\beta_k^2, \ldots\} & \text{if } i \leq p+1, \\ \{0, \overset{i-1}{\ldots}, 0, 1, 1+\beta_1^2, 1, 1+\beta_1^2\beta_2^2, \ldots, 1+\sum_{k=1}^{p}\beta_k^2, 1+\sum_{k=1}^{p}\beta_k^2, \ldots\} & \text{if } i > p+1, \end{cases}$$

where the sums are taken as zero if the upper limit of summation is zero. Note that these sequences do not satisfy condition (i) in Assumption 4 as, even if each sequence reaches its limit after a finite number of terms, the index of the term where the limit is reached diverges with i. In other words, we can dominate the sequence, but not by a summable one. However, for all $i \in \mathbb{N}$ the series are non-decreasing so condition (ii) is satisfied and we have

$$\sigma^2 \Lambda_{ii} = \begin{cases} 1 + \sum_{k=1}^{i-1} \beta_k^2 & \text{if } i \leq p+1 \\ 1 + \sum_{k=1}^{p} \beta_k^2 & \text{if } i > p+1. \end{cases}$$

Then, $\sigma^2 \lambda_0 = 1 + \sum_{k=1}^{p} \beta_k^2$. The sequences outside the main diagonal show a similar behavior, but they are not positive in general. As, following the same reasoning, they do not satisfy condition (i) in Assumption 4, we force them to satisfy condition (ii). For any $0 < r \leq P$, we have

$$\sigma^2 \left\{ \left(U^{(n)} \right)^{-1}_{i\, i+r} \right\}_{n \in \mathbb{N}} =$$

$$\begin{cases} \{-\beta_r + \sum_{k=1}^{i-(r+1)} \beta_k \beta_{k+r}, -\beta_r + \sum_{k=1}^{i-(r+1)} \beta_k \beta_{k+r}, \ldots\} & \text{if } i \leq p+1, \\ \{0, \overset{i-1}{\ldots}, 0, -\beta_r + \beta_1 \beta_{1+r}, -\beta_r + \beta_1 \beta_{1+r} + \beta_2 \beta_{2+r}, \ldots, & \\ -\beta_r + \sum_{k=1}^{p-r} \beta_k \beta_{k+r}, -\beta_r + \sum_{k=1}^{p-r} \beta_k \beta_{k+r}, \ldots\} & \text{if } i > p+1. \end{cases} \tag{D.24}$$

For these sequences to satisfy condition (ii) we need them to be non-decreasing or non-increasing. For $P \leq 2$ this is always satisfied but, for $P > 2$, we need to require all the β_k to have the same sign. In that case, condition (ii) holds and we have

$$\sigma^2 \Lambda_{i\, i+r} = \begin{cases} -\beta_r + \sum_{k=1}^{i-(r+1)} \beta_k \beta_{k+r} & \text{if} i \leq p+1, \\ -\beta_r + \sum_{k=1}^{p-r} \beta_k \beta_{k+r} & \text{if} i > p+1, \end{cases}$$

and, consequently, $\sigma^2 \lambda_r = -\beta_r + \sum_{k=1}^{p-r} \beta_k \beta_{k+r}$. As the sequence $\{\lambda_r\}_{r=1}^{\infty}$ is non-zero for for a finite number of terms (due to the bandedness of the inverse matrix), its sum converges and Assumption 4 is satisfied. □

Proof of Lemma 6.3.3. We start by rewriting the sum in (6.37) as a sum along each diagonal. Using the symmetry of $(\mathbf{U}^{(n)})^{-1}$ we have,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l,s=1}^{n} \left(U^{(n)} \right)^{-1}_{ls} \mathbb{1}\{\mu_l^{(n)} = \theta_k\} \mathbb{1}\{\mu_s^{(n)} = \theta_{k'}\}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n-1} \sum_{i=1}^{n-r} \left(U^{(n)} \right)^{-1}_{i\, i+r} \mathbb{1}\{\mu_i^{(n)} = \theta_k\} \mathbb{1}\{\mu_{i+r}^{(n)} = \theta_{k'}\} \tag{D.25}$$

$$+ \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n-1} \sum_{i=1}^{n-r} \left(U^{(n)} \right)^{-1}_{i\, i+r} \mathbb{1}\{\mu_{i+r}^{(n)} = \theta_k\} \mathbb{1}\{\mu_i^{(n)} = \theta_{k'}\} \tag{D.26}$$

$$+ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(U^{(n)} \right)^{-1}_{i\, i} \mathbb{1}\{\mu_i^{(n)} = \theta_k\} \mathbb{1}\{\mu_i^{(n)} = \theta_{k'}\}, \tag{D.27}$$

where (D.25),(D.26) and (D.27) are respectively the sums along all the superdiagonals, subdiagonals and along the main diagonal. Let's detail the general reasoning that we use to show that the three quantities converge. Let $\{a_i^{(n)}\}_{i\in\mathbb{N}}$ be a double sequence such that $\lim_{n\to\infty} a_i^{(n)} = a_i \in \mathbb{R}$, and let $\{b_i^{(n)}\}_{i\in\mathbb{N}}$ be a binary Cesàro summable double sequence, i.e. such that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} b_i^{(n)} = b$ and $b_i^{(n)} \in \{0,1\}$ for all $i, n \in \mathbb{N}$. Let's first show that, if $\{a_i^{(n)}\}_{n\in\mathbb{N}}$ satisfies any of the conditions (i) or (ii), and the sequence $\{a_i^{(1)} - a_i\}_{i=1}^{\infty} \in \ell_1$, we can write

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i^{(n)}\, b_i^{(n)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i\, b_i^{(n)}. \tag{D.28}$$

First, note that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i^{(n)}\, b_i^{(n)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left(a_i^{(n)} - a_i\right)\, b_i^{(n)} + \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i\, b_i^{(n)}. \tag{D.29}$$

Therefore, it suffices to show that the first term in (D.29) is zero to have (D.28). Using Hölder's inequality, we have

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\left|\sum_{i=1}^{n} \left(a_i^{(n)} - a_i\right)\, b_i^{(n)}\right| &\leq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left|\left(a_i^{(n)} - a_i\right) b_i^{(n)}\right| \\ &\leq \lim_{n\to\infty} \left(\sum_{i=1}^{n} \left(a_i^{(n)} - a_i\right)^2\right)^{\frac{1}{2}} \lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{n} b_i^{(n)}\right)^{\frac{1}{2}}.\end{aligned}$$

On one hand,

$$\lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{n} b_i^{(n)}\right)^{\frac{1}{2}} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} b_i^{(n)}\right)^{\frac{1}{2}} = 0.$$

On the other hand, let us show that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \left(a_i^{(n)} - a_i\right)^2 = 0 \tag{D.30}$$

if $\{a_i^{(n)}\}_{n\in\mathbb{N}}$ satisfies any of the conditions (i) or (ii). If $\{a_i^{(n)}\}_{n\in\mathbb{N}}$ satisfies (i), the sequence $\{(a_i^{(n)} - a_i)^2\}_{n\in\mathbb{N}}$ is dominated by the sequence $\{\alpha_i^2\}_{i\in\mathbb{N}}$, which is summable as $\ell_1 \subset \ell_2$. Then, (D.28) holds following the Dominated Convergence Theorem [311, Theorem 9.20]. If $\{a_i^{(n)}\}_{n\in\mathbb{N}}$ is non-increasing, then $a_i^{(n+1)} - a_i \leq a_i^{(n)} - a_i$ implies $(a_i^{(n+1)} - a_i)^2 \leq (a_i^{(n)} - a_i)^2$ and $\tilde{a}_i^{(n)} := (a_i^{(n)} - a_i)^2$ is a non-increasing and non-negative sequence. Similarly, if $\{a_i^{(n)}\}_{n\in\mathbb{N}}$ is non-decreasing, then $a_i^{(n+1)} - a_i \geq a_i^{(n)} - a_i$ implies $(a_i^{(n+1)} - a_i)^2 \leq (a_i^{(n)} - a_i)^2$ and $\tilde{a}_i^{(n)}$ is again a non-increasing and non-negative sequence. Then, the sequence $z_i^{(n)} :=$

$\tilde{a}_i^{(1)} - \tilde{a}_i^{(n)}$ is non-negative and non-decreasing. Thus, following the Monotone Convergence Theorem [311, Theorem 8.5], we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} z_i^{(n)} = \lim_{n\to\infty} \sum_{i=1}^{n} (a_i^{(1)} - a_i)^2, \tag{D.31}$$

which implies (D.30) if the limit in the right side of (D.31) exists and is finite. This is guaranteed if we ask the sequence $\{a_i^{(1)} - a_i\}_{i=1}^{\infty}$ to be summable. This always holds in our case as we can arbitrarily define the entries $\left(U^{(n)}\right)^{-1}_{i\,i+r}$ for $i > n$. Consequently, if we write $\{\left(U^{(1)}\right)^{-1}_{i\,i+r}\}_{i=1}^{\infty} = \{\left(U^{(1)}\right)^{-1}_{1\,1+r}, \Lambda_{2\,2+r}, \Lambda_{3\,3+r}, \ldots\}$, the sequence $\{\left(U^{(1)}\right)^{-1}_{i\,i+r} - \Lambda_{i\,i+r}\}_{i=1}^{\infty}$ is trivially summable. This proves (D.28).

Now, if we have that $\lim_{i\to\infty} a_i = a$, let us show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i\, b_i^{(n)} = ab. \tag{D.32}$$

First, let separate the sum in (D.32) as

$$\frac{1}{n} \sum_{i=1}^{n} a_i\, b_i^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (a_i - a)\; b_i^{(n)} + \frac{a}{n} \sum_{i=1}^{n} b_i^{(n)}. \tag{D.33}$$

The right term tends to ab when $n \to \infty$. Let's show that the first term tends to zero. For any $i_0 \in \mathbb{N}$, we can write

$$\left| \frac{1}{n} \sum_{i=1}^{n} (a_i - a)\; b_i^{(n)} \right| \leq \left| \frac{1}{n} \sum_{i=1}^{i_0-1} (a_i - a)\; b_i^{(n)} \right| + \left| \frac{1}{n} \sum_{i=i_0}^{n} (a_i - a)\; b_i^{(n)} \right| \tag{D.34}$$

$$\leq \sup_{i<i_0} |a_i - a| \frac{1}{n} \sum_{i=1}^{i_0-1} b_i^{(n)} + \sup_{i\geq i_0} |a_i - a| \frac{1}{n} \sum_{i=i_0}^{n} b_i^{(n)} \leq \frac{C}{n} + \sup_{i\geq i_0} |a_i - a| \frac{1}{n} \sum_{i=i_0}^{n} b_i^{(n)}, \tag{D.35}$$

where C is a real constant. Then, following the definition of limit, when can choose i_0 as the one such that for all $i \geq i_0$ we have $|a_i - a| \leq \frac{1}{n}$. Therefore,

$$\left| \frac{1}{n} \sum_{i=1}^{n} (a_i - a)\; b_i^{(n)} \right| \leq \frac{C}{n} + \frac{1}{n^2} \sum_{i=i_0}^{n} b_i^{(n)}, \tag{D.36}$$

which tends to zero when $n \to \infty$ using that $\{b_i^{(n)}\}_i \in \mathbb{N}$ has Cesàro sum b. Thus, we have (D.32). As the sequences $(U^{(n)})^{-1}_{i\,i+r}$ have limits $\Lambda_{i\,i+r}$ when $i \to \infty$, following Assumption 3, and the products of indicator functions are Cesàro summable thanks to Assumptions 2 and 3, we can use (D.28) and (D.32) to rewrite the three limits in (D.25),

(D.26), (D.27) as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mathbb{1}\{\mu_l^{(n)} = \theta_k\} \mathbb{1}\{\mu_s^{(n)} = \theta_{k'}\}$$

$$= \lim_{n\to\infty} \sum_{r=1}^{n-1} \lambda_r \left(\pi^r_{kk'} + \pi^r_{k'k}\right) + \lambda_0 \pi_k \delta_{kk'} = 2(\lambda - \lambda_0)\pi_k \pi_{k'} + \lambda_0 \pi_k \delta_{kk'}, \tag{D.37}$$

where the last limit is derived following the same reasoning as to prove (D.32). This concludes the proof. □

Proof of Proposition 6.3.2. We start by proving the element-wise convergence in probability of (6.28). More precisely, we show that

$$\hat{\Sigma}^{(n)}_{ij} \xrightarrow{p} \Sigma_{ij} + \lambda_0 \sum_{k=1}^{K^*} \pi_k \left(\theta_{ki} - \tilde{\theta}_i\right)\left(\theta_{kj} - \tilde{\theta}_j\right), \tag{D.38}$$

for all $i, j \in \{1, \dots, p\}$, where $\hat{\Sigma}^{(n)}_{ij}$ is the ij entry of $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ and we have defined $\tilde{\theta}_i = \sum_{k=1}^{K^*} \pi_k \theta_{ki}$. Recall that all the quantities in (D.38) have been defined in Assumptions 2 and 4. To prove (D.38), it suffices to show, following the same reasoning as in the proof of [104, Lemma C.1], that

$$\lim_{n\to\infty} \mathbb{E}\left(\hat{\Sigma}^{(n)}_{ij}\right) = \Sigma_{ij} + \lambda_0 \sum_{k=1}^{K^*} \pi_k \left(\theta_{ki} - \tilde{\theta}_i\right)\left(\theta_{kj} - \tilde{\theta}_j\right) \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Var}\left(\hat{\Sigma}^{(n)}_{ij}\right) = 0. \tag{D.39}$$

Indeed, (D.39) implies convergence in mean of $\hat{\Sigma}^{(n)}_{ij}$ towards the limit of its expectation and, following Markov's inequality, convergence in probability. Let start by rewriting $\hat{\Sigma}^{(n)}_{ij}$. Following (6.30), we can write

$$\begin{aligned}\hat{\Sigma}^{(n)}_{ij} = &\frac{1}{n-1} \sum_{l,s=1}^{n} X^{(n)}_{li} X^{(n)}_{js} \left(U^{(n)}\right)^{-1}_{ls} - \frac{1}{n-1} \bar{X}^{(n)}_j \sum_{l,s=1}^{n} X^{(n)}_{li} \left(U^{(n)}\right)^{-1}_{ls} \\ &- \frac{1}{n-1} \bar{X}^{(n)}_i \sum_{l,s=1}^{n} X^{(n)}_{sj} \left(U^{(n)}\right)^{-1}_{ls} + \frac{1}{n-1} \bar{X}^{(n)}_i \bar{X}^{(n)}_j \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls}.\end{aligned} \tag{D.40}$$

For simplicity, we denote as $A^{(n)}_{ij}$, $B^{(n)}_{ij}$, $C^{(n)}_{ij}$ and $D^{(n)}_{ij}$ the four terms in (D.40) respectively. First, let us derive their asymptotic expectations.

$$\begin{aligned}\mathbb{E}\left(A^{(n)}_{ij}\right) &= \frac{1}{n-1} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mathbb{E}\left(X^{(n)}_{li} X^{(n)}_{sj}\right) \\ &= \frac{1}{n-1} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mu^{(n)}_{li} \mu^{(n)}_{sj} + \frac{\Sigma_{ij}}{n-1} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} U^{(n)}_{sl} \\ &= \sum_{k,k'=1}^{K^*} \frac{1}{n-1} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mathbb{1}\{\mu_l^{(n)} = \theta_k\} \mathbb{1}\{\mu_s^{(n)} = \theta_{k'}\} \theta_{ki} \theta_{k'j} + \frac{n}{n-1} \Sigma_{ij}.\end{aligned}$$

Using Lemma 6.3.3, we have

$$\lim_{n\to\infty} \mathbb{E}\left(A_{ij}^{(n)}\right) = 2(\lambda-\lambda_0)\sum_{k=1}^{K^*}\pi_k\theta_{ki}\sum_{k=1}^{K^*}\pi_k\theta_{kj} + \lambda_0\sum_{k=1}^{K^*}\pi_k\theta_{ki}\theta_{kj} + \Sigma_{ij}. \tag{D.41}$$

Then,

$$\mathbb{E}\left(B_{ij}^{(n)}\right) = \frac{1}{n(n-1)}\sum_{l,s,r=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mathbb{E}\left(X_{li}^{(n)}\,X_{rj}^{(n)}\right)$$

$$= \frac{1}{n(n-1)}\sum_{l,s,r=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mu_{li}^{(n)}\mu_{rj}^{(n)} + \frac{\Sigma_{ij}}{n-1} = \frac{1}{n}\sum_{r=1}^{n}\mu_{rj}^{(n)}\frac{1}{n-1}\sum_{l,s}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mu_{li}^{(n)} + \frac{\Sigma_{ij}}{n-1}$$

$$= \sum_{k=1}^{K^*}\frac{1}{n}\sum_{r=1}^{n}\mathbb{1}\{\mu_r^{(n)}=\theta_k\}\theta_{kj}\sum_{k=1}^{K^*}\frac{1}{n-1}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mathbb{1}\{\mu_l^{(n)}=\theta_k\}\theta_{ki} + \frac{\Sigma_{ij}}{n-1}.$$

Using the same reasoning as to prove Lemma 6.3.3, we have

$$\lim_{n\to\infty}\frac{1}{n-1}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\mathbb{1}\{\mu_l^{(n)}=\theta_k\} = (2(\lambda-\lambda_0)+\lambda_0)\pi_k.$$

This, together with Assumption 2, yields

$$\lim_{n\to\infty}\mathbb{E}\left(B_{ij}^{(n)}\right) = \lim_{n\to\infty}\mathbb{E}\left(C_{ij}^{(n)}\right) = (2(\lambda-\lambda_0)+\lambda_0)\sum_{k=1}^{K^*}\pi_k\theta_{kj}\sum_{k=1}^{K^*}\pi_k\theta_{ki}, \tag{D.42}$$

where $B_{ij}^{(n)}$ and $C_{ij}^{(n)}$ have the same expectation by symmetry. Finally,

$$\mathbb{E}\left(D_{ij}^{(n)}\right) = \frac{1}{n^2(n-1)}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\sum_{r,r'=1}^{n}\mathbb{E}\left(X_{ri}^{(n)}\,X_{r'j}^{(n)}\right)$$

$$\frac{1}{n-1}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1}\left[\frac{1}{n^2}\sum_{r,r'=1}^{n}\mu_{ri}^{(n)}\mu_{r'j}^{(n)} + \frac{\Sigma ij}{n^2}\sum_{r,r'=1}^{n}U_{rr'}^{(n)}\right].$$

Using the same reasoning as to prove Lemma 6.3.3, we have

$$\lim_{n\to\infty}\frac{1}{n-1}\sum_{l,s=1}^{n}\left(U^{(n)}\right)_{ls}^{-1} = 2(\lambda-\lambda_0)+\lambda_0. \tag{D.43}$$

Moreover, we state that

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{l,s=1}^{n}U_{ls}^{(n)} = 0. \tag{D.44}$$

We prove (D.44) at the end of the proof. This claim, together with (D.43) and Assumption 2, yields

$$\lim_{n\to\infty}\mathbb{E}\left(D_{ij}^{(n)}\right) = (2(\lambda-\lambda_0)+\lambda_0)\sum_{k=1}^{K^*}\pi_k\theta_{ki}\sum_{k=1}^{K^*}\pi_k\theta_{kj}. \tag{D.45}$$

Consequently, following (D.41), (D.42) and (D.45), we have

$$\lim_{n\to\infty} \mathbb{E}\left(\hat{\Sigma}_{ij}^{(n)}\right) = \Sigma_{ij} + \lambda_0 \left[\sum_{k=1}^{K^*} \pi_k \theta_{ki}\theta_{kj} - \sum_{k=1}^{K^*} \pi_k \theta_{ki} \sum_{k=1}^{K^*} \pi_k \theta_{kj}\right]$$
$$= \Sigma_{ij} + \lambda_0 \sum_{k=1}^{K^*} \pi_k \left(\theta_{ki} - \tilde{\theta}_i\right)\left(\theta_{kj} - \tilde{\theta}_j\right). \tag{D.46}$$

This is the first statement in (D.39). To prove the second one, we show that the variance of each term in (D.40) tends to zero. To do so, we need the explicit form of the non-centered 4-th moments of a Gaussian distribution. More precisely, if $X_1, \ldots, X_4$ are four Gaussian random variables with $\mathbb{E}(X_i) = \mu_i$ and $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, for $i, j \in \{1, \ldots, 4\}$, we need the explicit form of the quantity

$$\mathbb{E}\left(X_1\, X_2\, X_3\, X_4\right) - \mathbb{E}\left(X_1\, X_2\right)\, \mathbb{E}\left(X_3\, X_4\right). \tag{D.47}$$

The first term can be derived using the moment generating function of a 4-dimensional normal distribution

$$M_{(X_1,\ldots,X_4)}(t_1, \ldots, t_4) = \exp\left(\sum_{i=1}^{4} \mu_i\, t_i + \frac{1}{2} \sum_{i,j=1}^{n} \sigma_{ij}\, t_i\, t_j\right),$$

and computing

$$\mathbb{E}\left(X_1\, X_2\, X_3\, X_4\right) = \left.\frac{\partial M_{(X_1,\ldots,X_4)}(t_1, \ldots, t_4)}{\partial\, t_1 \cdots \partial\, t_4}\right|_0.$$

Doing so, and using $\mathbb{E}(X_i\, X_j) = \mu_i\mu_j + \sigma_{ij}$, we can derive

$$\mathbb{E}\left(X_1\, X_2\, X_3\, X_4\right) - \mathbb{E}\left(X_1\, X_2\right)\, \mathbb{E}\left(X_3\, X_4\right) =$$
$$\sigma_{13}\sigma_{24} + \sigma_{14}\sigma_{23} + \mu_1\mu_4\sigma_{23} + \mu_1\mu_3\sigma_{24} + \mu_2\mu_3\sigma_{14} + \mu_2\mu_4\sigma_{13}. \tag{D.48}$$

We are ready to prove that $\mathrm{Var}\left(\hat{\Sigma}_{ij}^{(n)}\right)$ tends to zero. First, using $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, we have

$$\mathrm{Var}\left(A_{ij}^{(n)}\right) = \frac{1}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{sl}^{-1} \left(U^{(n)}\right)_{kr}^{-1} \left[\mathbb{E}\left(X_{li}\, X_{sj}\, X_{ri}\, X_{kj}\right) -\right.$$
$$\left.\mathbb{E}\left(X_{li}\, X_{sj}\right) \mathbb{E}\left(X_{ki}\, X_{rj}\right)\right]. \tag{D.49}$$

Using (D.48), we can separate (D.49) into the following six terms:

$$\operatorname{Var}\left(A_{ij}^{(n)}\right) = \frac{\Sigma_{ii}\Sigma_{jj}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{lk}^{(n)}\, U_{sr}^{(n)} \tag{D.50}$$

$$+\frac{\Sigma_{ij}^2}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{lr}^{(n)}\, U_{sk}^{(n)} \tag{D.51}$$

$$+\frac{\Sigma_{jj}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{sr}^{(n)}\, \mu_{li}^{(n)}\, \mu_{ki}^{(n)} \tag{D.52}$$

$$+\frac{\Sigma_{ij}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{sk}^{(n)}\, \mu_{li}^{(n)}\, \mu_{rj}^{(n)} \tag{D.53}$$

$$+\frac{\Sigma_{ij}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{lr}^{(n)}\, \mu_{ki}^{(n)}\, \mu_{sj}^{(n)} \tag{D.54}$$

$$+\frac{\Sigma_{ii}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{lk}^{(n)}\, \mu_{sj}^{(n)}\, \mu_{rj}^{(n)}. \tag{D.55}$$

Each of these terms tend to zero when $n \to \infty$. For (D.50), we have

$$\begin{aligned}\frac{\Sigma_{ii}\Sigma_{jj}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{lk}^{(n)}\, U_{sr}^{(n)} &= \frac{\Sigma_{ii}\Sigma_{jj}}{(n-1)^2} \sum_{l,s,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} U_{sr}^{(n)}\, \delta_{lr} \\ = \frac{\Sigma_{ii}\Sigma_{jj}}{(n-1)^2} \sum_{l,s=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} U_{sl}^{(n)} = \frac{\Sigma_{ii}\Sigma_{jj}}{(n-1)^2} \sum_{l=1}^{n} \delta_{ll} &= \frac{n}{(n-1)^2}\Sigma_{ii}\Sigma_{jj} \underset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

Identically we can show that (D.51) tends to zero. For (D.52), we have

$$\begin{aligned}&\frac{\Sigma_{jj}}{(n-1)^2} \sum_{l,s,k,r=1}^{n} \left(U^{(n)}\right)_{ls}^{-1} \left(U^{(n)}\right)_{kr}^{-1} U_{sr}^{(n)}\, \mu_{li}^{(n)}\, \mu_{ki}^{(n)} \\ = \frac{\Sigma_{jj}}{(n-1)^2} \sum_{l,k,r=1}^{n} \left(U^{(n)}\right)_{kr}^{-1} \delta_{lr}\, \mu_{li}^{(n)}\, \mu_{ki}^{(n)} &= \frac{\Sigma_{jj}}{(n-1)^2} \sum_{l,k=1}^{n} \left(U^{(n)}\right)_{kl}^{-1} \mu_{li}^{(n)}\, \mu_{ki}^{(n)} \\ = \sum_{r,r'=1}^{K^*} \frac{\Sigma_{jj}}{(n-1)^2} \sum_{l,k=1}^{n} \left(U^{(n)}\right)_{kl}^{-1} \mathbb{1}\{\mu_l^{(n)} = \theta_r\}\, \mathbb{1}\{\mu_k^{(n)} = \theta_{r'}\}\, \mu_{li}^{(n)}\, \mu_{ki}^{(n)}\, \theta_{ri}\, \theta_{r'i} &\underset{n\to\infty}{\longrightarrow} 0,\end{aligned}$$

where the limit is derived using Lemma 6.3.3. The same reasoning is used to show that (D.53), (D.54) and (D.55) tend to zero when $n \to \infty$. Therefore, we have $\lim_{n\to\infty} \operatorname{Var}\left(A_{ij}^{(n)}\right) = 0$. The same strategy, together with (D.43) and (D.44), is used to show that $\lim_{n\to\infty} \operatorname{Var}\left(B_{ij}^{(n)}\right) = \lim_{n\to\infty} \operatorname{Var}\left(C_{ij}^{(n)}\right) = \lim_{n\to\infty} \operatorname{Var}\left(D_{ij}^{(n)}\right) = 0$. Consequently, we have (D.38). Note that the sum in (D.38) can be written as the ij term of a matrix. Indeed, we have

$$\hat{\Sigma}_{ij}^{(n)} - \Sigma_{ij} \xrightarrow{p} \lambda_0 \left(\Theta^T \operatorname{diag}(\pi_1, \ldots, \pi_{K^*})\, \Theta\right)_{ij}, \tag{D.56}$$

where Θ is a $p \times K^*$ matrix having as entries $\Theta_{ij} = \theta_{ij} - \tilde{\theta}_j$. As $\lambda_0, \pi_1, \dots, \pi_{K^*} \geq 0$, the matrix $\lambda_0(\Theta^T \operatorname{diag}(\pi_1, \dots, \pi_{K^*})\,\Theta)$ is positive semi-definite, so the entries of $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) - \mathbf{\Sigma}$ converge in probability to the entries of a positive semi-definite matrix. Note that, as both $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ and $\mathbf{\Sigma}$ are positive definite, the eigenvalues of their difference are real. Finally, since the eigenvalues depend continuously on the entries of the matrix, the eigenvalues of $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) - \mathbf{\Sigma}$ converge in probability to the eigenvalues of a positive semi-definite matrix, which are non-negative. Therefore, we have (6.36).

Let us conclude by showing (D.44). To do show, note that we can write,

$$1 = \frac{1}{n}\sum_{k,l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{lk} U^{(n)}_{ks} = \frac{2}{n}\sum_{s=1}^{n}\sum_{r=1}^{n-1}\sum_{i=1}^{n-r} \left(U^{(n)}\right)^{-1}_{i\,i+r} U^{(n)}_{i+r\,s} + \frac{1}{n}\sum_{s,i=1}^{n} \left(U^{(n)}\right)^{-1}_{i\,i} U^{(n)}_{i\,s}.$$

Using the same reasoning as in the proof of Lemma 6.3.3, we have

$$1 = 2\lim_{n\to\infty}\sum_{r=1}^{n-1} \lambda_r \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i,s=1}^{n} U^{(n)}_{i+r\,s}\right) + \lambda_0 \lim_{n\to\infty}\frac{1}{n}\sum_{i,s=1}^{n} U^{(n)}_{i\,s},$$

which diverges unless the third limit is finite, which implies (D.44). □

D.3 Proofs of Section 6.4

Proof of Theorem 6.4.1. As mentioned after Theorem 6.4.1, we omit the proof of (6.48) as it is identical to the one of (6.14). Here, we show that the p-values defined using a non-maximal conditioning set $E_{12}(\mathbf{X}) \subset M_{12}(\mathbf{X})$ as (6.47) control the selective type I error for clustering (6.6). First, note that we have

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12}) \leq \alpha \,\middle|\, E_{12}(\mathbf{X}) \cap T_{12}(\mathbf{X})\right) = \alpha \tag{D.57}$$

following (6.47), for any $\alpha \in (0, 1)$. For simplicity, we will denote

$$A = \mathbb{1}\left\{p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12}) \leq \alpha\right\}. \tag{D.58}$$

Then, following a similar reasoning as in the proof of [104, Theorem 1] and the tower property of conditional expectation, we can write

$$\begin{aligned}
\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(p_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}; E_{12}) \leq \alpha \,\middle|\, M_{12}(\mathbf{X})\right) &= \mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(A \,\middle|\, M_{12}(\mathbf{X})\right) \\
&= \mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left[\mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(A \,\middle|\, M_{12}(\mathbf{X}) \cap E_{12}(\mathbf{X}) \cap T_{12}(\mathbf{X})\right) \,\middle|\, M_{12}(\mathbf{X})\right] \\
= \mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left[\mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(A \,\middle|\, E_{12}(\mathbf{X}) \cap T_{12}(\mathbf{X})\right) \,\middle|\, M_{12}(\mathbf{X})\right] &= \mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left[\alpha \,\middle|\, M_{12}(\mathbf{X})\right] = \alpha,
\end{aligned}$$

where the third equality follows from the fact $E_{12}(\mathbf{X}) \subset M_{12}(\mathbf{X})$ and the last equality follows from (D.57). □

D.4 Simulations of Sections 6.5.1 and 6.5.2 for further clustering algorithms

D.4.1 Uniform p-values under a global null hypothesis

Figure D.1 is the counterpart of Figure 6.1 for k-means and HC with centroid, single and complete linkage. As mentioned in Section 6.5.1, the encountered empirical distributions for the p-values (6.13) match the one of a uniform random variable in all cases, excluding HC with complete linkage and dependence setting (c) (panel (i) in Figure D.1). We postulate that the slight deviation from uniformity is an artefact coming from the noise that appears when simulating independent samples from an auto-regressive process. To illustrate so, we simulated M samples of size $n = 10$ drawn from a univariate AR(1) process with $\sigma = 1$ and $\rho = 0.9$, concatenated the M samples into a sample of size nM and computed its auto-correlation. Results are presented in Figure D.2 for $M \in \{10^3, 5 \cdot 10^3, 10^4, 5 \cdot 10^4\}$. They show how, when M is not large enough, the observed auto-correlation at lags higher than n exceeds the confidence intervals, although the corresponding observations have been independently simulated. Consequently, either large sample sizes or number of simulations are required to reduce the noise, that make the simulated p-values in Figure D.1(i) deviate from perfect independence and thus prevent their ECDF to converge to the CDF of a uniform random variable. The same effect is illustrated in Figure D.3, where the ECDF of the p-values (6.13) is displayed after performing HC with average linkage in the setting of Section 6.5.1, for the dependence scenario (c) and different number of simulations $K \in \{200, 500, 1000, 2000\}$. In Figure D.3 we observe how increasing the number of iterations -and thus reducing the noise illustrated in Figure D.2- makes the computed ECDF approximate to the diagonal. As it is appreciated in Figure D.1, the encountered noise seems to have a more substantial effect when p-values are computed by Monte Carlo approximation. Note that this phenomenon does not contradict the fact that p-values are uniformly distributed under the global null, but shows that in some cases the noise effect prevents us from correctly simulating their distribution.

D.4.2 Super-uniform p-values for unknown $\boldsymbol{\Sigma}$

Figure D.4 is the counterpart of Figure 6.1 for k-means and HC with centroid, single and complete linkage. As mentioned in Section 6.5.2, the encountered p-values (6.13) are stochastically larger than a uniform random variable in all cases. Note that the empirical distribution for HC with complete linkage and dependence setting (c) (panel (i) in Figure D.1) shows a more severe separation from the diagonal. This is explained due to the noise effect discussed in Section D.4.1. Regarding the simulation for k-means clustering, a larger sample size was needed to illustrate a super-uniform null distribution. We set $n = 1000$ and $\delta = \{10, 12\}$ in that case. For computational speeding-up we chose $p = 2$.

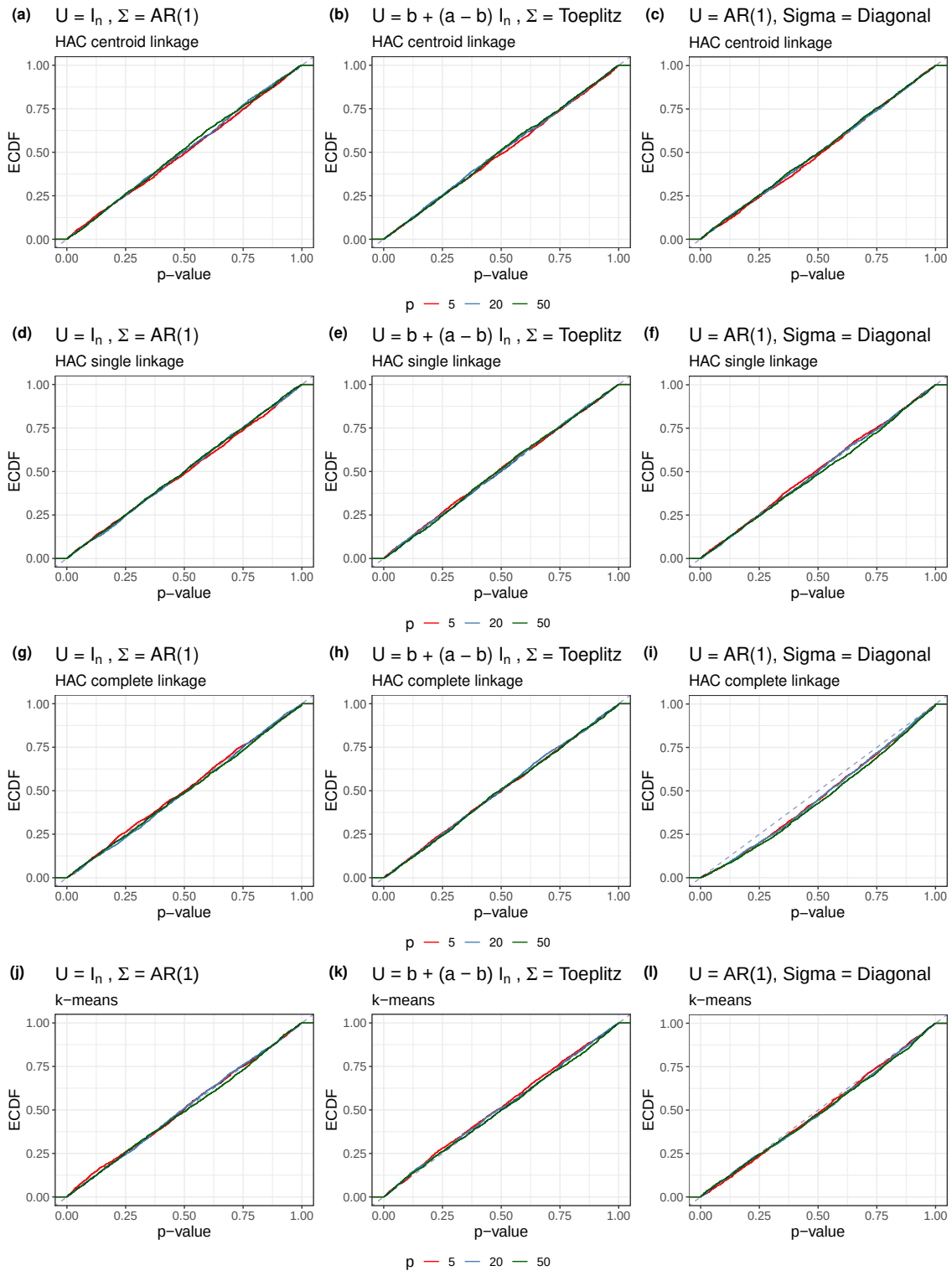


Figure D.1: Empirical cumulative distribution functions (ECDF) of p-values (6.13) with $\mathcal{C}$ being a hierarchical clustering algorithm (HC) with centroid (a-c), single (d-f) and complete (g-i) linkage and a k-means algorithm (j-l). The ECDF were computed from $K = 2000$ realizations of (6.2) under the three dependence settings (a), (b) and (c) with $\boldsymbol{\mu} = \mathbf{0}_{n\times p}$, $n = 100$ and $p \in \{5, 20, 50\}$.

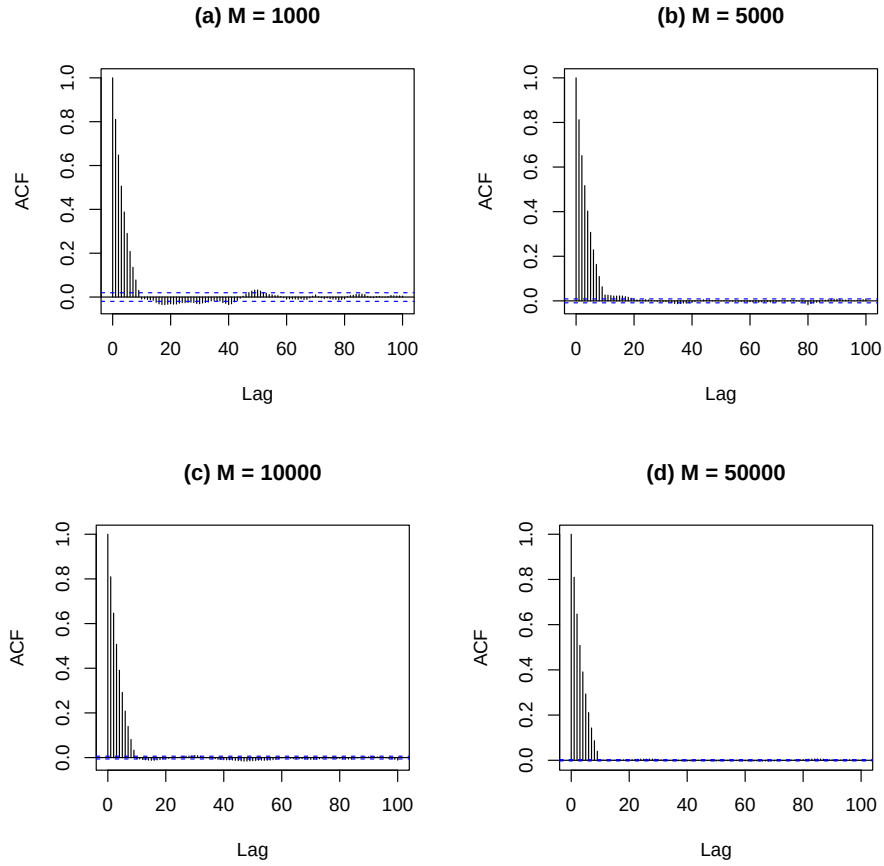


Figure D.2: Auto-correlation functions of M concatenated samples of size $n = 10$ drawn from an AR(1) process with $\sigma = 1$ and $\rho = 0.1$, as described in Section D.4.1.

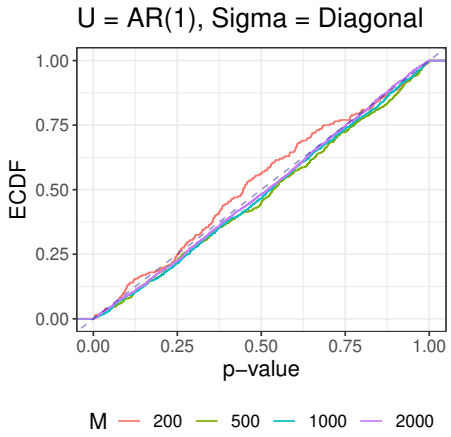


Figure D.3: Empirical cumulative distribution functions (ECDF) of p-values (6.13) computed from K iterations of hierarchical clustering with average linkage in the conditions described in Section D.4.1.

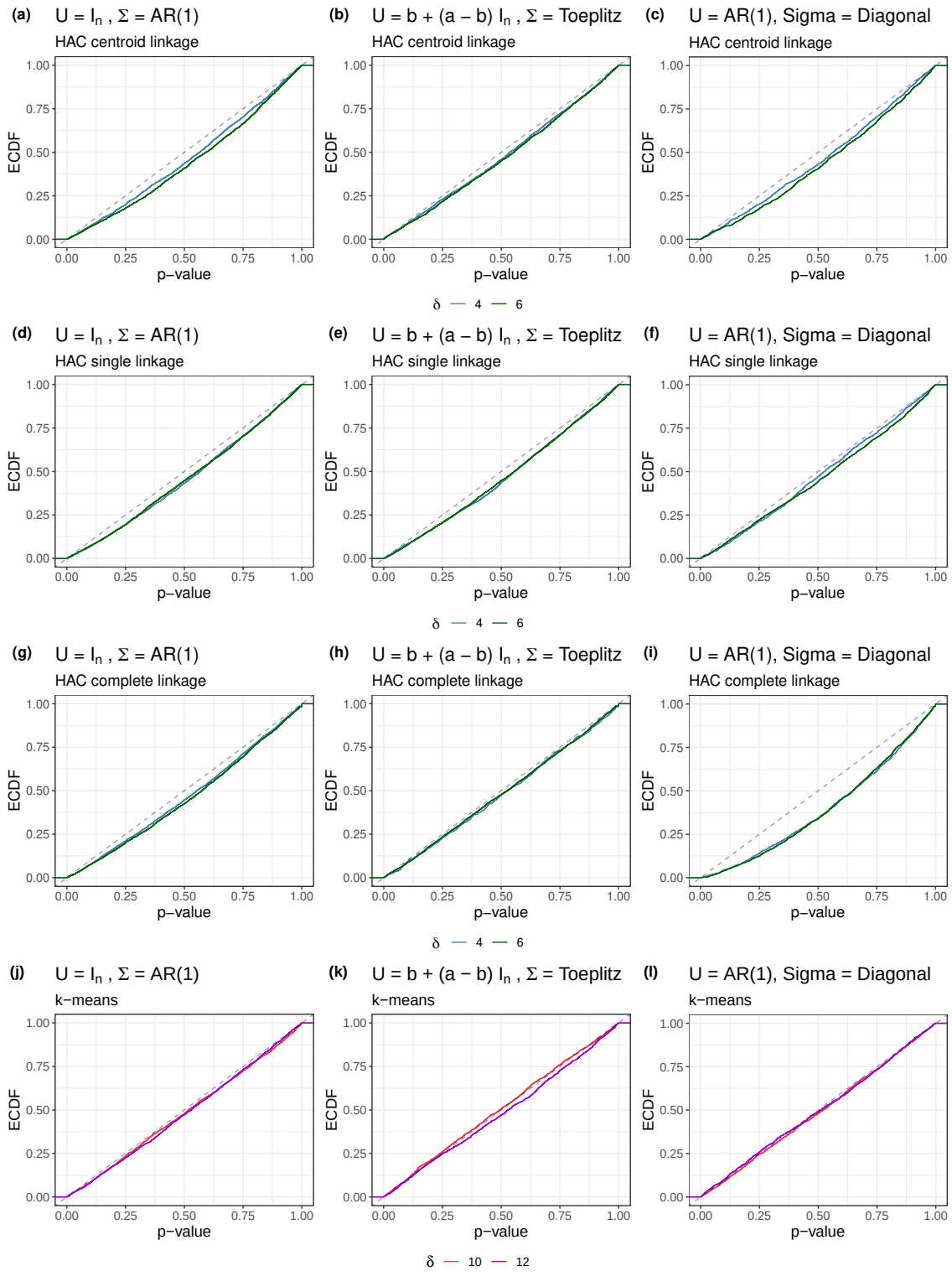


Figure D.4: Empirical cumulative distribution functions (ECDF) of p-values (6.13) with $\mathcal{C}$ being a hierarchical clustering algorithm with average linkage. The ECDF were computed from $K = 5000$ realizations of (6.2) under the three dependence settings (a), (b) and (c) with $n = 500$, $p = 10$ and $\boldsymbol{\mu}$ given by (6.50) with $\delta \in \{4, 6\}$. Only samples for which the null hypothesis held were kept, as described in Section 6.5.2.

Appendix E

Appendix of Chapter 7

Contents

E.1 UMAP and HDBSCAN algorithms

The Uniform Manifold Approximation and Projection (UMAP) algorithm was introduced in the very technical work [199], together with a more accessible and fully detailed documentation [198]. UMAP is a graph layout algorithm incorporating several theoretical foundations that provide it with a robust and well-established framework. Succintly, the UMAP algorithm builds a graph in the high dimensional space and then performs an optimization step to find the most similar graph in a lower dimension. UMAP begins by building balls centered at each point and connecting points whose corresponding balls overlap. This yields the representation of the dataset as a simplicial complex, that captures many of the main topological properties of the high-dimensional space [86]. To deal with the arbitrariness of the radius choice, the connections are made probabilistic and the edges of the graph are weighted. The resulting graphical representation is projected into a lower-dimensional space via a force-directed graph layout algorithm. The optimization procedure is similar to the one of t-SNE [294], but it effectively preserves a more substantial amount of global structure [310]. UMAP has found numerous applications in various domains, such as genetics [80, 82], single-cell [15, 272] or neuroimaging [211]. Besides, its popularity is steadily increasing due to its demonstrated empirical efficiency, especially in enhancing the performance of clustering methods [2].

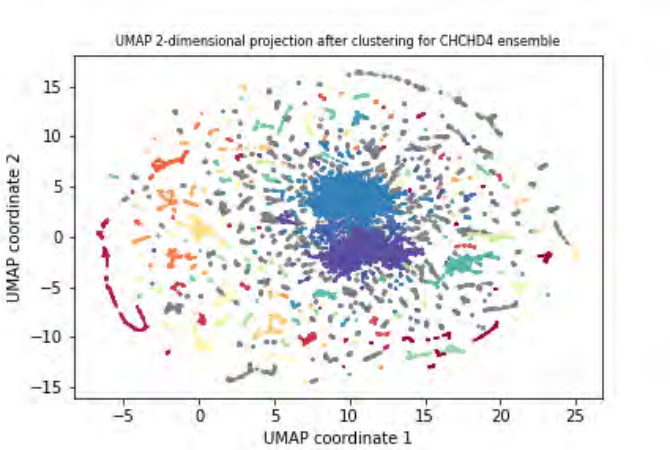
HDBSCAN [46] is a hierarchical version of the DBSCAN [89] clustering algorithm. It is a density-based method, so it performs better than classical distance-based techniques like k-means when clusters have arbitrary shapes and sizes, or in the presence of noise or outliers. The algorithm initially follows a similar approach to that of DBSCAN. It involves a density-based transformation of the space, akin to DBSCAN, and subsequently performs single linkage clustering on the transformed space. However, an alternative strategy is carried out to avoid the use of an epsilon value to define a cutoff level for the dendrogram, enabling the identification of the more stable or persistent clusters. Instead of the cutoff parameter, HDBSCAN needs the choice of the minimum cluster size, which is more intuitive and interpretable in practical scenarios. This, together with its remarkable computational efficiency, has made HDBSCAN a very popular algorithm often implemented in combination with dimensionality reduction techniques [7, 59, 82]. For a complete explanation of the algorithm details, we refer to the HDBSCAN documentation [197].

E.2 Results

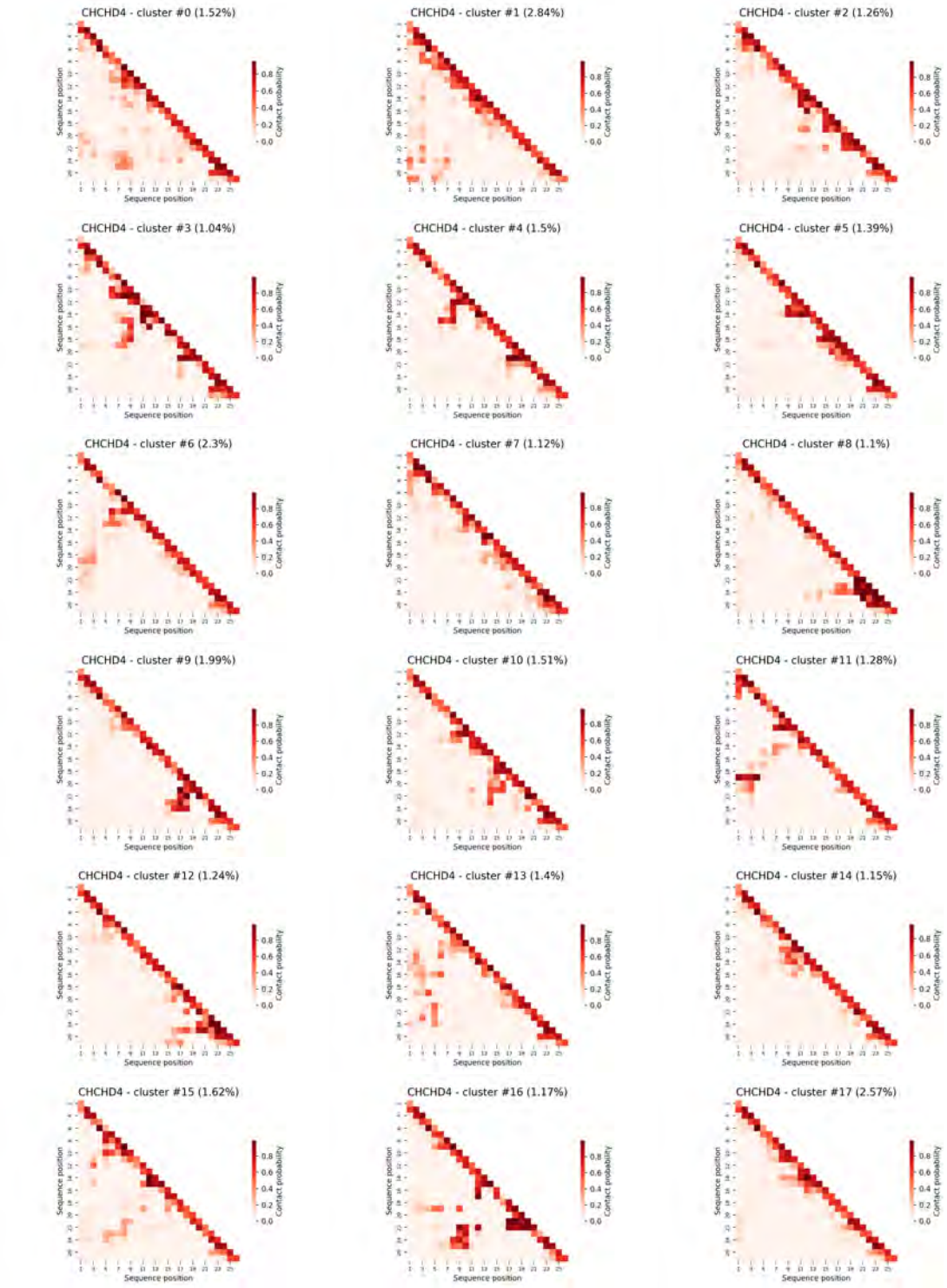
This Section presents the complete characterizations of the protein ensembles studied in Section 7.3. For each one of the examples, we first present the two-dimensional UMAP embedding of conformations featured by contact functions. Points are coloured according to the HDBSCAN classification, illustrating the overall distribution of the cluster occupancies. Then, the complete family of weighted ω-contact maps is presented for the ensemble. Finally, we show the secondary structure propensities for the conformations of each cluster, together with their average radius of gyration.

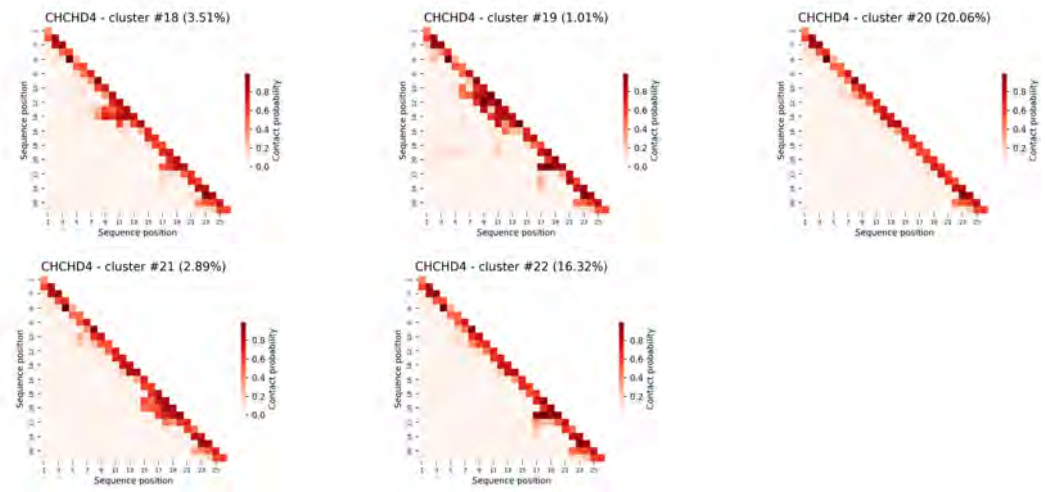
E.2.1 Complete characterization of CHCHD4

Two-dimensional UMAP projection

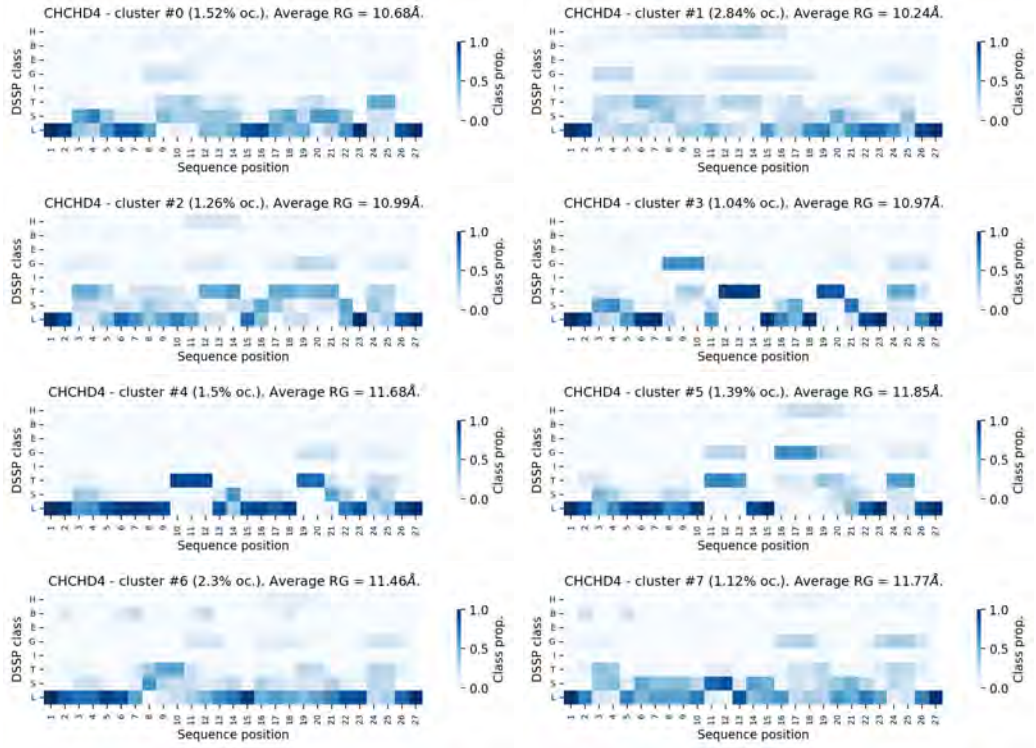


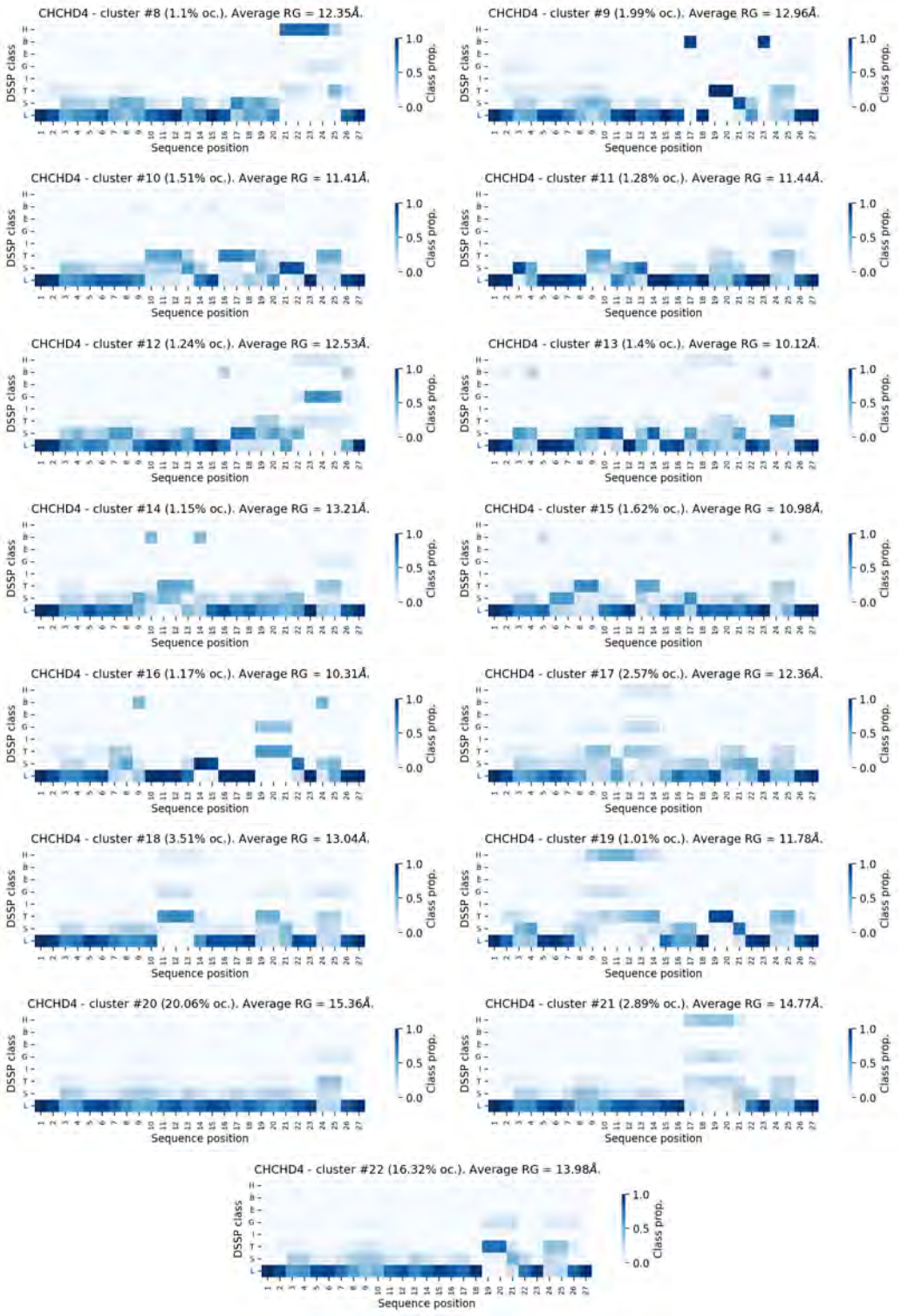
Complete family of weighted ω-contact maps





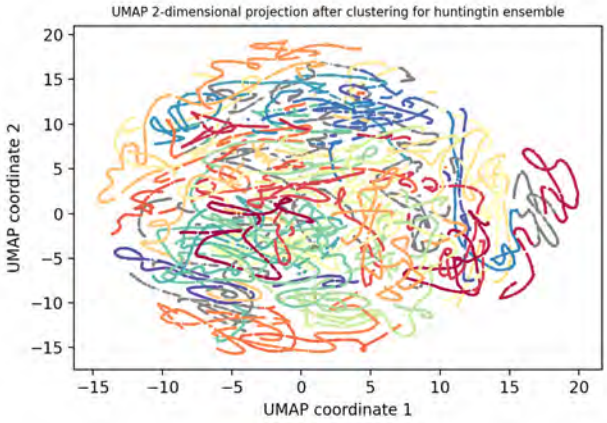
Secondary structure propensities and average radii of gyration



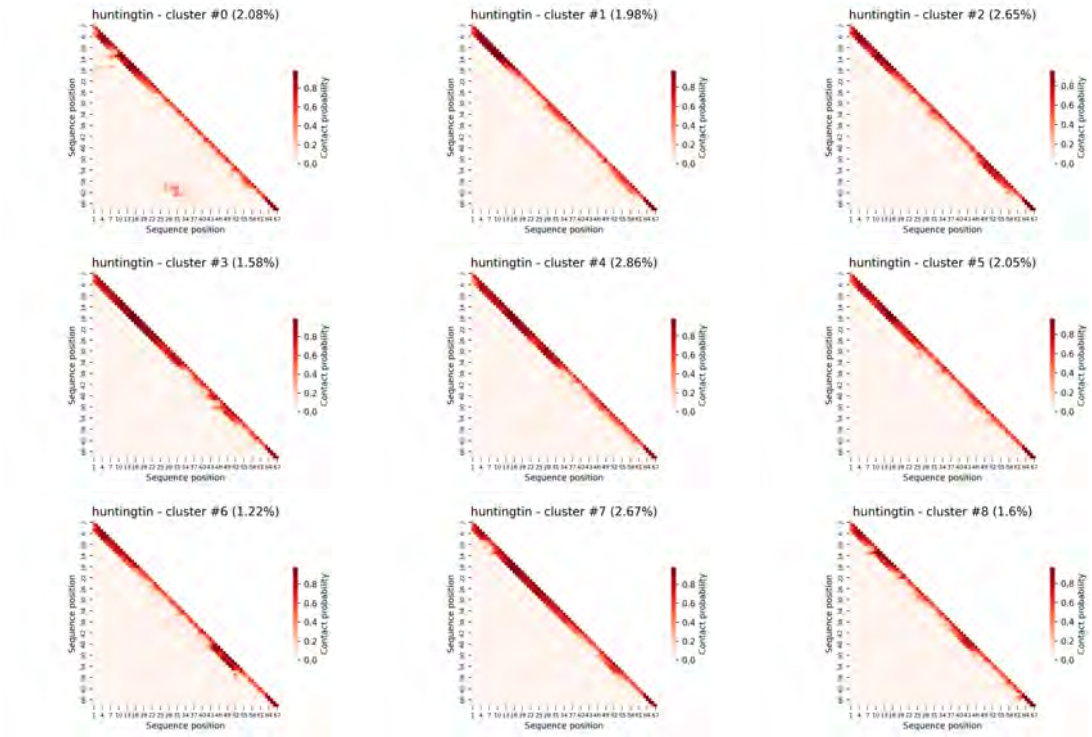

CHCHD4 - cluster #8 (1.1% oc.). Average RG = 12.35Å.
CHCHD4 - cluster #9 (1.99% oc.). Average RG = 12.96Å.
CHCHD4 - cluster #10 (1.51% oc.). Average RG = 11.41Å.
CHCHD4 - cluster #11 (1.28% oc.). Average RG = 11.44Å.
CHCHD4 - cluster #12 (1.24% oc.). Average RG = 12.53Å.
CHCHD4 - cluster #13 (1.4% oc.). Average RG = 10.12Å.
CHCHD4 - cluster #14 (1.15% oc.). Average RG = 13.21Å.
CHCHD4 - cluster #15 (1.62% oc.). Average RG = 10.98Å.
CHCHD4 - cluster #16 (1.17% oc.). Average RG = 10.31Å.
CHCHD4 - cluster #17 (2.57% oc.). Average RG = 12.36Å.
CHCHD4 - cluster #18 (3.51% oc.). Average RG = 13.04Å.
CHCHD4 - cluster #19 (1.01% oc.). Average RG = 11.78Å.
CHCHD4 - cluster #20 (20.06% oc.). Average RG = 15.36Å.
CHCHD4 - cluster #21 (2.89% oc.). Average RG = 14.77Å.
CHCHD4 - cluster #22 (16.32% oc.). Average RG = 13.98Å.
DSSP class
Sequence position
Class prop.

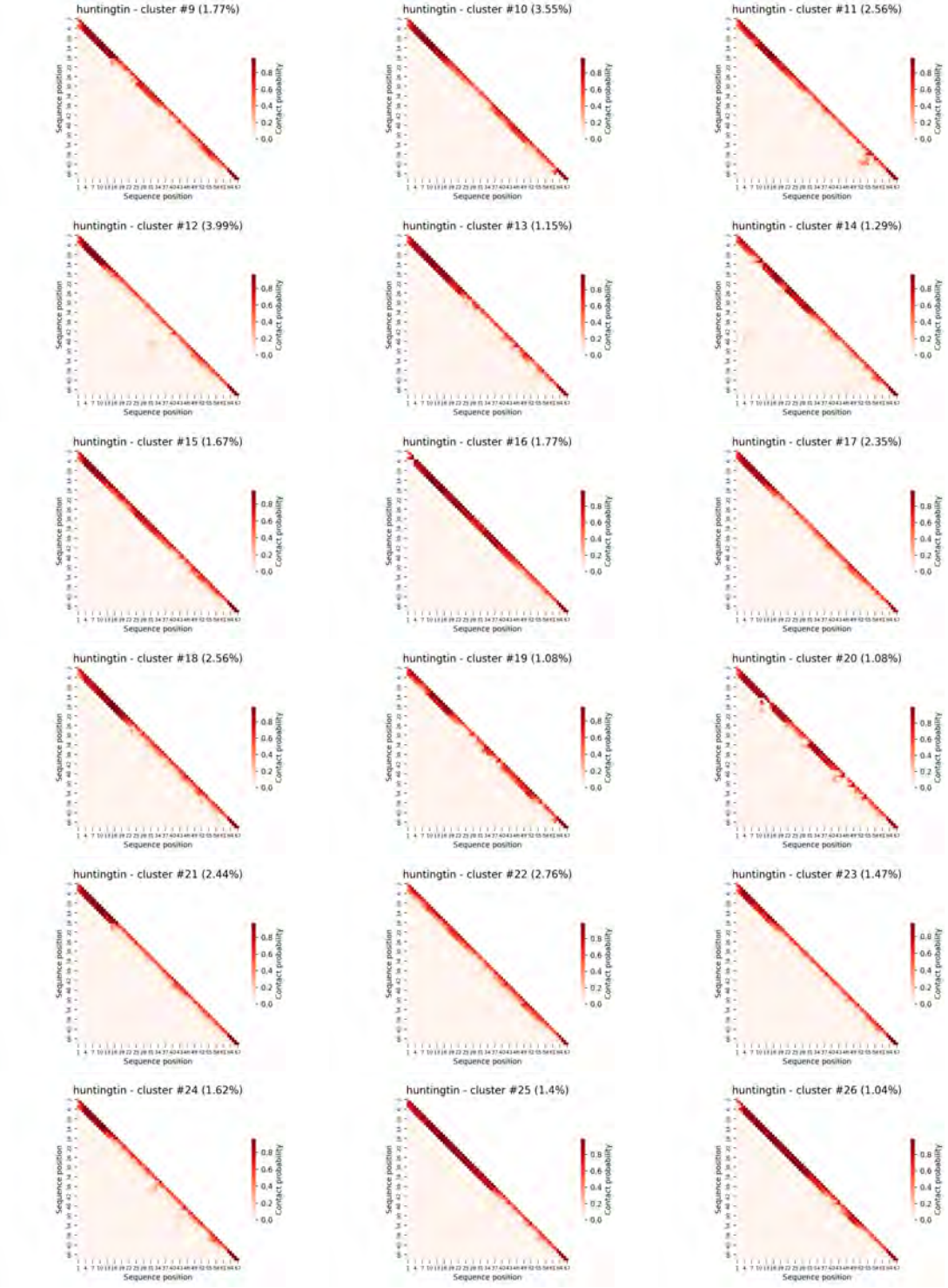
E.2.2 Complete characterization of Huntingtin

Two-dimensional UMAP projection

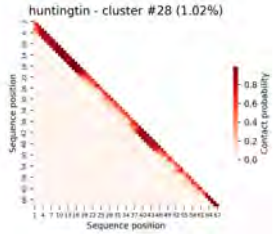


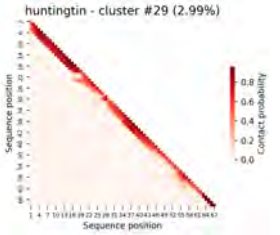
Complete family of weighted ω-contact maps

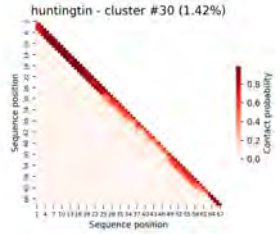


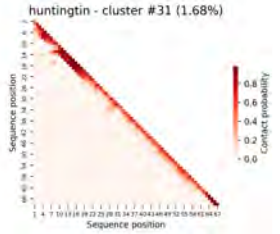

huntingtin - cluster #9 (1.77%)
huntingtin - cluster #10 (3.55%)
huntingtin - cluster #11 (2.56%)
huntingtin - cluster #12 (3.99%)
huntingtin - cluster #13 (1.15%)
huntingtin - cluster #14 (1.29%)
huntingtin - cluster #15 (1.67%)
huntingtin - cluster #16 (1.77%)
huntingtin - cluster #17 (2.35%)
huntingtin - cluster #18 (2.56%)
huntingtin - cluster #19 (1.08%)
huntingtin - cluster #20 (1.08%)
huntingtin - cluster #21 (2.44%)
huntingtin - cluster #22 (2.76%)
huntingtin - cluster #23 (1.47%)
huntingtin - cluster #24 (1.62%)
huntingtin - cluster #25 (1.4%)
huntingtin - cluster #26 (1.04%)
Sequence position
Contact probability


huntingtin - cluster #27 (2.37%)
Sequence position
Sequence position
Contact probability

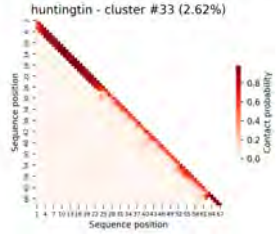

huntingtin - cluster #28 (1.02%)
Sequence position
Sequence position
Contact probability

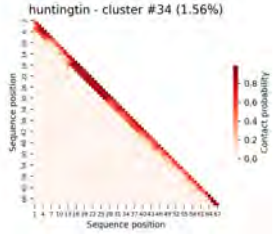

huntingtin - cluster #29 (2.99%)
Sequence position
Sequence position
Contact probability

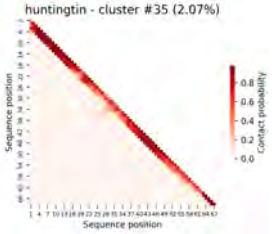

huntingtin - cluster #30 (1.42%)
Sequence position
Sequence position
Contact probability

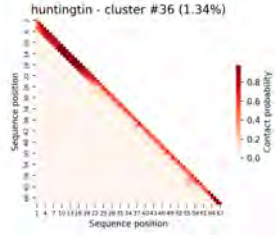

huntingtin - cluster #31 (1.68%)
Sequence position
Sequence position
Contact probability

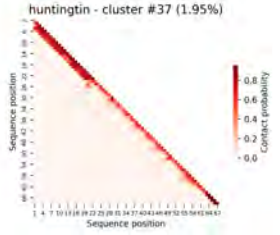

huntingtin - cluster #32 (2.93%)
Sequence position
Sequence position
Contact probability

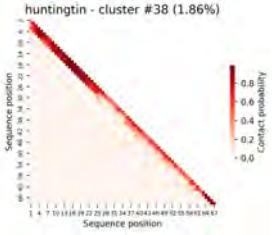

huntingtin - cluster #33 (2.62%)
Sequence position
Sequence position
Contact probability

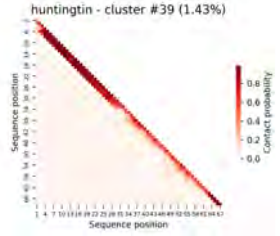

huntingtin - cluster #34 (1.56%)
Sequence position
Sequence position
Contact probability

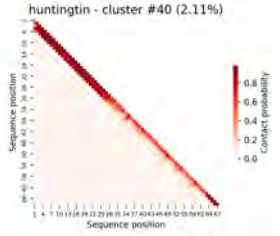

huntingtin - cluster #35 (2.07%)
Sequence position
Sequence position
Contact probability

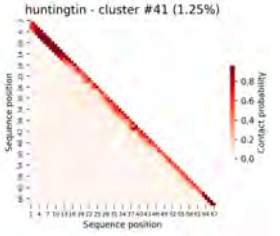

huntingtin - cluster #36 (1.34%)
Sequence position
Sequence position
Contact probability


huntingtin - cluster #37 (1.95%)
Sequence position
Sequence position
Contact probability


huntingtin - cluster #38 (1.86%)
Sequence position
Sequence position
Contact probability

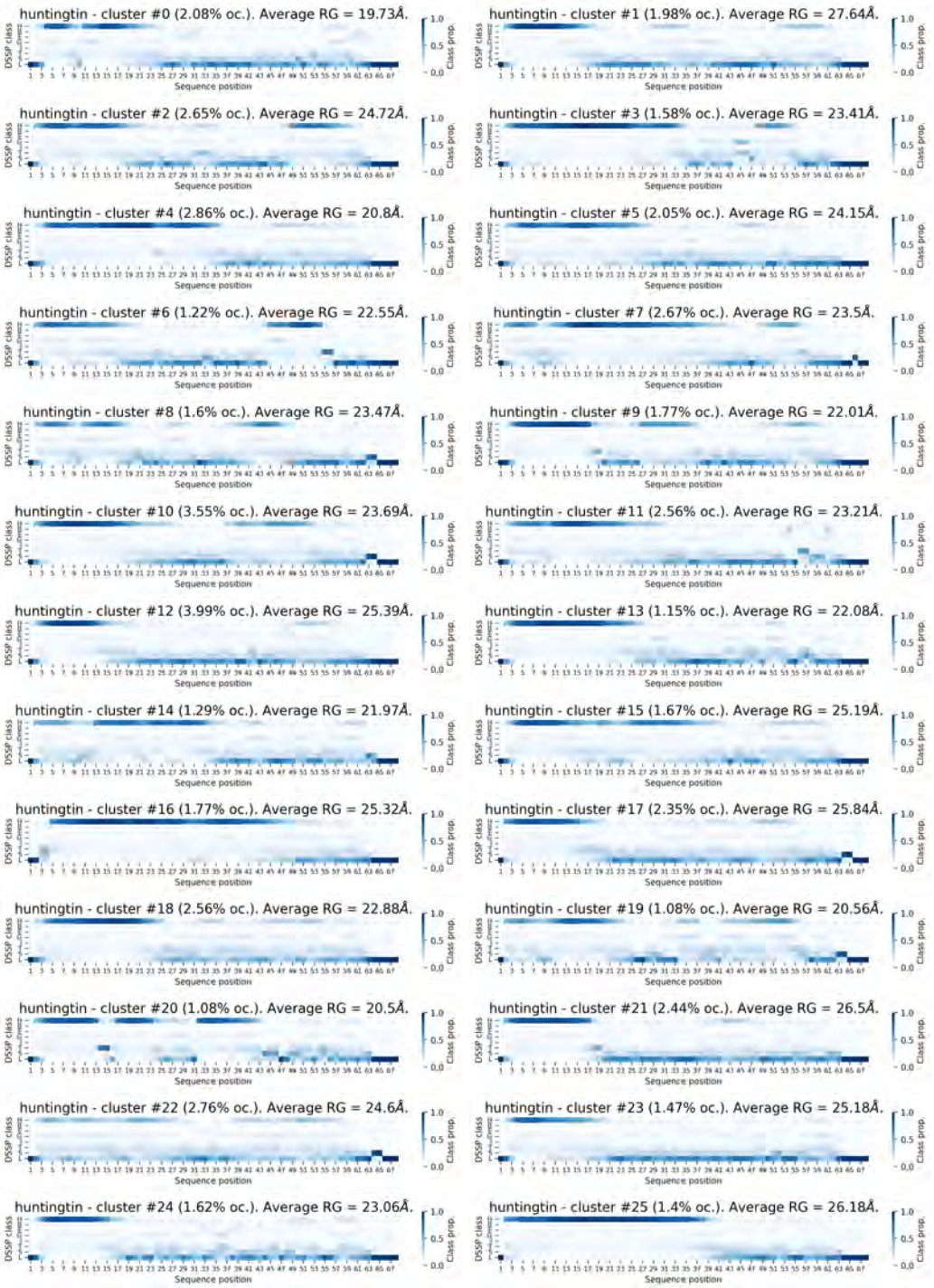

huntingtin - cluster #39 (1.43%)
Sequence position
Sequence position
Contact probability


huntingtin - cluster #40 (2.11%)
Sequence position
Sequence position
Contact probability


huntingtin - cluster #41 (1.25%)
Sequence position
Sequence position
Contact probability


huntingtin - cluster #42 (2.36%)
Sequence position
Sequence position
Contact probability

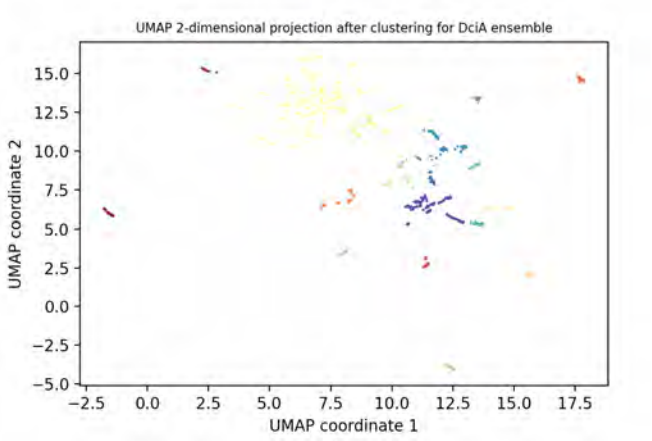
Secondary structure propensities and average radii of gyration



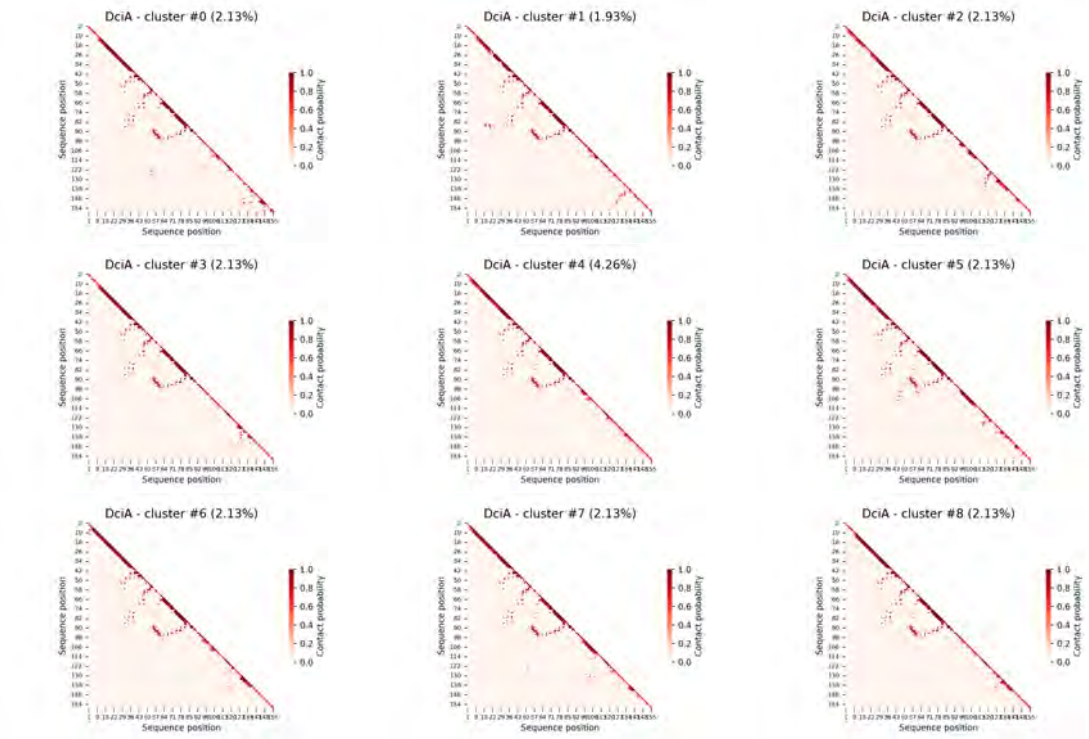

huntingtin - cluster #26 (1.04% oc.). Average RG = 25.6Å.
huntingtin - cluster #27 (2.37% oc.). Average RG = 24.55Å.
huntingtin - cluster #28 (1.02% oc.). Average RG = 21.47Å.
huntingtin - cluster #29 (2.99% oc.). Average RG = 24.16Å.
huntingtin - cluster #30 (1.42% oc.). Average RG = 22.56Å.
huntingtin - cluster #31 (1.68% oc.). Average RG = 22.56Å.
huntingtin - cluster #32 (2.93% oc.). Average RG = 23.73Å.
huntingtin - cluster #33 (2.62% oc.). Average RG = 20.86Å.
huntingtin - cluster #34 (1.56% oc.). Average RG = 22.75Å.
huntingtin - cluster #35 (2.07% oc.). Average RG = 22.97Å.
huntingtin - cluster #36 (1.34% oc.). Average RG = 26.95Å.
huntingtin - cluster #37 (1.95% oc.). Average RG = 22.75Å.
huntingtin - cluster #38 (1.86% oc.). Average RG = 20.55Å.
huntingtin - cluster #39 (1.43% oc.). Average RG = 23.28Å.
huntingtin - cluster #40 (2.11% oc.). Average RG = 24.75Å.
huntingtin - cluster #41 (1.25% oc.). Average RG = 21.73Å.
huntingtin - cluster #42 (2.36% oc.). Average RG = 26.34Å.
DSSP class
Sequence position
Class prop.

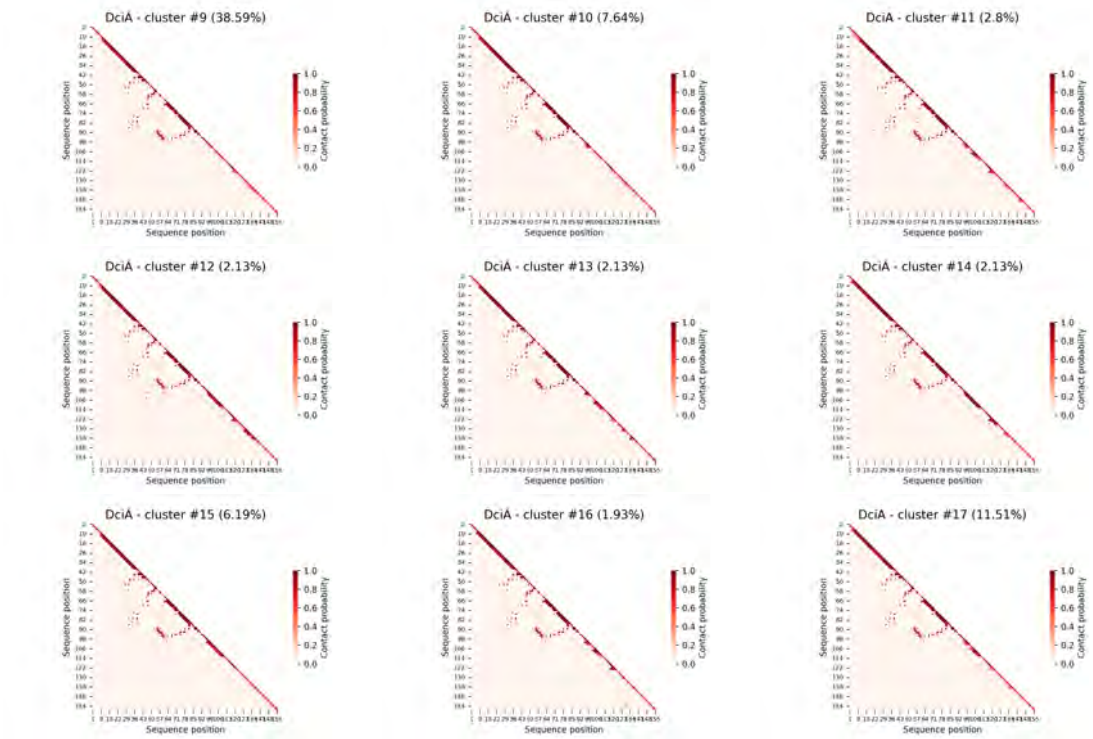
E.2.3 Complete characterization of DciA

Two-dimensional UMAP projection

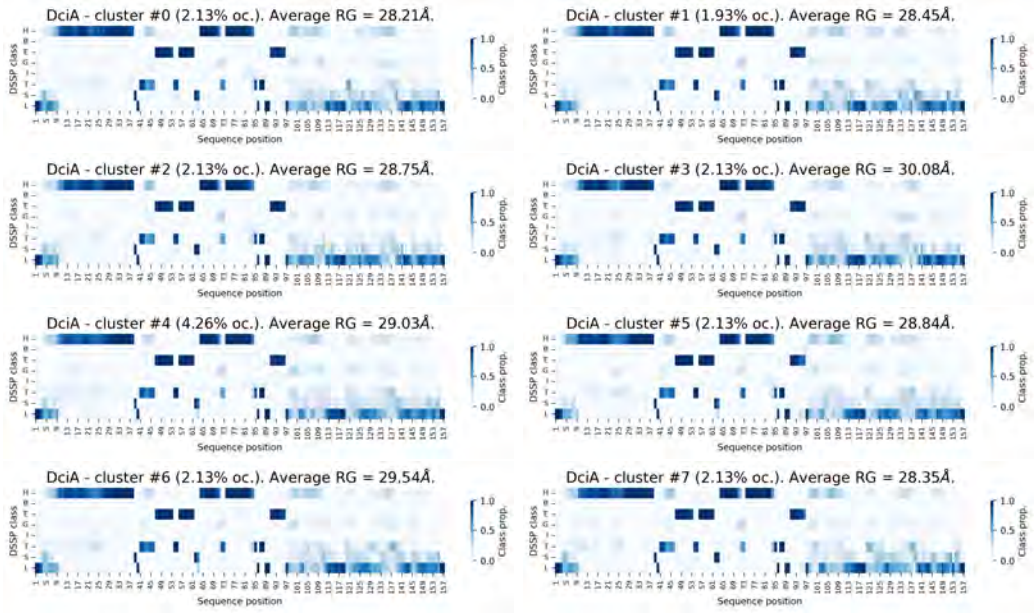


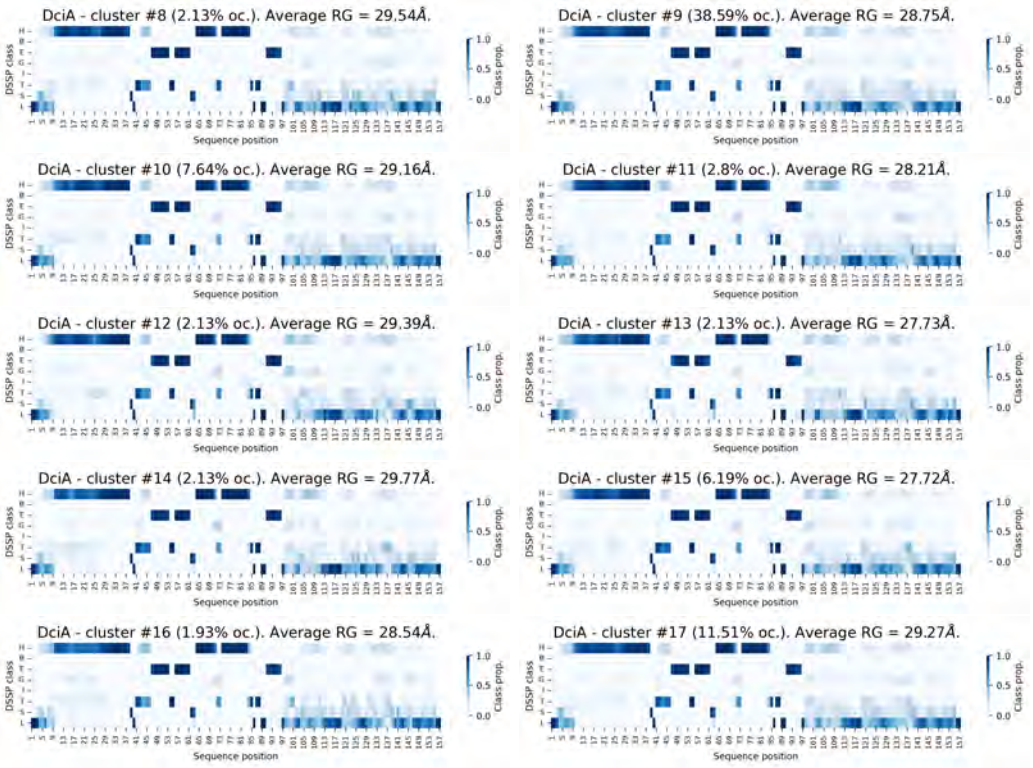
Complete family of weighted ω-contact maps





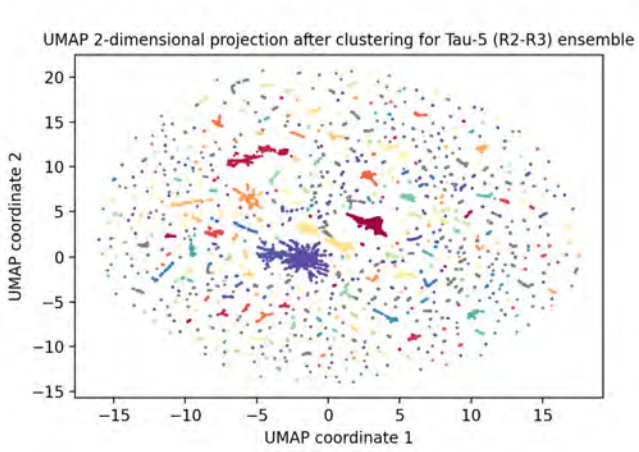
Secondary structure propensities and average radii of gyration



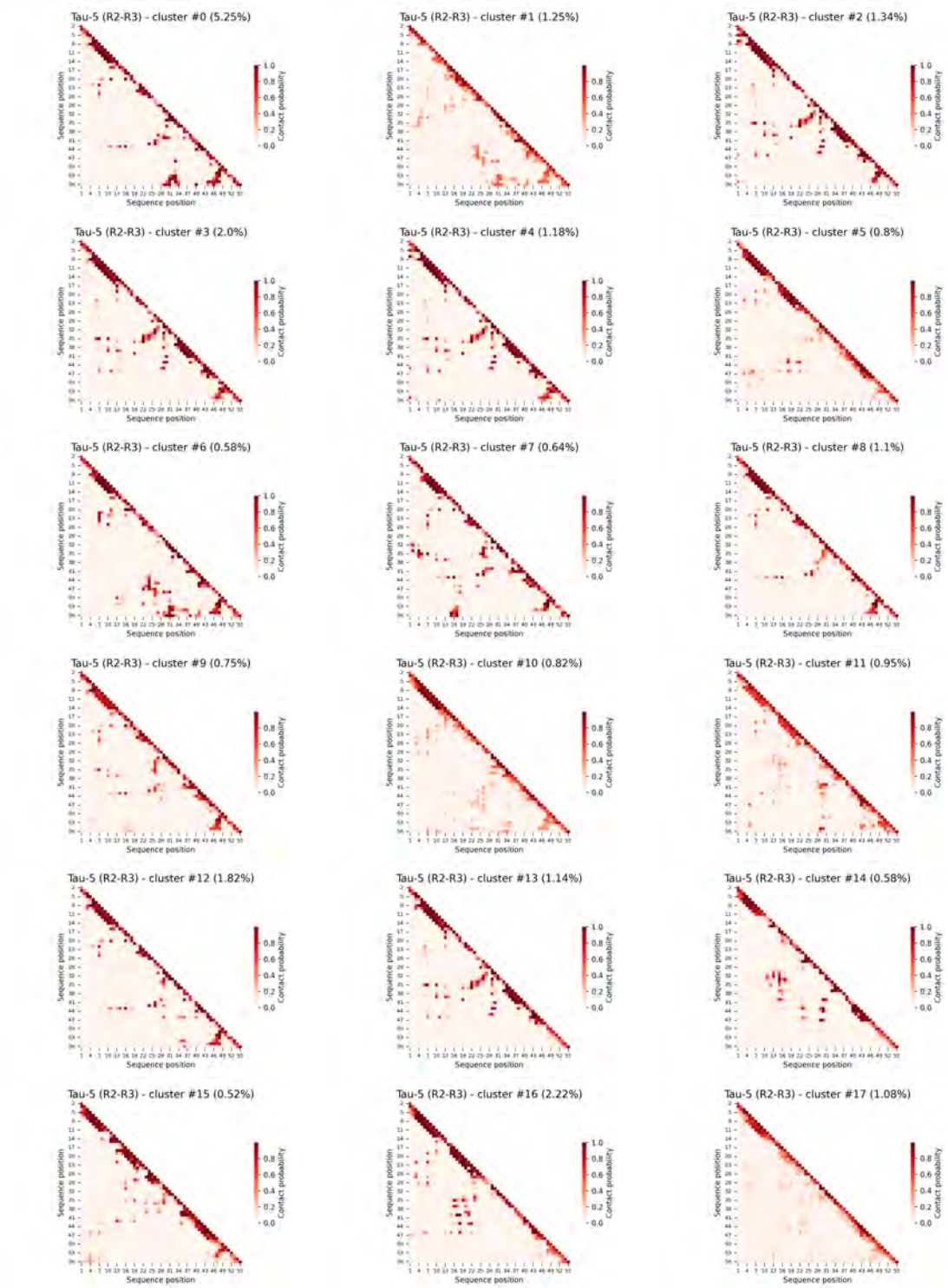


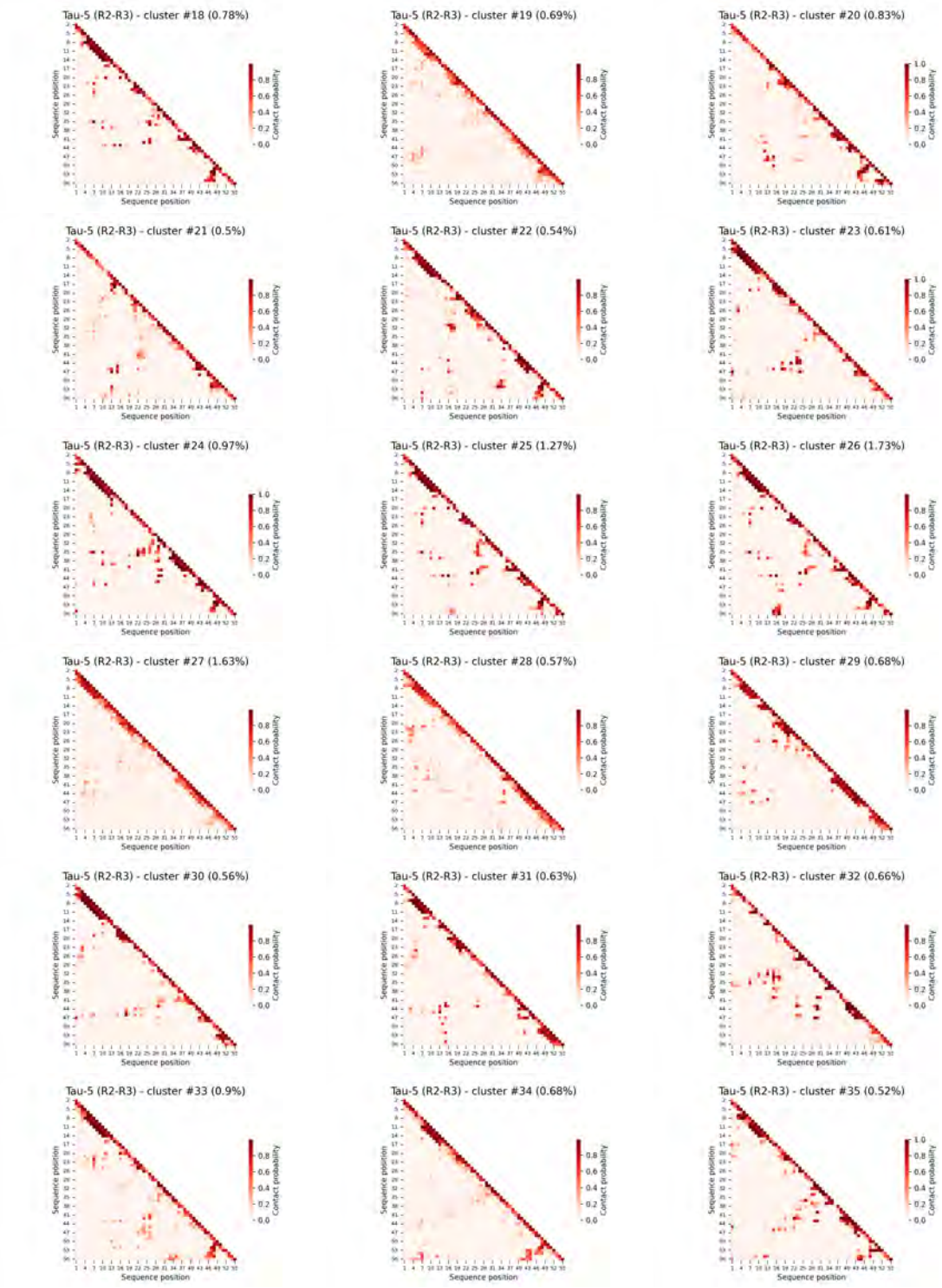
E.2.4 Complete characterization of Tau-5$_{\mathrm{R2-R3}}$

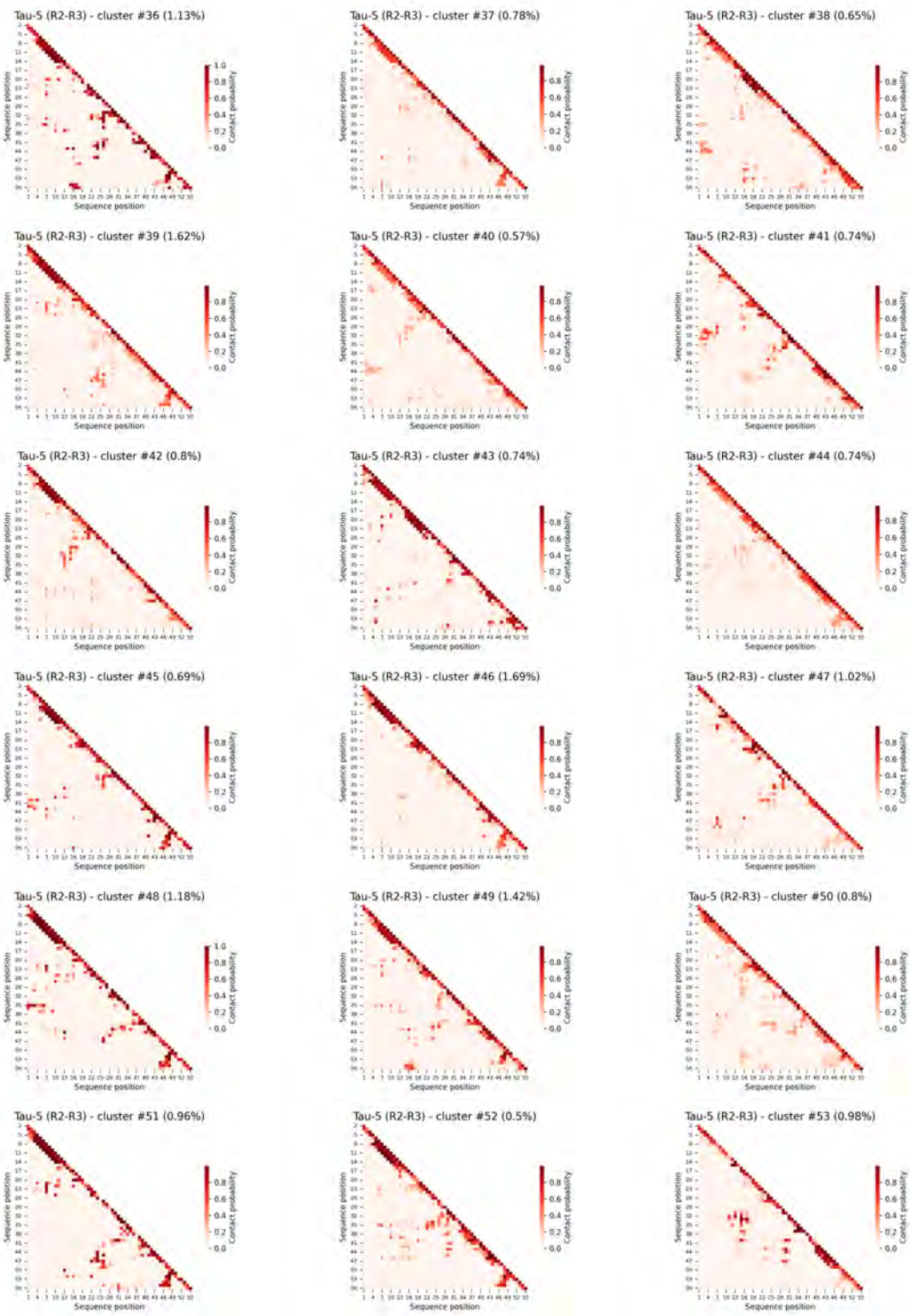
Two-dimensional UMAP projection

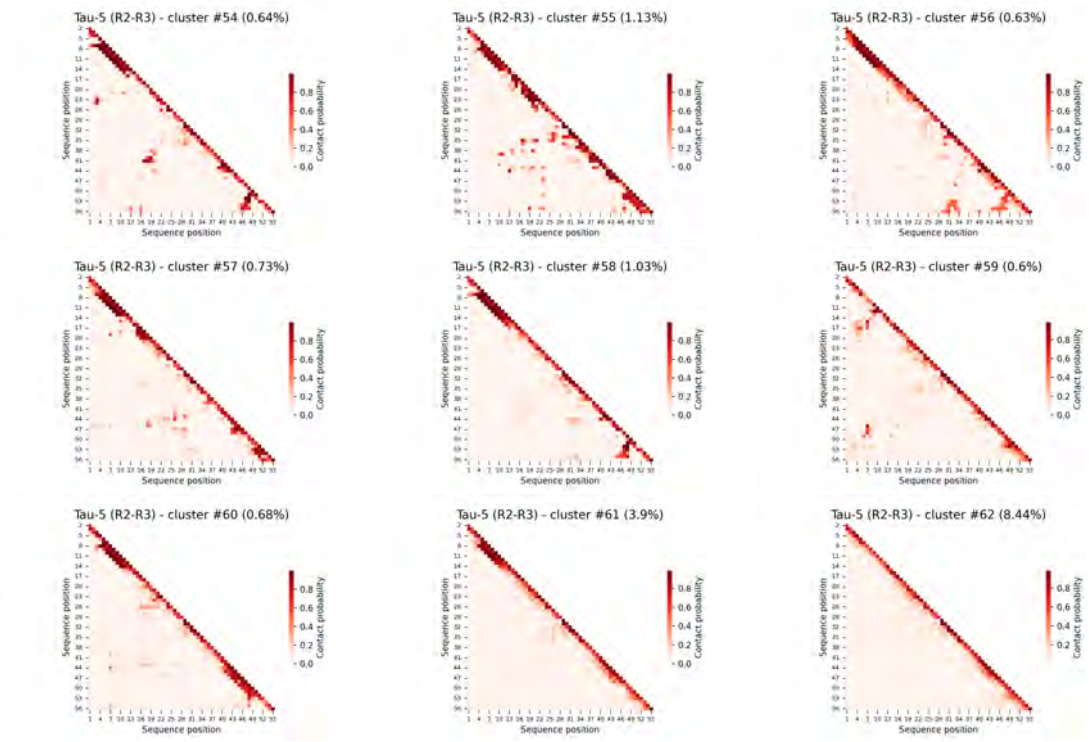


Complete family of weighted ω-contact maps

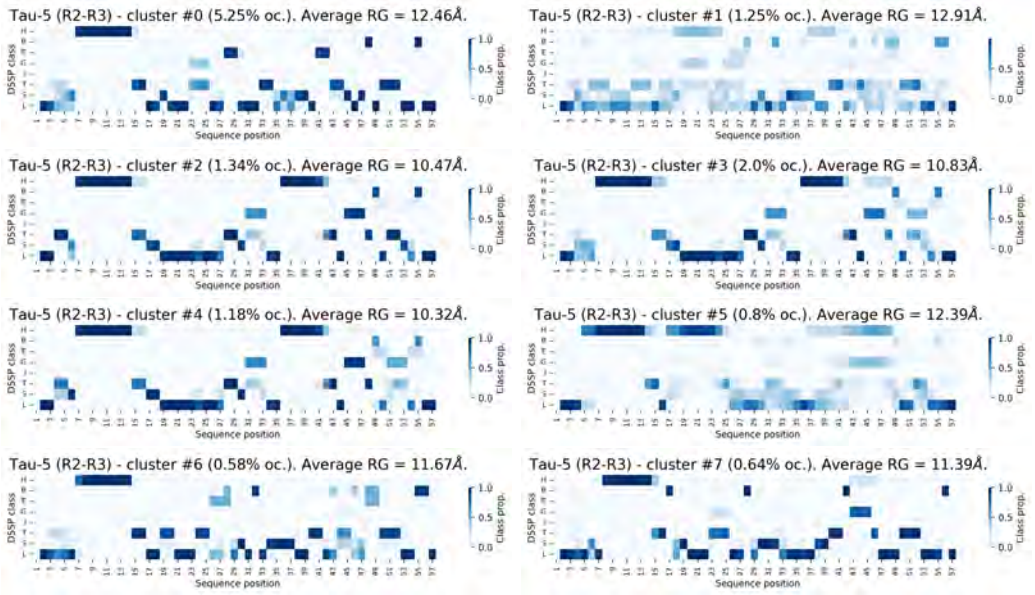


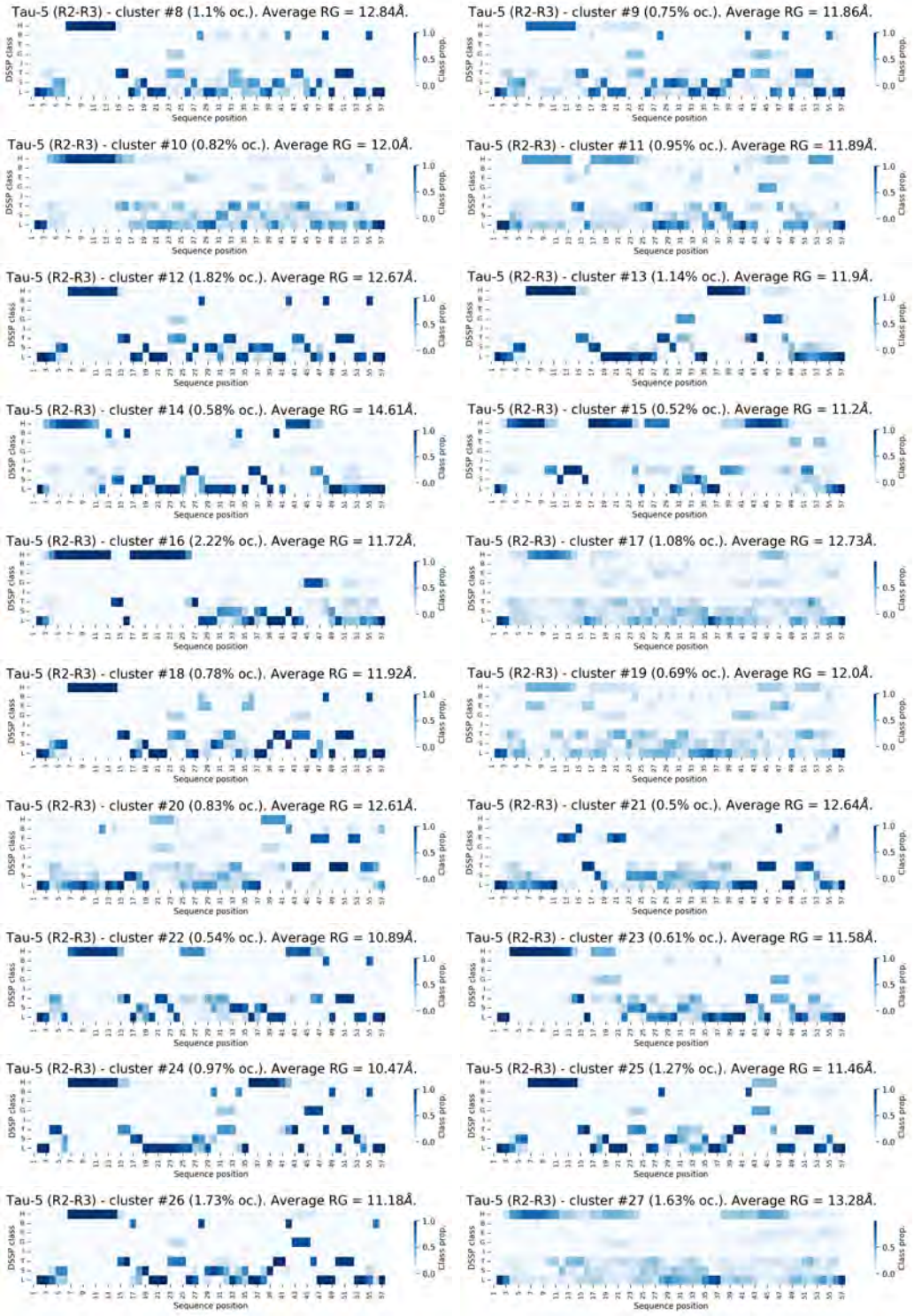

Tau-5 (R2-R3) - cluster #18 (0.78%)
Tau-5 (R2-R3) - cluster #19 (0.69%)
Tau-5 (R2-R3) - cluster #20 (0.83%)
Tau-5 (R2-R3) - cluster #21 (0.5%)
Tau-5 (R2-R3) - cluster #22 (0.54%)
Tau-5 (R2-R3) - cluster #23 (0.61%)
Tau-5 (R2-R3) - cluster #24 (0.97%)
Tau-5 (R2-R3) - cluster #25 (1.27%)
Tau-5 (R2-R3) - cluster #26 (1.73%)
Tau-5 (R2-R3) - cluster #27 (1.63%)
Tau-5 (R2-R3) - cluster #28 (0.57%)
Tau-5 (R2-R3) - cluster #29 (0.68%)
Tau-5 (R2-R3) - cluster #30 (0.56%)
Tau-5 (R2-R3) - cluster #31 (0.63%)
Tau-5 (R2-R3) - cluster #32 (0.66%)
Tau-5 (R2-R3) - cluster #33 (0.9%)
Tau-5 (R2-R3) - cluster #34 (0.68%)
Tau-5 (R2-R3) - cluster #35 (0.52%)
Sequence position
Contact probability

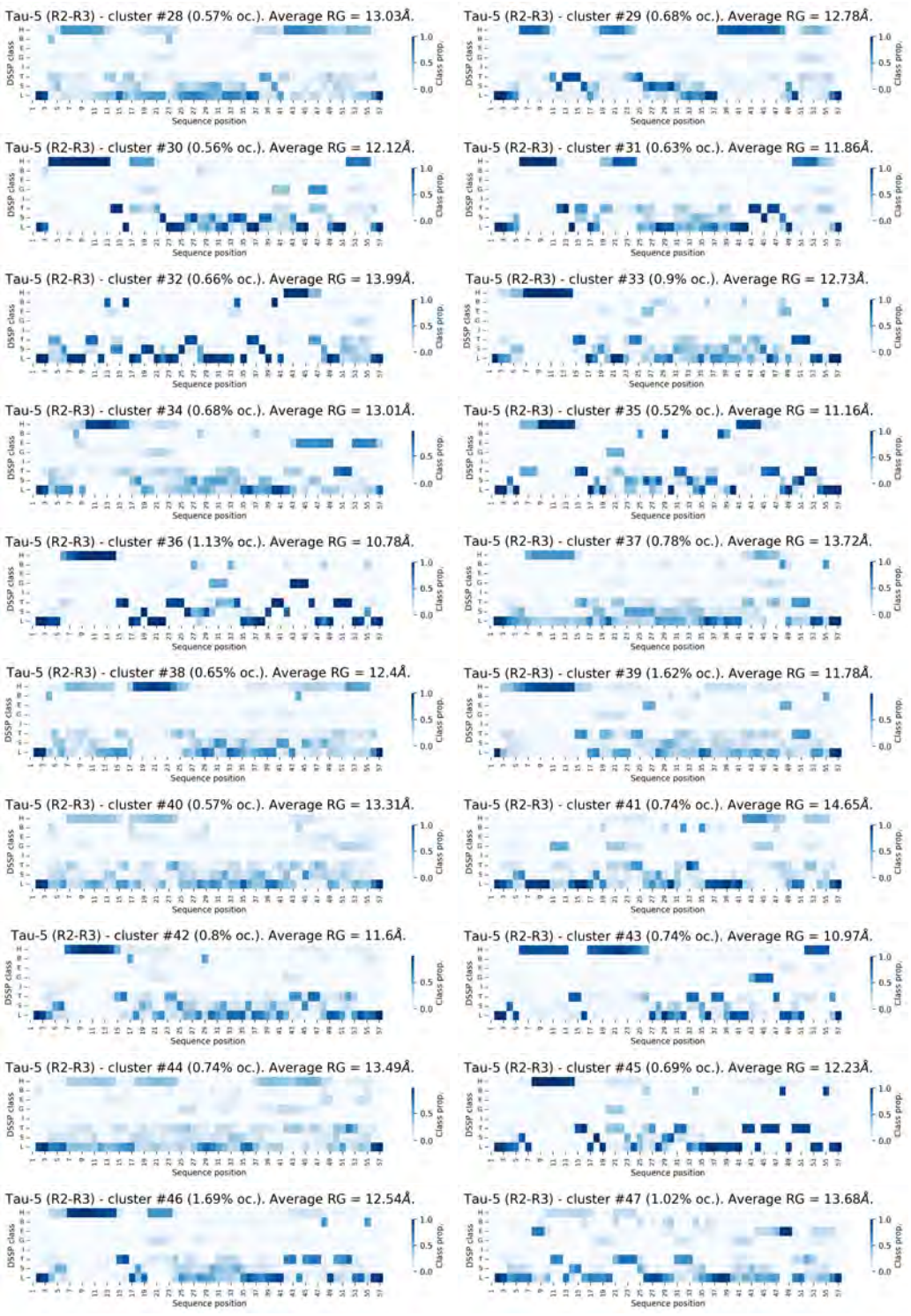

Tau-5 (R2-R3) - cluster #36 (1.13%)
Tau-5 (R2-R3) - cluster #37 (0.78%)
Tau-5 (R2-R3) - cluster #38 (0.65%)
Tau-5 (R2-R3) - cluster #39 (1.62%)
Tau-5 (R2-R3) - cluster #40 (0.57%)
Tau-5 (R2-R3) - cluster #41 (0.74%)
Tau-5 (R2-R3) - cluster #42 (0.8%)
Tau-5 (R2-R3) - cluster #43 (0.74%)
Tau-5 (R2-R3) - cluster #44 (0.74%)
Tau-5 (R2-R3) - cluster #45 (0.69%)
Tau-5 (R2-R3) - cluster #46 (1.69%)
Tau-5 (R2-R3) - cluster #47 (1.02%)
Tau-5 (R2-R3) - cluster #48 (1.18%)
Tau-5 (R2-R3) - cluster #49 (1.42%)
Tau-5 (R2-R3) - cluster #50 (0.8%)
Tau-5 (R2-R3) - cluster #51 (0.96%)
Tau-5 (R2-R3) - cluster #52 (0.5%)
Tau-5 (R2-R3) - cluster #53 (0.98%)
Sequence position
Contact probability

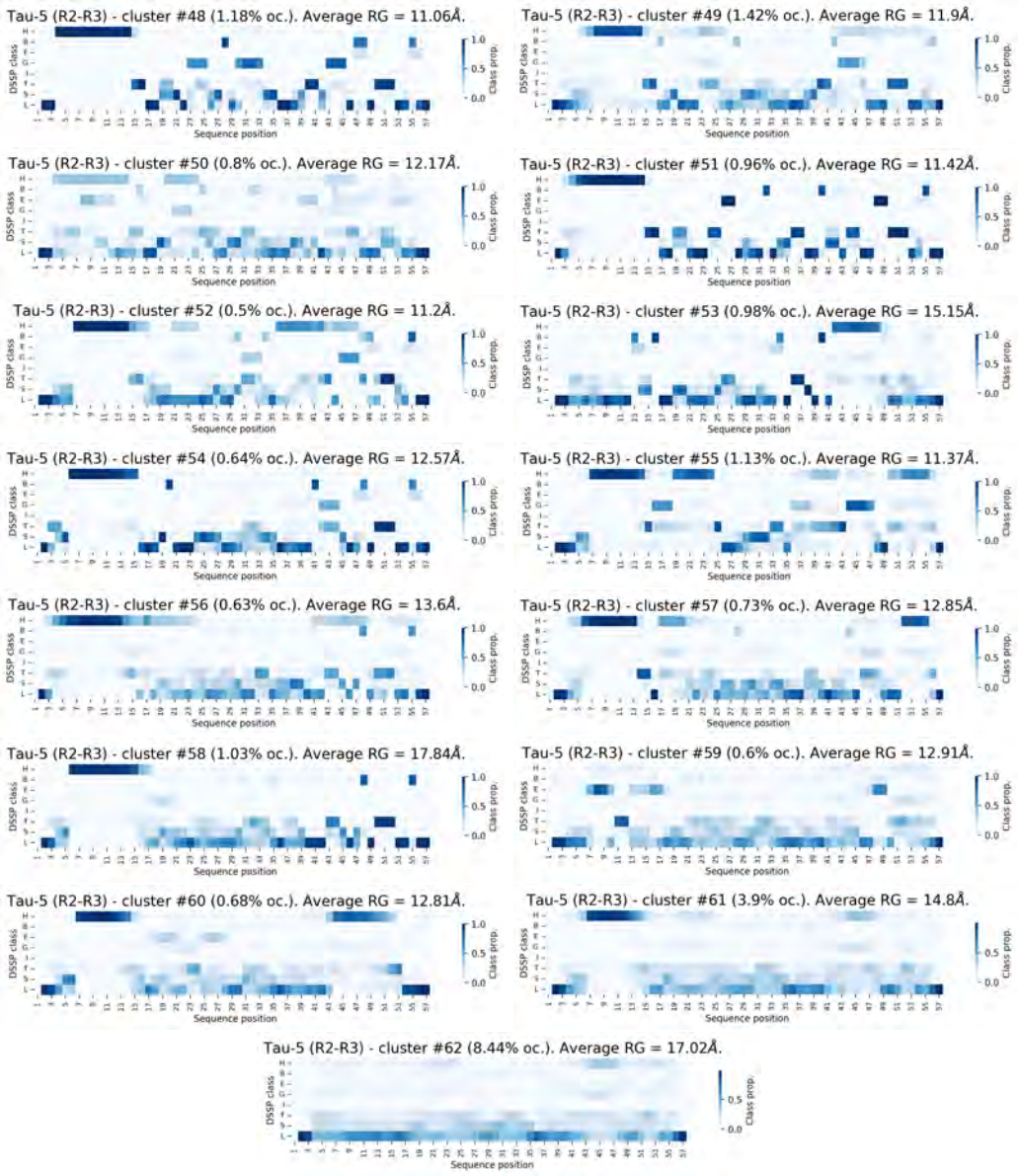


Secondary structure propensities and average radii of gyration




Tau-5 (R2-R3) - cluster #8 (1.1% oc.). Average RG = 12.84Å.
Tau-5 (R2-R3) - cluster #9 (0.75% oc.). Average RG = 11.86Å.
Tau-5 (R2-R3) - cluster #10 (0.82% oc.). Average RG = 12.0Å.
Tau-5 (R2-R3) - cluster #11 (0.95% oc.). Average RG = 11.89Å.
Tau-5 (R2-R3) - cluster #12 (1.82% oc.). Average RG = 12.67Å.
Tau-5 (R2-R3) - cluster #13 (1.14% oc.). Average RG = 11.9Å.
Tau-5 (R2-R3) - cluster #14 (0.58% oc.). Average RG = 14.61Å.
Tau-5 (R2-R3) - cluster #15 (0.52% oc.). Average RG = 11.2Å.
Tau-5 (R2-R3) - cluster #16 (2.22% oc.). Average RG = 11.72Å.
Tau-5 (R2-R3) - cluster #17 (1.08% oc.). Average RG = 12.73Å.
Tau-5 (R2-R3) - cluster #18 (0.78% oc.). Average RG = 11.92Å.
Tau-5 (R2-R3) - cluster #19 (0.69% oc.). Average RG = 12.0Å.
Tau-5 (R2-R3) - cluster #20 (0.83% oc.). Average RG = 12.61Å.
Tau-5 (R2-R3) - cluster #21 (0.5% oc.). Average RG = 12.64Å.
Tau-5 (R2-R3) - cluster #22 (0.54% oc.). Average RG = 10.89Å.
Tau-5 (R2-R3) - cluster #23 (0.61% oc.). Average RG = 11.58Å.
Tau-5 (R2-R3) - cluster #24 (0.97% oc.). Average RG = 10.47Å.
Tau-5 (R2-R3) - cluster #25 (1.27% oc.). Average RG = 11.46Å.
Tau-5 (R2-R3) - cluster #26 (1.73% oc.). Average RG = 11.18Å.
Tau-5 (R2-R3) - cluster #27 (1.63% oc.). Average RG = 13.28Å.
DSSP class
Sequence position
Class prop.


Tau-5 (R2-R3) - cluster #28 (0.57% oc.). Average RG = 13.03Å.
Tau-5 (R2-R3) - cluster #29 (0.68% oc.). Average RG = 12.78Å.
Tau-5 (R2-R3) - cluster #30 (0.56% oc.). Average RG = 12.12Å.
Tau-5 (R2-R3) - cluster #31 (0.63% oc.). Average RG = 11.86Å.
Tau-5 (R2-R3) - cluster #32 (0.66% oc.). Average RG = 13.99Å.
Tau-5 (R2-R3) - cluster #33 (0.9% oc.). Average RG = 12.73Å.
Tau-5 (R2-R3) - cluster #34 (0.68% oc.). Average RG = 13.01Å.
Tau-5 (R2-R3) - cluster #35 (0.52% oc.). Average RG = 11.16Å.
Tau-5 (R2-R3) - cluster #36 (1.13% oc.). Average RG = 10.78Å.
Tau-5 (R2-R3) - cluster #37 (0.78% oc.). Average RG = 13.72Å.
Tau-5 (R2-R3) - cluster #38 (0.65% oc.). Average RG = 12.4Å.
Tau-5 (R2-R3) - cluster #39 (1.62% oc.). Average RG = 11.78Å.
Tau-5 (R2-R3) - cluster #40 (0.57% oc.). Average RG = 13.31Å.
Tau-5 (R2-R3) - cluster #41 (0.74% oc.). Average RG = 14.65Å.
Tau-5 (R2-R3) - cluster #42 (0.8% oc.). Average RG = 11.6Å.
Tau-5 (R2-R3) - cluster #43 (0.74% oc.). Average RG = 10.97Å.
Tau-5 (R2-R3) - cluster #44 (0.74% oc.). Average RG = 13.49Å.
Tau-5 (R2-R3) - cluster #45 (0.69% oc.). Average RG = 12.23Å.
Tau-5 (R2-R3) - cluster #46 (1.69% oc.). Average RG = 12.54Å.
Tau-5 (R2-R3) - cluster #47 (1.02% oc.). Average RG = 13.68Å.
DSSP class
Sequence position
Class prop.


Tau-5 (R2-R3) - cluster #48 (1.18% oc.). Average RG = 11.06Å.
Tau-5 (R2-R3) - cluster #49 (1.42% oc.). Average RG = 11.9Å.
Tau-5 (R2-R3) - cluster #50 (0.8% oc.). Average RG = 12.17Å.
Tau-5 (R2-R3) - cluster #51 (0.96% oc.). Average RG = 11.42Å.
Tau-5 (R2-R3) - cluster #52 (0.5% oc.). Average RG = 11.2Å.
Tau-5 (R2-R3) - cluster #53 (0.98% oc.). Average RG = 15.15Å.
Tau-5 (R2-R3) - cluster #54 (0.64% oc.). Average RG = 12.57Å.
Tau-5 (R2-R3) - cluster #55 (1.13% oc.). Average RG = 11.37Å.
Tau-5 (R2-R3) - cluster #56 (0.63% oc.). Average RG = 13.6Å.
Tau-5 (R2-R3) - cluster #57 (0.73% oc.). Average RG = 12.85Å.
Tau-5 (R2-R3) - cluster #58 (1.03% oc.). Average RG = 17.84Å.
Tau-5 (R2-R3) - cluster #59 (0.6% oc.). Average RG = 12.91Å.
Tau-5 (R2-R3) - cluster #60 (0.68% oc.). Average RG = 12.81Å.
Tau-5 (R2-R3) - cluster #61 (3.9% oc.). Average RG = 14.8Å.
Tau-5 (R2-R3) - cluster #62 (8.44% oc.). Average RG = 17.02Å.
DSSP class
Sequence position
Class prop.

Conclusion and final remarks

The works presented in this thesis provide statistical methods for comparing and characterizing conformational ensembles of highly flexible proteins. The overall approach is constructed from a perspective that places the intrinsic probabilistic nature of these systems at the forefront, relying on mathematical techniques that capture the complex variability of conformational spaces and, whenever possible, provide statistical evidence about the true structural nature of the proteins under consideration. Furthermore, the methods presented here are strictly non-parametric, that is, they do not require the assumption of any specific model. We hope that this type of developments, which avoid reducing distributions to averages, will become a standard approach in the structural analysis of disordered proteins. Their utility and effectiveness in capturing their conformational complexity have been clearly demonstrated. We would also like to highlight the relevance of statistical inference in this type of problems. While descriptive methods are informative, their robustness diminishes if they are not provided with guarantees regarding the population behavior of the studied systems. In many areas of application, statistical guarantees are often overlooked, usually due to the theoretical complexity of their appropriate implementation. However, we believe that assessing the significance of the obtained conclusions is essential for a proper understanding of the observed phenomena. In this regard, the testing methods presented in Chapters 3 and 6 are aimed to take a first step in performing inference on protein structures, which we believe should be further pursued in future research. These methods, although motivated by open problems in Structural Biology, are generally applicable to problems in other fields of research and have also purely theoretical interest. We conclude by outlining several directions for further research.

Inference at the global scale

The presented techniques for the local structural analysis of flexible proteins are provided with statistical guarantees (Part I). Extending statistical inference to the global scale is a very challenging task due to the mathematical complexity of the underlying theory. A very remarkable contribution would be to provide WASCO (Chapter 5) with evidence about how significant are the differences between the three-dimensional global structural descriptors. This would require testing the equality of distributions supported on the three-dimensional Euclidean space, using the Wasserstein distance as a test statis-

tic. Characterizing the null distribution of the Wasserstein distance when the dimension of the ground space is higher than one is a non-trivial open problem, which would yield the definition of goodness-of-fit tests applicable to a large variety of practical problems. Regarding ensemble characterization, performing post-clustering inference on the output of WARIO (Chapter 7) would constitute an enriching contribution. Nevertheless, the work presented in Chapter 6 already represents a relevant progress in adapting this theory to more realistic problems, allowing for arbitrary dependence structures between observations and features. Incorporating this technique to WARIO would require its extension to non-Gaussian random variables and more sophisticated clustering algorithms that integrate dimensionality reduction. Although this represents a mathematically complex endeavor, it would greatly advance the applicability of selective inference in real-world problems.

Broaden the applicability to more complex systems

The comparison and characterization methods have been conceived for conformational ensembles at all-atom resolution. However, atomistic simulations frequently face computational limitations, due to the extensive spatial and temporal scales involved in thermodynamic and kinetic phenomena [305, 38]. This is sometimes overcome by reducting the dimensionality of the system and simulating a simplified representation of the molecule through the so-called coarse-grained models. However, the basis of the global structural descriptors defined here are the residue-specific reference frames built at every amino acid along the backbone (recall Section 1.2.2), that are defined using the all-atom coordinates. These frames are essential to have access to the residue-residue relative positions and orientations that integrate the global structural analyses. The adaptation of such reference systems to the coarse-grained framework, ensuring that orientation is properly accounted for, presents an intriguing avenue for future research. This approach would pave the way for the comparison and characterization of ensembles in more complex systems, such as proteins with longer sequences, or multi-domain proteins. The frames presented in Section 1.2.2 may also be adapted to more complex frameworks where, instead of reducing resolution, dimension is increased. This might be of interest for the study of the global structure of RNA molecules, whose role on the translation of the DNA genetic information into proteins has been proved essential [103]. Note that once a reference system at every sequence unit is properly defined, the adaptation of WASCO is straightforward. The same applies for the comparison of angular distributions, where the statistical tests presented in Chapter 3 might be extended to the flat torus of general dimension.

Furthermore, the computational optimization of the software implementing WASCO and WARIO would be an important contribution as it would enhance their applicability to larger systems. In this regard, an essential point would be to deal with the inherent computational complexity of the existing optimal transport solvers, especially for non-Euclidean spaces where the transportation cost is given as an $n \times m$ matrix, with n and m being the sample sizes. This aspect represented a great challenge in this thesis when considering the incorporation of orientation into comparison methods like WASCO,

since defining descriptors in non-Euclidean spaces with higher dimensions renders the computation of Wasserstein distances unfeasible in practice.

Integration to Machine-Learning techniques

The development of structure prediction methods has become highly relevant in recent years, especially with the arrival of AlphaFold [147]. However, as we pointed out in Chapter 1, their applicability within the context of flexible proteins remains questionable. A natural extension of the methods presented in this thesis is their integration into Machine Learning (ML) techniques for predicting the structure of intrinsically disordered proteins. These algorithms require two fundamental elements. The first one is a loss function that quantifies the differences between the objects to be predicted. A potential choice might be the overall metric defined from the output of WASCO (Chapter 5), which is a well-adapted distance to quantify discrepancies between pairs of IDP ensembles. The second requirement is a compact representation of the objects under study which can be fed into neural networks. The raw data produced by generative models are too complex and hinder this task. Instead, the weighted families of contact maps presented in Chapter 7 provide a compact ensemble characterization whose integration into ML models might be well-suited.

References

[1] A. A. Adzhubei, M. J. Sternberg, and A. A. Makarov. Polyproline-II helix in proteins: Structure and function. *Journal of Molecular Biology*, 425(12):2100–2132, June 2013.

[2] M. Allaoui, M. L. Kherfi, and A. Cheriet. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In *Lecture Notes in Computer Science*, pages 317–325. Springer International Publishing, 2020.

[3] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 06 2017.

[4] M. AlQuraishi. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, 65:1–8, Dec. 2021.

[5] L. Ambrosio, M. Goldman, and D. Trevisan. On the quadratic random matching problem in two-dimensional domains. *Electronic Journal of Probability*, 2021.

[6] R. J. Anderson, Z. Weng, R. K. Campbell, and X. Jiang. Main-chain conformational tendencies of amino acids. *Proteins*, 60(4):679–689, 2005.

[7] R. Appadurai, J. K. Koneru, M. Bonomi, P. Robustelli, and A. Srivastava. Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. *Journal of Chemical Theory and Computation*, June 2023.

[8] R. Appadurai, J. K. Koneru, M. Bonomi, P. Robustelli, and A. Srivastava. Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. *Journal of Chemical Theory and Computation*, June 2023.

[9] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[10] F. Avbelj and R. L. Baldwin. Origin of the neighboring residue effect on peptide backbone conformation. *Proc. Natl. Acad. Sci. U.S.A*, 101(30):10967–10972, 2004.

[11] S. Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer International Publishing, 2014.

[12] L. Banci, I. Bertini, C. Cefaro, S. Ciofi-Baffoni, A. Gallo, M. Martinelli, D. P. Sideris, N. Katrakili, and K. Tokatlidis. MIA40 is an oxidoreductase that catalyzes oxidative protein folding in mitochondria. *Nature Structural & Molecular Biology*, 16(2):198–206, Feb. 2009.

[13] A. Barozet, P. Chacón, and J. Cortés. Current approaches to flexible loop modeling. *Curr. Res. Struct. Biol.*, 3:187–191, 2021.

[14] M. S. Bartlett. An Inverse Matrix Adjustment Arising in Discriminant Analysis. *The Annals of Mathematical Statistics*, 22(1):107 – 111, 1951.

[15] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, Dec. 2018.

[16] A. Belloni, V. Chernozhukov, I. Fernandez-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[17] C. Belta and V. Kumar. Euclidean metrics for motion generation on SE(3). *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 216(1):47–60, Jan. 2002.

[18] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[19] A. Berg, O. Kukharenko, M. Scheffner, and C. Peter. Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers. *PLoS Comput Biol*, 14(11):e1006589, 2018.

[20] H. M. Berman. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, Jan. 2000.

[21] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proc. Natl. Acad. Sci. U.S.A*, 102(47):17002–17007, 2005.

[22] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc*, 129(17):5656–5664, 2007.

[23] P. Berthet and J.-C. Fort. Weak convergence of empirical wasserstein type distances, 2019. arXiv:1911.02389v1.

[24] D. Bertsekas. *Network Optimization: Continuous and Discrete Methods*. Athena scientific optimization and computation series. Athena Scientific, 1998.

[25] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. 01 1998.

[26] R. B. Best. Computational and theoretical advances in studies of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 42:147–154, 2017. Folding and binding • Proteins: Bridging theory and experiment.

[27] M. R. Betancourt and J. Skolnick. Local propensities and statistical potentials of backbone dihedral angles in proteins. *J. Mol. Biol.*, 342(2):635–649, 2004.

[28] S. Bhattacharya and X. Lin. Recent advances in computational protocols addressing intrinsically disordered proteins. *Biomolecules*, 9(4):146, Apr 2019.

[29] P. Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

[30] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999.

[31] E. Black, S. Yeom, and M. Fredrikson. FlipTest. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2020.

[32] C. E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Libreria internazionale Seeber, 1936.

[33] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51, 01 2014.

[34] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement Interpolation Using Lagrangian Mass Transport. *ACM Transactions on Graphics (SIGGRAPH ASIA 2011)*, 30(6), 2011.

[35] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci. U.S.A*, 105(26):8932–8937, 2008.

[36] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Pure and Applied Mathematics. Academic Press, London, 1975.

[37] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[38] R. Bradley and R. Radhakrishnan. Coarse-grained models for protein-cell membrane interactions. *Polymers*, 5(3):890–936, July 2013.

[39] C. Brändén and J. Tooze. *Introduction to Protein Structure (2nd ed.)*. Garland Science, New York, 1998.

[40] D. Braun, G. Wider, and K. Wuethrich. Sequence-corrected ^{15}N "random coil" chemical shifts. *J. Am. Chem. Soc.*, 116(19):8466–8469, 1994.

[41] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[42] P. Brézellec, I. Vallet-Gely, C. Possoz, S. Quevillon-Cheruel, and J.-L. Ferat. DciA is an ancestral replicative helicase operator essential for bacterial replication initiation. *Nature Communications*, 7(1), Nov. 2016.

[43] R. Brüschweiler. Efficient RMSD measures for the comparison of two molecular ensembles. *Proteins*, 50(1):26–34, 2003.

[44] F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar. Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular Cell*, 61(3):341–351, 2016.

[45] A. R. Camacho-Zarco, S. Kalayil, D. Maurin, N. Salvi, E. Delaforge, S. Milles, M. R. Jensen, D. J. Hart, S. Cusack, and M. Blackledge. Molecular basis of host-adaptation interactions between influenza virus polymerase PB2 subunit and ANP32a. *Nature Communications*, 11(1), July 2020.

[46] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg, 2013.

[47] F. Cazals, T. Dreyfus, D. Mazauric, C.-A. Roth, and C. H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J Comput Chem*, 36(16):1213–1231, 2015.

[48] M. Chan-Yao-Chong, S. Marsin, S. Quevillon-Cheruel, D. Durand, and T. Ha-Duong. Structural ensemble and biological activity of dcia intrinsically disordered region. *Journal of Structural Biology*, 212(1):107573, 2020.

[49] J.-M. Chandonia, N. K. Fox, and S. E. Brenner. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res.*, 47(D1):D475–D481, 11 2018.

[50] Y. Chen, S. Jewell, and D. Witten. More powerful selective inference for the graph fused lasso. *Journal of Computational and Graphical Statistics*, 32(2):577–587, 2023.

[51] Y. T. Chen and D. M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

[52] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8(1), Apr. 2007.

[53] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017.

[54] E. Chevallier and N. Guigui. A bi-invariant statistical model parametrized by mean and covariance on rigid motions. *Entropy*, 22(4):432, Apr. 2020.

[55] M.-K. Cho, H.-Y. Kim, P. Bernado, C. O. Fernandez, M. Blackledge, and M. Zweckstetter. Amino acid bulkiness defines the local conformations and dynamics of natively unfolded α-synuclein and tau. *J. Am. Chem. Soc.*, 129(11):3032–3033, 2007.

[56] C. Chothia. PRINCIPLES THAT DETERMINE THE STRUCTURE OF PROTEINS. *Annual Review of Biochemistry*, 53(1):537–572, June 1984.

[57] D. Clementel, A. D. Conte, A. M. Monzon, G. F. Camagni, G. Minervini, D. Piovesan, and S. C. E. Tosatto. RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Research*, 50(W1):W651–W656, May 2022.

[58] I. Clerc, A. Sagar, A. Barducci, N. Sibille, P. Bernadó, and J. Cortés. The diversity of molecular interactions involving intrinsically disordered proteins: A molecular modeling perspective. *Computational and Structural Biotechnology Journal*, 19:3817–3828, 2021.

[59] A. Conev, M. M. Rigo, D. Devaurs, A. F. Fonseca, H. Kalavadwala, M. V. de Freitas, C. Clementi, G. Zanatta, D. A. Antunes, and L. E. Kavraki. EnGens: a computational framework for generation and analysis of representative protein conformational ensembles. *Briefings in Bioinformatics*, 24(4):bbad242, 07 2023.

[60] D. Cordero-Erausquin. Sur le transport de mesures périodiques. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 329(3):199 – 202, 1999.

[61] Y. Coudène. Measurable partitions and σ-algebras. In *Universitext*, pages 155–163. Springer London, 2016.

[62] P. M. Crespo and J. Gutierrez-Gutierrez. On the elementwise convergence of continuous functions of hermitian banded toeplitz matrices. *IEEE Transactions on Information Theory*, 53(3):1168–1176, 2007.

[63] J. A. Cuesta and C. Matrán. Notes on the Wasserstein metric in Hilbert spaces. *The Annals of Probability*, pages 1264–1276, 1989.

[64] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[65] J. Cárcamo, A. Cuevas, and L.-A. Rodríguez. Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli*, 26(3):2143 – 2175, 2020.

[66] S. A. Dames, R. Aregger, N. Vajpai, P. Bernado, M. Blackledge, and S. Grzesiek. Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J. Am. Chem. Soc.*, 128(41):13508–13514, 2006.

[67] J. Damjanovic, J. M. Murphy, and Y.-S. Lin. Catboss: Cluster analysis of trajectories based on segment splitting. *J Chem Inf Model*, 61(10):5066–5081, 2021.

[68] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, Nov. 2012.

[69] L. de Lara, A. González-Sanz, N. Asher, L. Risser, and J.-M. Loubes. Transport-based counterfactual models, 2023. arXiv:2108.13025.

[70] C. M. Deane, F. H. Allen, R. Taylor, and T. L. Blundell. Carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. *Protein Eng. Des. Sel.*, 12(12):1025–1028, 1999.

[71] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society*, 127(2):476–477, Dec. 2004.

[72] E. del Barrio, J. A. Cuesta-Albertos, C. Matran, and J. M. Rodriguez-Rodriguez. Tests of goodness of fit based on the l2-wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[73] E. del Barrio, A. González-Sanz, and M. Hallin. A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *Journal of Multivariate Analysis*, page 104671, 2020.

[74] E. del Barrio, A. González-Sanz, and J.-M. Loubes. Central limit theorems for general transportation costs, 2021. arXiv:2102.06379v2.

[75] E. del Barrio, A. González-Sanz, and J.-M. Loubes. Central limit theorems for semidiscrete wasserstein distances, 2022. arXiv:2202.06380.

[76] E. del Barrio, P. Gordaliza, and J.-M. Loubes. A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8, 12 2019.

[77] E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926 – 951, 2019.

[78] J. Delon, J. Salomon, and A. Sobolevski. Fast transport optimization for monge cost on the circle. *SIAM Journal on Applied Mathematics*, 70(7/8):2239–2258, 2010.

[79] S. Demko, W. F. Moss, and P. Smith. Decay rates for inverses of band matrices. *Mathematics of Computation*, 43:491–499, 1984.

[80] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, 15(11):e1008432, Nov. 2019.

[81] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. A review of UMAP in population genetics. *Journal of Human Genetics*, 66(1):85–91, Oct. 2020.

[82] A. Diaz-Papkovich, S. Zabad, C. Ben-Eghan, L. Anderson-Trocmé, G. Femerling, V. Nathan, J. Patel, and S. Gravel. Topological stratification of continuous genetic variation in large biobanks. July 2023. bioRxiv 2023.07.06.548007.

[83] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, 11(1), Mar. 2020.

[84] P. Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

[85] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[86] S. Eilenberg and N. Steenrod. *Foundations of Algebraic Topology*. Princeton University Press, Dec. 1952.

[87] C. A. Elena-Real, A. Sagar, A. Urbanek, M. Popovic, A. Morató, A. Estaña, A. Fournet, C. Doucet, X. L. Lund, Z. D. Shi, L. Costa, A. Thureau, F. Allemand, R. E. Swenson, P. E. Milhiet, R. Crehuet, A. Barducci, J. Cortés, D. Sinnaeve, N. Sibille, and P. Bernadó. The structure of pathogenic huntingtin exon 1 defines the bases of its aggregation propensity. *Nature Structural and Molecular Biology*, 30(3):309–320, 2023.

[88] A. Estaña, N. Sibille, E. Delaforge, M. Vaisset, J. Cortés, and P. Bernadó. Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. *Structure*, 27(2):381–391.e2, 2019.

[89] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[90] V. Estivill-Castro. Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75, June 2002.

[91] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering, Design and Selection*, 12(1):15–21, Jan. 1999.

[92] G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 03 1987.

[93] H. J. Feldman and C. W. Hogue. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins: Structure, Function, and Genetics*, 46(1):8–23, Dec. 2001.

[94] H.-P. Fink. Structure analysis by small-angle x-ray and neutron scattering. *Acta Polymerica*, 40(3):224–224, Mar. 1989.

[95] E. Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, Oct. 1894.

[96] M. Fischer, S. Horn, A. Belkacemi, K. Kojer, C. Petrungaro, M. Habich, M. Ali, V. Küttner, M. Bien, F. Kauff, J. Dengjel, J. M. Herrmann, and J. Riemer. Protein import and oxidative folding in the mitochondrial intermembrane space of intact mammalian cells. *Molecular Biology of the Cell*, 24(14):2160–2170, July 2013.

[97] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection, 2017. arXiv:1410.2597.

[98] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *J Mach Learn Res*, 22(78):1–8, 2021.

[99] P. J. Flory and M. Volkenstein. Statistical mechanics of chain molecules. *Biopolymers*, 8(5):699–700, 1969.

[100] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, 2015.

[101] G. Freitag, C. Czado, and A. Munk. A nonparametric test for similarity of marginals—with applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference*, 137(3):697–711, 2007. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri.

[102] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113 – 161, 1996.

[103] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*, 20(8):474–489, June 2019.

[104] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 0(0):1–11, 2022.

[105] J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs. *J. Mol. Biol.*, 198(3):425–443, 1987.

[106] J.-F. Gibrat, B. Robson, and J. Garnier. Influence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. *Biochemistry*, 30(6):1578–1586, 1991.

[107] J. González-Delgado, A. González-Sanz, J. Cortés, and P. Neuvial. Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology. *Electronic Journal of Statistics*, 17(1):1547 – 1586, 2023.

[108] R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[109] S. R. Griffiths-Jones, G. J. Sharman, A. J. Maynard, and M. S. Searle. Modulation of intrinsic ϕ,ψ propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a β-hairpin peptide. *J. Mol. Biol.*, 284(5):1597–1609, 1998.

[110] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. arXiv:2203.05794.

[111] J. Gu, T. Zhang, C. Wu, Y. Liang, and X. Shi. Refined contact map prediction of peptides based on GCN and ResNet. *Frontiers in Genetics*, 13, Apr. 2022.

[112] J. J. Güven, N. Molkenthin, S. Mühle, and A. S. J. S. Mey. What geometrically constrained models can tell us about real-world protein contact maps. *Physical Biology*, 20(4):046004, May 2023.

[113] J. Habchi, L. Mamelli, H. Darbon, and S. Longhi. Structural disorder within henipavirus nucleoprotein and phosphoprotein: From predictions to experimental assessment. *PLoS ONE*, 5(7):e11684, July 2010.

[114] M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328 – 1371, 2021.

[115] E. Hangen, O. Féraud, S. Lachkar, H. Mou, N. Doti, G. M. Fimia, N. vy Lam, C. Zhu, I. Godin, K. Muller, A. Chatzi, E. Nuebel, F. Ciccosanti, S. Flamant, P. Bénit, J.-L. Perfettini, A. Sauvat, A. Bennaceur-Griscelli, K. S.-L. Roux, P. Gonin, K. Tokatlidis, P. Rustin, M. Piacentini, M. Ruvo, K. Blomgren, G. Kroemer, and N. Modjtahedi. Interaction between AIF and CHCHD4 regulates respiratory chain biogenesis. *Molecular Cell*, 58(6):1001–1014, June 2015.

[116] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14):2403–2410, Dec. 2018.

[117] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100, 1979.

[118] H. HARTLEY. Origin of the word 'protein'. *Nature*, 168(4267):244–244, Aug. 1951.

[119] J. M. B. Haslbeck and D. U. Wulff. Estimating the number of clusters via a corrected clustering instability. *Computational Statistics*, 35(4):1879–1894, May 2020.

[120] B. Hess. Convergence of sampling in protein simulations. *Physical Review E*, 65(3):031910, 2002.

[121] B. Hivert, D. Agniel, R. Thiébaut, and B. P. Hejblum. Post-clustering difference testing: valid inference and practical considerations. *arXiv preprint arXiv:2210.13172*, 2022.

[122] B. K. Ho and R. Brasseur. The ramachandran plots of glycine and pre-proline. *BMC Struct. Biol.*, 5:14, 2005.

[123] S. Hofmann, U. Rothbauer, N. Mühlenbein, K. Baiker, K. Hell, and M. F. Bauer. Functional and mutational characterization of human MIA40 acting during import into the mitochondrial intermembrane space. *Journal of Molecular Biology*, 353(3):517–528, Oct. 2005.

[124] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138, 1993.

[125] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[126] S. HOLMES and W. Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2018.

[127] R. Horn and C. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2013.

[128] S. Hovmöller, T. Zhou, and T. Ohlson. Conformations of amino acids in proteins. *Acta Crystallogr. D*, 58(5):768–776, 2002.

[129] J.-R. Huang, V. Ozenne, M. R. Jensen, and M. Blackledge. Direct prediction of nmr residual dipolar couplings from the primary sequence of unfolded proteins. *Angew. Chem. Int. Ed.*, 52(2):687–690, 2013.

[130] J. Huihui and K. Ghosh. An analytical theory to describe sequence-specific inter-residue distance profiles for polyampholytes and intrinsically disordered proteins. *The Journal of Chemical Physics*, 152(16):161102, 2020.

[131] J. Huihui and K. Ghosh. Intrachain interaction topology can identify functionally similar intrinsically disordered proteins. *Biophysical Journal*, 120(10):1860–1868, 2021.

[132] S. Hundrieser, M. Klatt, and A. Munk. The statistics of circular optimal transport. In *Forum for Interdisciplinary Mathematics*, pages 57–82. Springer Nature Singapore, 2022.

[133] S. Hundrieser, M. Klatt, T. Staudt, and A. Munk. A unifying approach to distributional limits for empirical optimal transport. *arXiv preprint*, 2022.

[134] S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity, 2022. arXiv:2202.10434.

[135] V. G. Ivancevic and T. T. Ivancevic. *Applied Differential Geometry*. WORLD SCIENTIFIC, May 2007.

[136] D. A. Jacques and J. Trewhella. Small-angle scattering for structural biology-expanding the frontier while avoiding the pitfalls. *Protein Science*, 19(4):642–657, Jan. 2010.

[137] J. Janin and S. J. Wodak. Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology*, 42:21–78, 1983.

[138] G. Janson, G. Valdes-Garcia, L. Heo, and M. Feig. Direct generation of protein conformational ensembles via machine learning. 2022. bioRxiv:2022.06.18.496675.

[139] M. R. Jensen, M. Zweckstetter, J. rong Huang, and M. Blackledge. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chemical Reviews*, 114(13):6632–6660, Apr. 2014.

[140] S. Jephthah, F. Pesce, K. Lindorff-Larsen, and M. Skepö. Force field effects in simulations of flexible peptides with varying polyproline II propensity. *J Chem Theory Comput*, 17(10):6634–6646, 2021.

[141] S. Jewell, P. Fearnhead, and D. Witten. Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104, Apr. 2022.

[142] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U.S.A*, 102(37):13099–13104, 2005.

[143] A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28):9691–9702, 2005.

[144] C. Jocobi and C. W. Borchardt. De investigando ordine systematis aequationum differentialium vulgarium cujuscunque. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1865:297 – 320.

[145] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006, Nov. 2014.

[146] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021.

[147] J. M. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. A. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[148] S. Jung, K. Park, and B. Kim. Clustering on the torus by conformal prediction. *The Annals of Applied Statistics*, 15(4):1583 – 1603, 2021.

[149] Y.-S. Jung, K.-I. Oh, G.-S. Hwang, and M. Cho. Neighboring residue effects in terminally blocked dipeptides: Implications for residual secondary structures in intrinsically unfolded/disordered proteins. *Chirality*, 26(9):443–452, 2014.

[150] E. A. Kabat and T. T. Wu. The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: Comparison of predicted and experimental determination of β-sheets in concanavalin a. *Proc. Natl. Acad. Sci. U.S.A*, 70(5):1473–1477, 1973.

[151] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[152] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[153] H. S. Kang, N. A. Kurochkina, and B. Lee. Estimation and use of protein backbone angle probabilities. *J. Mol. Biol.*, 229(2):448–460, 1993.

[154] L. Kantorovich and G. S. Rubinstein. On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13:52–59, 1958.

[155] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, Sept. 2002.

[156] K. Kasahara, H. Terazawa, T. Takahashi, and J. Higo. Studies on molecular dynamics of intrinsically disordered proteins and their fuzzy complexes: A mini-review. *Computational and Structural Biotechnology Journal*, 17:712–720, 2019.

[157] A. Kessel and N. Ben-Tal. *Introduction to Proteins*. Chapman and Hall/CRC, Mar. 2018.

[158] F. Klein, E. E. Barrera, and S. Pantano. Assessing SIRAH's capability to simulate intrinsically disordered proteins and peptides. *Journal of Chemical Theory and Computation*, 17(2):599–604, Jan. 2021.

[159] M. Knott and R. B. Best. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *The Journal of Chemical Physics*, 140(17):175102, 05 2014.

[160] M. H. J. Koch, P. Vachette, and D. I. Svergun. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Quarterly Reviews of Biophysics*, 36(2):147–227, May 2003.

[161] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A*, 101(34):12491–12496, 2004.

[162] M. Krzeminski, J. A. Marsh, C. Neale, W.-Y. Choy, and J. D. Forman-Kay. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*, 29(3):398–399, Dec. 2012.

[163] S. Kullback. An application of information theory to multivariate analysis. *The Annals of Mathematical Statistics*, pages 88–102, 1952.

[164] I. Kuntz, G. Crippen, P. Kollman, and D. Kimelman. Calculation of protein tertiary structure. *Journal of Molecular Biology*, 106(4):983–994, 1976.

[165] J. M. Lalmansingh, A. T. Keeley, K. M. Ruff, R. V. Pappu, and A. S. Holehouse. Soursop: A python package for the analysis of simulations of intrinsically disordered proteins. *bioRxiv*, 2023.

[166] U. Lang and V. Schroeder. Kirszbraun's theorem and metric spaces of bounded curvature. *Geometric & Functional Analysis GAFA*, 7:535–560, 1997.

[167] T. Lazar, M. Guharoy, W. Vranken, S. Rauscher, S. J. Wodak, and P. Tompa. Distance-based metrics for comparing conformational ensembles of intrinsically disordered proteins. *Biophys J*, 118(12):2952–2965, 2020.

[168] T. Lazar, E. Martínez-Pérez, F. Quaglia, A. Hatos, L. B. Chemes, J. A. Iserte, N. A. Méndez, N. A. Garrone, T. E. Saldaño, J. Marchetti, A. J. V. Rueda, P. Bernadó, M. Blackledge, T. N. Cordeiro, E. Fagerberg, J. D. Forman-Kay, M. S. Fornasari, T. J. Gibson, G.-N. W. Gomes, C. C. Gradinaru, T. Head-Gordon, M. R. Jensen, E. A. Lemke, S. Longhi, C. Marino-Buslje, G. Minervini, T. Mittag, A. M. Monzon, R. V. Pappu, G. Parisi, S. Ricard-Blum, K. M. Ruff, E. Salladini, M. Skepö, D. Svergun, S. D. Vallet, M. Varadi, P. Tompa, S. C. E. Tosatto, and D. Piovesan. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Research*, 49(D1):D404–D411, Dec. 2020.

[169] J. Ledoux and L. Tchertanov. Does generic cyclic kinase insert domain of receptor tyrosine kinase KIT clone its native homologue? *International Journal of Molecular Sciences*, 23(21):12898, Oct. 2022.

[170] J. Ledoux, A. Trouvé, and L. Tchertanov. The inherent coupling of intrinsically disordered regions in the multidomain receptor tyrosine kinase KIT. *International Journal of Molecular Sciences*, 23(3):1589, Jan. 2022.

[171] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), jun 2016. arXiv:1311.6238.

[172] E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

[173] J. Leiner, B. Duan, L. Wasserman, and A. Ramdas. Data fission: splitting a single data point, 2021.

[174] W. Li, J. E. Cerise, Y. Yang, and H. Han. Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology*, 15(04):1750017, Aug. 2017.

[175] A. Liljas, L. Liljas, J. Piskur, G. Lindblom, P. Nissen, and M. Kjeldgaard. *Textbook Of Structural Biology*. World Scientific Publishing, Singapore, 2009.

[176] K. Lindorff-Larsen and J. Ferkinghoff-Borg. Similarity measures for protein ensembles. *PLoS One*, 4(1):1–13, 01 2009.

[177] K. Lindorff-Larsen and J. Ferkinghoff-Borg. Similarity measures for protein ensembles. *PLoS ONE*, 4(1):e4203, Jan. 2009.

[178] K. Lindorff-Larsen and B. B. Kragelund. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J Mol Biol*, 433(20):167196, 2021.

[179] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *Journal of the American Chemical Society*, 126(10):3291–3299, Feb. 2004.

[180] J. Liu, H. Tan, and B. Rost. Loopy proteins appear conserved in evolution. *Journal of Molecular Biology*, 322(1):53–64, Sept. 2002.

[181] K. Liu, J. Markovic, and R. Tibshirani. More powerful post-selection inference, with application to the lasso, 2018.

[182] S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3):437–450, 2003.

[183] J. R. López-Blanco and P. Chacón. KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics*, 35(17):3013–3019, 2019.

[184] P. Mahalanobis. On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India (Calcutta)*, (2):44–55, 1936.

[185] V. N. Maiorov and G. M. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol*, 235(2):625–634, 1994.

[186] V. N. Maiorov and G. M. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol*, 235(2):625–634, 1994.

[187] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps, 2021.

[188] T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *arXiv preprint*, 2021.

[189] K. V. Mardia, C. C. Taylor, and G. K. Subramaniam. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63(2):505–512, 2007.

[190] S. Marsin, Y. Adam, C. Cargemel, J. Andreani, S. Baconnais, P. Legrand, I. L. de la Sierra-Gallay, A. Humbert, M. Aumont-Nicaise, C. Velours, F. Ochsenbein, D. Durand, E. L. Cam, H. Walbott, C. Possoz, S. Quevillon-Cheruel, and J.-L. Ferat. Study of the DnaB:DciA interplay reveals insights into the primary mode of loading of the bacterial replicative helicase. *Nucleic Acids Research*, 49(11):6569–6586, June 2021.

[191] A. J. M. Martin, M. Vidotto, F. Boscariol, T. D. Domenico, I. Walsh, and S. C. E. Tosatto. RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, 27(14):2003–2005, Apr. 2011.

[192] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen. Comparing molecular dynamics force fields in the essential subspace. *PLoS One*, 10(3):e0121114, 2015.

[193] R. McCann. Polar factorization of maps on riemannian manifolds. *GAFA, Geom. funct. anal.*, 11:589–608, 2001.

[194] R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80:309–323, 1995.

[195] C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

[196] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528–1532, 2015.

[197] L. McInnes. How hdbscan works, 2016. `https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html`. Accessed: 2023-07-21.

[198] L. McInnes. How umap works, 2018. `https://umap-learn.readthedocs.io/en/latest/how_umap_works.html`. Accessed: 2023-07-21.

[199] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. arXiv:1802.03426.

[200] G. Mena and J. Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, 2019.

[201] D. Mercadante, F. Gräter, and C. Daday. CONAN: A tool to decode dynamical information from molecular interaction maps. *Biophysical Journal*, 114(6):1267–1273, Mar. 2018.

[202] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[203] B. Milorey, H. Schwalbe, N. O'Neill, and R. Schweitzer-Stenner. Repeating aspartic acid residues prefer turn-like conformations in the unfolded state: Implications for early protein folding. *J. Phys. Chem. B*, 125(41):11392–11407, 2021.

[204] B. Milorey, R. Schweitzer-Stenner, B. Andrews, H. Schwalbe, and B. Urbanc. Short peptides as predictors for the structure of polyarginine sequences in disordered proteins. *Biophys. J.*, 120(4):662–676, 2021.

[205] N. Miolane, N. Guigui, H. Zaatiti, C. Shewmake, H. Hajri, D. Brooks, A. L. Brigant, J. Mathe, B. Hou, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, Y. Cabanes, T. Gerald, P. Chauchat, B. Kainz, C. Donnat, S. Holmes, and X. Pennec. Introduction to geometric learning in python with geomstats. In *Proceedings of the Python in Science Conference*. SciPy, 2020.

[206] A. Mittal, N. Lyle, T. S. Harmon, and R. V. Pappu. Hamiltonian switch metropolis monte carlo simulations for improved conformational sampling of intrinsically disordered regions tethered to ordered domains of proteins. *Journal of Chemical Theory and Computation*, 10(8):3550–3562, June 2014.

[207] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[208] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, 12(4):345–364, 1992.

[209] A. Munk and C. Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 1998.

[210] J. R. Munkres. *Topology*. Prentice Hall, Inc., 2 edition, Jan. 2000.

[211] B. Mwangi, T. S. Tian, and J. C. Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244, Sept. 2013.

[212] M. H. Newton, J. Rahman, R. Zaman, and A. Sattar. Enhancing protein contact map prediction accuracy via ensembles of inter-residue distance predictors. *Computational Biology and Chemistry*, 99:107700, 2022.

[213] K. Nishikawa, T. Ooi, Y. Isogai, and N. Saitô. Tertiary structure of proteins. i. representation and computation of the conformations. *Journal of the Physical Society of Japan*, 32(5):1331–1337, 1972.

[214] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, and M. Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *Journal of the American Chemical Society*, 131(49):17908–17918, Nov. 2009.

[215] V. Ntranos, L. Yi, P. Melsted, and L. Pachter. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nature Methods*, 16(2):163–166, Jan. 2019.

[216] K.-I. Oh, Y.-S. Jung, G.-S. Hwang, and M. Cho. Conformational distributions of denatured and unstructured proteins are similar to those of 20 x 20 blocked dipeptides. *J. Biomol. NMR*, 53:25–41, 2012.

[217] K.-I. Oh, K.-K. Lee, E.-K. Park, Y. Jung, G.-S. Hwang, and M. Cho. A comprehensive library of blocked dipeptides reveals intrinsic backbone conformational propensities of unfolded proteins. *Proteins*, 80(4):977–990, 2012.

[218] M. Orešič and D. Shalloway. Specific correlations between relative synonymous codon usage and protein secondary structure11edited by g. von heijne. *Journal of Molecular Biology*, 281(1):31–48, 1998.

[219] V. Ozenne, F. Bauer, L. Salmon, J.-r. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, and M. Blackledge. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 2012.

[220] F. Pagani, M. Raponi, and F. E. Baralle. Synonymous mutations in cftr exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences*, 102(18):6368–6372, 2005.

[221] G. A. Papoian. Proteins with weakly funneled energy landscapes challenge the classical structure–function paradigm. *Proceedings of the National Academy of Sciences*, 105(38):14237–14238, Sept. 2008.

[222] R. V. Pappu, R. Srinivasan, and G. D. Rose. The flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *Proc. Natl. Acad. Sci. U.S.A*, 97(23):12565–12570, 2000.

[223] F. C. Park. Distance metrics on the rigid-body motions with applications to mechanism design. *Journal of Mechanical Design*, 117(1):48–54, Mar. 1995.

[224] R. Pearce and Y. Zhang. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Current Opinion in Structural Biology*, 68:194–207, June 2021.

[225] C. J. Penkett, C. Redfield, I. Dodd, J. Hubbard, D. L. McBay, D. E. Mossakowska, R. A. Smith, C. M. Dobson, and L. J. Smith. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J. Mol. Biol.*, 274(2):152–159, 1997.

[226] X. Pennec and V. Arsigny. Exponential barycenters of the canonical cartan connection and invariant means on lie groups. In *Matrix Information Geometry*, pages 123–166. Springer Berlin Heidelberg, Aug. 2012.

[227] N. C. Petroni. Taking rational numbers at random, 2019. arXiv:1908.06944v1.

[228] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[229] D. Phillips. British biochemistry, past and present. In *London Biochem. Soc. Symp.*, page 11. Academic Press, 1970.

[230] B. Phipson and G. K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, 9(1), 2010.

[231] S. Piana, J. L. Klepeis, and D. E. Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*, 24:98–105, Feb. 2014.

[232] A. Platzer. Visualization of SNPs with t-SNE. *PLoS ONE*, 8(2):e56883, Feb. 2013.

[233] M. Punta and B. Rost. PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, May 2005.

[234] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics*, 40(3):191–285, Aug. 2007.

[235] J. P. Quirk and R. Saposnik. Admissibility and Measurable Utility Functions*. *The Review of Economic Studies*, 29(2):140–146, 02 1962.

[236] J. Rabin, J. Delon, and Y. Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41, 01 2009.

[237] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7(1):95–99, 1963.

[238] G. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23:283–437, 1968.

[239] A. Ramdas, N. Garcia, and M. Cuturi. On wasserstein two sample testing and related families of nonparametric tests. *Entropy*, 19, 09 2015.

[240] S. Rao, L. Mondragón, B. Pranjic, T. Hanada, G. Stoll, T. Köcher, P. Zhang, A. Jais, A. Lercher, A. Bergthaler, D. Schramek, K. Haigh, V. Sica, M. Leduc, N. Modjtahedi, T.-P. Pai, M. Onji, I. Uribesalgo, R. Hanada, I. Kozieradzki, R. Koglgruber, S. J. Cronin, Z. She, F. Quehenberger, H. Popper, L. Kenner, J. J. Haigh, O. Kepp, M. Rak, K. Cai, G. Kroemer, and J. M. Penninger. AIF-regulated oxidative phosphorylation supports lung cancer development. *Cell Research*, 29(7):579–591, May 2019.

[241] S. Rao and M. G. Rossmann. Comparison of super-secondary structures in proteins. *J Mol Biol*, 76(2):241–256, 1973.

[242] S. Rao and M. G. Rossmann. Comparison of super-secondary structures in proteins. *J Mol Biol*, 76(2):241–256, 1973.

[243] D. G. Rasines and G. A. Young. Splitting strategies for post-selection inference. *Biometrika*, 12 2022. asac070.

[244] I. A. Rata, Y. Li, and E. Jakobsson. Backbone statistical potential from local sequence-structure interactions in protein loops. *J. Phys. Chem. B*, 114(5):1859–1869, 2010.

[245] C. Reinhardt, G. Arena, K. Nedara, R. Edwards, C. Brenner, K. Tokatlidis, and N. Modjtahedi. AIF meets the CHCHD4/mia40-dependent mitochondrial import pathway. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1866(6):165746, June 2020.

[246] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[247] E. Roquain. Type i error rate control for testing many hypotheses: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38, 2011.

[248] D. Rosenbaum, M. Garnelo, M. Zielinski, C. Beattie, E. Clancy, A. Huber, P. Kohli, A. W. Senior, J. Jumper, C. Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. 2021. arXiv:2106.14108.

[249] A. A. Rosenberg, A. Marx, and A. M. Bronstein. Codon-specific ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. *Nature Communications*, 13(1), may 2022.

[250] M. G. Rossman and A. Liljas. Recognition of structural domains in globular proteins. *Journal of Molecular Biology*, 85(1):177–181, May 1974.

[251] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov. 1987.

[252] K. M. Ruff, R. V. Pappu, and A. S. Holehouse. Conformational preferences and phase behavior of intrinsically disordered low complexity sequences: insights from multiscale simulations. *Current Opinion in Structural Biology*, 56:1–10, 2019.

[253] A. Sagar, C. M. Jeffries, M. V. Petoukhov, D. I. Svergun, and P. Bernadó. Comment on the optimal parameters to derive intrinsically disordered protein conformational ensembles from small-angle X-ray scattering data using the ensemble optimization method. *J Chem Theory Comput*, 17(4):2014–2021, 2021.

[254] M. N. Sanches, K. Knapp, A. B. Oliveira, P. G. Wolynes, J. N. Onuchic, and V. B. P. Leite. Examining the ensembles of amyloid-β monomer variants and their propensities to form fibers using an energy landscape visualization method. *The Journal of Physical Chemistry B*, 126(1):93–99, Dec. 2021.

[255] C. Sander. In structural aspects of recognition and assembly in biological macromolecules (sussman, j., traub, w. & yonath, a., eds). *Balaban Int. Science Services, Philadelphia*, pages 183–196, 1981.

[256] F. Santambrogio. Optimal transport for applied mathematicians. calculus of variations, pdes and modeling. 2015.

[257] F. Saudou and S. Humbert. The Biology of Huntingtin, 2016.

[258] R. Saunders and C. M. Deane. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research*, 38(19):6719–6728, 06 2010.

[259] D. Schuhmacher, B. Bähre, C. Gottschlich, V. Hartmann, F. Heinemann, and B. Schmitzer. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, 2020. R package version 0.12-2.

[260] R. Schweitzer-Stenner and S. E. Toal. Anticooperative nearest-neighbor interactions between residues in unfolded peptides and proteins. *Biophys. J.*, 114(5):1046–1057, 2018.

[261] L. Serrano. Comparison between the ψ distribution of the amino acids in the protein database and nmr data indicates that amino acids have various ψ propensities in the random coil conformation. *J. Mol. Biol.*, 254(2):322–333, 1995.

[262] M. Serrurier, F. Mamalet, A. Gonzalez-Sanz, T. Boissin, J.-M. Loubes, and E. del Barrio. Achieving robustness in classification using optimal transport with hinge regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 505–514, June 2021.

[263] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, Oct. 2007.

[264] J.-E. Shea, R. B. Best, and J. Mittal. Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Current Opinion in Structural Biology*, 67:219–225, 2021.

[265] Y. Shen, J. Roche, A. Grishaev, and A. Bax. Prediction of nearest neighbor effects on backbone torsion angles and nmr scalar coupling constants in disordered proteins. *Protein Sci.*, 27(1):146–158, 2018.

[266] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, 255(3):494–506, 1996.

[267] L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson. The concept of a random coil: Residual structure in peptides and denatured proteins. *Fold. Des.*, 1(5):R95–R106, 1996.

[268] M. Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

[269] M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019.

[270] T. Staudt, S. Hundrieser, and A. Munk. On the uniqueness of kantorovich potentials. *arXiv preprint*, 2022.

[271] T. Steiner. The hydrogen bond in the solid state. *Angewandte Chemie International Edition*, 41(1):48–76, Jan. 2002.

[272] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

[273] D. I. Svergun and M. H. J. Koch. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics*, 66(10):1735–1782, Sept. 2003.

[274] M. B. Swindells, M. W. MacArthur, and J. M. Thornton. Intrinsic ϕ and ψ propensities of amino acids, derived from the coil regions of known structures. *Nat. Struct. Mol.*, 2:596–603, 1995.

[275] S. Tanaka and H. A. Scheraga. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences*, 72(10):3802–3806, Oct. 1975.

[276] G. Tesei and K. Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Research Europe*, 2:94, Jan. 2023.

[277] G. Tesei, T. K. Schulze, R. Crehuet, and K. Lindorff-Larsen. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences*, 118(44), Oct. 2021.

[278] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, Dec. 1953.

[279] M. Tiberti, E. Papaleo, T. Bengtsen, W. Boomsma, and K. Lindorff-Larsen. ENCORE: software for quantitative ensemble comparison. *PLoS Comput Biol*, 11(10):e1004415, 2015.

[280] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. Dunbrack. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput. Biol.*, 6(4):e1000763, 2010.

[281] S. E. Toal, N. Kubatova, C. Richter, V. Linhard, H. Schwalbe, and R. Schweitzer-Stenner. Randomizing the unfolded state of peptides (and proteins) by nearest neighbor interactions between unlike residues. *Chem. Eur. J.*, 21(13):5173–5192, 2015.

[282] P. Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, Oct. 2002.

[283] W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, 43(D1):D364–D368, 2014.

[284] W. F. Trench. Asymptotic distribution of the spectra of a class of generalized kac–murdock–szegö matrices. *Linear Algebra and its Applications*, 294(1):181–192, 1999.

[285] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun. Advanced ensemble modelling of flexible macromolecules using x-ray solution scattering. *IUCrJ*, 2(2):207–217, Feb. 2015.

[286] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ*, 2(2):207–217, 2015.

[287] A. Urbanek, M. Popovic, A. Morató, A. Estaña, C. A. Elena-Real, P. Mier, A. Fournet, F. Allemand, S. Delbecq, M. A. Andrade-Navarro, J. Cortés, N. Sibille, and P. Bernadó. Flanking Regions Determine the Structure of the Poly-Glutamine in Huntingtin through Mechanisms Common among Glutamine-Rich Human Proteins. *Structure*, 28(7):733–746.e5, 2020.

[288] V. N. Uversky, V. Davé, L. M. Iakoucheva, P. Malaney, S. J. Metallo, R. R. Pathak, and A. C. Joerger. Pathological unfoldomics of uncontrolled chaos: Intrinsically disordered proteins and human diseases. *Chemical Reviews*, 114(13):6844–6879, May 2014.

[289] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, 41(3):415–427, 2000.

[290] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18(5):343–384, 2005.

[291] A. W. Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[292] A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[293] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, Apr. 2014.

[294] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[295] A. Vandenbon and D. Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature Communications*, 11(1), Aug. 2020.

[296] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, Nov. 2021.

[297] A. P. Verbyla. A note on the inverse covariance matrix of the autoregressive process1. *Australian Journal of Statistics*, 27(2):221–224, 1985.

[298] C. Villani. *Topics in Optimal Transportation*. American mathematical society, Providence, Rhode Island, 2003.

[299] C. Villani. *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg, 2008.

[300] A. Villié, P. Veber, Y. D. Castro, and L. Jacob. Neural networks beyond explainability: Selective inference for sequence motifs. *Transactions on Machine Learning Research*, 2023.

[301] A. Vitalis and R. V. Pappu. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry*, 30(5):673–699, Apr. 2009.

[302] A. Vullo, I. Walsh, and G. Pollastri. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, 7(1), Mar. 2006.

[303] P. Walters. *An Introduction to Ergodic Theory*. Graduate texts in mathematics. Springer, New York, NY, Oct. 2000.

[304] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):e1005324, Jan. 2017.

[305] W. Wang and R. Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials*, 5(1), Dec. 2019.

[306] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019.

[307] R. Wilder. *Introduction to the Foundations of Mathematics: Second Edition*. Dover Books on Mathematics. Dover Publications, Incorporated, 2012.

[308] J. Wise. The autocorrelation function and the spectral density function. *Biometrika*, 42(1/2):151–159, 1955.

[309] K.-P. Wu, D. S. Weinstock, C. Narayanan, R. M. Levy, and J. Baum. Structural reorganization of α-synuclein at low pH observed by NMR and REMD simulations. *Journal of Molecular Biology*, 391(4):784–796, Aug. 2009.

[310] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, and X. Chen. A comparison for dimensionality reduction methods of single-cell rna-seq data. *Frontiers in Genetics*, 12, 2021.

[311] J. Yeh. *Real Analysis*. World Scientific, 3rd edition, 2014.

[312] C. Yuan, H. Chen, and D. Kihara. Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics*, 13(1), Nov. 2012.

[313] R. Zallot, N. Oberg, and J. A. Gerlt. The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry*, 58(41):4169–4182, Sept. 2019.

[314] M. H. Zaman, M.-Y. Shen, R. Berry, K. F. Freed, and T. R. Sosnick. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. *J. Mol. Biol.*, 331(3):693–711, 2003.

[315] M. Zefran, V. Kumar, and C. Croke. On the generation of smooth three-dimensional rigid body motions. *IEEE Transactions on Robotics and Automation*, 14(4):576–589, 1998.

[316] O. Zhang, M. Haghighatlari, J. Li, J. M. C. Teixeira, A. Namini, Z.-H. Liu, J. D. Forman-Kay, and T. Head-Gordon. Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. 2022. arXiv:2206.12667.

[317] W. Zheng, Y. Li, C. Zhang, R. Pearce, S. M. Mortuza, and Y. Zhang. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1149–1164, Aug. 2019.

[318] S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Trans Pattern Anal Mach Intell*, 28(6):917–929, 2006.

[319] J. Zhu, X. Salvatella, and P. Robustelli. Trajectories and Code from "Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states" Zhu et al. 2022 (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7120845. 2022.

[320] J. Zhu, X. Salvatella, and P. Robustelli. Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states. *Nature Communications*, 13(1), Oct. 2022.

[321] S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules*, 10(1):1–9, 1977.

Appendix F

Introduction en français

F.1 Le désordre intrinsèque des protéines

Les protéines sont des molécules essentielles dans tous les organismes vivants. Elles jouent un rôle central dans la majorité des processus biologiques, opérant au niveau moléculaire, cellulaire et de l'organisme. Le terme *protéine* a été introduit pour la première fois par le chimiste suédois Jöns Jacob Berzelius dans une lettre adressée au chimiste néerlandais Gerardus Johannes Mulder en 1838 [118]:

"Or je présume que l'oxyde organique, qui est la base de la fibrine et de l'albumine (et auquel il faut donner un nom particulier p. ex. protéine) est composé d'un radical ternaire, combiné avec de l'oxygène dans quelqu'un de ses rapports simples que la nature inorganique nous présente."

Cette lettre a marqué le début d'un long voyage que la biologie structurale a entrepris pour comprendre la structure de ces macromolécules et les relier à leurs fonctions cruciales aux niveaux supérieurs du monde vivant. Bien entendu, ce voyage est allé de pair avec les progrès technologiques qui ont permis la détermination expérimentale de la structure des protéines. Après les premières techniques de cristallographie aux rayons X, la cryo-microscopie électronique (cryo-EM) et la résonance magnétique nucléaire (RMN) ont constitué une avancée majeure pour la reconstruction d'une seule particule, permettant de résoudre la structure tridimensionnelle d'une macromolécule à l'échelle atomique [175]. Ces progrès ont continuellement repoussé les limites de la résolution réalisable et ont permis l'observation de structures de taille et de complexité croissantes. Plus nous pouvons observer et mieux nous pouvons observer, plus les approches et les perspectives qui permettent de déchiffrer ce que nous voyons sont riches. Grâce à la biologie structurelle, nous sommes en mesure de rendre visibles des objets à l'échelle subatomique et d'adopter le principe "voir, c'est croire". Cependant, *comprendre* ce que nous voyons nécessite l'implication d'une famille diversifiée de domaines de connaissance, dans laquelle, avec la reconnaissance récente de l'importance du désordre, les mathématiques doivent prendre part.

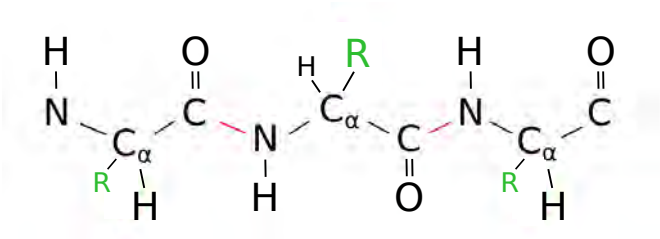


Figure F.1: Représentation simplifiée d'un polypeptide. Les liaisons peptidiques, les atomes du squelette et les chaînes latérales sont marqués respectivement en rouge, noir et vert.

F.1.1 La structure et la fonction des protéines

Une protéine est une macromolécule constituée de résidus d'acides aminés liés par des liaisons peptidiques. Ce type de molécule polymère est également appelé polypeptide. Un acide aminé est une molécule composée d'un atome de carbone (α-carbone) attaché à un groupe carboxyle (-COOH), un groupe amine (-NH_2), un hydrogène et une chaîne latérale variable, également appelée radical. Une liaison peptidique est une double liaison entre le carbone du groupe carboxyle d'un résidu et l'azote du groupe amine suivant. Une représentation simplifiée d'un polypeptide est présentée dans la figure F.1. Notons que la formation de liaisons peptidiques permet de distinguer deux parties principales dans la protéine. D'une part, la séquence d'atomes d'azote, de carbone α, d'hydrogène, de carbone et d'oxygène, appelée *squelette*, représentée en noir dans la Figure F.1. D'autre part, les *chaînes latérales*, c'est-à-dire la famille des différents radicaux liés à chaque α-carbone, illustrée en vert dans la Figure F.1. Les chaînes latérales déterminent les propriétés physico-chimiques des acides aminés et constituent l'empreinte digitale de la protéine.

La séquence des résidus d'acides aminés est appelée *structure primaire* (Figure F.2a). Par souci de simplicité, nous appellerons également la structure primaire *séquence*. Bien que l'on connaisse environ 500 acides aminés naturels, seuls 20 d'entre eux se retrouvent dans les protéines. Cela donne déjà une idée de la complexité du monde dans lequel nous nous plongeons, car il est possible de concevoir jusqu'à 20^L protéines avec une longueur de séquence L. Pour des protéines de 100 acides aminés, cela implique d'envisager jusqu'à 10^{130} séquences possibles dans un univers contenant 10^{82} atomes. L'impossibilité de connaître toutes les protéines met en évidence la nécessité de stratégies intelligentes pour comprendre leur comportement sur la base des informations disponibles. La biologie structurale cherche à y parvenir en décryptant les mécanismes qui régissent la transformation de la structure primaire en la forme tridimensionnelle de la protéine, que nous appellerons généralement *structure*. Ce processus est connu sous le nom de *folding*. Au cours du processus de repliement, certaines parties de la séquence adoptent des éléments de *structure secondaire* relativement stables et bien définis, les plus représentatifs étant les hélices α et les feuilles β Figure F.2b). L'arrangement spatial de ces éléments, qui sont reliés par des tours et des bobines, forme la *structure tertiaire* (Figure F.2c). Pour une introduction

(a) (b) (c)

Figure F.2: Structure primaire (a), secondaire (b) et tertiaire (c) de la protéine. En (b,c), les hélices α sont représentées en rouge et les feuilles β sont marquées en jaune.

approfondie à la structure des protéines, nous nous référons à [175, 271, 1, 56].

Les protéines remplissent de nombreuses fonctions qui sont étroitement liées à leurs propriétés structurelles et dynamiques [175]. Par exemple, les enzymes catalysent divers types de réactions chimiques. D'autres protéines jouent le rôle de protéines nutritives et de stockage, cruciales pour la croissance et la survie des graines dans de nombreuses plantes. D'autres encore permettent la contraction des cellules, lient et transportent des substances, agissent comme des protéines structurelles pour donner aux cellules une forme définie et régulent divers processus cellulaires. Toutes ces fonctions dépendent souvent directement de la structure repliée de la protéine, également connue sous le nom d'"état natif". L'état natif n'est pas une conformation fixe, mais plutôt un ensemble d'états accessibles que la protéine peut adopter, en fonction de facteurs tels que les conditions du solvant et la température. Les énergies de l'état natif des protéines qui se replient dans une structure 3D bien définie présentent des minima globaux stables. Cependant, de nombreuses protéines ne correspondent pas à cette description et présentent des paysages énergétiques relativement plats avec de multiples minima locaux. Ces protéines, connues sous le nom de *protéines intrinsèquement désordonnées* (IDP), sont dans une forme constante, changeant et passant d'un état à l'autre [139]. L'ensemble de ces conformations est appelé *ensemble* de protéines. L'absence d'un état d'équilibre nécessite de réadapter les techniques classiques traditionnellement utilisées pour étudier la relation structure-fonction, en s'ouvrant à de nouveaux paradigmes qui permettent de comprendre la richesse fonctionnelle conférée par leur variabilité structurale.

F.1.2 Protéines intrinsèquement désordonnées: chute du paradigme structure-fonction

Jusqu'à la fin du 20e siècle, la grande majorité de la communauté scientifique soutenait le paradigme dit *structure-fonction*: une protéine fonctionnelle nécessite une structure stable et bien définie. En outre, les interactions protéine-protéine dépendent de la complémentation précise des surfaces. Les modèles classiques, tels que le "lock-key" proposée par le lauréat du prix Nobel Emil Fischer en 1894, s'inscrivent dans ce cadre [95]. En affirmant que les protéines non structurées sont dénaturées, les protéines intrinsèquement désordonnées remettent en cause ce paradigme [282]. En effet, bien qu'elles soient dépourvues de

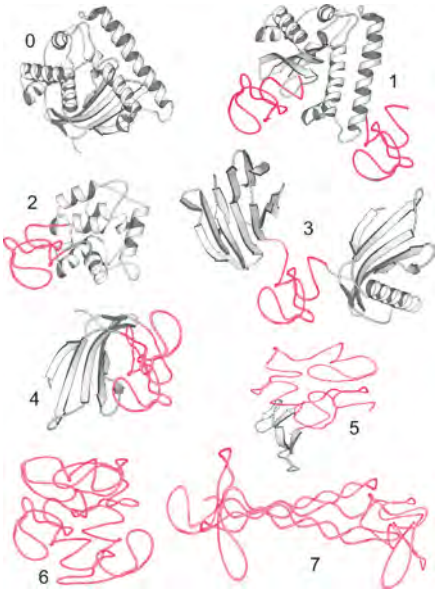


Figure F.3: Figure 1 dans [290]. Différents niveaux d'ordre (gris) et de désordre (rouge): (0) pas de désordre, (1) terminaisons désordonnées, (2) lien désordonné, (3) boucle désordonnée, (4) domaine désordonné, (5) protéine désordonnée avec une certaine structure résiduelle, (6) protéine entièrement désordonnée, en grande partie effondrée et (7) protéine entièrement désordonnée et étendue.

structure secondaire et tertiaire stable de manière isolée, les PDI remplissent une grande diversité de fonctions biologiques en exploitant leur flexibilité intrinsèque [293]. De plus, les IDP peuvent mal fonctionner dans certaines circonstances, telles que des mutations ou des conditions environnementales inappropriées. Ce phénomène peut induire des maladies graves, notamment des cancers, des maladies cardiovasculaires ou neurodégénératives [288]. Tout ceci justifie la pertinence fonctionnelle des protéines désordonnées et la nécessité de réadapter le paradigme structure-fonction pour intégrer la variabilité structurale.

Au cours des 20 dernières années, la biologie structurale a intégré le désordre dans l'étude des protéines. Cela a conduit à passer de la dichotomie rigide vs. non structuré à la prise en compte du désordre comme un continuum. En effet, la plupart des protéines ne sont ni totalement ordonnées ni totalement désordonnées, mais contiennent des régions ordonnées et désordonnées dans des proportions différentes. rapports [290]. C'est ce qu'illustre la figure F.3, extraite de [290]. Le passage au continuum a également eu un impact sur l'étude des paysages énergétiques, qui considèrent désormais les profils faiblement entonnoirs comme une transition entre les minima énergétiques profonds et les paysages accidentés des systèmes très désordonnés [221]. Comme nous pouvons le constater, le désordre se manifeste avec une intensité et un placement séquentiel variables. En outre, l'échec du repliement est codé par la structure primaire, car les IDP

présentent des propriétés séquentielles uniques. Certaines d'entre elles sont le biais de composition [289], qui se manifeste par des résidus à faible hydrophobicité (pour un faible compactage) et à charge nette élevée (pour une forte répulsion charge-charge), un faible contenu de structure secondaire prédite [180] ou une forte variabilité de séquence (faible conservation) [113]. Cela met en évidence le fait que le processus de repliement est fortement lié à la séquence de la protéine. Le décryptage de cette relation est essentiel pour le développement de méthodes de prédiction de la structure à partir de la séquence des protéines, qui ont connu un essor avec l'arrivée de l'apprentissage profond. C'est le cas du célèbre algorithme AlphaFold [147], qui peut prédire la structure des protéines avec une précision au niveau atomique. La base de données AlphaFold sur la structure des protéines [296] contient une structure prédite pour presque toutes les protéines du protéome de nombreux organismes qui ont été totalement ou partiellement séquencés, y compris les protéines comportant des régions intrinsèquement désordonnées (IDR). Cependant, les IDR présentent de faibles valeurs de la métrique de confiance AlphaFold, appelée test de différence de distance locale prédite (pLDDT), ce qui signifie une faible confiance dans les prédictions structurelles et donc des descriptions inexactes de ces régions désordonnées (*a priori*). Par conséquent, l'étude des protéines désordonnées nécessite des approches alternatives qui combinent intelligemment des méthodes expérimentales, théoriques et informatiques. Un aperçu de l'état de l'art est présenté dans la section suivante.

F.1.3 Approches existantes pour modéliser les IDP

L'accès aux données expérimentales est sans doute la caractéristique la plus pertinente qui différencie l'étude des protéines ordonnées de celle des PDI. Il existe un contraste frappant entre les deux mondes en ce qui concerne la quantité de structures connues expérimentalement. La Protein Data Bank (PDB) [20] est une base de données librement accessible qui contient plus de 200 000 structures expérimentales de protéines repliées. Son équivalent pour les systèmes désordonnés est la Protein Ensemble Database [168], une base de données en libre accès qui comprend des données IDP, mais qui contient 280 entrées à ce jour. Dans ce contexte, les données expérimentales ne peuvent pas fournir d'informations précises sur chacune des conformations individuelles de l'ensemble, mais seulement des mesures moyennes. C'est pourquoi les données expérimentales IDP sont utiles en tant que *contrainte à la simulation*. En effet, l'étude des protéines désordonnées en tant qu'ensembles conformationnels est largement régie par des techniques de simulation et de modélisation souvent calibrées avec des données expérimentales. Dans la section suivante, nous présentons un bref aperçu des deux grandes familles qui intègrent les méthodes de génération d'ensembles. Ces techniques n'étant pas l'objet de cette thèse, l'aperçu ne présente que quelques-unes des contributions les plus pertinentes au sein d'un champ d'étude vaste et diversifié. Pour un aperçu plus complet de la littérature existante, nous nous référons aux reviews [58, 26, 252, 28, 264, 156].

Méthodes de génération d'ensembles

La première catégorie de méthodes informatiques vise à générer des ensembles représentatifs de conformations par une exploration efficace de l'espace conformationnel. Le terme "efficace" s'explique par le fait qu'il n'est pas possible d'inspecter au hasard l'espace d'état complet. En fait, une exploration efficace incorpore des informations dérivées de structures déterminées expérimentalement, optimisant ainsi la procédure de calcul. La méthode basée sur la connaissance la plus distinctive est Flexible-Meccano (FM) [219], qui construit chaque conformation en assemblant séquentiellement des unités de plan peptidique à l'aide d'une bibliothèque de bobines spécifiques aux résidus obtenue à partir de structures cristallographiques. Avec Flexible-Meccano, TraDES [93] est une autre technique d'échantillonnage stochastique populaire. Les conformations produites par ces méthodes sont validées par leur ajustement aux données expérimentales, à l'aide d'outils informatiques tels que ENSEMBLE [162], ASTEROIDS [214] ou EOM [22, 285]. Ces techniques utilisent des paramètres mesurables par RMN ou des données de diffusion des rayons X aux petits angles (SAXS). Bien que sa résolution soit plus faible, le SAXS est capable de récupérer des informations structurelles et dynamiques globales sur les macromolécules biologiques, y compris celles qui ne peuvent pas se cristalliser, comme l'IDP [94, 136, 160, 234, 273]. Cependant, des approches comme Flexible-Meccano ne parviennent pas à capturer les éléments de structure secondaire qui impliquent plusieurs résidus consécutifs dans l'IDP [21, 142]. Cette limitation a été surmontée dans [88] en affinant la calibration expérimentale à l'aide d'une vaste bibliothèque de fragments à trois résidus.

La deuxième grande famille de méthodes utilise des modèles physiques pour échantillonner l'espace conformationnel, en simulant le comportement dynamique de la PDI. La technique prééminente dans ce contexte est la simulation dynamique moléculaire (MD), qui résout les équations du mouvement de Newton pour recréer l'évolution temporelle du système [155, 231]. Bien que capables de représenter convenablement l'espace d'état de la PDI, les MD présentent un inconvénient majeur qui réside dans leur coût de calcul excessif lorsqu'elles sont appliquées à des molécules de grande taille. En effet, le rayon de giration important présenté par l'IDP par rapport aux protéines repliées, ainsi que leurs fluctuations inhérentes, font considérablement augmenter la taille de la boîte de simulation contenant la protéine et les molécules d'eau. Une solution pour traiter ce type de systèmes est l'utilisation de modèles à gros grains, qui fournissent une représentation plus simpliste de la protéine mais permettent une investigation plus large de l'espace d'état [158, 68, 159]. En outre, la précision des techniques basées sur la MD dépend fortement des champs de force et des modèles de solvatation utilisés, dont la détermination pour les protéines flexibles est un domaine de recherche très actif [140, 301]. Les performances des méthodes MD peuvent également être renforcées par l'intégration de données expérimentales pour restreindre l'exploration de l'espace conformationnel [71, 179, 309]. Des approches hybrides remarquables ont également été proposées, réalisant des simulations MD avec des potentiels dérivés de l'apprentissage automatique, tels que CALVADOS [277, 276]. Les

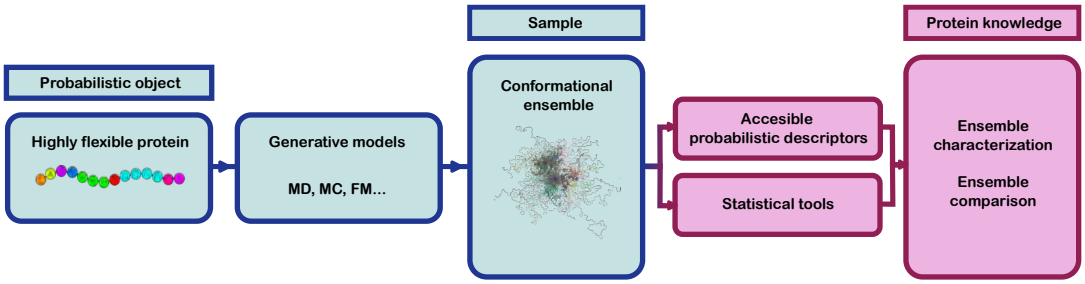


Figure F.4: Les ensembles conformationnels sont conçus comme des échantillons d'états IDP produits par des modèles génératifs basés sur la connaissance et la physique. Ces données sont caractérisées par des descripteurs probabilistes et analysées à l'aide d'outils statistiques, dans le but de caractériser et de comparer avec précision les ensembles récupérés.

méthodes de Monte Carlo (MC) constituent une alternative physique à la MD, parmi lesquelles on peut citer le schéma de Metropolis de la chaîne de Markov [202], sa variante Hamiltonian Switch Metropolis Monte Carlo [206] adaptée à l'étude de l'IDR, ou ABSINTH [301], une approche MC intermédiaire entre les modèles à gros grains et les modèles à tous les atomes.

Méthodes de caractérisation et de comparaison des ensembles

Les méthodes présentées ci-dessous visent à produire des représentations d'ensemble de protéines désordonnées. En effet, la grande majorité des contributions méthodologiques dans l'étude des protéines flexibles se concentrent sur la compensation du manque de données expérimentales par la simulation d'ensembles conformationnels. Ici, une réflexion naturelle s'impose: une fois que nous sommes capables de générer des ensembles IDP de taille indéterminée, quelle est la suite? Que faisons-nous de toutes ces données? Comment transformer les résultats des modèles génératifs en représentations concises et interprétables qui permettent de comprendre la relation séquence-structure dans la PDI? Plus succinctement: comment pouvons-nous *caractériser* et *comparer* des ensembles de protéines hautement flexibles? La seule réponse possible à ces questions est de recourir à des techniques conçues pour traiter la variabilité des objets intrinsèquement désordonnés: Les probabilités et les statistiques. Les systèmes aléatoires doivent être décrits comme des objets probabilistes, et les échantillons tirés de ces systèmes doivent être analysés à l'aide de techniques statistiques. Cette idée est schématisée dans la Figure F.4, qui situe la contribution de cette thèse (en violet) par rapport à l'état de l'art précédemment décrit (en bleu).

L'objectif principal de cette thèse est de fournir l'étape suivante naturelle aux techniques de modélisation IDP, en rendant les résultats des modèles génératifs interprétables avec des garanties statistiques. Pour ce faire, il est essentiel de définir des méthodes qui

permettent une *caractérisation* et une *comparaison* compactes et interprétables des ensembles de PDI. Les applications qui motivent ces objectifs sont nombreuses et diverses. Parmi elles, on peut citer l'analyse a posteriori des performances des modèles génératifs, y compris leur comparaison avec les données expérimentales, la comparaison relative des champs de force et des modèles de solvatation ou l'évaluation de l'effet des restrictions expérimentales. D'autres applications remarquables sont l'évaluation de l'effet des mutations de séquence, qui peut conduire à l'incorporation de la PDI dans le domaine de la conception des protéines, ou la définition des fonctions de perte et des descripteurs compacts requis pour le développement d'algorithmes d'apprentissage automatique. Ce dernier point serait déterminant pour l'extension des méthodes de prédiction de structure aux séquences avec IDR.

Le développement de méthodes pour la caractérisation et la comparaison d'ensembles désordonnés prend de plus en plus d'importance en biologie structurale, avec de nombreuses contributions remarquables ces dernières années [167, 7, 59, 57]. Ces études, qui seront discutées plus en détail dans les chapitres composant ce manuscrit, apportent des contributions intéressantes et innovantes. Cependant, nous pensons qu'elles n'ont pas encore intégré de manière productive la nature probabiliste des protéines flexibles. Nous proposons ici de nous attaquer à notre objectif en plaçant la variabilité structurelle au centre et en concevant les PDI comme des objets intrinsèquement probabilistes qui doivent être analysés à l'aide des techniques statistiques les plus appropriées. Nous détaillons cette stratégie dans la section suivante.

F.2 La nature probabiliste inhérente des protéines flexibles

La stratégie que nous présentons pour capturer la nature probabiliste intrinsèque des protéines flexibles consiste à (*i*) définir les descripteurs structurels par des distributions de probabilité supportées sur des espaces bien adaptés et (*ii*) caractériser et comparer ces distributions avec des techniques statistiques appropriées, en fournissant des garanties statistiques sur le comportement de la population lorsque c'est possible. Nous proposons d'appliquer cette stratégie à la fois au niveau local (échelle des acides aminés) et au niveau global (séquence entière). Les descripteurs structurels locaux et globaux sont définis dans la section F.2.2. Ensuite, dans la section F.2.3, nous présentons les principaux outils statistiques qui seront utilisés pour les caractériser et les comparer. Tout d'abord, nous rappelons quelques concepts essentiels de la théorie des probabilités et définissons la notation qui sera utilisée tout au long du manuscrit. Ces notions sont présentées dans la section suivante, qui peut être omise par les lecteurs moins intéressés par les aspects mathématiques.

F.2.1 Contexte et notation

Cette section rassemble les notations clés et les définitions qui seront supposées tout au long du manuscrit. Des notations spécifiques supplémentaires seront présentées dans chaque chapitre. Nous commençons par définir les espaces de probabilité, c'est-à-dire des espaces de mesure où la mesure de l'ensemble entier est égale à un [29].

Definition F.2.1 (Espace de probabilité)**.** *Soit Ω un ensemble non vide et Σ une Σ-algèbre, c'est-à-dire un ensemble de sous-ensembles de Ω tel que*

(i) $\Omega \in \Sigma$*,*

(ii) Toute union dénombrable d'éléments de Σ est également dans Σ,

(iii) Le complément de chaque élément de Σ est dans Σ.

Si $\mathbb{P} : \Sigma \longrightarrow [0,1]$ est tel que $\mathbb{P}(\Omega) = 1$ et dénombrablement additif, c'est-à-dire que $\mathbb{P}(\cup_{i\in\mathbb{N}} E_i) = \sum_{i\in\mathbb{N}} \mathbb{P}(E_i)$ pour toute collection dénombrable d'ensembles disjoints par paire $\{E_i\}_{i\in\mathbb{N}} \subset \Sigma$, alors $\mathbb{P}$ est appelé un mesure de probabilité sur Σ et le triplet $(\Omega, \Sigma, \mathbb{P})$ est appelé un espace de probabilité.

Soit $(\Omega, \Sigma, \mathbb{P})$ un espace de probabilité, $\mathcal{E}$ un espace topologique [210] et $\mathcal{T}$ l'algèbre σ générée par la topologie de $\mathcal{E}$ [29]. Soit $f : \Omega \longrightarrow \mathcal{E}$ une fonction mesurable, c'est-à-dire telle que $f^{-1}(O) \in \Sigma$ pour tout $O \in \mathcal{T}$. La mesure *push-forward* de $\mathbb{P}$ par f est l'application $f_{\#}\mathbb{P} : \mathcal{T} \longrightarrow [0,1]$ telle que $f_{\#}\mathbb{P}(O) = (\mathbb{P} \circ f^{-1})(O)$ pour tout $O \in \mathcal{T}$. Cette transformation est la clé pour définir la distribution de probabilité d'une variable aléatoire.

Definition F.2.2 (Variable aléatoire, distribution)**.** *Soit $(\Omega, \Sigma, \mathbb{P})$ un espace de probabilité et $\mathcal{E}$ un espace topologique. Une variable aléatoire est une fonction mesurable $X : \Omega \longrightarrow \mathcal{E}$. La mesure de $\mathbb{P}$ par X, notée $P_X := X_{\#}\mathbb{P}$, est appelée la distribution (de probabilité) de X, ou la loi de X.*

Pour toute variable aléatoire X définie sur un espace de probabilité $(\Omega, \Sigma, \mathbb{P})$ et prenant des valeurs dans un espace topologique $\mathcal{E}$, nous définissons le *support* de sa distribution comme étant l'ensemble fermé

$$\text{supp}(P_X) = \{x \in \mathcal{E} \,:\, P_X(U_x) > 0 \text{ pour tout } U_x \text{ voisinage de x}\}.$$

Nous désignerons par $\mathcal{P}(\mathcal{E})$ l'ensemble de toutes les distributions de probabilité supportées sur $\mathcal{E}$, c'est-à-dire dont le support est un sous-ensemble de $\mathcal{E}$. On notera que le terme *distribution* fait référence à une variable aléatoire et que le terme *mesure* opère directement sur un espace de probabilité. Cependant, lorsque Ω est un espace topologique, nous pouvons prendre $\Omega = \mathcal{E}$, $\Sigma = \mathcal{T}$, X l'application identité et parler directement de $\mathbb{P}$ comme d'une *distribution*. Comme tous les espaces considérés ici seront dotés d'une topologie, nous utiliserons indifféremment les termes *distribution* ou *mesure* tout au long du manuscrit, en omettant le push-forward vers l'avant de la variable aléatoire lorsque le

contexte est clair. Ainsi, nous ferons également référence à $\mathcal{P}(\mathcal{E})$ comme étant l'ensemble des *mesures* de probabilité supportées sur $\mathcal{E}$.

Le concept de variable aléatoire peut être étendu au cas où son espace image est un produit cartésien d'espaces topologiques équipés de la topologie du produit [210]. Cette construction peut être considérée comme la combinaison de deux variables aléatoires définies sur le même espace de probabilité.

Definition F.2.3 (Distribution jointe, marginales)**.** *Soit $(\Omega, \Sigma, \mathbb{P})$ un espace de probabilité, $\mathcal{E}$ un espace topologique et $X, Y : \Omega \longrightarrow \mathcal{E}$ deux variables aléatoires. La distribution jointe de X et Y est la distribution de probabilité de la variable aléatoire*

$$\begin{aligned}(X,Y)\colon \Omega &\longrightarrow \mathcal{E}\times\mathcal{E}\\ \omega &\longmapsto (X(\omega),Y(\omega)),\end{aligned}$$

c'est-à-dire la mesure $P_{XY} = (X,Y)_{\#}\mathbb{P} \in \mathcal{P}(\mathcal{E}\times\mathcal{E})$. Les mesures P_X et P_Y sont appelées distributions marginales de P_{XY}.

En rendant implicite la "push forward" des variables aléatoires, pour une paire de mesures $P, Q \in \mathcal{P}(\mathcal{E})$, nous dénotons par $\Pi(P,Q)$ l'ensemble des distributions de probabilité ayant P et Q comme marginales. Nous pouvons écrire $\Pi(P,Q)$ plus précisément comme suit

$$\Pi(P,Q) = \{\gamma \in \mathcal{P}(\mathcal{E}\times\mathcal{E}) \,:\, p^x_{\#}\gamma = P, \quad p^y_{\#}\gamma = Q\}, \quad \forall\, P, Q \in \mathcal{P}(\mathcal{E}),$$

où $p^x, p^y : \mathcal{E}\times\mathcal{E} \longrightarrow \mathcal{E}$ sont tels que $p^x(x,y) = x$ et $p^y(x,y) = y$ pour tous $x, y \in \mathcal{E}$. Les éléments de $\Pi(P,Q)$ sont également appelés *couplages*.

Dans les problèmes pratiques, il est souvent impossible de connaître la véritable distribution sous-jacente P d'une variable aléatoire X. Au lieu de cela, nous avons généralement accès à un *échantillon* de X. Plus précisément, nous définissons un *échantillon* de X comme une famille de variables aléatoires indépendantes $X_1, \ldots, X_n$ identiquement distribuées comme X (c'est-à-dire dont la distribution de probabilité est P). En pratique, nous observons une *réalisation* de $X_1, \ldots, X_n$, c'est-à-dire l'image des variables aléatoires en n points $\omega_1, \ldots, \omega_n \in \Omega$. Les réalisations sont généralement désignées en lettres minuscules par $x_1, \ldots, x_n$, où $x_i = X_i(\omega_i)$ pour tout $i \in \{1, \ldots, n\}$. Les échantillons permettent d'obtenir des informations statistiquement significatives sur la population grâce à la *mesure empirique* de P, définie ci-dessous.

Definition F.2.4. *Soit $(\Omega, \Sigma, \mathbb{P})$ un espace de probabilité, $\mathcal{E}$ un espace topologique et $\mathcal{T}$ l'algèbre σ générée par la topologie de $\mathcal{E}$. Soit $X : \Omega \longrightarrow \mathcal{E}$ une variable aléatoire avec une distribution P et $X_1, \ldots, X_n$ un échantillon de X, pour $n \in \mathbb{N}$. La mesure empirique de P est la mesure de probabilité $P_n : \mathcal{T} \longrightarrow [0,1]$ satisfaisant*

$$P_n(E) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in E\} \quad \forall\, E \in \mathcal{T}, \tag{F.1}$$

où $\mathbb{1}\{A\}$ désigne la fonction indicatrice de l'événement A.

Notons que l'écriture de $X_i \in E$ comme dans (F.1) est un abus de notation, et formellement nous devrions la remplacer par $\{\omega \in \Omega : X_i(\omega) \in E\}$. Par souci de clarté, la notation largement acceptée dans (F.1) sera conservée. Les mesures empiriques sont le principal outil permettant d'*inférer* le comportement des mesures de population en exploitant uniquement les informations fournies par un échantillon. Ce point est décrit plus en détail dans la section F.2.3. En outre, l'étude des mesures empiriques fournit une analyse bien fondée du comportement des distributions sous-jacentes lorsque P_n converge presque sûrement [29, Théorème 6.1] et uniformément [29, Théorème 20.6] vers P au fur et à mesure que n croît jusqu'à l'infini. Nous renvoyons à [29] pour une introduction approfondie à la théorie des mesures et aux probabilités. Pour se familiariser avec les principaux concepts de la théorie de la topologie, que nous utiliserons également, nous nous référons à [210].

F.2.2 Descripteurs structuraux probabilistes

La première étape de notre stratégie consiste à définir des descripteurs structuraux appropriés qui intègrent autant d'informations que possible sur la variabilité conformationnelle des protéines flexibles. Nous le ferons en considérant des distributions de probabilité bien adaptées à la structure locale et globale du système et, surtout, dont les variables aléatoires correspondantes fournissent des *réalisations accessibles*. En d'autres termes, nous cherchons à définir des observables aléatoires qui peuvent être *mesurés* en pratique sur des modèles de protéines.

Descripteurs structuraux locaux

L'investigation des propriétés structurales et dynamiques des protéines au niveau local implique principalement l'analyse des angles dièdres du squelette, ϕ et ψ, des résidus d'acides aminés individuels le long de la séquence [39, 175]. Une illustration pour trois acides aminés consécutifs est présentée dans la Figure F.5. L'examen des valeurs autorisées et de la distribution statistique de (ϕ, ψ) a fait l'objet d'études depuis plus d'un demi-siècle, commençant par le travail fondateur de Ramachandran *et al.* [237, 238]. L'analyse des angles (ϕ, ψ) dans les chaînes de polypeptides a de nombreuses applications, telles que la validation et l'affinement de structures déterminées par des techniques biophysiques [208, 182], le développement de modèles ou de fonctions de notation pour la prédiction et la conception de la structure des protéines [106, 153, 27, 35, 244, 280] ou l'étude des états dénaturés des protéines globulaires [267, 142] et des protéines intrinsèquement désordonnées [265, 88]. Alors que les valeurs de (ϕ, ψ) sont physiquement restreintes pour les protéines qui adoptent une structure tridimensionnelle stable, elles présentent une grande variabilité pour les protéines intrinsèquement désordonnées (IDP). Par conséquent, nous sommes amenés à considérer la paire (ϕ, ψ) comme une variable aléatoire prenant des valeurs sur le tore plat bidimensionnel $\mathbb{T}^2$, qui est le produit cartésien d'une paire de cercles unitaires. Une définition technique de $\mathbb{T}^2$ est présentée au chapitre 3, où nous

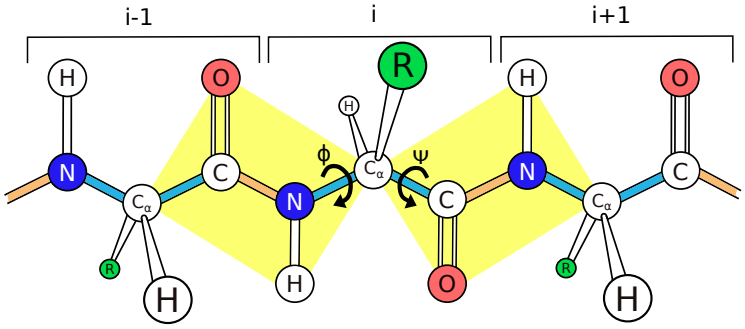


Figure F.5: La détermination des angles de torsion ϕ et ψ. Chaque liaison peptidique (en orange) comporte six atomes dans un plan (en jaune), qui décrivent entièrement la conformation de l'acide aminé i-ème. L'angle ϕ (resp. ψ) détermine la rotation du squelette polypeptidique autour de la liaison N-C_α (resp. C_α-C).

analysons également ses propriétés géométriques et topologiques fondamentales. Comme $\mathbb{T}^2$ peut en effet être doté d'une topologie, nous sommes en mesure de considérer l'ensemble $\mathcal{P}(\mathbb{T}^2)$ des distributions de probabilité supportées sur $\mathbb{T}^2$. En conséquence, pour un acide aminé donné, ses angles dièdres (ϕ, ψ) seront associés à un élément de $\mathcal{P}(\mathbb{T}^2)$, que nous définirons comme le *descripteur structural local* du résidu d'acide aminé.

Definition F.2.5 (Descripteur structural local)**.** *Soit (ϕ, ψ) les angles dièdres aléatoires d'un résidu d'acide aminé. Son descripteur structural local est défini comme la distribution de probabilité de (ϕ, ψ), qui est un élément de $\mathcal{P}(\mathbb{T}^2)$.*

Comme mentionné précédemment, nous cherchons à considérer des descripteurs structuraux qui peuvent être "mesurés", ou en d'autres termes, dont les distributions de probabilité empiriques sont faciles à calculer. C'est le cas des angles (ϕ, ψ), qui peuvent être déterminés expérimentalement avec une grande résolution pour les protéines rigides ou connus lorsque les conformations sont simulées avec les méthodes présentées dans la Section F.1.3. Cela nous conduit à définir le *descripteur structural local empirique* d'un résidu d'acide aminé comme la distribution de probabilité empirique de son descripteur structural local.

Descripteurs structuraux globaux

La description structurelle d'une séquence entière est une tâche plus complexe. Bien que certaines méthodes expérimentales telles que la cristallographie aux rayons X et la cryo-EM, ainsi que des modèles génératifs, soient capables de fournir les coordonnées de tous les atomes de la protéine (pour les protéines structurées), ces coordonnées ne peuvent pas être comparées entre différentes conformations car elles ne se réfèrent pas à un système de référence absolu dans lequel tous les états peuvent être exprimés. De plus, la structure d'un état est invariante sous les transformations de corps rigides ou, de

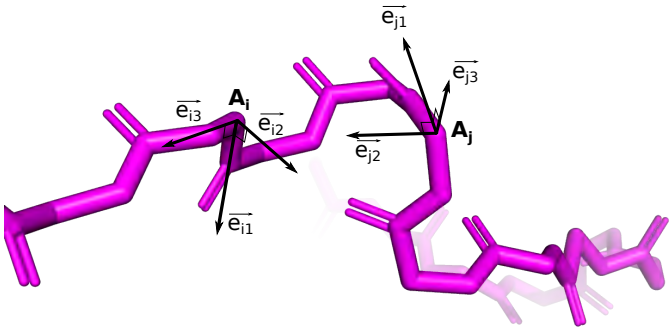


Figure F.6: Illustration des systèmes de référence $\mathcal{F}_i$ et $\mathcal{F}_j$ construits au niveau des résidus i-ème et j-ème.

manière équivalente, sous le changement de base dans l'espace vectoriel euclidien. Par conséquent, décrire la structure globale en utilisant directement les coordonnées de tous les atomes nous conduirait à recourir à la relation d'équivalence suivante[7]. Si n_a désigne le nombre d'atomes dans la séquence, deux éléments x et $y \in \mathbb{R}^{3n_a}$ sont équivalents, notés $x \sim y$, si et seulement s'ils sont égaux jusqu'à une transformation de corps rigide près. Notez que, en effet, $\sim$ est une relation d'équivalence car l'espace des mouvements de corps rigides est - entre autres - un groupe. Un tel espace est appelé le groupe spécial euclidien de trois dimensions et est généralement noté $SE(3)$ [135]. Ensuite, nous pourrions définir un descripteur structural global pour l'ensemble de la séquence comme une distribution de probabilité supportée sur l'espace quotient $\mathbb{R}^{3n_a}/\sim$. Bien que mathématiquement stimulant, recourir à $\mathcal{P}(\mathbb{R}^{3n_a}/\sim)$ pour définir des descripteurs structuraux est inutilement compliqué et frôle la pédanterie mathématique. Pour capturer la structure de l'ensemble de la séquence, nous proposons de construire un référentiel à chaque résidu d'acide aminé en utilisant les atomes du squelette. Une illustration pour une paire de résidus est présentée dans la Figure F.6.

Soit L la longueur de la séquence et A_i le i-ème acide aminé, pour $i \in \{1, \ldots, L\}$. En utilisant les coordonnées des atomes C, C_α et N du i-ème résidu, nous pouvons définir un système de référence qui tient compte de la configuration géométrique du squelette au niveau du i-ème résidu. L'origine du repère de référence est fixée aux coordonnées de l'atome C_β, c'est-à-dire le premier atome de la chaîne latérale (voir Figure F.1) pour les résidus non-glyciniques. Pour les glycines, nous plaçons l'origine aux coordonnées de l'atome C_α. Si nous notons $\mathcal{F}i = \{\overrightarrow{e_{i1}}, \overrightarrow{e_{i2}}, \overrightarrow{e_{i3}}\}$ le i-ème système de référence, la structure globale de l'ensemble est décrite par L cadres de référence $\mathcal{F}_1, \ldots, \mathcal{F}_L$. Notez que chaque $\mathcal{F}_i$ peut être formalisé comme un élément de $SE(3)$. Outre la complexité de la

[7]Une relation binaire $\sim$ sur un ensemble $\mathcal{X}$ est dite une *relation d'équivalence* si elle est réflexive, symétrique et transitive. L'ensemble de tous les éléments de $\mathcal{X}$ qui sont équivalents à $x \in \mathcal{X}$ est appelé la *classe d'équivalence* de x. L'ensemble des classes d'équivalence de tous les éléments de $\mathcal{X}$ est appelé l'ensemble quotient de $\mathcal{X}$ par $\sim$, noté $\mathcal{X}/\sim$ [307].

comparaison de cadres entre différentes conformations, qui peut être résolue dans certains cas, il convient de noter que s'appuyer sur $SE(3)$ est extrêmement complexe et nécessite la manipulation de la géométrie riemannienne. Bien que cet espace soit largement utilisé en robotique [17, 315, 223], son application dans ce contexte reste une tâche excessivement compliquée et impraticable, en raison, par exemple, de la non-unicité de ses courbes géodésiques, ce qui entrave le calcul des distances [226]. Bien que certaines contributions remarquables traitant de distributions de probabilité et de statistiques dans $SE(3)$ aient été récemment proposées [54, 205], nous préférons définir des descripteurs euclidiens de la famille $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ qui permettent leur comparaison directe entre les états et un calcul efficace.

Deux stratégies différentes seront suivies en fonction de la nécessité de comparer ou de caractériser les ensembles. La première reposera sur la mise en correspondance de la famille $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ avec un produit cartésien d'espaces euclidiens tridimensionnels. Cette transformation aboutira à la définition du *descripteur structural global tridimensionnel* de l'ensemble.

Definition F.2.6 (Descripteur structural global tridimensionnel)**.** *Soit L la longueur de la séquence de la protéine et $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ la famille de systèmes de référence construits à chaque résidu d'acide aminé. Soit*

$$\begin{aligned} T_{\mathbb{R}^3} \colon SE(3) \times \overset{L}{\cdots} \times SE(3) &\longrightarrow \mathbb{R}^3 \times \overset{L(L-1)/2}{\cdots} \times \mathbb{R}^3 \\ (\mathcal{F}_1, \ldots \mathcal{F}_L) &\longmapsto \left(\overrightarrow{R}_{11}, \ldots, \overrightarrow{R}_{(L-1)L}\right) \end{aligned} \tag{F.2}$$

la transformation qui mappe la famille de cadres de référence à toutes les positions relatives de chaque paire de résidus le long de la séquence. Plus précisément, $\overrightarrow{R}_{ij}$ représente les coordonnées de l'origine de $\mathcal{F}_j$ par rapport à $\mathcal{F}_i$, pour chaque $i < j$. Le descripteur structural global tridimensionnel de l'ensemble est défini comme le $L(L-1)/2$-uplet

$$\left(P_{11}, \ldots, P_{(L-1)L}\right) \in \mathcal{P}(\mathbb{R}^3) \times \overset{L(L-1)/2}{\cdots} \times \mathcal{P}(\mathbb{R}^3), \tag{F.3}$$

où $P_{ij} \in \mathcal{P}(\mathbb{R}^3)$ est la distribution de probabilité de $\overrightarrow{R}_{ij}$, pour chaque $i < j$.

En effet, la définition d'un repère de référence à chaque acide aminé permet la détermination de la position relative de chaque paire de résidus. Ces positions seront des variables aléatoires prenant des valeurs dans $\mathbb{R}^3$ et leurs distributions de probabilité (F.3) serviront de descripteurs structuraux globaux de l'ensemble des protéines. Notez également que les réalisations de chaque $\overrightarrow{R}ij$ sont comparables entre les conformations. En effet, la transformation (F.2) convertit la configuration structurelle de l'ensemble de la séquence en un ensemble de descripteurs euclidiens tridimensionnels qui ne dépendent pas des coordonnées absolues fournies en entrée. En d'autres termes, les réalisations de $\overrightarrow{R}ij$ sont accessibles et comparables, permettant la définition du *descripteur structural global tridimensionnel empirique* de l'ensemble en tant que famille des contreparties empiriques de (F.3).

Une approche différente sera choisie lorsque l'objectif est de caractériser les ensembles de protéines. Dans ce cas, la famille de cadres de référence sera mise en correspondance avec un produit cartésien d'intervalles réels. Au lieu d'analyser toutes les positions relatives des paires d'acides aminés, nous tiendrons désormais compte des interactions entre résidus qui apparaissent dans chaque conformation de protéine.

Definition F.2.7 (Descripteur structural global unidimensionnel)**.** *Soit L la longueur de la séquence de protéine et $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ la famille de cadres de référence construits à chaque résidu d'acide aminé. Soit*

$$\begin{aligned} T_{[0,1]} \colon SE(3) \times \overset{L}{\cdots} \times SE(3) &\longrightarrow [0,1] \times \overset{L(L-1)/2}{\cdots} \times [0,1] \\ (\mathcal{F}_1, \ldots \mathcal{F}_L) &\longmapsto \left(\omega_{11}^C, \ldots, \omega_{(L-1)L}^C\right) \end{aligned} \tag{F.4}$$

la transformation qui mappe la famille de cadres de référence à un vecteur d'éléments dans $[0,1]$, agissant comme un proxy pour l'interaction entre les résidus i et $j > i$. Le descripteur structural global unidimensionnel de l'ensemble est défini comme le $L(L-1)/2$-uplet

$$\left(P_{11}^C, \ldots, P_{(L-1)L}^C\right) \in \mathcal{P}([0,1]) \times \overset{L(L-1)/2}{\cdots} \times \mathcal{P}([0,1]), \tag{F.5}$$

où $P_{ij}^C \in \mathcal{P}([0,1])$ est la distribution de probabilité de ω_{ij}^C, pour chaque $i < j$.

Les quantités ω_{ij}^C tiennent compte du contact entre les acides aminés aux positions i et $j > i$. Ces variables seront conçues comme une extension de la notion classique binaire de contact, qui est basée sur un seuil universel pour la distance euclidienne [213, 275, 229]. Dans ce cas, la transformation (F.4) transformera la famille $\{\mathcal{F}_1, \ldots, \mathcal{F}_L\}$ en un tuple de valeurs qui varieront de manière continue dans $[0,1]$ et dépendront non seulement de l'identité des acides aminés en interaction et de leur position dans la séquence, mais aussi de l'orientation relative de $\mathcal{F}_i$ et $\mathcal{F}_j$. Une fois de plus, ces quantités sont comparables entre les conformations et permettent la définition du *descripteur structural global unidimensionnel empirique* sous forme du vecteur de distributions de probabilité empiriques de (F.5).

F.2.3 Outils statistiques pour comparer et caractériser les ensembles

La structure des ensembles de protéines sera décrite par les distributions de probabilité présentées dans les Définitions F.2.5, F.2.6 et F.2.7. La prochaine étape consiste à trouver des outils statistiques qui capturent le plus fidèlement possible la variabilité de ces distributions et fournissent des résultats compacts, clairs et interprétables tenant compte de la variabilité conformationnelle des protéines flexibles et des changements dans leur structure secondaire. Parfois, des outils appropriés à cette fin ont déjà été trouvés dans la littérature, correspondant même à des méthodes standard utilisées en biostatistique et en mathématiques appliquées de manière générale. Cependant, les problèmes de comparaison et de caractérisation des ensembles soulèvent naturellement des questions auxquelles il n'y avait

pas encore de réponse méthodologique. Dans ce cas, des contributions théoriques applicables dans un contexte plus général ont été proposées. Il convient de noter que, plutôt que l'étude de problèmes statistiques, la motivation de cette thèse est de fournir des réponses à des problèmes ouverts en biologie structurale. Néanmoins, cela a naturellement conduit au développement de certaines techniques mathématiques qui peuvent être d'intérêt d'un point de vue plus large. Les familles de méthodes statistiques utilisées dans cette thèse sont décrites dans les sections suivantes. Nous omettrons les détails sur le premier groupe car il comprend des techniques standard couramment utilisées en biologie structurale et en biostatistique. Nous porterons davantage d'attention aux suivantes, dont l'application en biologie structurale est plus novatrice et où nos contributions méthodologiques ont été faites.

Clustering sur un espace de basse dimension (utilisé dans le chapitre 7)

La réduction de dimension est une technique largement utilisée en biostatistique en raison de la dimension intrinsèquement élevée des données biologiques. La plupart des applications de cette théorie sont liées aux domaines très actifs de la neuroimagerie [211], de la biologie cellulaire unique [15, 272] ou de la génétique [83, 81], entre autres. Ici, nous nous concentrerons sur les algorithmes de réduction de dimension non linéaires, qui ont montré des performances empiriques efficaces pour identifier les structures sous-jacentes dans des données complexes [80, 81, 15, 174, 232, 83]. En particulier, nous utiliserons l'algorithme Uniform Manifold Approximation and Projection (UMAP) [199]. Ce choix est motivé par sa capacité à préserver la topologie en haute dimension des données et à révéler efficacement la structure des populations [81, 83, 15]. Depuis quelque temps, la combinaison d'algorithmes de réduction de dimension non linéaires avec des techniques de clustering est devenue une procédure standard pour détecter les structures révélées par la projection en basse dimension et les classer en groupes bien définis. L'utilisation de cette stratégie est étayée par son efficacité empirique réussie [82, 2, 15, 83]. Dans cette thèse, nous proposons de projeter des données en haute dimension dans un espace UMAP en basse dimension et d'appliquer l'algorithme HDBSCAN [46] sur cet espace, que nous considérons comme l'une des techniques basées sur la densité les plus sophistiquées. Les principes fondamentaux des algorithmes UMAP et HDBSCAN sont expliqués dans l'Annexe E.1.

Le transport optimal (utilisé dans les chapitres 3-5)

Le Transport Optimal (TO) est une théorie mathématique qui a gagné une importance considérable ces dernières années en raison de son applicabilité efficace et polyvalente. En particulier, la popularité du TO a augmenté grâce à son intégration dans les techniques d'apprentissage automatique, notamment dans le cadre des réseaux générateurs [9], de la robustesse [262] ou de l'équité [76, 69, 31], entre autres. À quelques exceptions notables près [47, 19, 248, 67, 107], le TO n'a pas été largement utilisé en biologie structurale. Dans cette thèse, nous proposons de nous appuyer sur le TO pour rendre compte des différences

entre les descripteurs structuraux globaux et locaux. Commençons par introduire les principaux concepts de cette théorie.

Le Transport Optimal est un cas spécifique de *transport de masse*, qui est le problème général de faire correspondre deux distributions de probabilité P et Q définies sur un espace polonais $\mathcal{X}$, c'est-à-dire un espace topologique séparable et complètement métrisable [210]. À noter que l'espace euclidien de dimension arbitraire est polonais, tout comme le tore plat bidimensionnel $\mathbb{T}^2$, comme cela a été montré dans le chapitre 3. Par conséquent, cette théorie est applicable aux distributions qui composent les descripteurs structuraux locaux et globaux[8] définis dans la Section F.2.2. Le problème du transport de masse vise à sélectionner un couplage dans $\Pi(P, Q)$, c'est-à-dire une distribution de probabilité conjointe ayant P et Q comme marginales.

Un couplage peut être vu comme une correspondance aléatoire, faisant correspondre chaque instance du support de P à éventuellement plusieurs contreparties du support de Q avec des poids de probabilité. Cette transformation peut également être comprise comme une reconfiguration de la *masse de probabilité* de P pour récupérer celle de Q. Plus visuellement, on pourrait penser à chaque distribution marginale comme un tas de sable sur $\mathcal{X}$. Un couplage est un plan de transport transformant un tas en l'autre, qui spécifie comment déplacer chaque masse élémentaire de sable de la première distribution pour récupérer la seconde. Un couplage est dit *déterministe* si chaque instance de P est appariée à une unique instance de Q. Dans ce cas, le couplage est localisé sur le graphe d'une application (P-presque sûrement unique[9]) $T : \mathcal{E} \longrightarrow \mathcal{E}$ qui envoie P vers Q, c'est-à-dire que $T_{\#}P = Q$. Nous notons par $\mathcal{T}(P, Q)$ l'ensemble des applications mesurables qui envoient P vers Q.

Le Transport Optimal est devenu un outil populaire pour définir de tels couplages en sélectionnant ceux qui sont optimaux d'une certaine manière. Cette théorie remonte à Monge [207] qui, en 1781, a défini les applications de TO comme des fonctions qui transforment P en Q avec un effort minimal selon une fonction de coût positive $c : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^+$. Formellement, les applications de transport sont définies comme les solutions de

$$\inf_{T \in \mathcal{T}(P,Q)} \int_{\mathcal{X}} c(x, T(x))\, dP(x). \tag{Monge}$$

Un défi mathématique surgit de la contrainte de la push-forward, ce qui rend le problème irréalisable dans de nombreux scénarios généraux, en particulier lorsque les distributions P et Q ne sont pas absolument continues par rapport à la mesure de Lebesgue [29] ou ont un nombre d'atomes déséquilibré. Cette complication a motivé la relaxation du problème de TO appelée relaxation de Kantorovich, introduite par Kantorovich et Rubinshtein en

[8] Comme les sous-ensembles fermés d'espaces polonais sont également polonais, cela s'applique également aux distributions composant les descripteurs structuraux unidimensionnels locaux introduits dans la Définition F.2.7.

[9] C'est-à-dire que s'il existe une autre correspondance $T' \neq T$ dont le graphe localise le même couplage, elle ne diffère de T que sur un ensemble O avec $P(O) = 0$.

1958 [154]:

$$\inf_{\gamma \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\, d\gamma(x,y). \qquad \text{(Kantorovich)}$$

Les solutions à l'équation (Kantorovich) sont des couplages optimaux (en général non déterministes) entre P et Q par rapport au coût c. Contrairement aux cartes de TO, elles existent sous des hypothèses très légères, comme la non-négativité du coût [299]. Notez que, puisqu'un opérateur de push-forward peut être identifié avec un couplage, l'ensemble des solutions admissibles de l'équation (Monge) est inclus dans l'ensemble des solutions admissibles de l'équation (Kantorovich).

Les solutions de (Kantorovich) nous intéressent particulièrement car elles définissent une distance dans $\mathcal{P}(\mathcal{X})$ [299]. Plus précisément, pour $p \geq 1$, la valeur optimale

$$\mathcal{W}p(P,Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \int \mathcal{X} \times \mathcal{X} c^p(x,y)\, d\gamma(x,y) \right)^{\frac{1}{p}} \tag{F.6}$$

est appelée la distance de Wasserstein p entre P et Q, et elle représente le coût minimum de transport nécessaire pour reconfigurer la masse de P afin de récupérer la masse de Q. Notez que la distance de Wasserstein est capable d'intégrer la géométrie de l'espace sous-jacent $\mathcal{X}$ si la fonction de coût est choisie, par exemple, comme la distance géodésique sur $\mathcal{X}$. Cela en fait une métrique bien adaptée pour capturer la variabilité de l'espace conformationnel et comparer de manière appropriée une paire de descripteurs structuraux.

Nous concluons en présentant comment résoudre (Kantorovich) lorsque, en pratique, nous n'avons accès qu'aux équivalents empiriques de P et Q, notés P_n et Q_m pour $n, m \in \mathbb{N}$. Ce scénario correspond à la version *discrète* du problème de Kantorovich, où les points de l'échantillon tiré de P sont envoyés aux points de l'échantillon tiré de Q avec des probabilités données par une matrice $n \times m$, que nous identifions avec le couplage dans (Kantorovich). Soient X et Y deux variables aléatoires ayant respectivement P et Q comme distributions de probabilité, $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons de X et Y et $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_m)$ deux réalisations de tels échantillons. La version discrète de (Kantorovich) correspond à résoudre

$$\inf_{M \in U(P_n, Q_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j)\, M_{ij}, \tag{F.7}$$

où $U(P_n, Q_m)$ est l'ensemble des matrices réelles $n \times m$ $M = (M_{ij})_{ij}$ telles que $\sum_{i=1}^{n} M_{ij} = m^{-1}$ et $\sum_{j=1}^{m} M_{ij} = n^{-1}$. Une fois de plus, la distance de Wasserstein p entre les mesures empiriques P_n et Q_m est donnée par

$$\mathcal{W}p(P_n, Q_m) = \left(\inf_{M \in U(P_n, Q_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} c^p(X_i, Y_j)\, M_{ij} \right)^{\frac{1}{p}}. \tag{F.8}$$

Notez que (F.6) est un nombre réel positif tandis que (F.8) est une variable aléatoire, car elle dépend des échantillons tirés de P et de Q. Heureusement, la distance de Wasserstein

empirique susmentionnée (F.8) présente de solides garanties statistiques. En particulier, elle converge faiblement vers la distance entre les mesures réelles (F.6) à mesure que n et m augmentent jusqu'à l'infini sous des hypothèses légères [299, Corollaire 6.9]. Cela justifie l'utilisation de (F.8) pour prendre en compte les différences entre les descripteurs structuraux locaux et globaux en calculant la distance de Wasserstein entre leurs équivalents empiriques. Pour ces applications pratiques, nous fixerons $p = 2$ en raison des propriétés statistiques bien connues associées au coût quadratique, en particulier l'unicité de la solution de (Kantorovich) sous des hypothèses légères [298, Théorème 2.12]. Pour une compréhension approfondie des propriétés mathématiques de la distance de Wasserstein et du problème du transport optimal, nous vous renvoyons à [299].

La résolution du problème d'optimisation (F.7) est devenue un autre domaine de recherche étendu. Comme la fonction objectif et les contraintes sont linéaires dans les variables d'intérêt, la formulation discrète du problème de Kantorovich est un programme linéaire. Par conséquent, il peut être résolu avec une grande variété d'outils algorithmiques issus de la programmation linéaire et de l'optimisation combinatoire. Parmi eux, nous pouvons souligner l'algorithme classique du Simplexe en réseau [24], qui est implémenté dans les solveurs OT les plus courants [98, 259]. Une autre stratégie populaire est celle des méthodes de Dual Ascent [144], notamment le célèbre algorithme hongrois [25]. Le principal défi rencontré lors de la manipulation de solutions empiriques de transport optimal réside dans leur complexité computationnelle élevée et leurs besoins en mémoire. La résolution de (F.7) demande généralement $O((n+m)nm\log(n+m))$ opérations informatiques. De plus, pour des fonctions de coût non standard, une matrice de coefficients $C_{ij} = c(x_i, y_j)$ de taille $n \times m$ doit être stockée. Dans les problèmes pratiques, notamment dans les applications d'apprentissage automatique, il est courant de considérer des schémas de régularisation entropique, qui peuvent réduire la complexité computationnelle à $O(nm)$ opérations [64]. Cependant, ces approximations ne résolvent pas les problèmes de mémoire et perdent les propriétés mathématiques et statistiques qui motivent l'utilisation de la distance de Wasserstein dans le contexte inférentiel. Pour une introduction moins technique à l'OT et une analyse approfondie des aspects computationnels discutés ci-dessous, nous renvoyons à [228].

Test d'hypothèse (utilisé dans les chapitres 2-6)

Les tests d'hypothèses, tout comme l'estimation, constituent l'autre pilier fondamental de l'inférence statistique. L'objectif de ces tests est de déterminer, en se basant sur les informations collectées dans l'échantillon, si une certaine hypothèse concernant une caractéristique de la population doit être rejetée ou non. En résumé, tester une hypothèse revient à déterminer si elle est "compatible" avec ce qui est observé dans l'échantillon. Plus précisément, cela implique de comparer la validité de deux énoncés complémentaires sur la population. L'un d'entre eux est appelé l'hypothèse nulle (H_0), tandis que l'autre est appelé l'hypothèse alternative (H_1). Il convient de noter que les tests statistiques ne sont pas symétriques par rapport à H_0 et H_1 dans le sens où ils ne choisissent pas l'hypothèse

la plus plausible en se basant sur l'échantillon. Au lieu de cela, leur objectif est simplement de déterminer s'il existe suffisamment de preuves pour rejeter ce que H_0 affirme. Par conséquent, le test ne conclut jamais que l'hypothèse nulle est vraie, mais plutôt qu'il n'y a pas suffisamment de preuves pour la rejeter. Formellement, nous pouvons définir un test statistique comme une partition mesurable [61, Definition 15.1] de l'espace des échantillons.

Definition F.2.8 (Test d'hypothèse)**.** *Soit* $(\Omega, \Sigma, \mathbb{P})$ *un espace de probabilité. Un test d'hypothèse est une partition mesurable de* Ω *en deux régions, à savoir la région critique -ou de rejet- (C.R.), qui comprend les instances qui conduisent au rejet de* H_0*, et son complémentaire, la région alternative (A.R.), composée des résultats qui ne conduisent pas au rejet de* H_0*. Le test est caractérisé par la fonction indicatrice de la région critique, également appelée fonction de test,* $\pi \colon \Omega \longrightarrow \{0, 1\}$ *où*

$$\pi(\omega) = \begin{cases} 1 & si\ \omega \in \text{C.R.} \\ 0 & si\ \omega \in \text{A.R.} \end{cases} \tag{F.9}$$

Le potentiel des tests d'hypothèses réside dans les garanties statistiques qu'ils offrent en ce qui concerne la partition (F.9). Plus précisément, les fonctions de test sont conçues pour garantir le contrôle de ce qu'on appelle l'*erreur de type I*, c'est-à-dire la probabilité de rejeter H_0 lorsqu'elle est vraie.

Definition F.2.9 (Erreur de type I)**.** *Soit* $(\Omega, \Sigma, \mathbb{P})$ *un espace probabiliste et* $\pi : \Omega \longrightarrow \{0, 1\}$ *une fonction de test. L'erreur de type I de* π *est définie comme la probabilité de rejeter* H_0 *lorsqu'elle est vraie, c'est-à-dire*

$$\mathbb{P}\left(\{\omega \in \Omega \,:\, \pi(\omega) = 1 \,|\, H_0\}\right) = \mathbb{P}_{H_0}\left(\text{C.R.}\right).$$

La procédure de construction d'un test d'hypothèse commence par fixer une limite supérieure pour l'erreur de type I, appelée *niveau de signification*, et choisir, parmi tous les tests qui la contrôlent, celui qui détecte le plus efficacement les fausses hypothèses nulles, ou en d'autres termes, le plus *puissant*.

Definition F.2.10 (Puissance)**.** *Soit* $(\Omega, \Sigma, \mathbb{P})$ *un espace probabiliste et* $\pi : \Omega \longrightarrow \{0, 1\}$ *une fonction de test. La puissance de* π *est définie comme la probabilité de rejeter* H_0 *lorsqu'elle est fausse, c'est-à-dire*

$$\mathbb{P}\left(\{\omega \in \Omega \,:\, \pi(\omega) = 1 \,|\, H_1\}\right) = \mathbb{P}_{H_1}\left(\text{C.R.}\right).$$

Il convient de noter que la construction d'un test d'hypothèse repose sur le choix d'une "bonne" région critique. Pour ce faire, les considérations suivantes doivent être prises en compte:

(*i*) Les divergences par rapport à l'hypothèse nulle sont recherchées, de sorte que la région critique doit inclure des valeurs de l'échantillon qui sont peu probables de se produire sous H_0, même si elles sont possibles,

(*ii*) En général, la région critique est déterminée avant d'analyser les résultats expérimentaux (bien que cela ne soit pas toujours vrai, comme discuté dans la section suivante),

(*iii*) La région critique est généralement exprimée en termes d'une statistique, c'est-à-dire une fonction mesurable à valeurs réelles de l'échantillon, appelée *statistique de test*. Elle mesure les divergences entre les échantillons dans la région critique et l'hypothèse nulle. La distribution de la statistique de test sous H_0 est utilisée pour garantir le niveau de signification.

Il convient de noter que la fonction de test (F.9) fournit une sortie binaire « rejet vs non-rejet » concernant H_0. Cependant, que l'hypothèse nulle soit rejetée ou non, il est généralement intéressant de "mesurer la distance" entre le résultat de l'échantillon et H_0. Cela donne lieu au concept de p-valeur.

Definition F.2.11 (p-valeur)**.** *Soit $(\Omega, \Sigma, \mathbb{P})$ un espace probabiliste, π une fonction de test définie sur Ω dont la statistique de test est une fonction d'un échantillon $X_1, \ldots, X_n$. La p-valeur associée à une réalisation $(x_1, \ldots, x_n)$ est le plus petit niveau de signification auquel l'hypothèse nulle H_0 est rejetée par π.*

Cette valeur agit comme un proxy pour la plausibilité de l'échantillon sous l'hypothèse nulle. Si la p-valeur est grande, cela signifie que nous travaillons avec des échantillons ayant une forte probabilité de se produire si H_0 est vrai. Dans ce cas, il n'y a pas suffisamment de preuves contre la nullité et H_0 ne devrait pas être rejetée. Cependant, le rejet devrait être choisi si la p-valeur est petite. Nous tenons à souligner que la p-valeur peut -et doit- être considérée comme un indicateur quantitatif de la "crédibilité" de l'hypothèse nulle. En conséquence, avec les précautions appropriées, les p-valeurs calculées dans les mêmes conditions (par exemple, égalité des tailles d'échantillons) sont quantitativement comparables et fournissent un indicateur correct des échantillons qui contredisent le plus la validité de H_0 au sein d'une famille de réalisations. Ce point sera essentiel dans les travaux présentés ici. De plus, il convient de noter que la p-valeur peut également être utilisée pour vérifier efficacement la bonne définition d'un test d'hypothèse. Conformément à sa définition, une p-valeur est statistiquement valide si et seulement si elle est super-uniforme sous H_0. Une variable aléatoire réelle X est dite super-uniforme si sa fonction de répartition cumulative (CDF) est majorée par celle de la distribution uniforme, c'est-à-dire:

$$\mathbb{P}(X \leq x) \leq x \text{ pour tout } x \text{ dans } [0, 1]$$

(voir, par exemple, [172, Section 3.3]). De plus, plus la distribution des p-valeurs sous l'hypothèse nulle est proche de $U[0, 1]$, plus le test correspondant est puissant. Vérifier la super-uniformité des p-valeurs sous H_0 est essentiel pour garantir l'adéquation du test statistique correspondant.

Dans cette thèse, nous nous concentrons principalement sur un cas particulier de test statistique, connu sous le nom de test de bonne adéquation à deux échantillons. En bref, il

vise à évaluer si deux distributions de probabilité sont identiques. Plus précisément, pour deux mesures P et Q prises en charge sur un espace polonais $\mathcal{X}$, l'objectif est de tester les hypothèses nulle et alternative suivantes:

$$H_0 : P = Q \qquad \text{vs.} \qquad H_1 : P \neq Q. \tag{F.10}$$

Notez que, dans ce cadre, nous testons uniquement l'égalité de P et Q, indépendamment de l'identité de ces distributions. La question clé ici est le choix d'une statistique de test appropriée qui tienne compte adéquatement des différences entre P et Q et dont la distribution nulle est connue[10]. Les approches les plus couramment utilisées pour tester (F.10) sont principalement définies pour les mesures prises en charge sur la ligne réelle (par exemple, les statistiques de Kolmogorov-Smirnov et de Wilcoxon). Cependant, tester l'égalité de distributions prises en charge sur des espaces plus généraux est un problème beaucoup moins étudié, et cela revêt une grande importance dans notre objectif de comparer correctement les descripteurs structuraux locaux. Les distributions de probabilité rendant compte de la variabilité conformationnelle de la protéine à l'échelle des acides aminés (cf. la Définition F.2.5) sont prises en charge sur le tore plat bidimensionnel, qui est un espace non euclidien. En particulier, une statistique de test définie pour comparer des distributions dans $\mathcal{P}(\mathbb{T}^2)$ doit être adaptée à la périodicité de leur support. C'est pourquoi la distance de Wasserstein (F.6) s'avère être une mesure appropriée pour comparer les mesures sur $\mathbb{T}^2$ si la distance géodésique sur un tel espace est choisie comme fonction de coût. Dans cette thèse, nous proposons deux approches pour définir des tests de bonne adéquation à deux échantillons dans $\mathcal{P}(\mathbb{T}^2)$ en utilisant la distance de Wasserstein comme statistique de test. Cela fournira une *évidence statistique* des divergences entre les descripteurs structuraux locaux ou, en d'autres termes, de la signification statistique des changements dans la structure protéique locale.

Inférence post-sélection (utilisée dans le chapitre 6)

Lorsque l'on effectue une enquête statistique, un modèle pour les données doit être préalablement spécifié. Ce modèle peut être la distribution sous-jacente de l'échantillon, les variables qui expliquent un résultat donné ou une hypothèse à tester. Dans un contexte plus classique, le modèle est défini avant la collecte des données. Cela peut être le cas si, par exemple, les observations suivent une loi physique connue. Cependant, dans des applications plus réalistes, l'inférence est effectuée sur un modèle qui est choisi *à partir des données*. Un exemple simple est de tester la signification des caractéristiques sélectionnées par un modèle de régression dont les coefficients ont été obtenus à partir des données. Dans ce cas, les hypothèses nulles à tester, c'est-à-dire les *questions* auxquelles l'inférence doit répondre, *dépendent des données*. Si les mêmes données sont utilisées pour l'étape de test subséquente, les garanties statistiques ne sont pas assurées. Ce phénomène est similaire au surajustement dans les tâches de prédiction. Adapter les statistiques inférentielles

[10]Nous faisons généralement référence à la distribution de toute variable aléatoire sous H_0 comme sa *distribution nulle*.

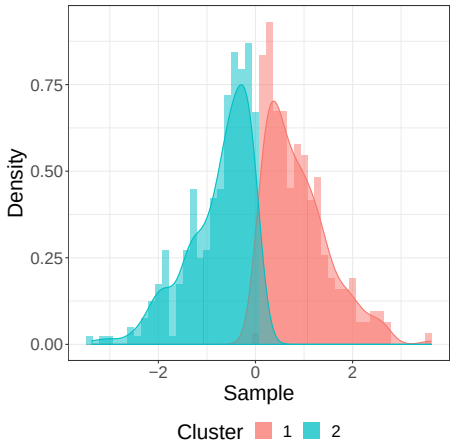


Figure F.7: Distributions empiriques de deux groupes d'observations obtenues après avoir simulé un échantillon de taille $n = 500$ prélevé dans la distribution univariée $\mathcal{N}(0, 1) + \mathcal{U}(-0, 2, 0, 2)$ et l'avoir classifié en deux groupes à l'aide d'un algorithme de k-means. Les couleurs indiquent les classes obtenues par le pipeline. Cette figure est adaptée de [121, Figure 1].

au cadre où le choix du modèle est guidé par les données est l'objectif de l'inférence *post-sélection*. La pertinence de ce domaine a considérablement augmenté ces dernières années en raison de son utilité dans de nombreuses disciplines telles que l'inférence causale [16], la régression linéaire en haute dimension [171] ou les réseaux neuronaux [300], entre autres. Ici, nous nous concentrons sur les *tests d'hypothèses sélectifs*, qui abordent la définition de tests statistiques lorsque les hypothèses nulles sont choisies à partir des données. Les bases de cette théorie ont été introduites de manière rigoureuse dans [97].

Dans cette thèse, notre objectif est de réaliser une inférence après la phase de regroupement en testant les différences entre les moyennes des groupes. Les algorithmes de regroupement visent à classer les observations en un certain nombre de classes, qui est généralement déterminé directement ou indirectement. Les résultats de ces méthodes sont très sensibles aux paramètres requis par chaque technique, et la partition résultante de l'espace peut fortement varier pour un jeu de données donné. Bien qu'il existe des critères pour optimiser le réglage de ces paramètres [278, 251, 119], il convient de souligner qu'en général, il n'y a pas de classification vraie sous-jacente [90]. Les algorithmes de regroupement cherchent à trouver des classes qui représentent de manière compacte la distribution de l'ensemble de données, mais ces groupes ne sont pas inhérents à la population en général et ne servent que de description de l'échantillon. L'inférence après la phase de regroupement vise à éclairer cette question en fournissant des preuves statistiques des véritables différences entre les groupes. La pertinence de ce problème peut être facilement illustrée en simulant une variable aléatoire gaussienne unidimensionnelle avec un bruit uniforme

et en demandant à l'algorithme classique des k-means [117] de trouver deux groupes. Le résultat est présenté dans la Figure F.7. Si, comme c'est le cas en pratique, la distribution sous-jacente est inconnue, il est difficile d'évaluer visuellement si les deux groupes correspondent à deux populations différentes. Si nous essayons de répondre à cette question en omittant que l'hypothèse nulle a été choisie en regardant les données, c'est-à-dire en suivant les résultats d'un algorithme de regroupement, la conclusion sera trompeuse. En effet, un test Z classique renvoie une p-valeur inférieure à la précision machine ($5.87 \cdot 10^{-67}$), conduisant à un rejet fort de l'hypothèse de moyennes de groupe égales. Si nous effectuons une inférence sélective, l'approche présentée dans [51] pour tester les différences de moyennes de groupe renvoie une p-valeur égale à 0.84, ce qui est cohérent avec la configuration réelle (ici connue, car les données ont été simulées).

Notre motivation ici est de fournir des garanties statistiques sur les algorithmes de regroupement couramment utilisés pour caractériser les ensembles de protéines. En effet, définir une partition de l'espace conformationnel en caractérisant les états par des descripteurs classiques est devenu une technique standard [7, 59, 170, 169]. Cependant, ces caractérisations manquent de preuves statistiques sur les véritables différences entre les groupes, que nous considérons essentielles pour interpréter correctement leur résultat.

F.3 Plan de la thèse

Le principal objectif de cette thèse est de définir des méthodes statistiques pour la caractérisation et la comparaison appropriées des ensembles conformationnels de protéines flexibles. Dans la Section F.2.2, nous avons introduit les descripteurs probabilistes de la variabilité structurelle de ces systèmes, et dans la Section F.2.3, nous avons détaillé les méthodes statistiques que nous utiliserons pour les analyser. Il ne nous reste qu'une question à aborder: comment? La réponse à cette question est développée tout au long des chapitres qui composent cette thèse. Le manuscrit est divisé en deux parties principales, consacrées à l'analyse structurale à l'échelle locale et globale, respectivement. À l'intérieur de chaque partie, nous exposons les stratégies mises en place pour déployer les méthodes de la Section F.2.3 sur les descripteurs de la Section F.2.2, et présentons les techniques résultantes pour la caractérisation et la comparaison. Nous fournissons maintenant un bref aperçu de ces méthodes à travers un plan de ce manuscrit, où les principaux résultats et contributions sont mis en évidence.

Disponibilité du logiciel et reproductibilité

Les méthodes de caractérisation et de comparaison présentées dans cette thèse ont été mises à disposition de la communauté. Pour garantir la reproductibilité, le code implémentant toutes les analyses statistiques ainsi que les données ont été partagés de manière équitable. Les liens sont spécifiés dans chaque chapitre.

F.3.1 Analyse structurale locale des ensembles de protéines (Partie I)

La première partie de la thèse est consacrée à l'analyse structurale des ensembles au niveau local. Plus précisément, nous étudierons les distributions de probabilité des angles dièdres (ϕ, ψ) qui définissent les descripteurs structuraux locaux. Cette partie se compose de trois chapitres:

- Dans le chapitre 2, nous évaluons l'effet des acides aminés voisins sur la distribution de (ϕ, ψ), montrant que les identités des résidus gauches et droits doivent être prises en compte *simultanément* pour décrire les structures locales. Cela définit les fragments de trois acides aminés (tripeptides) comme la brique unitaire pour analyser la conformation des protéines localement.

- Le chapitre 3 introduit deux approches pour réaliser des tests de correspondance de deux échantillons sur $\mathcal{P}(\mathbb{T}^2)$, en utilisant la théorie du transport optimal et la distance de Wasserstein. Ces méthodes seront l'outil principal pour prendre en compte les différences statistiquement significatives entre les descripteurs structuraux locaux. Nous illustrons également leur utilité pour rejeter l'hypothèse de la paire isolée de Flory [99].

- Enfin, dans le chapitre 4, nous présentons une application moins triviale des méthodes introduites dans le chapitre 3, à savoir l'évaluation de l'effet du codon traduit sur les distributions de (ϕ, ψ). Ce chapitre fait suite au travail de Rosenberg *et al.* [249], où le même problème a été analysé mais en utilisant une méthodologie inadaptée.

Interdépendance entre les effets des voisins les plus proches (Chapitre 2)

L'analyse structurale des ensembles conformationnels, tant au niveau local que global, doit reposer sur une base solide concernant la manière dont la séquence influence la structure des acides aminés. L'hypothèse de la paire isolée de Flory [99], qui affirme que les angles (ϕ, ψ) d'un résidu donné sont indépendants de l'identité de ses voisins, a déjà été réfutée par la communauté dans de nombreuses études [40, 225, 66, 260, 149, 204] (bien que, comme le montre le chapitre 3, aucune de ces approches ne fournisse de preuve statistique de son rejet). Cependant, une question importante reste sans réponse: les effets des voisins de gauche et de droite sont-ils indépendants? En d'autres termes, la structure locale d'une protéine peut-elle être décrite à partir de deux fragments d'acides aminés (dipeptides), ou l'unité de base devrait-elle être le tripeptide? Des réponses contradictoires ont été proposées à ce sujet [109, 27, 129, 244], mais aucune d'entre elles n'a été basée sur une méthodologie solide qui fournit des garanties statistiques sur la distribution de (ϕ, ψ). Ici, nous visons à *tester* l'indépendance des effets des voisins.

Soit C l'identité d'un résidu d'acide aminé et L, R les identités de ses voisins de gauche et de droite dans la séquence, respectivement. À partir des travaux [280, 244], nous pouvons montrer que le descripteur structurale locale donné par le tripeptide complet,

noté $P(\phi, \psi, |, L, C, R)$, peut être obtenu à partir des descripteurs donnés par les dipeptides de gauche et de droite comme

$$P(\phi, \psi \,|\, L, C, R) = \frac{P(\phi, \psi \,|\, L, C), P(\phi, \psi \,|\, C, R)}{S, P(\phi, \psi \,|\, C)}, \tag{F.11}$$

où S est une constante de normalisation, si et seulement si l'hypothèse suivante est vérifiée

$$L \text{ et } R \text{ sont indépendants étant donné } C \text{ et } (\phi, \psi). \tag{F.12}$$

Notez que (F.11) revient à affirmer que les influences des voisins de gauche et de droite sur la distribution de (ϕ, ψ) peuvent être considérées indépendamment pour la reconstruire. Comme il s'agit d'une déclaration équivalente, nous proposons de tester (F.12) à l'aide d'un test d'indépendance classique χ^2 [172]. Méthodologiquement, une approche appropriée pour conditionner $\{C, \phi, \psi\}$ est proposée, consistant principalement à discrétiser intelligemment $\mathbb{T}^2$ et à effectuer un test par subdivision et valeur de C. Ensuite, tous les p-valeurs sont corrigés pour la multiplicité [125] et stratifiés par l'identité du résidu. Nos résultats démontrent de manière indéniable les effets couplés des voisins de gauche et de droite, indiquant qu'ils ne peuvent pas être considérés indépendamment les uns des autres. De plus, nous montrons que l'ampleur de l'interdépendance, mesurée en termes de p-valeurs, est affectée par les propriétés physico-chimiques des voisins les plus proches et l'origine structurelle des données. Ces observations représentent une étape fondamentale vers la compréhension des relations séquence-structure dans les peptides et les protéines.

Tests à deux échantillons pour comparer les structures locales (Chapitre 3)

L'investigation structurale des angles (ϕ, ψ) implique de quantifier l'amplitude attendue des effets structuraux associés aux changements locaux dans la séquence. Dans ce contexte, la définition d'une distance appropriée entre les distributions sur $\mathbb{T}^2$, dont la signification statistique peut être évaluée, est essentielle. Dans ce chapitre, nous visons à tester les hypothèses

$$H_0 \,:\, P = Q \qquad \text{vs.} \qquad H_1 \,:\, P \neq Q, \tag{F.13}$$

pour une paire de descripteurs structuraux locaux $P, Q \in \mathcal{P}(\mathbb{T}^2)$. Comme précédemment mentionné, les distributions seront comparées à l'aide de la distance de Wasserstein 2 (F.6), qui intègre la géométrie sous-jacente de l'espace conformationnel au niveau local. Comme l'étude du Transport Optimal dans $\mathbb{T}^2$ n'a pas encore été complètement abordée, nous commençons par étendre les principaux résultats de cette théorie au tore plat de dimension arbitraire, noté $\mathbb{T}^d$. En particulier, nous montrons l'unicité sous des hypothèses modérées de la solution de (Kantorovich) dans $\mathcal{P}(\mathbb{T}^d)$, et nous dérivons un Théorème Central Limite (TCL) pour les fluctuations du coût de transport empirique. Nous justifions pourquoi le TCL proposé n'est pas adapté pour définir un test asymptotique de bonté d'ajustement pour (F.13), ce qui motive l'exploration d'approches alternatives.

La stratégie pour définir un test à deux échantillons pour (F.13) doit reposer sur le rejet de l'hypothèse nulle lorsque la distance de Wasserstein entre les contreparties empiriques de P et Q est "trop grande" ou, en d'autres termes, trop improbable sous H_0. Si nous notons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons indépendants identiquement distribués selon P et Q respectivement, cela conduit à la définition de la région critique suivante

$$C.R. = \{(x_1, \ldots, x_n; y_1, \ldots, y_m) \, : \, \mathcal{W}_2^2(P_n, Q_m) \geq c_{nm}(\alpha)\}, \tag{F.14}$$

où x_i (resp. y_j) désigne une réalisation de X_i (resp. Y_j) pour $i = 1, \ldots, n$ (resp. $j = 1, \ldots, m$). Le seuil $c_{nm}(\alpha) > 0$ doit être choisi pour garantir le contrôle de l'erreur de type I au niveau $\alpha \in [0, 1]$ grâce à la distribution nulle de la statistique du test. Cependant, la connaissance de la distribution de $W_2^2(P_n, Q_m)$ sous H_0 demeure un problème ouvert et non trivial, en particulier lorsque l'espace de base a une dimension supérieure à un. Cette difficulté intrinsèque nous a conduit à rechercher des approches alternatives qui exploitent les scénarios où une statistique de test basée sur la distance de Wasserstein à distribution connue peut être définie. Nous proposons deux stratégies à cet égard, que nous décrivons dans les paragraphes suivants.

La première approche consiste à contourner le problème de dimension en projetant P_n et Q_m sur les géodésiques fermées unidimensionnelles de $\mathbb{T}^2$, qui sont des spirales fermées isomorphes au cercle S^1 [36]. Ensuite, nous définissons une statistique de test basée sur la distance de Wasserstein 2 pour comparer les mesures sur S^1 dont nous pouvons dériver la distribution nulle. Nous montrons que cette distribution ne dépend pas des identités de P et Q sous H_0 ou, en d'autres termes, que la statistique définie est *sans distribution* sous l'hypothèse nulle. Cela permet de définir un test de bonne adéquation à deux échantillons pour une paire de projections géodésiques de P_n et Q_m. La stratégie pour définir une valeur de p pour le test bidimensionnel est de (i) choisir une famille de géodésiques fermées sélectionnées aléatoirement sur $\mathbb{T}^2$, (ii) pour chaque géodésique, projeter P_n et Q_m et obtenir une valeur de p pour l'égalité de leurs projections, et enfin, (iii) agréger toutes les p-valeurs par agrégation de Bonferroni [32], obtenant ainsi une p-valeur bien définie pour (F.13). Nous concluons en montrant la cohérence du test sous des alternatives fixes, c'est-à-dire que sa puissance tend vers un lorsque les tailles d'échantillon n et m tendent vers l'infini.

La deuxième approche vise à comparer directement les descripteurs structuraux dans l'espace bidimensionnel. En raison de l'incapacité à construire un test exact ou asymptotique basé sur (F.14), nous proposons de trouver une borne supérieure pour sa p-valeur associée. Il convient de noter qu'une borne supérieure d'une p-valeur donne des tests d'hypothèses statistiquement valides. En effet, si la borne supérieure est inférieure au niveau de signification, il en ira de même pour la vraie -et inconnue- p-valeur et le rejet avec un contrôle de l'erreur de type I sera assuré. En d'autres termes, cela revient à avoir des p-valeurs Super-Uniformes. Cependant, il en résulte une perte de puissance. En résumé, l'idée est de d'abord majorer les écarts de la statistique par rapport à son espérance, puis de montrer que, sous l'hypothèse nulle, cette espérance présente un taux

de convergence rapide vers zéro. Cela conduit à la définition d'un test à deux échantillons qui est *asymptotiquement consistant au niveau* α [292, Definition 14.2]. Cela signifie que les garanties statistiques sont assurées à la limite $n, m \to \infty$, de sorte que, en pratique, le test peut être effectué pour des tailles d'échantillons importantes. Cependant, la conservativité de cet essai à l'échantillon fin devient avantageuse dans le contexte de notre application, le rendant complémentaire à l'approche de projection géodésique.

Une analyse numérique est réalisée pour illustrer l'efficacité relative des deux approches et les comparer à d'autres techniques alternatives de la littérature. De plus, nous démontrons leur pertinence pour détecter les différences entre les descripteurs structuraux locaux, en utilisant une base de données structurales de fragments de trois résidus extraits de structures de protéines à haute résolution déterminées expérimentalement [88] pour réfuter l'hypothèse de la paire isolée de Flory.

L'effet du codon sur la structure locale (Chapitre 4)

Nous concluons la première partie du manuscrit en présentant une application pratique des méthodes introduites dans le Chapitre 3. De nombreux processus biologiques, tels que l'épissage de l'ARN, les taux de traduction et le repliement des protéines, ont démontré la pertinence de l'utilisation de codons synonymes [220, 44]. Bien que la relation entre les codons synonymes et la structure secondaire des protéines traduites ait été largement étudiée [218, 258], Rosenberg *et al.* [249] ont adopté une approche plus détaillée en évaluant l'impact de l'identité des codons sur la distribution des angles dièdres (ϕ, ψ) au sein des éléments de structure secondaire. Leur travail visait à déterminer s'il existe des différences significatives lorsque des codons synonymes sont utilisés, grâce à la mise en œuvre d'un test statistique pour des mesures sur $\mathcal{P}(\mathbb{T}^2)$. Cependant, leur méthodologie statistique est formellement incorrecte, jetant le doute sur les résultats obtenus.

Dans ce chapitre, nous démontrons d'abord que les p-valeurs définies dans [249] sont statistiquement invalides en prouvant que leur distribution n'est pas Super-Uniforme sous l'hypothèse nulle (voir Section F.2.3). De plus, nous montrons que ces p-valeurs sont très conservatrices pour de grandes valeurs de la statistique, entraînant un nombre important de faux négatifs et ignorant ainsi des différences substantielles qui pourraient apparaître entre les descripteurs structuraux locaux. De plus, la procédure de tests multiples utilisée dans [249] échoue à contrôler le taux de découverte de fausses hypothèses (FDR) car elle nécessite que les p-valeurs soient Super-Uniformes sous l'hypothèse nulle [247]. Les inexactitudes techniques de cette étude nous ont incités à étudier l'effet du codon avec des outils statistiques appropriés. C'est la motivation qui nous a poussés à mettre en œuvre les méthodes présentées dans le Chapitre 3 pour tester correctement les différences significatives entre les descripteurs structuraux locaux spécifiques aux codons.

Nos résultats confirment l'influence du codon sur les distributions de (ϕ, ψ), mais diffèrent de ceux de [249] en ce qui concerne la force de signification des différences selon le type de structure secondaire. De plus, nous avons évalué l'impact de la classification structurelle et du contexte de la séquence locale sur ces résultats. Les résultats ont révélé

que les effets spécifiques aux codons présentent des niveaux de signification similaires dans différentes régions de $\mathbb{T}^2$. Cependant, ces effets peuvent être plus prononcés pour des types de structure secondaire spécifiques, tels que les feuillets β par rapport aux hélices α. De plus, les résultats suggèrent que les effets de codons synonymes sont amplifiés lorsqu'on prend en compte le contexte de la séquence locale, ce qui va dans le sens des conclusions du Chapitre 2.

F.3.2 Analyse structurale globale des ensembles de protéines (Partie II)

La deuxième partie du manuscrit est consacrée à l'analyse structurale des protéines flexibles au niveau global. Nous utilisons les descripteurs structuraux globaux définis dans la Section F.2.2 pour développer des outils statistiques permettant de comparer et de caractériser les ensembles conformationnels, tout en fournissant aux techniques de regroupement classiques des garanties statistiques. Cette partie comprend trois chapitres, décrits ci-dessous.

- Le chapitre 5 présente WASCO, un outil statistique basé sur la distance de Wasserstein pour comparer les ensembles conformationnels de protéines hautement flexibles. L'idée principale de WASCO est d'utiliser la distance de Wasserstein pour comparer les descripteurs structuraux globaux tridimensionnels (Définition F.2.6), en intégrant également l'information au niveau local grâce aux techniques présentées dans le chapitre 3. Nous démontrons l'utilité de la méthode pour comparer différents champs de force au sein de simulations de dynamique moléculaire ou pour évaluer l'effet de l'affinement avec des données expérimentales.

- Le chapitre 6 est consacré à l'étude de l'inférence post-regroupement lorsque les données présentent des structures de dépendance arbitraires entre les caractéristiques et les observations. Ce travail, qui constitue l'extension naturelle du cadre de [104, 51], offre aux techniques de regroupement classiques la caractérisation d'ensemble avec des garanties statistiques sur les véritables différences entre les groupes de conformations obtenus.

- Le chapitre 7 présente WARIO, une caractérisation des ensembles conformationnels basée sur les contacts. La méthode adapte les cartes de contacts classiques qui caractérisent les structures repliées au cadre des ensembles, en caractérisant une protéine flexible à travers une famille pondérée de cartes de contacts, construites à partir des descripteurs structuraux globaux unidimensionnels (Définition F.2.7). L'applicabilité de WARIO est illustrée par la caractérisation de plusieurs ensembles conformationnels de protéines intrinsèquement désordonnées (IDP).

Un outil basé sur la distance de Wasserstein pour comparer les ensembles de protéines (Chapitre 5)

La comparaison des ensembles conformationnels est un problème essentiel en biologie structurale. Lorsqu'il s'agit d'ensembles de protéines hautement flexibles, les outils existants proposés dans la littérature se basent sur des descripteurs moyennés sur l'ensemble des conformations [167, 131]. Cependant, réduire des distributions complexes à leur moyenne entraîne généralement une perte considérable d'informations et masque des caractéristiques pertinentes qui pourraient distinguer les systèmes. Dans ce chapitre, nous présentons une technique de comparaison qui intègre l'ensemble de la variabilité de l'espace conformationnel et utilise la distance de Wasserstein pour tenir compte des différences entre les descripteurs probabilistes complets.

L'idée principale de WASCO est de calculer la distance de Wasserstein entre les descripteurs structuraux globaux tridimensionnels de deux ensembles (Définition F.2.6). Plus précisément, pour chaque paire de résidus aux positions $i < j$ dans la séquence, la méthode calcule la distance $\mathcal{W}_{ij} = \mathcal{W}_2(P^A_{ij;n}, P^B_{ij;m})$, où $P^A_{ij;n}$ (resp. $P^B_{ij;m}$) désigne le composant ij du descripteur global structural empirique de l'ensemble A (resp. B). La quantité $\mathcal{W}_{ij}$ est la distance entre la distribution de la position relative des résidus i, j des deux ensembles. La même idée est appliquée pour comparer tous les descripteurs structuraux locaux (Définition F.2.5) pour chaque résidu le long de la séquence. Si $P^A_{i;n}$ (resp. $P^B_{i;m}$) désigne le i-ème composant du descripteur local structural empirique de l'ensemble A (resp. B), WASCO calcule les quantités $\mathcal{W}_i = \mathcal{W}_2(Pi; n^A, Pi; m^B)$, auxquelles nous associons la borne supérieure de la valeur p introduite au Chapitre 3. Remarquez que cette formulation permet une représentation claire et compacte des résultats sous la forme d'une matrice triangulaire, ayant comme entrées les quantités $\mathcal{W}_{ij}$ dans le triangle inférieur et les distances $\mathcal{W}_i$ le long de la diagonale. En combinant toutes les différences structurales aux niveaux local et global dans la même représentation, on peut clairement mettre en évidence les différences spécifiques aux résidus les plus pertinentes et évaluer la relation entre les variations des distributions (ϕ, ψ) au niveau des acides aminés et les désaccords structuraux à l'échelle de la séquence entière.

La variabilité dans les structures expérimentales et simulées entraîne des incertitudes et du bruit statistique qui peuvent biaiser considérablement l'estimation de la distance. Par conséquent, les différences calculées entre les descripteurs structuraux globaux et locaux sont corrigées pour filtrer ce bruit et mettre en évidence les désaccords pertinents entre les ensembles. Cette correction est effectuée en estimant et en supprimant les différences *intra-ensemble*, c'est-à-dire les équivalents des quantités $\mathcal{W}_{ij}$ et $\mathcal{W}_i$ calculées entre des échantillons indépendants du même ensemble. Nous utilisons également les différences intra-ensemble pour définir un score final qui permet d'interpréter quantitativement les distances de Wasserstein calculées en utilisant le bruit comme référence. De plus, nous définissons une distance globale qui tient compte de la différence entre tous les descripteurs structuraux globaux et locaux (c'est-à-dire une distance dans leur espace produit), en agrégeant correctement les quantités $\mathcal{W}_{ij}$ et $\mathcal{W}_i$ après la correction du bruit.

Nous démontrons l'utilité de WASCO pour comparer des ensembles de conformation (*i*) produits à partir de simulations de dynamique moléculaire utilisant différentes forces, et (*ii*) avant et après le raffinement avec des données expérimentales SAXS. Nous montrons également l'applicabilité de la méthode pour évaluer la convergence des simulations de dynamique moléculaire et discutons d'autres applications potentielles telles que dans les approches basées sur l'apprentissage automatique. Un des avantages de cet outil est sa mise en œuvre conviviale sous forme de cahier Jupyter, qui a été rendue disponible à la communauté.

Inférence post-clustering sous dépendance (Chapitre 6)

Une stratégie courante pour caractériser les ensembles de protéines consiste à partitionner l'espace conformationnel en mettant en œuvre des algorithmes de regroupement [7, 59, 170, 169]. Cependant, comme discuté dans la Section F.2.3, la sortie des techniques de regroupement manque d'interprétabilité en raison de leur grande sensibilité aux paramètres du pipeline et de l'absence d'une classification sous-jacente véritable. Ce problème peut être résolu en ayant recours à la théorie de l'inférence post-regroupement, qui fournit des garanties statistiques concernant les différences entre les moyennes des groupes. Les techniques mathématiques qui permettent un tel test sélectif dépendent fortement de l'algorithme de regroupement et de la distribution des données. Récemment, le travail fondateur de Gao et al. [104] a introduit le cadre permettant de réaliser une inférence après un regroupement hiérarchique lorsque les observations sont indépendantes et identiquement distribuées en tant que variables aléatoires gaussiennes de dimension p avec une matrice de covariance sphérique. Cela correspond au modèle de matrice normal suivant [127]:

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \sigma^2\mathbf{I}_p), \tag{F.15}$$

où $\mathbf{X}$ est une matrice de dimensions $n \times p$ dont les lignes sont des vecteurs de caractéristiques dans $\mathbb{R}^p$. Les moyennes des vecteurs gaussiens p-dimensionnels sont données par les lignes de la matrice $n \times p$ $\boldsymbol{\mu}$, la matrice identité $n \times n$ tient compte de l'indépendance des observations, et $\sigma\mathbf{I}_p$ est la matrice de covariance des caractéristiques pour chaque ligne. En vertu de (F.15), les auteurs de l'article [104] ont défini une p-valeur qui contrôle l'erreur de type I sélective, c'est-à-dire la probabilité de rejeter l'égalité des moyennes des groupes *à condition que ces groupes aient été trouvés*. De plus, les auteurs ont montré qu'une surestimation asymptotique de σ permet de contrôler asymptotiquement l'erreur de type I sélective, fournissant un estimateur approprié qui peut être utilisé en pratique. Traiter l'estimation des paramètres compatible avec le contrôle de l'erreur de type I sélective est une contribution très remarquable et novatrice, qui avait été négligée dans les travaux précédents pertinents dans le domaine [173, 243]. Récemment, cette approche a été adaptée à la classification k-means dans [51] et au cadre au niveau des caractéristiques, c'est-à-dire l'identification des variables contenant un vrai signal, dans [121].

Bien que les contributions de [104, 51] soient très pertinentes d'un point de vue statis-

tique, leur application à des problèmes réalistes reste limitée. En effet, le modèle (F.15) suppose que les variables sont indépendantes et ont une variance égale, ce qui est très improbable dans la pratique. En particulier, les descripteurs couramment utilisés dans les techniques de regroupement appliquées aux structures de protéines sont généralement les distances euclidiennes entre tous les atomes C_α le long de la séquence. Même si ces quantités peuvent être considérées comme des variables aléatoires gaussiennes p-dimensionnelles, leur forte corrélation empêche l'hypothèse de (F.15). De plus, les conceptions peuvent présenter une dépendance temporelle lorsqu'elles sont générées avec des approches basées sur la physique, telles que les simulations MD. Par conséquent, un modèle admettant des structures de dépendance entre les variables et les observations est requis dans ce cadre. Dans ce chapitre, nous étendons le cadre présenté dans [104, 51] au modèle matriciel normal général

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}), \tag{F.16}$$

où $\mathbf{U}$ code la dépendance entre les observations et $\boldsymbol{\Sigma}$ la covariance entre les caractéristiques. Nous définissons une valeur p qui contrôle l'erreur de type I sélective sous (F.16) pour les algorithmes de regroupement hiérarchique et k-means. De plus, nous généralisons la surestimation de σ au cadre matriciel, montrant qu'il existe un ordre partiel - le soi-disant ordre partiel de Loewner [127] - dans l'espace des matrices hermitiennes pour lesquelles la surestimation asymptotique de $\boldsymbol{\Sigma}$ assure le contrôle asymptotique de l'erreur de type I sélective. Nous fournissons également un estimateur de $\boldsymbol{\Sigma}$ qui peut être utilisé en pratique sous certaines hypothèses, que nous montrons être satisfaites pour plusieurs modèles courants de dépendance entre les observations. En plus d'illustrer les performances numériques du test présenté avec des simulations sur des données synthétiques, nous montrons comment la méthode fournit des garanties statistiques après le regroupement de données réelles provenant d'ensembles de protéines dont les conformations sont caractérisées par des descripteurs gaussiens.

Une caractérisation basée sur les contacts des ensembles de protéines (Chapitre 7)

La dernière contribution de la thèse est une méthode pour caractériser les ensembles conformationnels de protéines hautement flexibles. Les méthodes existantes dans la littérature peuvent être classées en deux grandes familles: les approches basées sur le regroupement et les approches basées sur la moyenne. Les premières proposent une idée intéressante qui intègre la nature probabiliste des ensembles désordonnés. De plus, à la suite du travail présenté au chapitre 6, ces méthodes peuvent être dotées de garanties statistiques concernant les différences entre les groupes. En utilisant les distances C_α-C_α pour caractériser les conformations, les états sont souvent comparés en utilisant la déviation quadratique moyenne (RMSD) [241, 185]. Classer les états de cette manière a tendance à regrouper les conformations en fonction des bonnes alignements, c'est-à-dire dont les structures sont globalement similaires. Cette approche, qui a du sens pour les ensembles conformationnels de protéines ordonnées/structurées, ne fournit pas de caractérisations appropriées ici en

raison de la grande variabilité conformationnelle du système: les conformations des protéines IDP ne s'alignent pas bien. Par conséquent, forcer un tel alignement ne donne pas de partitions claires et interprétables de l'espace conformationnel. D'autre part, les approches basées sur la moyenne réduisent considérablement la variabilité spatiale et masquent les caractéristiques structurales pertinentes mais peu fréquentes du système. L'inadéquation de toutes ces méthodes est illustrée au chapitre7.

Nous proposons d'aborder la question de la manière de caractériser correctement un ensemble désordonné en revenant aux origines de la caractérisation structurale: les cartes de contact. Les cartes de contact et de distance ont servi d'outils principaux pour caractériser la structure des protéines rigides [229, 213, 275], démontrant leur aptitude à détecter les domaines structuraux [250, 164, 255, 137]. Elles consistent en une matrice triangulaire binaire $(C_{ij})_{ij}$, où $C_{ij} = 1$ si la distance euclidienne entre les atomes C_α de i-ème et le j-ème résidu est inférieure à un seuil donné, et $C_{ij} = 0$ sinon. Bien qu'elles se révèlent être des outils très utiles pour caractériser des structures rigides, leur extension naïve aux ensembles conformationnels, consistant à estimer les probabilités de contact en moyennant les contacts binaires sur l'ensemble des conformations, perd les motifs de contact en dehors de la diagonale qui apparaissent pour des ensembles de conformations à faible occupation. Nous croyons que les informations sur les contacts doivent rester la clé pour caractériser la variabilité conformationnelle des protéines flexibles, mais l'extension basée sur la moyenne doit être remplacée par une approche plus intelligente pour dévoiler la complexité structurale des protéines IDP. Le message que nous proposons est clair: utilisez les contacts, *mais faites clustering d'abord.*

Le chapitre 7 présente WARIO, une caractérisation basée sur les contacts des ensembles de protéines hautement flexibles. Cette méthode exploite le potentiel des cartes de contact en intégrant intelligemment le comportement statistique des systèmes désordonnés. Pour ce faire, nous utilisons d'abord un algorithme de regroupement bien adapté qui révèle comment les interactions résidu-résidu se manifestent à travers la dynamique de la protéine. Pour ce faire, nous caractérisons les conformations par les descripteurs structuraux globaux unidimensionnels (Définition F.2.7), c'est-à-dire par une fonction continue prenant des valeurs dans $[0, 1]$ qui fait office de proxy pour l'interaction entre chaque paire de résidus. Cette fonction intègre les informations de séquence et l'orientation relative entre les acides aminés en interaction, que nous montrons être cruciale pour prendre correctement en compte la formation de motifs structuraux locaux. Ensuite, chaque groupe de conformations est décrit par sa configuration de contact représentative. En bref, un ensemble conformationnel est caractérisé par une *famille pondérée de cartes de contact*, tenant compte de sa diversité structurale à travers un ensemble de motifs de contact qui apparaissent avec une fréquence donnée le long des fluctuations conformationnelles de la protéine. Nous illustrons l'utilité de WARIO à travers quatre exemples de protéines flexibles, et nous le comparons aux approches classiques de clustering qui utilisent les distances pour caractériser les conformations.