WARIO: Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins

Javier González-Delgado[1,2], Pau Bernadó[3], Pierre Neuvial[2] and Juan Cortés[1]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*

[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, Toulouse, France.*

[3] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, Montpellier, France.*

Abstract

Characterizing the structure of flexible proteins, particularly within the realm of intrinsic disorder, presents a formidable challenge due to their high conformational variability. Currently, their structural representation relies on (possibly large) conformational ensembles derived from a combination of experimental and computational methods. The detailed structural analysis of these ensembles is a difficult task, for which existing tools have limited effectiveness. This study proposes an innovative extension of the concept of contact maps to the ensemble framework, incorporating the intrinsic probabilistic nature of disordered proteins. Within this framework, a conformational ensemble is characterized through a weighted family of contact maps. To achieve this, conformations are first described using a refined definition of contact that appropriately accounts for the geometry of the inter-residue interactions and the sequence context. Representative structural features of the ensemble naturally emerge from the subsequent clustering of the resulting contact-based descriptors. Importantly, transiently-populated structural features are readily identified within large ensembles. The performance of the method is illustrated by several use cases and compared with other existing approaches, highlighting its superiority in capturing relevant structural features of highly flexible proteins. An open-source implementation of the method is provided together with an easy-to-use Jupyter notebook, available at https://gitlab.laas.fr/moma/WARIO.

1 Introduction

The function of numerous proteins is intricately linked to their dynamic structure, which often displays significant conformational variability. In particular, intrinsically disordered proteins/regions (IDPs/IDRs) represent an extreme example of this phenomenon [1–4]. Nevertheless, the conformational characterization of highly-flexible proteins or regions remains a challenge. Currently, structural ensembles of disordered proteins, such as those deposited in the Protein Ensemble Database (PED) [5], are defined by a (possibly large) set of atomistic models, which are very hard to analyze, and heavy in terms of storage and manipulation. This is even truer for ensembles derived from molecular dynamics (MD) simulations. Structural analyses of these ensembles are often reduced to very simple descriptors, such as the radius of gyration or the relative solvent accessibility, which provide very limited structural insights and that are not necessarily related with their function. Moreover, these descriptors are averaged values over the whole ensemble, ignoring the information about their distribution. The presence of transiently-populated secondary structural elements and interactions between distant residues in the sequence are more relevant structural descriptors. However, their identification in large atomistic ensembles is often hampered by their reduced population. New descriptors are therefore needed to represent large conformational ensembles in a compact and meaningful way.

For globular proteins, contact and distance maps have become fundamental tools for define their 3D fold [6–8], demonstrating their suitability to identify structural domains [9–11]. More recently, contact

maps have proven key for the development of machine-learning-based approaches for structure prediction [12–14]. A naive extension of contact and distance maps to conformational ensembles, which involves estimating contact probabilities by averaging binary contacts for every conformation, has been used to describe interaction propensities in ordered systems [15–18]. However, in the presence of structural disorder, this approach is not appropriate. More specifically, contacts between residues that are far apart in the sequence, which may be structurally or functionally important but occur with low probability, are undetectable from these representations. Similarly, scarcely populated structural motifs are diluted in the average contact/distance maps. This phenomenon is illustrated in Figure 3(a), which displays the average contact map for a conformational ensemble of a 27-residue long intrinsically disordered region in CHCHD4, one of the proteins used as an example in this study (see Results section). This representation only highlights contacts around the diagonal of the matrix, while long-range contacts that appear at low frequency remain undetected. Consequently, the characterization of conformational ensembles on the basis of contacts represents a non-trivial task that requires novel approaches integrating the statistical nature of flexible proteins.

In order to overcome the above-described limitations, we have developed a novel strategy that, while exploiting the power of contact maps, is adapted to the structural variability of highly-flexible proteins. More precisely, we propose to characterize a conformational ensemble by a *weighted family of contact maps*, representing its structural diversity through a set of short- and long-range contact patterns that appear at a given frequency. This is done by first applying a well-suited clustering algorithm that unravels the underlying conformational variability of the protein and then characterizing such distribution through its representative network of contacts.

Clustering conformations of highly-flexible proteins is a challenging problem since their conformational space can be considered as a high-dimensional manifold with non-Euclidean geometry. In this regard, non-linear dimensionality reduction algorithms such as t-SNE [19] or UMAP [20] are very attractive tools for disentangling features embedded in high-dimensional data [21–24]. Besides, their incorporation into clustering algorithms has shown remarkable empirical efficiency [25–29]. This idea has been successfully exploited for analyzing ensembles of conformations produced by MD simulations in recent studies [30, 31]. However, in these works, conformations were usually featured by descriptors such as atom coordinates or backbone torsion angles, and compared using root-mean-square deviation (RMSD) [32, 33], whose suitability to compare unfolded conformations is questionable. We propose to use contacts also to feature conformations prior to clustering. However, unlike current approaches based on an arbitrary threshold, contacts are defined as a continuous weight function that acts as a proxy of the interaction between residue pairs. Importantly, the weight depends on the amino-acid types, their separation in the sequence, their Euclidean distance and their relative orientation. We show that the appropriate combination of these parameters in the contact definition is crucial for the detection of transient structural features within large ensembles. Then, clustering can be performed on the conformational space featured with contact-based information using HDBSCAN [34], passing through a low-dimensional UMAP [20] projection. In addition to the contact pattern, several descriptors associated to each cluster can be derived, such as secondary structure propensities, average radius of gyration and end-to-end distance. The pipeline with the steps of the method, which we named WARIO, is illustrated in Figure 1.

This original approach provides a compact and meaningful representation of the conformational ensembles of flexible proteins, from which functionally important structural features can be easily identified. Through the analysis of several long MD trajectories, we demonstrate that WARIO outperforms current approaches in identifying and characterizing local structures and long-range interactions occurring at extremely low frequencies within ensembles. In addition to WARIO's unique ability to disentangle large

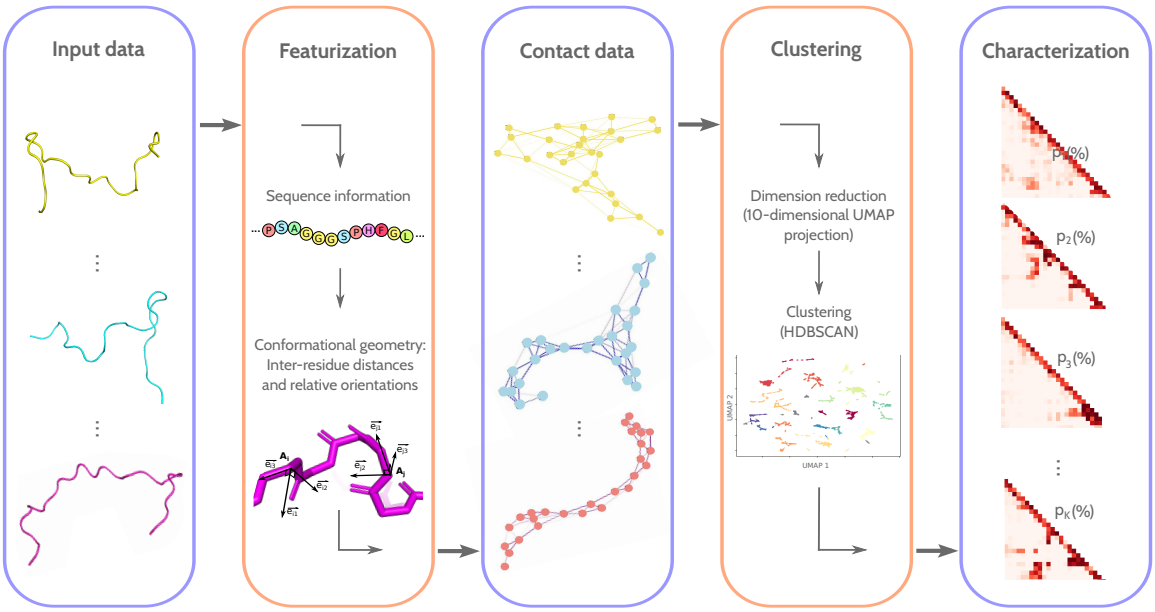

Figure 1: Overview of WARIO pipeline implementation. The method takes a conformational ensemble as input. For each conformation, inter-residue distances are computed, considering the sequence and the relative orientation of residue pairs. Using this information, a proxy for the inter-residue contacts of the conformation is calculated, and the resulting values are used as structural descriptors. Subsequently, the conformations are classified using the aforementioned descriptors through a clustering algorithm that incorporates a projection into a low-dimensional space. Finally, each cluster is represented by an average contact map describing the inter-residue interactions within the corresponding group, along with their frequency within the ensemble. This weighted family of contact maps characterizes the conformational ensemble.

conformational ensembles of disordered proteins, the proposed structural representation strategy has a great potential in machine-learning (ML) pipelines.

2 Methods

This section presents our approach, WARIO, for the characterization of conformational ensembles of (highly-)flexible proteins. The first subsection describes how conformations are featured by a continuous contact function that integrates sequence and geometric information. Then, explanations are provided about how clustering is performed on the featured data. This allows for the definition of a weighted family of contact maps characterizing the conformational ensemble. Finally, the implementation in Python of the complete pipeline as an easy-to-use Jupyter notebook will be briefly presented.

2.1 Description of intramolecular contact as a sequence and orientation-dependent continuous function

Conventionally, a contact between a pair of amino acids is defined as a binary indicator when the Euclidean distance between their C_α (or C_β) atoms is less than a certain threshold, typically set between 6Å and 12Å [35]. This indicator is universal for every pair of residues regardless of their identities, positions in the sequence or relative orientation. However, it is known that these parameters mediate the

interactions between residues. Indeed, when looking at how Euclidean contact distances are distributed in experimentally-determined high-resolution structures, we observed that they concentrate around distance values that are strongly dependent on the amino acid identities and their relative position in the sequence (from now on, we will use the term *range* to designate the relative position in the sequence). Furthermore, interacting residues present preferred relative orientations that clearly manifest for short-range contacts. Consequently, a precise contact descriptor must integrate both sequence and geometric information, and avoid universal binary indicators that, as we show here, yield a substantial loss of structural information.

In this work, we redefine *contact* as a continuous function taking values in the range $[0, 1]$ that integrates sequence information and the relative orientation between the interacting residues. To do so, we followed the steps briefly explained below. Details are provided in Section S1 of the Supplementary Information (SI). The contact function was defined based on the analysis of 15,177 experimentally-determined high-resolution (< 2Å) structures of protein domains extracted from the SCOPe 2.07 release [36], which we will refer to as the *structural database*. The first step corresponds to the identification of Euclidean contact distance maxima in the structural database, which depend on the identity and range of the two residues. These maxima are used to define the so-called *Euclidean contact interval*, which represents the range of Euclidean distance values for which the interaction between residues begins to be meaningful. We observed that, for Euclidean distances below its upper limit, preferred orientations clearly stand out in the structural database. Once again, they are dependent on the residue identities and range, and they are more clearly observed for short-range contacts. These preferred orientations need to be combined with the Euclidean distance in a suitable way. In this respect, we ask the relative orientation between a pair of residues: (i) to be neglected for large values of the Euclidean distance, and (ii) to contribute to enhance the proxy for contact only if it is close to the specific preferred orientation of the residue pair. Conditions (i) and (ii) yield the definition of the so-called *relative pose distance*, which equals the

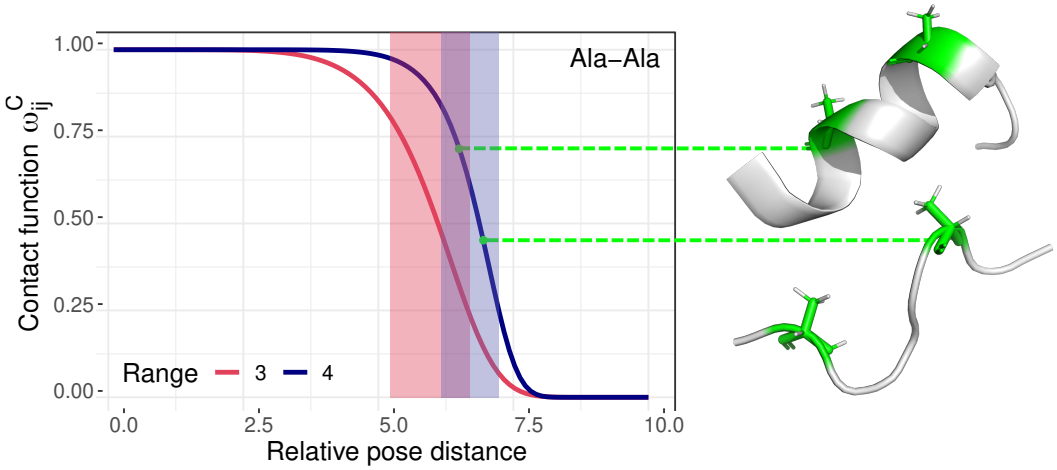

Figure 2: Contact function for Ala-Ala residue pairs at ranges 3 and 4. The growth of each curve is concentrated within the so-called contact interval, which is marked by a band colored red and blue for ranges 3 and 4 respectively. When inter-residue Euclidean distances remain equal, the contact function yields higher values for relative orientations that more closely resemble the preferred ones observed in the structural database. On the right, two Ala-Ala pairs at range 4 with equal C_β-C_β distances of 7.4Å. The configuration on the top has a higher contact value, indicating its closer alignment to the preferred orientation of both alanine residues at range 4.

Euclidean distance when it has large values, or if the contact is long range, and progressively reduces the Euclidean distance as the inter-residue relative orientation approaches its preferred one.

The relative pose distance is explicitly constructed as a continuous function combining the Euclidean distance and the deviation from the preferred orientation. Its functional form is parameterized by the identities and the range of the corresponding residue pair. Finally, the relative pose distance is transformed into a proxy for contact taking values in $[0, 1]$, which we refer to as *contact function*. The contact function is a decreasing function of the relative pose distance, parameterized by the sequence information. The growth of the curve is concentrated within a specific interval, which was analogously determined through the analysis of the relative pose distance distribution in the structural database. This is illustrated with an example in Figure 2, where the contact function is depicted for Ala-Ala residue pairs at two different range values.

The redefinition of contact as a continuous function in $[0, 1]$, depending on the relative position, orientation and sequence information, proves to be essential for an appropriate characterization of the structural dynamics of flexible proteins. This is detailed in Section S2 of the SI.

2.2 Clustering pipeline and ensemble characterization

Instead of merely averaging contacts across conformations, which results in a substantial loss of valuable information, WARIO employs a different approach. It first disentangles the underlying distribution by clustering. WARIO then represents the resulting sub-ensembles as a weighted collection of contact maps, each of them capturing a distinct structural feature.

The clustering method applied in WARIO relies on the contact function defined above. Conformations are 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 is described by the $n \times L(L-1)/2$ matrix:

$$\mathbf{W}_C = \begin{pmatrix} \omega^C_{12;1} & \cdots & \omega^C_{ij;1} & \cdots & \omega^C_{(L-1)L;1} \\ \omega^C_{12;2} & \cdots & \omega^C_{ij;2} & \cdots & \omega^C_{(L-1)L;2} \\ \vdots & & \vdots & & \vdots \\ \omega^C_{12;n} & \cdots & \omega^C_{ij;n} & \cdots & \omega^C_{(L-1)L;n} \end{pmatrix}, \tag{1}$$

where $\omega^C_{ij;k}$ denotes the value of the contact function for residues $i, j \in \{1, \ldots, L\}$ in 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 has previously done in related methods such as RING [18, 37]. Here, the set of nodes is given by the set of residues and every pair of residues i,j is linked by an edge with a weight $\omega^C_{ij;k}$. This procedure is depicted in Figure 1. It should be noted, however, that the graphical representation is merely an alternative visualization of the data, and that our methodology does not rely on graph theory.

The clustering method performed on the contact function matrix (1) is based on the combination of a dimensionality reduction technique with an efficient clustering algorithm, similarly to state-of-the-art approaches [30, 31]. Here, we opt for UMAP [20] to first embed the data (1) into a 10-dimensional space. This choice is motivated by its ability to preserve the topology of the high-dimensional data and efficiently reveal population structure [22, 29]. We then apply the HDBSCAN clustering algorithm [34] to the embedding, which we consider 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 number of classes. This is suitable for our implementation, as the practitioner might have more intuition of 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. Details on UMAP and HDBSCAN are provided in SI.

Once the clustering is performed, each class is characterized through a cluster-specific contact map. Let K be the number of retrieved clusters and $\mathcal{C}_k \subset \{1, \dots, n\}$ be the subset of conformations constituting the k-th 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 (1), we define the k-th *cluster-specific ω-contact map* as the $(L-1) \times (L-1)$ matrix:

$$\overline{W}_{\mathcal{C}_k} = \left(\frac{1}{|\mathcal{C}_k|} \sum_{l \in \mathcal{C}_k} \omega_{ij;l}^{C} \right)_{ij} \quad \text{for } i < j \in \{1, \dots, L\}, \tag{2}$$

where $|\mathcal{C}_k|$ denotes the cardinality of $\mathcal{C}_k$. The matrix (2) is the average of all the rows in (1) that belong to the k-th cluster, represented in a matrix form. Its entries are the cluster averages of the contact function values for every pair of residues along the sequence, and it accounts for the contact patterns that dominate the cluster. A weight p_k can be assigned to each matrix (2) based on the cluster occupancy proportion:

$$p_k = \frac{|\mathcal{C}_k|}{n}. \tag{3}$$

This allows us to define the *ensemble characterization* as the K-tuple of weighted cluster-specific ω-contact maps:

$$\mathcal{E} = \left(\left(\overline{W}_{\mathcal{C}_1}, p_1 \right), \dots, \left(\overline{W}_{\mathcal{C}_K}, p_K \right) \right). \tag{4}$$

The representation (4) provides a compact characterization of how residue-residue interactions are distributed within the ensemble.

Note that, thanks to the proportions (3), it is easy to extract a representative set of conformations by sampling from the distribution

$$P_{\text{rep}}(\mathcal{E}) = p_1 \mathcal{U}(\mathcal{C}_1) + \cdots + p_K \mathcal{U}(\mathcal{C}_K), \tag{5}$$

where $\mathcal{U}(\mathcal{C}_k)$ denotes the discrete uniform distribution on $\mathcal{C}_k$, for $k \in \{1, \dots, K\}$. Since the HDBSCAN algorithm might not classify every conformation, the proportions (3) must be normalized to one before sampling from (5).

Each cluster of conformations can be analyzed *a posteriori* on the basis of additional descriptors that can provide further information for the practitioner's needs. Here, we propose to evaluate the secondary structure propensities based on the structural classification provided by DSSP [38] and to compute the cluster average radius of gyration. More specific descriptors can be easily added to the post-clustering analysis, using methods implemented in tools such as SOURSOP [39].

2.3 The Jupyter Notebook

WARIO has been implemented in Python, and can be executed through an easy-to-use Jupyter Notebook. The open-source code, along with detailed installation and user instructions, is available at: https://gitlab.laas.fr/moma/WARIO. The notebook takes a conformational ensemble as input and returns the ensemble characterization defined in (4). The data featurization is performed at a first stage,

allowing the user to adjust the resolution of the clustering algorithm afterwards. Then, clustering and post-processing are performed, and the results are displayed and saved.

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 [40], one multiframe .pdb file, or a folder containing one .pdb file per conformation. Users can also choose to characterize sequence segments instead of the entire sequence. Details are provided in the notebook documentation. It should be noted that the current implementation of WARIO requires an all-atom representation of the protein backbone.

The main output of WARIO is given through a weighted set of ω-contact maps depicting the interaction patterns that characterize each cluster. Plots with cluster-specific DSSP propensities and the average radius of gyration are also provided. The notebook allows to export the clusters of conformations in the same format as the input ensemble. These files can be used for further analysis of the retrieved contact patterns and for the calculation of other structural descriptors corresponding to the practitioner's needs.

3 Results

We have used WARIO to characterize ensembles of three highly-flexible proteins containing different levels of structure. We applied the pipeline described in Section 2 to ensembles extracted from long MD trajectories. Details of these simulations can be found in the original articles, cited below. Through these examples, we demonstrate the ability of our approach to localize scarcely populated structural patterns, including secondary structural elements and transient long-range contacts. We also compared WARIO with other clustering approaches to highlight its unique ability to cluster structural patterns that often remain unidentified by other strategies.

3.1 Characterization of the N-terminal region of CHCHD4

We use the intrinsically disordered Coiled-Coil-Helix-Coiled-Coil-Helix Domain Containing 4 (CHCHD4) to illustrate WARIO's ability to identify structural features in large ensembles. This protein plays a crucial role in the import of intermembrane space-targeted proteins [41, 42]. Only the structure of the folded domain of CHCHD4 (residues 45-109) has been experimentally resolved [43]. However, the interaction with most of its clients exclusively involves the intrinsically disordered N-terminal region (27 residues) [44], which is the fragment analysed here.

A conformational ensemble of the disordered N-terminal region of CHCHD4, encompassing $n = 100050$ conformations, was generated from 50 independent MD trajectories of 200 ns each (Mazzanti *et al.*, unpublished). Data are publicly available [45]. Using a minimum cluster size of 1% of the total number of conformations, WARIO identified 23 clusters with different levels of occupancy. The two most populated clusters encompassed approximately 20% and 16% of the conformations, while the remaining 21 clusters only represented 1-3% of them. The overall cluster distribution can be visualized through the projection to a 2-dimensional UMAP space (see Section S4.1 in the SI). 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 the SI. The average ω-contact map of the two most populated clusters showed the presence of some local structuring at the C-terminus of the chain and a complete absence of long-range contacts (Figure 3(b,c)). Interestingly, all the remaining, low occupied conformational clusters presented more specific structural features (Figure 3(d-f)). For instance, two clusters containing 1.5% and 1.1% of the population presented a turn from residues 6 to 15 and a short α-helix, respectively (panels (d)

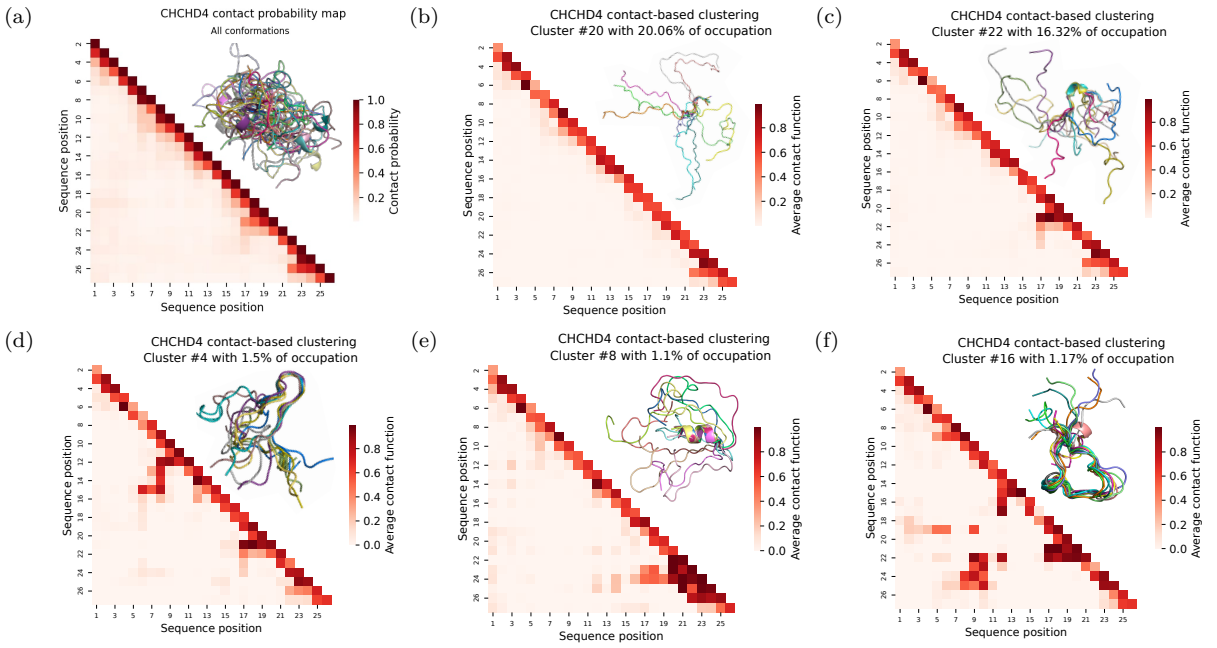

Figure 3: Structural characterization of CHCHD4. (a) Contact probability map for the conformational ensemble of CHCHD4. Each contact probability is estimated as the proportion of contacts between a pair of residues, considering a 8Å distance threshold between the C_β atoms (C_α for glycine). In the upper triangle, 10 randomly selected conformations from the cluster. (b-f) Cluster-specific ω-contact maps (2) for five clusters of CHCHD4. Panels (b) and (c) correspond to the two most populated groups encompassing 20.06% and 16.32% of the conformations. In each upper triangle, 10 randomly selected conformations from the corresponding cluster and aligned at residues exhibiting off-diagonal contact patterns. Note that the cluster numbering is arbitrary and it is not related with its population.

and (e) in Figure 3). Indeed, the clusters provided by WARIO group together conformations that exhibit the same secondary structure motifs. As discussed in Section S2 of the SI, this is due to the proper incorporation of relative orientation in the definition of inter-residue contacts. Another low populated (1.17%) cluster displayed a well-defined long-range contact between the central and the C-terminal region of the peptide (Figure 3(f)). When analysing the average radius of gyration for all the identified clusters, a large difference was observed between the two most populated ones, with values of 15.36Å and 13.98Å, respectively, and the others, presenting values around 10-12Å. This observation substantiates the presence of long-range contacts in the majority of the low-populated clusters. This analysis demonstrates the ability of WARIO to identify and localize scarcely populated structural patterns from large ensembles.

3.2 Comparison of WARIO with other clustering approaches

We compared the results obtained using WARIO for CHCHD4 with those provided by two existing approaches based on pairwise distances and inter-residue Lennard-Jones contact energies. In distance-based methods, structural data are featured by Euclidean distances between residue pairs. This metric combined with dimensionality reduction and a clustering algorithm has been previously used to characterize flexible proteins [30, 31]. Concretely, we implemented the UMAP + HDBSCAN pipeline on the structural data featured with pairwise Euclidean distances between all C_β atoms (C_α for glycines) to characterize the CHCHD4 MD ensemble. This strategy retrieved 10 clusters, among which one contained the 67% of conformations. Note that WARIO retrieved 23 clusters for the same ensemble and that the two most

occupied clusters contained approximately 20% and 16% of the conformations. Figure 4(a-c) displays the average distance maps for the three most occupied clusters, together with 30 conformations randomly drawn from each cluster and aligned using all residues. As the UMAP + HDBSCAN pipeline is fed with all the pairwise distances, clusters tend to group conformations having similar global shapes and do not necessarily group them according to the presence of structural motifs or long-range contacts. As a consequence, the structural clusters yield much broader contact maps when compared to the results yielded by WARIO. Therefore, distance-based methods do not seem adapted to identify scarcely populated states diluted in a conformationally diverse ensemble.

The use of an inter-residue Lennard-Jones (LJ) interaction potential to feature individual conformations has been recently reported [30]. The capacity to capture interactions within the chain makes this strategy similar to our continuous contact function. In order to implement the LJ potential, we repeated the same strategy as in the previous distance-based analysis but featuring each conformation $k \in \{1, \ldots, n\}$ by the vector

$$\left(V_{12;k}, \ldots, V_{ij;k}, \ldots, V_{L(L-1);k}\right), \tag{6}$$

where $V_{ij;k}$ is the inter-residue LJ contact energy between residues i and j of the k-th conformation. The explicit form of the interaction potential is given in [46, Eq. 1-3].

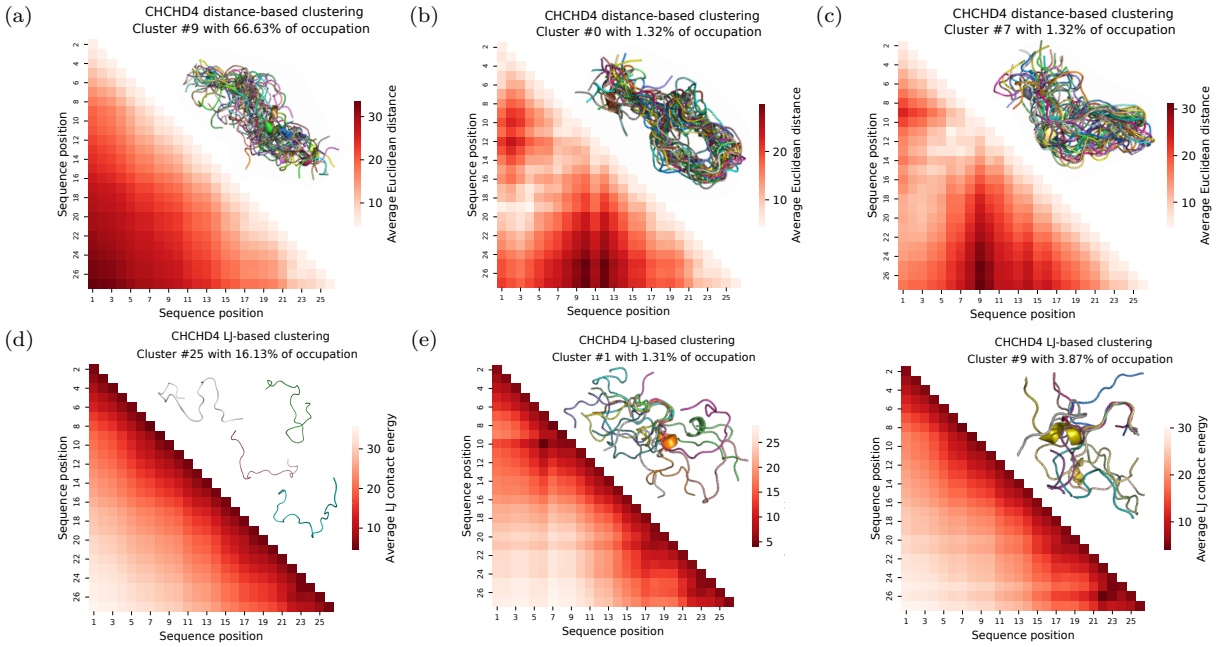

Figure 4: Comparison of WARIO with other clustering approaches. (a-c): CHCHD4 cluster-specific average distance maps after implementing the UMAP + HDBSCAN pipeline to the set of conformations featured by all Euclidean inter-residue distances. In each upper triangle, 30 randomly selected CHCHD4 conformations from the corresponding cluster and aligned at all residues are displayed. (d-f) CHCHD4 cluster-specific Lennard-Jones contact maps after implementing the UMAP + HDBSCAN pipeline to the set of conformations featured by all inter-residue interaction potentials (6). In each upper triangle of (e,f), 10 randomly selected CHCHD4 conformations from the corresponding cluster and aligned at residues showing low average contact energy values. In the upper triangle of panel (a), corresponding to the most populated cluster, four non-aligned randomly selected conformations from the group are displayed. Note that the cluster numbering is arbitrary and it is not related with its population.

After classifying the LJ interaction matrices with the UMAP + HDBSCAN pipeline, one predominant cluster was retrieved containing around 16% of conformations, together with 25 other groups with populations ranging from 1% to 5% (Figure 4(d-f)), similarly to the number of clusters retrieved with WARIO. The regions of these maps displaying low energy values indicate pairs of residues that interact with each other at high frequency. Although this representation is more diffuse than that based on contact functions implemented in WARIO (see Figure 3), they still allow for the identification of cluster-specific interaction patterns. When looking at the most populated cluster (Figure 4(d)), an interaction map with low energy values near the diagonal that steadily increases towards the interior of the matrix was observed, and no local contact or long-range interaction could be identified. This contradicts the inter-residue interactions observed for some conformations of the cluster, as shown in Figure 4(d). The inspection of less populated clusters indicates that LJ-based interaction maps are more diffuse than the continuous-contact ones and that the derivation of a specific structural features from these maps is less straightforward (Figure 4(e)-(f)). In order to exemplify this last observation, we searched among the LJ-based clusters one presenting a helical motif at residues 21-24, as detected by WARIO (Figure 3(e)). For this, we identified three LJ maps presenting energy minima at the C-terminus (Figure 4(f), Figure S9(c,e)). However, the secondary structure analysis of these three clusters displayed a negligible α-helical propensity for residues 21-24 (Fig. S9(b,d,f)), indicating that the LJ-based contact description yields less structurally well-defined clusters.

3.3 Characterization of Huntingtin Exon-1 MD trajectory

The N-terminal region of huntingtin, the so-called exon-1, is the causative agent of Huntington's disease, a deadly neurodegenerative pathology [47]. This fragment contains a poly-glutamine tract, poly-Q, that is flanked by 17 amino acids (N17) and a proline-rich region at N- and C-termini, respectively. Importantly, when the length of the poly-Q exceeds a pathological threshold of 35, the protein forms large amyloidogenic aggregates in neurons that cause the pathology [48]. The structural changes occurring to the protein above this threshold have been the object of huge research efforts from experimental [49, 50] and theoretical perspectives [51]. In a recent study by integrating NMR chemical shifts, SAXS and computational modelling, huntingtin exon-1 (HTTExon-1) was described as an equilibrium of multiple partially helical states involving the N17 and increasing fragments of the poly-Q tract [52]. In the same study, a MD simulation of a HTTExon-1 encompassing the N17, a poly-Q tract with 46 glutamines and five prolines was also presented with the aim to decipher the mechanisms governing the structure of the protein. Details of this simulation can be found in the original publication [52]. Here, we have analysed this 20 microsecond trajectory ($n = 96000$) with WARIO in order to disentangle the large number of coexisting conformational states conforming this protein.

WARIO found a large number of low-populated clusters, $K = 43$, in the ensemble of HTTExon-1. Interestingly, all the 43 clusters presented comparable sizes, grouping 1-3% of the conformations when defining to 1% of n the minimum cluster size. The overall distribution of the weights (3) is illustrated by projecting (1) to a two-dimensional UMAP space (see SI). The ensemble characterization of the HTTExon-1 trajectory identified a family of structural clusters displaying a systematic helix extension along the poly-Q tract. Among the 43 clusters, very few long-range contacts emerged from the analysis and the clustering is mainly governed by the short-range helical motifs, which appear close to the diagonal. Indeed, these structurally consecutive clusters can be connected by monitoring the number of residues involved in the α-helix, which naturally emerge from the cluster-specific ω-contact maps and their DSSP secondary structure analysis. This is illustrated for a group of representative clusters in Figure 5(a) (see SI for the complete characterization). Note that the detection of the different steps along the secondary

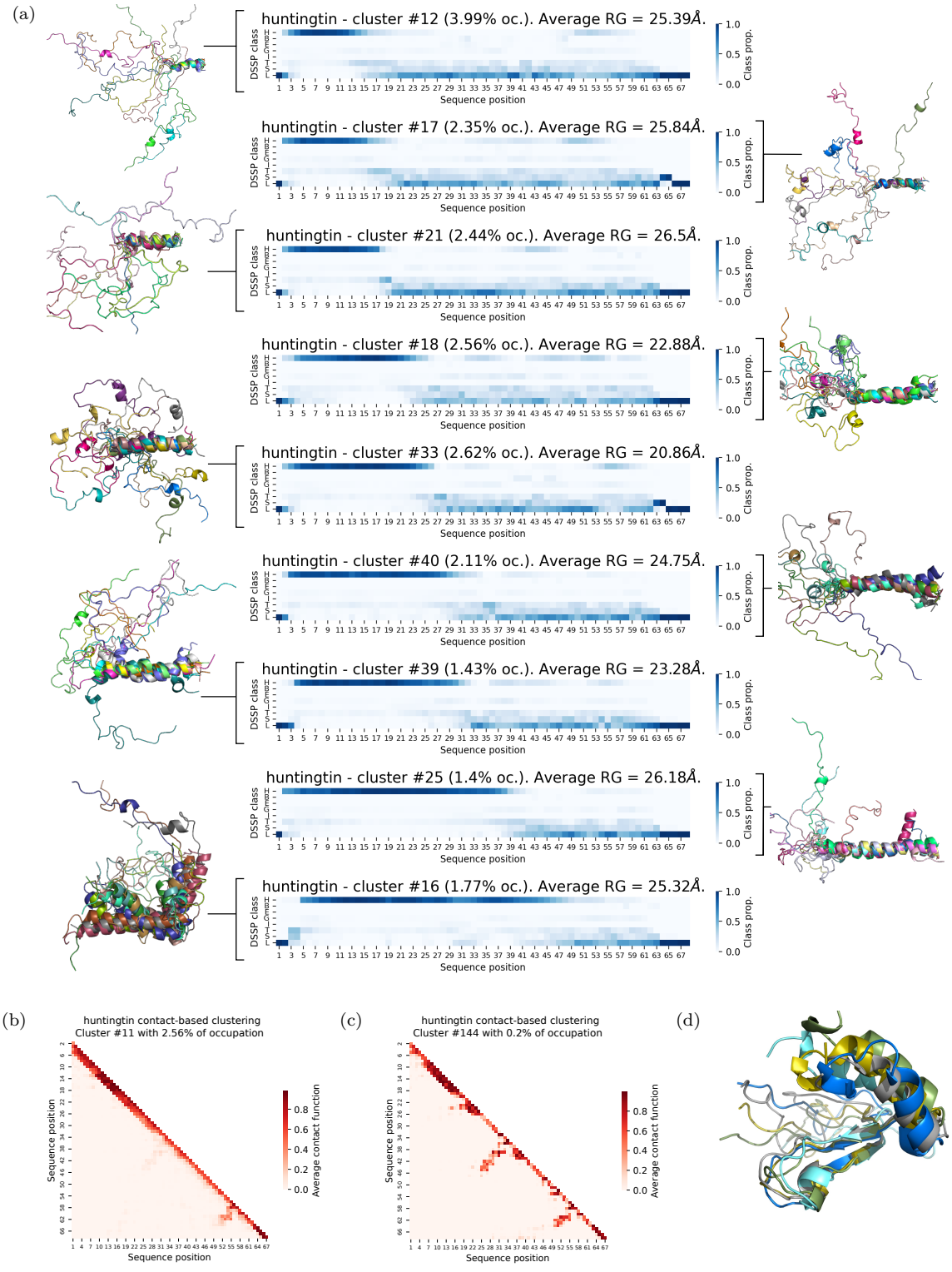

Figure 5: Structural analyses if HTTExon-1 MD trajectory. (a) Nine HTTExon-1 clusters represented by their DSSP secondary structure propensities and ordered by increasing number of residues involved in the α-helix. Together with each DSSP matrix, ten conformations randomly selected from the group are depicted and aligned at the α-helix. (b, resp. c) Cluster-specific ω-contact map for the 11-th (resp. 144-th) cluster of HTTExon-1 characterization after setting to the 1% (resp. 0.1%) of n as the minimum cluster size. (d) Ten conformations randomly selected from the 144-th cluster and aligned at the extended sheet structure around residue 35. Note that the cluster numbering is arbitrary and it is not related with its population.

structure formation is not hampered by the rest of the chain, which preserves an important level of disorder, underlining the power of WARIO to focus on the relevant molecular interactions governing the structure.

We also used the example of HTTExon-1 to assess the effect of the clustering resolution, which can be calibrated by selecting the minimum cluster size, for the detection of extremely low populated states in ensembles. Note that, in some cases, the detection of structural motifs or contact patterns might be essential, but they might be unnoticed if the resolution of the analysis is not high enough. For instance, the cluster-specific ω-contact map for the 11-th cluster, presented in Figure 5(b) and corresponding to the 2.56% of conformations, displays a network of long-range interactions between residues ∼50-55 and ∼60-65, and a much less well-defined contact pattern around residue 35, which could correspond to a β-sheet. This last feature is barely appreciated as its corresponding contact function average value remains around 0.1-0.2. Moreover, it is unclear whether the two structural motifs can simultaneously occur in the trajectory. This can be confirmed by refining the clustering at higher resolution and evaluating the partitioning of the 11-th cluster conformations in the new analysis. We repeated the algorithm by setting the minimum cluster size to 0.1% of the total number of conformations. With this new threshold, WARIO retrieved 440 clusters. Although this representation can be too detailed for some purposes, it was useful to extract specific contact patterns that might be hidden inside a broader classification. During the inspection of the new set of clusters, we identified cluster 144, containing 0.2% of the conformations, whose ω-contact map clearly indicated the previously alluded β-sheet contact pattern (Figure 5(c,d)). Importantly, this cluster was the only one presenting this structural pattern, which also appeared simultaneously to the C-terminal long-range interaction. These results show that WARIO is able to detect and group the 192 conformations presenting this particular contact pattern among the 96000 of the ensemble, highlighting its sensitivity to scarcely populated intramolecular contacts.

3.4 Characterization of TAR DNA-binding protein 43 (TDP-43)

TAR DNA-binding protein 43 (TDP-43) is a multidomain protein involved in key RNA regulatory processes and it is commonly associated with various neurodegenerative disorders, including amyotrophic lateral sclerosis and frontotemporal dementia [53]. Multiple investigations suggest that TDP-43 is capable of experiencing liquid-liquid phase separation (LLPS) in vitro and to integrate into biological condensates [54]. The domain architecture of TDP-43 consists of a folded N-terminal domain (NTD, res. 1–78), two RNA-recognition motif domains (RRM1/2, res. 104–263) connected by a short linker, and a disordered C-terminal region (CTD, res. 264–414) that contains an evolutionary conserved region (CR, res. 319–341) known to populate a transient α-helical structure that is key for droplet formation [55]. The role of the different protein domains in the intra- and intermolecular interactions that give rise to LLPS is not yet fully understood [56]. In this context, multi-scale simulations are a suitable approach to rationalize experimental results [57, 58].

In a recent study, all-atom MD simulations of TDP-43 (414 residues) at two different ionic strengths (100 and 300 mM NaCl) were analyzed to unravel the interdomain interactions governing phase separation [58]. Through an average contact map, in the original study, a complex interaction network involving linkers and the disordered C-terminal domain was identified. Here, we applied WARIO to the ensembles generated by both all-atom MD simulations, showing that the method is capable of finely disentangling the interdomain interactions that occur across the system. For each NaCl concentration, the trajectory was formed by the concatenation of three independent replicas with 5000 conformations each. Details on the MD simulations can be found in [58]. Note that, due to the large size of the system, the dimension of the contact-based feature space is $p = L(L-1)/2 = 85491$. The complete characterization of TDP-43

at 100 mM NaCl is included in the SI. Average distance representations used in the original study, while capable of detecting the most prevalent contacts, loose information about the correlation between different interactions, hampering the identification of these occurring simultaneously. The WARIO analysis of the low ionic strength trajectory is used here to exemplify the interaction network patterns. Two cluster-specific contact maps are presented to illustrate this point (Figure 6). The first one shows a simultaneous interaction of the folded N-terminal domain NTD with both C-terminal disordered regions (IDR1, res. 263-318; IDR2, res. 342-414), as well as an intradomain interaction between IDR1 and IDR2 (Figure 6(a)). In the second one, IDR2 interacts simultaneously with NTD, IDR1, the first linker (L1, res. 79-103) and the conserved region (CR, res. 319-341) of the C-terminal domain (Figure 6(b)). Interestingly, this last configuration induces the formation of an anti-parallel extended sheet in IDR2 involving residues 342-343 and 373-374 (Figure 6(c)). By inspecting all clusters derived from the trajectory at 100 mm NaCl, this motif only appears when the above-described long range interaction pattern of the 27th cluster is present. These concomitant interactions could not be identified in the original analyses.

The accurate clustering of conformations provided by WARIO enables the comparison between ensembles obtained with different parameters to identify the structural features that change between them. To exemplify this point, we have compared the clusters derived from the trajectories obtained at both ionic strengths. First, we focus on an interaction that remained unaltered with the salt concentration according to the original study: the long-range contact between the first linker L1 and the second RNA-recognition motif domain RRM2 (Figure 6(d)). At 100 mM NaCl, this interaction was clearly observed within 9 clusters containing 16.32% of conformations. Besides, L1-RRM2 contacts were coupled with other inter-domain interactions. Concretely, clusters 17 (lower triangle of Figure 6(d)) and 19 (comprising 3.99% of conformations altogether) presented simultaneous NTD-IDR1, NTD-IDR2, IDR1-IDR2, RRM1-IDR1 and RRM1-L2 interactions. Clusters 39, 40 and 43 with 5.25% of conformations presented L2-RRM1 interactions (see SI for the complete family of cluster-specific ω-contact maps at both ionic strengths). However, the L1-RRM2 interaction at 300 mM NaCl was uncoupled to any other long-range contact, appearing as the only inter-domain interaction within the corresponding clusters (see e.g. the upper triangle in panel (e) of Figure 6). This observation suggests that the ionic strength increase breaks some intramolecular interactions, while keeping some of them unaltered.

Interestingly, in the original study, slight perturbations in the interaction network involving disordered regions and linkers were detected for TDP-43 when the salt concentration was increased [58]. For instance, the interaction between IDR1 and IDR2 was reported to disappear at high ionic strength. The inspection of the contact maps of the 300 mM trajectory clusters, indicates that this contact is observed for some conformations, although its population diminishes (from 27.25% at 100 mM to 15.01% at 300 mM), underlining the higher sensitivity of our approach with respect to the averaged contact map (Figure 6(e)). This feature of WARIO is also evidenced for the contact between residues 353-355 of IDR1 and 285-287 of IDR2 identified with WARIO (see e.g. cluster 9 in the lower triangle of Figure 6(e)), which was not detected in the original study.

4 Discussion

The method presented in this work provides a compact and meaningful characterization of conformational ensembles through a weighted family of contact maps. The idea of using a graph-based characterization built from contact information to investigate biomolecular ensembles has been previously proposed [18]. However, due to the enormous structural variability of highly-flexible proteins and to the sparsity of most long-range contacts, the average probability of residue-residue contacts within the ensemble is not a suitable structural descriptor. To account for the complex nature of the contact distribution, WARIO first

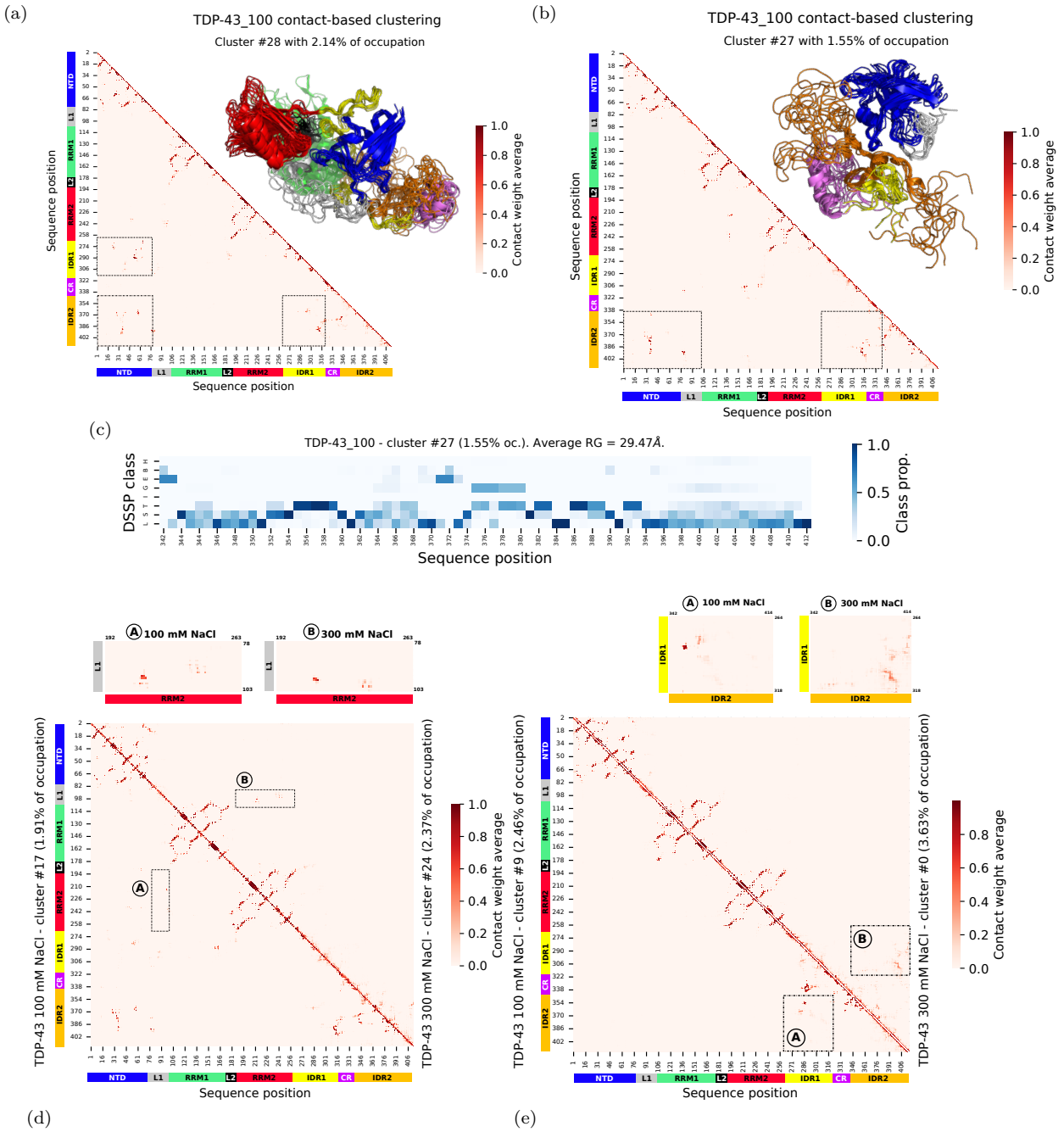

Figure 6: (a,b) Cluster-specific ω-contact maps for two clusters of the TDP-43 ensemble at 100 mM NaCl. In the upper triangles, 10 conformations randomly extracted from the corresponding cluster, aligned at residues that present prominent interactions. The N-terminal domain (NTD) is colored in blue, the two RNA recognition motif-containing domains (RRM1, RRM2) are colored in green and red respectively, the first and second linkers (L1, L2) are colored in gray and black respectively, the two intrinsically disordered regions (IDR1, IDR2) are colored in yellow and orange respectively and the conserved region (CR) is colored in violet. In (b), RRM1, RRM2 and L2 have been removed for easier visualization. (c) Average DSSP secondary structure propensities and average radius of gyration across the 27th cluster conformations, restricted to the IDR2 residues. (d,e) Two pairs of cluster-specific ω-contact maps comparing clusters presenting RRM1-L1 (d) or IDR1-IDR2 (e) inter-domain interactions, at both salt concentrations (100 mM lower triangle, 300 mM upper triangle).

unravels the most determinant interaction patterns that characterize the ensemble and then represents them as easily interpretable cluster-specific contact maps, with associated weights accounting for their population. A key point of this procedure is a novel definition of contact that integrates the chemical nature of the residues involved, their distance along the sequence and their relative orientation. Taking into account the relative orientation of nearby residues along the sequence is essential to correctly identify local structural motifs that can be scarcely populated in the ensemble, but that often are the anchoring points for large biomolecular assemblies where they can modulate the thermodynamics and kinetics of recognition events [59, 60]. The interest of integrating the relative orientation also in the case of long-range contacts is less clear. Indeed, in our analyses of high-resolution protein structures, we could not identify preferred orientations in long-range interactions. Despite this, in our examples, very clear contacts between residues far apart in the sequence were detected in multiple clusters. Importantly, WARIO has been shown to be able to cluster these conformations based primarily on the presence of these long-range contacts. This is possible thanks to the use of contact information to feature conformations instead of global descriptors based on atomic coordinates, which we showed to be less effective to derive structurally meaningful clusters.

It is important to emphasize that descriptors based on contact information hold particular significance for the investigation of disordered proteins. Indeed, they can be directly associated to experimental data reporting on local and global structural details obtained from Nuclear Magnetic Resonance (NMR) [61], Small-Angle X-ray Scattering (SAXS) [62], single molecule Förster Resonance Energy Transfer (sm-FRET) [63], Electronic Paramagnetic Resonance (EPR) [64] or to results of mutational studies [65]. In contrast, the use of atomic coordinates, the most standard descriptor for rigid protein structures, is less suitable in this context, since the experimental techniques providing such information, namely X-ray crystallography and cryo-electron microscopy, are not applicable to highly-flexible systems.

The proposed ensemble characterization approach relying on contact-based clustering is clearly defined and easy to interpret. 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 and unavoidable property of all clustering algorithms. In the vast majority of cases, it is difficult to provide a physical meaning for these hyper-parameters. In our pipeline, the minimum cluster size is easily interpretable as the resolution applied for the ensemble characterization (4). The smaller the size, the finer the classification will be and less frequent contact patterns will be detected, as illustrated in Figure 5(b-d). However, too high resolution could result in redundant group classifications, which would be more difficult to interpret. The choice of the clustering resolution should be made based on the practitioner's needs and its readjustment can be envisioned depending on the results of the clustering. It is important to emphasize that, in general, there is no "true number of clusters", as all classification algorithms aim at representing the diversity of the conformational states rather than revealing all co-existing groups. An effective solution to deal with the dependence on the minimum cluster size would be to incorporate to the pipeline statistical techniques providing evidence of the differences between the clusters obtained at different resolutions and evaluate whether several clusters can be merged into a larger one, or vice versa. This problem is a growing field of research in selective inference [66] and it is referred to as post-clustering inference. However, these methods are highly dependent on the type of algorithm used for clustering and on the interdependence of the observations and descriptors employed. Despite recent remarkable advances [67–69], their application to the evaluation of WARIO results remains to be explored.

In the present study, we have applied WARIO to single-chain trajectories, but its range of applications could be easily extended to study large biomolecular multi-chain complexes with different levels of disorder and ensembles containing several copies of the same or different molecules [70, 71]. Note however that the current implementation of WARIO operates in an all-atom representation of the protein backbone. This

is required for the definition of the residue-specific reference frame and, therefore, for the integration of relative orientation into the contact function. The adaptation of WARIO to coarse-grained models would be extremely valuable in the present context of continuous improvement of force-fields with the aim of investigating condensed states of phase separating systems [72, 73].

As illustrated through the above-presented examples, WARIO can be easily applied to analyze the structural behavior of highly-flexible protein from conformational ensembles produced by MD simulations or other sampling methods, and which are difficult to interpret without the help of this type of statistical tools. Furthermore, the results provided by WARIO can also help to understand structural effects of mutations or changes in the environment of the protein, as shown with the analysis of TDP-43 in two ionic strengths. Nevertheless, we believe that WARIO's greatest potential lies in its coupling with ML methods for the prediction of the conformational behaviour of disordered chains in solution. Some recent studies have shown the potential of ML method to predict structural properties of IDPs/IDRs directly from sequence [74, 75]. However, these approaches are based on extremely simple structural descriptors, such as the radius of gyration and the end-to-end distance, and therefore provide very limited insights into the conformational details at the residue level. Using more detailed descriptors, such as the contact maps proposed in this work, would enable the development of more accurate predictors and generative models for IDPs and IDRs.

Acknowledgements

We are grateful to Liuba Mazzanti, Tâp Ha-Duong, Priyesh Mohanty and Jeetain Mittal for the useful data they provided, as well as for their valuable discussions and comments.

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). The CBS is a member of the French Infrastructure for Integrated Structural Biology (FRISBI), a national infrastructure supported by the French National Research Agency (ANR-10-INBS-05).

References

[1] H. J. Dyson and P. E. Wright. "Intrinsically unstructured proteins and their functions". *Nature Reviews Molecular Cell Biology* 6 (2005), pp. 197–208.

[2] C. J. Oldfield and A. K. Dunker. "Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions". *Annual Review of Biochemistry* 83.1 (2014), pp. 553–584.

[3] 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 (2021), pp. 3817–3828.

[4] A. S. Holehouse and B. B. Kragelund. "The molecular basis for cellular function of intrinsically disordered protein regions". *Nature Reviews Molecular Cell Biology* 25 (2023), pp. 187–211.

[5] H. Ghafouri, T. Lazar, A. Del Conte, L. G. Tenorio Ku, PED-Consortium, P. Tompa, S. C. E. Tosatto, and A. M. Monzon. "PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins". *Nucleic Acids Researchearch* 52.D1 (2023), pp. D536–D544.

[6] D. C. Phillips. “British biochemistry, past and present”. *London Biochemical Society Symposia*. Academic Press. 1970, p. 11.

[7] 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 (1972), pp. 1331–1337.

[8] 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 (1975), pp. 3802–3806.

[9] M. G. Rossman and A. Liljas. “Recognition of structural domains in globular proteins”. *Journal of Molecular Biology* 85.1 (1974), pp. 177–181.

[10] I. Kuntz, G. Crippen, P. Kollman, and D. Kimelman. “Calculation of protein tertiary structure”. *Journal of Molecular Biology* 106.4 (1976), pp. 983–994.

[11] 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 (1983), pp. 21–78.

[12] 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 (2019), pp. 1149–1164.

[13] M. AlQuraishi. “Machine learning in protein structure prediction”. *Current Opinion in Chemical Biology* 65 (2021), pp. 1–8.

[14] 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 (2021), pp. 583–589.

[15] D. Mercadante, F. Gräter, and C. Daday. “CONAN: A Tool to Decode Dynamical Information from Molecular Interaction Maps”. *Biophysical Journal* 114.6 (2018), pp. 1267–1273.

[16] C. Yuan, H. Chen, and D. Kihara. “Effective inter-residue contact definitions for accurate protein fold recognition”. *BMC Bioinformatics* 13.1 (2012).

[17] 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 (2023), p. 046004.

[18] D. Clementel, A. Del 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 Researchearch* 50.W1 (2022), W651–W656.

[19] L. van der Maaten and G. Hinton. “Visualizing Data using t-SNE”. *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[20] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426.

[21] 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 (2019), e1008432.

[22] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. “A review of UMAP in population genetics”. *Journal of Human Genetics* 66.1 (2020), pp. 85–91.

[23] 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 (2017), p. 1750017.

[24] A. Platzer. "Visualization of SNPs with t-SNE". *PLOS ONE* 8.2 (2013), e56883.

[25] 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.* 2023. bioRxiv: 2023.07.06.548007.

[26] M. Allaoui, M. L. Kherfi, and A. Cheriet. "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study". *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 317–325.

[27] M. Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" (2022). arXiv: 2203.05794.

[28] 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 (2018), pp. 38–44.

[29] 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 (2020).

[30] 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* 19.14 (2023), pp. 4711–4727.

[31] 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 (2023), bbad242.

[32] S. Rao and M. G. Rossmann. "Comparison of super-secondary structures in proteins". *Journal of Molecular Biology* 76.2 (1973), pp. 241–256.

[33] V. N. Maiorov and G. M. Crippen. "Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins". *Journal of Molecular Biology* 235.2 (1994), pp. 625–634.

[34] R. J. G. B. Campello, D. Moulavi, and J. Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013, pp. 160–172.

[35] 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 (2022), p. 107700.

[36] 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 Research.* 47.D1 (2018), pp. D475–D481.

[37] 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 (2011), pp. 2003–2005.

[38] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22.12 (1983), pp. 2577–2637.

[39] 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". *Journal of Chemical Theory and Computation* 19.16 (2023), pp. 5609–5620.

[40] 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 (2015), pp. 1528–1532.

[41] 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 (2005), pp. 517–528.

[42] 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 (2013), pp. 2160–2170.

[43] 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 (2009), pp. 198–206.

[44] E. Hangen, O. Féraud, S. Lachkar, H. Mou, N. Doti, G. M. Fimia, N.-v. 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. Ser-Le 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 (2015), pp. 1001–1014.

[45] L. Mazzanti and T. Ha-Duong. "Conformational ensemble of the intrinsically disordered CHCHD4 N-terminal segment (N27)" (2024). DOI: 10.5281/zenodo.10777456.

[46] C. Clementi, M. Vendruscolo, A. Maritan, and E. Domany. "Folding Lennard-Jones proteins by a contact potential". *Proteins: Structure, Function, and Bioinformatics* 37.4 (1999), pp. 544–553.

[47] F. Saudou and S. Humbert. "The Biology of Huntingtin". 89.5 (2016), pp. 910–926.

[48] C. Zuccato, M. Valenza, and E. Cattaneo. "Molecular Mechanisms and Potential Therapeutical Targets in Huntington's Disease". *Physiological Reviews* 90.3 (2010), pp. 905–981.

[49] 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 (2020), 733–746.e5.

[50] J. B. I. Warner, K. M. Ruff, P. S. Tan, E. A. Lemke, R. V. Pappu, and H. A. Lashuel. "Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths". *Journal of the American Chemical Society* 139.41 (2017), pp. 14456–14469.

[51] H. Kang, F. X. Vázquez, L. Zhang, P. Das, L. Toledo-Sherman, B. Luan, M. Levitt, and R. Zhou. "Emerging β-Sheet Rich Conformations in Supercompact Huntingtin Exon-1 Mutant Structures". *Journal of the American Chemical Society* 139.26 (2017), pp. 8820–8827.

[52] 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 (2023), pp. 309–320.

[53] T. J. Cohen, V. M. Lee, and J. Q. Trojanowski. "TDP-43 functions and pathogenic mechanisms implicated in TDP-43 proteinopathies". *Trends in Molecular Medicine* 17.11 (2011), pp. 659–667.

[54] D. S. Protter and R. Parker. "Principles and Properties of Stress Granules". *Trends in Cell Biology* 26.9 (2016), pp. 668–679.

[55] A. E. Conicella, G. L. Dignon, G. H. Zerze, H. B. Schmidt, A. M. D'Ordine, Y. C. Kim, R. Rohatgi, Y. M. Ayala, J. Mittal, and N. L. Fawzi. "TDP-43 α-helical structure tunes liquid–liquid phase separation and function". *Proceedings of the National Academy of Sciences* 117.11 (2020), pp. 5883–5894.

[56] P. Mohanty, J. Shenoy, A. Rizuan, J. F. Mercado-Ortiz, N. L. Fawzi, and J. Mittal. "A synergy between site-specific and transient interactions drives the phase separation of a disordered, low-complexity domain". *Proceedings of the National Academy of Sciences U.S.A.* 120.34 (2023), e2305625120.

[57] H. I. Ingólfsson, A. Rizuan, X. Liu, P. Mohanty, P. C. Souza, S. J. Marrink, M. T. Bowers, J. Mittal, and J. Berry. "Multiscale simulations reveal TDP-43 molecular-level interactions driving condensation". *Biophysical Journal* 122.22 (2023), pp. 4370–4381.

[58] P. Mohanty, A. Rizuan, Y. C. Kim, N. L. Fawzi, and J. Mittal. "A complex network of interdomain interactions underlies the conformational ensemble of monomeric TDP-43 and modulates its phase behavior". *Protein Science* 33.2 (2024), e4891.

[59] P. Tompa, E. Schad, A. Tantos, and L. Kalmar. "Intrinsically disordered proteins: emerging interaction specialists". *Current Opinion in Structural Biology* 35 (2015), pp. 49–59.

[60] N. E. Davey. "The functional importance of structure in unstructured protein regions". *Current Opinion in Structural Biology* 56 (2019), pp. 155–163.

[61] S. Milles, N. Salvi, M. Blackledge, and M. R. Jensen. "Characterization of intrinsically disordered proteins and their dynamic complexes: From in vitro to cell-like environments". *Progress in Nuclear Magnetic Resonance Spectroscopy* 109 (2018), pp. 79–100.

[62] P. Bernadó and D. I. Svergun. "Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering". *Molecular BioSystems* 8 (1 2012), pp. 151–167.

[63] A. Chowdhury, D. Nettels, and B. Schuler. "Interaction Dynamics of Intrinsically Disordered Proteins from Single-Molecule Spectroscopy". *Annual Review of Biophysics* 52.1 (2023), pp. 433–462.

[64] G. Jeschke. "Conformational dynamics and distribution of nitroxide spin labels". *Progress in Nuclear Magnetic Resonance Spectroscopy* 72 (2013), pp. 42–60.

[65] K. Pounot, C. Piersson, A. K. Goring, F. Rosu, V. Gabelica, M. Weik, S. Han, and Y. Fichou. "Mutations in Tau Protein Promote Aggregation by Favoring Extended Conformations". *JACS Au* 4.1 (2024), pp. 92–100.

[66] W. Fithian, D. Sun, and J. Taylor. *Optimal Inference After Model Selection*. 2017. arXiv: 1410.2597.

[67] L. L. Gao, J. Bien, and D. Witten. "Selective Inference for Hierarchical Clustering". *Journal of the American Statistical Association* 119.545 (2024), pp. 332–342. DOI: 10.1080/01621459.2022.2116331.

[68] Y. T. Chen and D. M. Witten. "Selective inference for k-means clustering". *Journal of Machine Learning Research* 24.152 (2023), pp. 1–41.

[69] J. González-Delgado, J. Cortés, and P. Neuvial. *Post-clustering Inference under Dependency*. 2023. arXiv: 2310.11822.

[70] N. Galvanetto, M. T. Ivanović, A. Chowdhury, A. Sottini, M. F. Nüesch, D. Nettels, R. B. Best, and B. Schuler. "Extreme dynamics in a biomolecular condensate". *Nature* 619.7971 (2023), pp. 876–883.

[71] S. Guseva, V. Schnapka, W. Adamski, D. Maurin, R. W. H. Ruigrok, N. Salvi, and M. Blackledge. "Liquid–Liquid Phase Separation Modifies the Dynamic Properties of Intrinsically Disordered Proteins". *Journal of the American Chemical Society* 145.19 (2023), pp. 10548–10563.

[72] 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 (2021).

[73] A. Rizuan, N. Jovic, T. M. Phan, Y. C. Kim, and J. Mittal. "Developing Bonded Potentials for a Coarse-Grained Model of Intrinsically Disordered Proteins". *Journal of Chemical Information and Modeling* 62.18 (2022), pp. 4474–4485.

[74] G. Tesei, A. I. Trolle, N. Jonsson, J. Betz, F. E. Knudsen, F. Pesce, K. E. Johansson, and K. Lindorff-Larsen. "Conformational ensembles of the human intrinsically disordered proteome". *Nature* (2024).

[75] J. M. Lotthammer, G. M. Ginell, D. Griffith, R. J. Emenecker, and A. S. Holehouse. "Direct prediction of intrinsically disordered protein conformational properties from sequence". *Nature Methods* (2024).