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

Javier González-Delgado[1,2], Amin Sagar[3], Christophe Zanon[1], Kresten Lindorff-Larsen[4], 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.*

[4] *The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Denmark*

Abstract

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 work, 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.

1 Introduction

The comparison of protein structures is a crucial problem in structural biology. In the early works [1, 2], 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 [3]. More recently, Lindorff-Larsen and Ferkinghoff-Borg [4] 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.* [5] 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 [4] and [5] 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 [6] 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.* [7], 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 [8, 9]; 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 [10], 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 2.

In this work, 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 the Supplementary Information (SI), 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.

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?

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 $\mathbb{T}^2$ [11,12]. 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

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

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

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

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.

2.2 Accessing empirical probability distributions from sampled conformations

Estimating the local structural descriptor (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

$$(P_{1;n}^l, \ldots, P_{L;n}^l), \tag{3}$$

where each $P_{i,n}^l$, $i = 1, \ldots, L$, is the empirical probability distribution of P_i^l, and n is the number of conformations constituting the sample. Such empirical probability distributions are commonly represented through Ramachandran maps [13].

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.

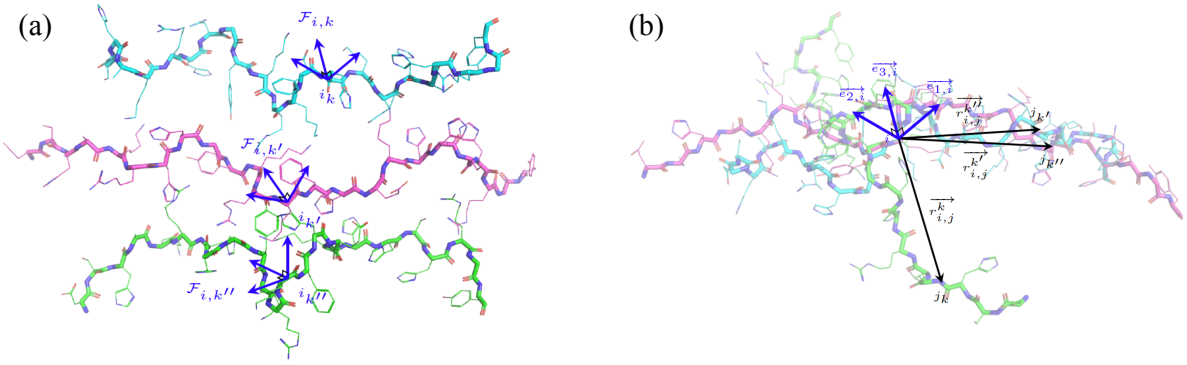

Figure 1: Illustration of how samples of global structural descriptors are obtained, for a pair of positions i, j along the sequence. In (a), the reference frame is built for every conformation at residue i. In (b), all the frames are superimposed using this reference frame. Then, for any $j \neq i$, the vectors $\overrightarrow{r_{i,j}}$ constitute a sample of $\overrightarrow{R_{i,j}}$.

The 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 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 (2) will be estimated by its *empirical* counterpart

$$(P^g_{1,2;n}, P^g_{1,3;n}, \ldots, P^g_{L-1,L;n}), \tag{4}$$

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 S8.

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 an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Integrating the geometry of the underlying space.

The use of metrics between probability distributions is not new in structural biology. For instance, Ting *et al.* [14] 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 [4, 15] 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 [10] and attractive empirical performance [16]. Informally, it represents the minimum transportation cost needed to reconfigure the mass of one probability distribution to recover the other. A more technical definition is provided in SI, and we refer to [16] 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 [17], robustness [18] or fairness [19], among others. With some notable exemptions [5, 20–23], Wasserstein distance has not been widely used in structural biology. More related to our work, in [5], Cazals *et al.* used Wasserstein distance to compare energy landscapes sampled from conformational ensembles. Recently, it was used in [23] to define statistical tests assessing 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 are given in SI.

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

$$(\mathcal{W}_1^{l,A,B}, \dots, \mathcal{W}_L^{l,A,B}) = \left(\mathcal{W}(P_{1;n_A}^{l,A}, P_{1;n_B}^{l,B}), \dots, \mathcal{W}(P_{L;n_A}^{l,A}, P_{L;n_B}^{l,B})\right), \tag{5}$$

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 (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 recently defined in [23]. The second of the introduced techniques 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 [23]. Consequently, together with the L-tuple (5) of distances comparing local structural descriptors, we are able to supply a L-tuple of p-values (corrected for multiplicity [24]) accounting for the statistical significance of the corresponding distances:

$$(p_1^{A,B}, \dots, p_L^{A,B}). \tag{6}$$

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 (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

$$(\mathcal{W}_{1,2}^{g,A,B}, \dots, \mathcal{W}_{L-1,L}^{g,A,B}) = \left(\mathcal{W}(P_{1,2;n_A}^{g,A}, P_{1,2;n_B}^{g,B}), \dots, \mathcal{W}(P_{L-1,L;n_A}^{g,A}, P_{L-1,L;n_B}^{g,B})\right), \tag{7}$$

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 (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 underlying mathematical theory when the ground space has dimension $d \geq 3$. Note that (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). 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.

2.4.1 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 [25, 26]. 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

$$(\widetilde{\mathcal{W}}_1^{l,A,B}, \ldots, \widetilde{\mathcal{W}}_L^{l,A,B}), \tag{8}$$

where each corrected distance is defined as

$$\widetilde{\mathcal{W}}_i^{l,A,B} = \left(\frac{1}{n_I} \sum_{s=1}^{n_I} \mathcal{W}_i^{l,A_s,B_s} - \frac{1}{2(n_I-1)} \sum_{s=2}^{n_I} \left(\mathcal{W}_i^{l,A_1,A_s} + \mathcal{W}_i^{l,B_1,B_s} \right) \right)_+, \quad \text{for all } i = 1, \ldots, L, \tag{9}$$

where, for any real number x, $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise. The first term in (9) is an average of n_I Wasserstein distances between n_I paired independent replicas of A and B. As it was shown in [27], 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 (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 (9) can be accordingly adapted. As it is illustrated in Section 3, the use of corrected distances (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.

2.4.2 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 [1, 2]. 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 (9), the score

$$\frac{\widetilde{\mathcal{W}}_i^{l,A,B}}{\mathcal{W}_{\text{intra}}^{l,A,B}} = \frac{\left(\mathcal{W}_{\text{inter}}^{l,A,B} - \mathcal{W}_{\text{intra}}^{l,A,B}\right)_+}{\mathcal{W}_{\text{intra}}^{l,A,B}}, \tag{10}$$

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.

2.4.3 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 [4]. 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 $d_1, \dots, d_L$ are L distances defined on L metric spaces $\mathcal{X}_1, \dots, \mathcal{X}_L$, 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,

$$\mathcal{OW}^{l,A,B} = \left(\sum_{i=1}^{L} \left(\mathcal{W}_i^{l,A,B}\right)^2\right)^{1/2} \tag{11}$$

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,

$$\mathcal{OW}^{g,A,B} = \left(\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \left(w_{ij}\mathcal{W}_{i,j}^{g,A,B}\right)^2\right)^{1/2}, \quad \text{with } w_{ij} > 0 \text{ for all } i, j \in \{1, \dots, L\}, \tag{12}$$

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 (12). This allows to consider distances between specific residue pairs as more relevant than the others when computing the overall discrepancy [28]. 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

$$w_{ij} = w(i,j) = \frac{1}{\tanh 1} \tanh\left(\left(\frac{|i-j|}{L-1}\right)^{\frac{1}{2}}\right), \tag{13}$$

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 uncertainty discussed in Section 2.4.1. 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 (11) (resp. (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.

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 2.4, containing global and local differences. Users can choose to correct the computed distances by uncertainty, as proposed in Section 2.4.1. 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 defined in Section 2.4.2. The overall score defined in Section 2.4.3, aggregating the corrected distances, is 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 (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), (7). The values for the discrepancies between the global structural descriptors (4) are provided in the lower triangle. The differences between the local structural descriptors (3) are displayed along the diagonal. Details on the interpretation of the matrix are given in Section S1.3 and illustrated in Figure S4. These guidelines are also presented in the software documentation.

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 (9), as independent replicas were available. The results are depicted through the score (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 (11) and (12), as discussed in Section 2.4.3. The weight function (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 (12), and not to depict distance values in the matrix representation, which correspond to the interpretable scale defined in Section 2.4.2. An additional analysis illustrating the application of WASCO to assess the convergence of MD simulations is included in SI.

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 [29] 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 [29]). 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 (9).

Figure 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 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 [15, 30].

Matrices returned by WASCO also allow a residue-specific analysis of the distances. In Figure 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

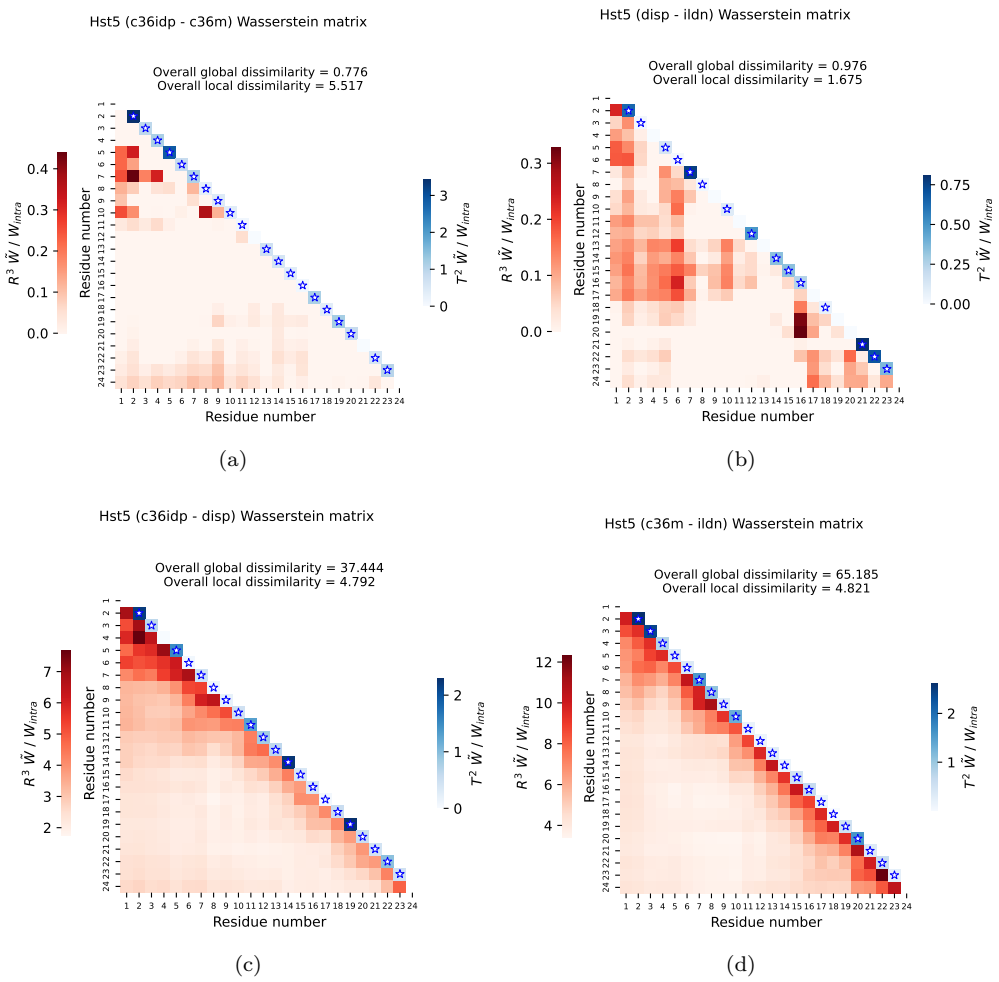

Figure 2: Comparison of Molecular Dynamics simulations of Hst5 ensemble using different force-fields. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (10), representing the relative difference between the inter-ensemble distances and the uncertainty.
The coefficients in the lower-triangle (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 (6) is smaller than $\alpha = 0.05$). Note the different scales used in the different plots.

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 SI.

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) [31] on the resulting ensemble. We first compared the Hst5 ensemble simulated with Flexible-Meccano [25, 32] with the refined one using previously reported SAXS data [33]. The results are presented in Figure 3. 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 [33]. Small but significant differences were observed at the central part of the

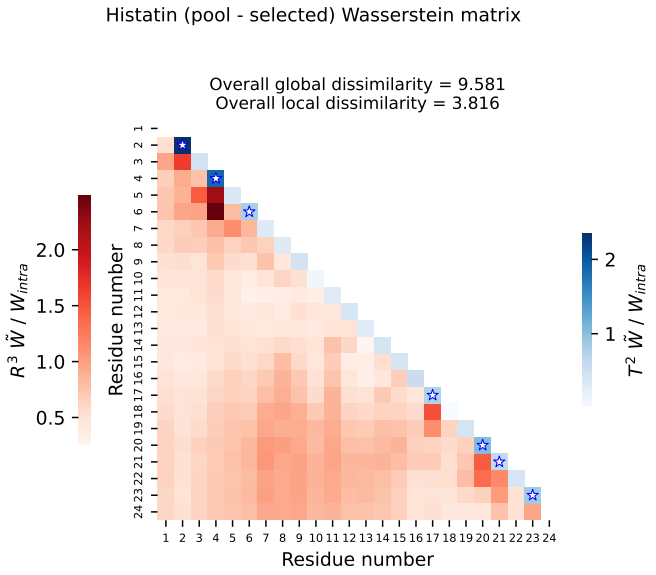

Figure 3: Comparison of Hst5 ensemble before and after filtering with experimental SAXS data. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower triangle (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 (6) is smaller than $\alpha = 0.05$).

peptide (from residues 6 to 13). Most probably, the SAXS-based refinement selected conformations with an extended central region to account for the overall expansion of peptide in the solution [31]. Moreover, we observed highly significant local distances that propagate 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 towards the C-terminus.

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 [26], 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 (Figures S9a and S9b), 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 3).

In a recent study, ENCORE was used to show that refined ensembles were closer to each other than different input ensembles [15]. This can also be illustrated using WASCO, by comparing the Hst5 ensembles generated in both directions before and after the filtering with SAXS data (Figures S9c and S9d). 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 S9c that disappeared after refinement.

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 [34] or secondary structure propensities [29]. 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 [7]. This is illustrated with an example in SI (Section S2.3). 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 [35–37]. 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.

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).

References

[1] Rao, S. and Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J Mol Biol* **76**, 2, 241–256.

[2] Maiorov, V. N. and Crippen, G. M. (1994). Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol* **235**, 2, 625–634.

[3] Brüschweiler, R. (2003). Efficient RMSD measures for the comparison of two molecular ensembles. *Proteins* **50**, 1, 26–34.

[4] Lindorff-Larsen, K. and Ferkinghoff-Borg, J. (2009). Similarity measures for protein ensembles. *PLoS One* **4**, 1 (01), 1–13.

[5] Cazals, F., Dreyfus, T., Mazauric, D., Roth, C.-A., and Robert, C. H. (2015). Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J Comput Chem* **36**, 16, 1213–1231.

[6] Huihui, J. and Ghosh, K. (2021). Intrachain interaction topology can identify functionally similar intrinsically disordered proteins. *Biophys J* **120**, 10, 1860–1868.

[7] Lazar, T., Guharoy, M., Vranken, W., Rauscher, S., Wodak, S. J., and Tompa, P. (2020). Distance-based metrics for comparing conformational ensembles of intrinsically disordered proteins. *Biophys J* **118**, 12, 2952–2965.

[8] Kullback, S. (1952). An application of information theory to multivariate analysis. *The Annals of Mathematical Statistics*, 88–102.

[9] Zhou, S. K. and Chellappa, R. (2006). From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Trans Pattern Anal Mach Intell* **28**, 6, 917–929.

[10] Villani, C. (2008). *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg.

[11] Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63**, 2, 505–512.

[12] Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci USA* **105**, 26, 8932–8937.

[13] Ramachandran, G., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 1, 95–99.

[14] Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M., and Dunbrack, R. (2010). Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput Biol* **6**, 4, e1000763.

[15] Tiberti, M., Papaleo, E., Bengtsen, T., Boomsma, W., and Lindorff-Larsen, K. (2015). ENCORE: software for quantitative ensemble comparison. *PLoS Comput Biol* **11**, 10, e1004415.

[16] Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* **11**, 5-6, 355–607.

[17] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds. Proceedings of Machine Learning Research, Vol. **70**. PMLR, 214–223.

[18] Serrurier, M., Mamalet, F., González-Sanz, A., Boissin, T., Loubes, J.-M., and del Barrio, E. (2020). Achieving robustness in classification using optimal transport with hinge regularization. arXiv:2006.06520v3.

[19] del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2019). 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*.

[20] Berg, A., Kukharenko, O., Scheffner, M., and Peter, C. (2018). Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers. *PLoS Comput Biol* **14**, 11, e1006589.

[21] Rosenbaum, D., Garnelo, M., Zielinski, M., Beattie, C., Clancy, E., Huber, A., Kohli, P., Senior, A. W., Jumper, J., Doersch, C., and others. (2021). Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. arXiv:2106.14108.

[22] Damjanovic, J., Murphy, J. M., and Lin, Y.-S. (2021). Catboss: Cluster analysis of trajectories based on segment splitting. *J Chem Inf Model* **61**, 10, 5066–5081.

[23] González-Delgado, J., González-Sanz, A., Cortés, J., and Neuvial, P. (2021). Two-sample goodness-of-fit tests on the flat torus based on wasserstein distance and their relevance to structural biology. arXiv:2108.00165.

[24] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Stat* **6**, 2, 65–70.

[25] Ozenne, V., Bauer, F., Salmon, L., Huang, J.-r., Jensen, M. R., Segard, S., Bernadó, P., Charavay, C., and Blackledge, M. (2012). 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.

[26] Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J., and Bernadó, P. (2019). Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. *Structure* **27**, 2, 381–391.e2.

[27] Sommerfeld, M., Schrieber, J., Zemel, Y., and Munk, A. (2019). Optimal transport: Fast probabilistic approximation with exact solvers. *J Mach Learn Res* **20**, 105, 1–23.

[28] Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**, 1, 123–138.

[29] Jephthah, S., Pesce, F., Lindorff-Larsen, K., and Skepö, M. (2021). Force field effects in simulations of flexible peptides with varying polyproline II propensity. *J Chem Theory Comput* **17**, 10, 6634–6646.

[30] Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W., and Lindorff-Larsen, K. (2015). Comparing molecular dynamics force fields in the essential subspace. *PLoS One* **10**, 3, e0121114.

[31] Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **129**, 17, 5656–5664.

[32] Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. H., and Blackledge, M. (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci USA* **102**, 47, 17002–17007.

[33] Sagar, A., Jeffries, C. M., Petoukhov, M. V., Svergun, D. I., and Bernadó, P. (2021). 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.

[34] Tria, G., Mertens, H. D. T., Kachala, M., and Svergun, D. I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2**, 2, 207–217.

[35] Lindorff-Larsen, K. and Kragelund, B. B. (2021). 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.

[36] Janson, G., Valdes-Garcia, G., Heo, L., and Feig, M. (2022). Direct generation of protein conformational ensembles via machine learning. bioRxiv:2022.06.18.496675.

[37] Zhang, O., Haghighatlari, M., Li, J., Teixeira, J. M. C., Namini, A., Liu, Z.-H., Forman-Kay, J. D., and Head-Gordon, T. (2022). Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. arXiv:2206.12667.