Supplementary Information

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*

S1 Methodology details

S1.1 Building a residue-specific reference frame

S1.1.1 Reference frame definition

We seek to define a reference frame that determines the global pose (position and orientation) of a given residue and that allows to describe the relative pose of other residues along the sequence. As we want this reference system to be universally defined (independently of the residue identity), we first define a virtual atom $\widetilde{C_\beta}$, which exists also for glycines. The position of $\widetilde{C_\beta}$ is an estimate of the position of the true C_β when it exists, but it is defined for every residue using only the atoms that are always present. Its definition allows the construction of a universal frame that locally represents the geometry of the backbone.

Let $\overrightarrow{C}$ and $\overrightarrow{N}$ be the vectors going from C_α to C and N atoms, respectively. If a C_β atom is present, let $\overrightarrow{C_\beta}$ denote the vector going from C_α to C_β. In such case, $\overrightarrow{C_\beta}$ can be determined using the vectors $\overrightarrow{C}$, $\overrightarrow{N}$ and $\overrightarrow{C} \times \overrightarrow{N}$ together with their angles with respect to $\overrightarrow{C_\beta}$, denoted θ_C, θ_N and θ_{CN} respectively. See Figure S1a for an illustration. This can be done by solving the following linear system, whose unknown variables are the three coordinates of C_β.

$$\begin{cases} \|\overrightarrow{N}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_N = \overrightarrow{N} \cdot \overrightarrow{C_\beta} \\ \|\overrightarrow{C}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_C = \overrightarrow{C} \cdot \overrightarrow{C_\beta} \\ \|\overrightarrow{C} \times \overrightarrow{N}\| \, \|\overrightarrow{C_\beta}\| \cos\theta_{CN} = (\overrightarrow{C} \times \overrightarrow{N}) \cdot \overrightarrow{C_\beta}. \end{cases} \tag{1}$$

To define a *universal* C_β, denoted $\widetilde{C_\beta}$, we will estimate fixed values for θ_N, θ_C and θ_{CN} from all non-glycine residues of a set of protein structures and *define* the $\widetilde{C_\beta}$ coordinates as the solution of (1), independently of the residue identity. Details on angles estimation are given in the following section. Consequently, for a given residue, the virtual atom $\widetilde{C_\beta}$ is determined from the coordinates of its C_α, N and C atoms. This allow us to define a reference system at each sequence position through the following three vectors, where $\overrightarrow{CN} = \overrightarrow{N} - \overrightarrow{C}$.

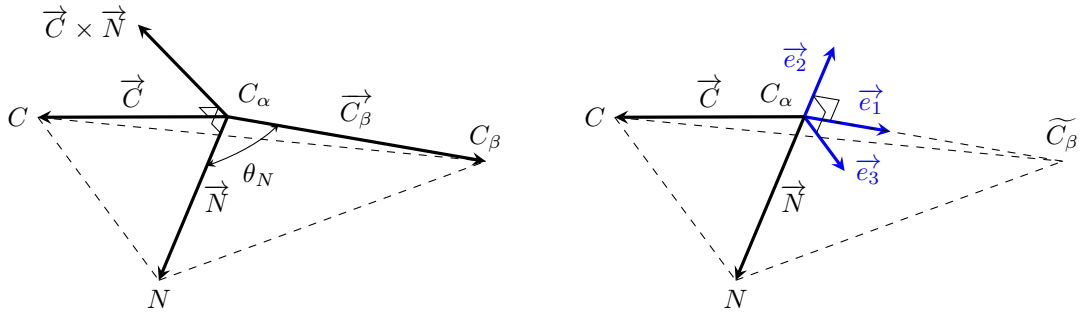


Figure S1: (a) Illustration of vectors and angles involved in the construction of the residue-specific reference frame. The vector $\overrightarrow{C_\beta}$ can be determined from vectors $\overrightarrow{C}$ and $\overrightarrow{N}$ together with the angles θ_N (the only depicted for simplicity), θ_C and θ_{CN}. (b) The three vectors $\{\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}\}$ defining the reference frame, built from the virtual atom $\widetilde{C_\beta}$ and vectors $\overrightarrow{C}$ and $\overrightarrow{N}$.

$$\begin{cases} \overrightarrow{e_1} = \overrightarrow{\widetilde{C}_\beta} / \| \overrightarrow{\widetilde{C}}_\beta \| \\ \overrightarrow{e_2} = \overrightarrow{CN} / \| \overrightarrow{CN} \| \times \overrightarrow{e_1} \\ \overrightarrow{e_3} = \overrightarrow{e_1} \times \overrightarrow{e_2}. \end{cases} \tag{2}$$

Once the reference system of the i-th residue, denoted $\mathcal{F}_i = \{\overrightarrow{e_{1,i}}, \overrightarrow{e_{2,i}}, \overrightarrow{e_{3,i}}\}$, has been built, its origin will be placed at the C_β atom when it exists, or at the C_α otherwise. This allows the computation of relative positions and distances with respect to C_β atoms for all non-glycine residues.

S1.1.2 Estimation of θ_C, θ_N and θ_{CN}

We estimated three fixed values for θ_C, θ_N and θ_{CN}, to be replaced in the linear system (1). After that, the vector $\overrightarrow{\widetilde{C}_\beta}$ is determined for each residue along the sequence by solving (1) after plugging in the corresponding coordinates of C_α, C and N atoms. As mentioned in Section S1.1, this allows the definition of a residue-specific reference frame, built independently of the residue identity.

To estimate the three angles, we used a set of 15177 experimentally-determined high-resolution structures of protein domains extracted from the SCOPe 2.07 release [1]. For each structure, θ_C, θ_N and θ_{CN} were computed and stored for every non-glycine residue. The three corresponding histograms, together with a kernel density estimate, are presented in Figure S2, for all residue types. The residue-specific counterparts of Figure S2 did not show important fluctuations from the overall densities. Therefore, for simplicity, we did not estimate three angles per residue type, but three universal values.

The three distributions of Figure S2 show that all the angle distributions are strongly concentrated around their kernel density maximum. Consequently, these values were chosen as an estimate of θ_C, θ_N and θ_{CN}. Due to the symmetry of the empirical distributions, choosing the mean would provide similar estimates. Figure S2 depicts the theoretical angle values under the hypothesis that C, N, C_β and H (when present) are the vertices of a regular tetrahedron, with C_α as its centroid. One could think of using these values as estimates, but the deviation from the experimental value of θ_{CN} is too high, showing how the fluctuations from the regular polyhedron are not homogeneous along its faces.

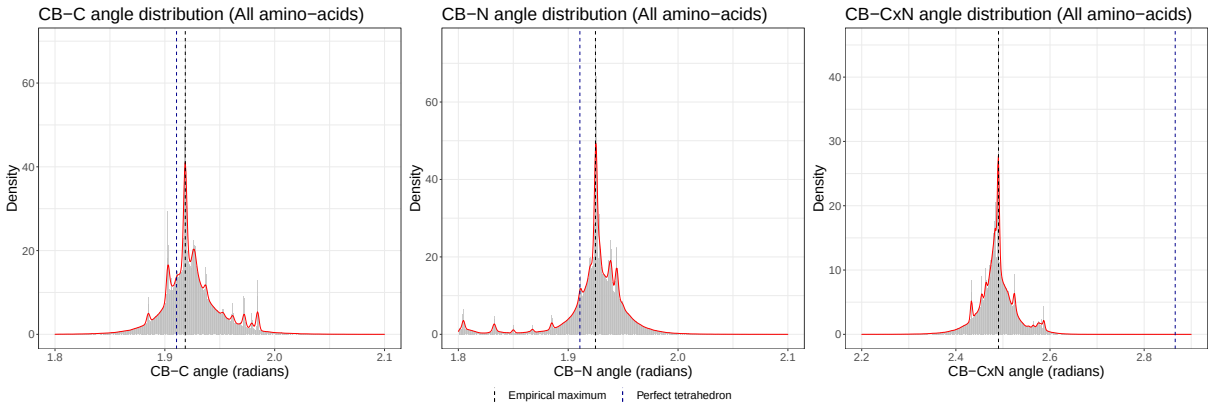


Figure S2: From left to right: empirical distributions of θ_C, θ_N and θ_{CN} respectively, extracted from a set of 15177 protein structures, considering all non-glycine residues. The red line corresponds to a kernel density estimate, whose maximum (vertical black dashed line) was used as angle estimate. The blue dashed line depicts the theoretical value of each angle under the hypothesis that the four atoms bound to the C_α form a regular tetrahedron.

S1.2 Wasserstein distance: definition and computation

S1.2.1 The Optimal Transport problem

Let P, Q be two probability distributions supported on an arbitrary[1] space $\mathcal{X}$. Let $\mathcal{U}(P, Q)$ denote the space of probability distributions supported on $\mathcal{X} \times \mathcal{X}$ having P and Q as marginals on $\mathcal{X}$. Finally, let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ be a cost function, usually a distance on $\mathcal{X}$. A probability distribution $\pi \in \mathcal{U}(P, Q)$ is said to be an *optimal transport plan for the cost* d^p between P and Q if it solves, for $p > 1$,

$$\mathcal{W}_p(P, Q) := \left(\inf_{\gamma \in \mathcal{U}(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} d^p(x, y) d\gamma(x, y) \right)^{\frac{1}{p}}. \tag{3}$$

The optimal value $\mathcal{W}_p(P, Q)$ is called the *p-Wasserstein distance* between P and Q. Indeed, $\mathcal{W}_p(P, Q)$ is a distance on the space of probability distributions supported on $\mathcal{X}$ [2]. The Wasserstein distance corresponds to the minimum transportation cost needed to reconfigure the mass of P to recover Q using the transport plan given by the minimizer π of (3). Note that, for a pair of measurable sets $A, B \subset \mathcal{X}$, $\pi(A \times B)$ represents the probability of sending to B the mass in A or, in other words, the proportion of the mass in A that must be sent to B. The optimization problem (3) is the continuous version of the so-called *Kantorovich problem*. When, instead of continuous probability distributions P, Q, we consider their empirical counterparts P_n, Q_m, built from a sample drawn from P and Q respectively, the problem (3) is rewritten in terms of matrices and can be easily solved in practice for small dimensions. The resulting optimal value is called the *empirical* p-Wasserstein distance. Under very mild assumptions, is a good approximation of (in the sense that it converges in probability to) the Wasserstein distance between the corresponding pair of continuous measures.

In our case, we set $\mathcal{X}$ to $\mathbb{R}^3$ and $\mathbb{T}^2$ for the global and local descriptors respectively, where the cost function is the geodesic distance d in such spaces. Note that this makes (3) integrate the geometry of the conformational space. We consider the 2-Wasserstein distance and refer to it simply as the Wasserstein distance.

[1] $\mathcal{X}$ is only required to be a *Polish* space, i.e. complete, separable and metric. This is the case for the spaces of interest in this work, namely $\mathbb{R}^d$ and $\mathbb{T}^d$, for any $d > 1$.

S1.2.2 Practical implementation

The Wasserstein distance can be easily computed from a pair of samples drawn from the corresponding probability distributions. However, a major drawback of the algorithms that compute the Wasserstein distance is their inability to handle large datasets ($\gtrsim 10^3$ points). The current implementations in Python [3] or R [4] only admit datasets with $\lesssim 5 \cdot 10^3$ points, which is usually not enough for conformational ensembles of IDPs. To the best of our knowledge, there are no existing algorithms that solve an OT problem for large sample sizes and that are easily implementable, considerably fast (which, in our case, is essential due to the large number of Wasserstein distances to compute), and that accept non-euclidean ground distances (like the distance in the torus).

Here, we propose an approximation method to "simplify" the input empirical distributions and compute the Wasserstein distance from a pair of smaller samples sizes. The efficiency of this approach in terms of error is illustrated via simulations on real protein data, but we provide no theoretical bounds. The proposed algorithm consists in clustering the original distribution and defining its clustered version as a discrete probability distribution supported on the set of clusters whose mass is given by the proportion of points assigned to each cluster. Then, the Wasserstein distance is computed between the pair of clustered distributions, whose samples have admissible sizes. The method is implemented for both local and global structural descriptors, which are empirical probability distributions supported on $\mathbb{T}^2$ and $\mathbb{R}^3$ respectively.

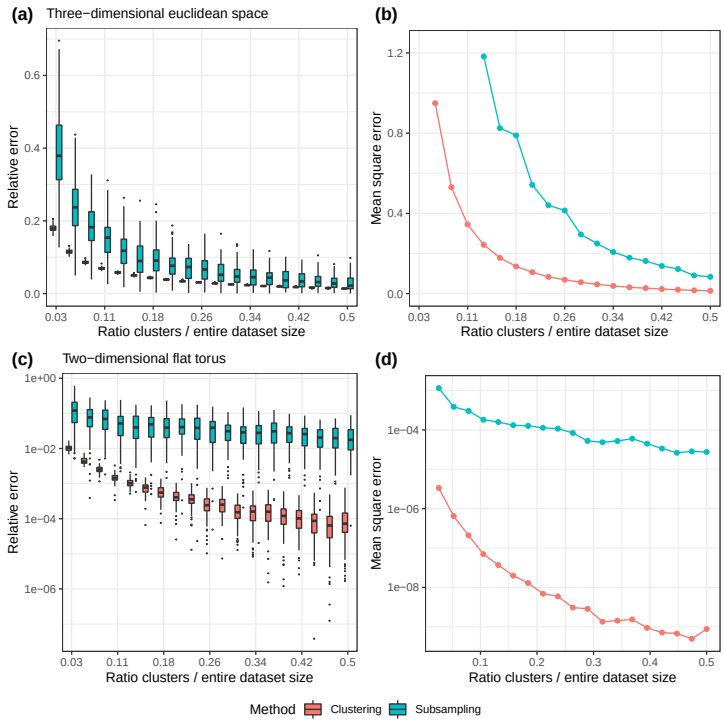


Figure S3: From left to right (columns): relative and mean square error estimates of the Wasserstein distance between the clustered distribution as an estimate of the Wasserstein distance between the original datasets. In abscissas, the proportion of the number of clusters with respect to the entire dataset size. The first row (a,b) corresponds to samples drawn from local structural descriptors (dihedral angles) and the second (c,d) to samples drawn from global structural descriptors (pairwise relative positions of residues).

The accuracy in terms of relative and mean-square error is presented in Figure S3. Note that the approximation algorithm has a considerably better performance when implemented for local structural descriptors, which was expected due to the boundedness of the corresponding ground space. Accuracy in $\mathbb{R}^3$ is slightly worse, as cloud points representing the relative position of residues are in general more disperse, and therefore the clustered distribution needs a larger number of centroids to better capture its variability. Nevertheless, we observe that, in both cases, the error estimates for a proportion of $\sim 10\%$ of clusters with respect to the entire dataset size (the proportion we will be using in practice) are acceptable for our practical purposes. To enrich the interpretation, we performed the same accuracy analysis but by computing the Wasserstein distance between subsamples drawn uniformly from the corresponding datasets. As shown in Figure S3, the effect of clustering significantly improves the quality of the approximation.

S1.3 The matrix representation

The result of the comparison analysis is represented through a matrix, $\mathbf{W}$. We will denote by $\mathcal{W}_{ij}$ the entries of $\mathbf{W}$, where $i, j \in \{1, \ldots, L\}$. The matrix will be lower triangular (i.e. $\mathcal{W}_{ij} = 0$ if $j > i$). Figure S4 illustrates the main elements of the matrix representation, which are described below.

1. The matrix is headed by a title describing the comparison, introduced by the user.

2,3. Below the title, the overall local and global discrepancies are depicted (equations (11) and (12) in the main text). By default, they are computed by aggregating and weighting the corrected distances as described in Section 2.4.3. These features can be modified by the user.

4,5. The matrix entries are represented using two independent color scales, for local and global differences. Both scales correspond to the score (10) defined in Section 2.4.2, which can be computed when several independent replicas of each ensemble are available. Otherwise, distances cannot be corrected by uncertainty and the scale will correspond to the (non-corrected) inter-ensemble local and global distances (equations (5) and (7) in the main text).

6. The entries $\mathcal{W}_{ij}$ for $i < j$ correspond to the scores (10) computed for the i, j-th global structural descriptors, i.e. the score comparing the relative position distribution of the i-th and j-th residues in the two ensembles. If no independent replicas are available, the entry corresponds to the i, j-th global distance in (7).

7. The entries $\mathcal{W}_{ii}$ correspond to the scores (10) computed for the i-th local structural descriptors, i.e. the score comparing the (ϕ, ψ) distribution of the i-th residue in the two ensembles. If no independent replicas are available, the entry corresponds to the i-th local distance in (5).

8. The entries $\mathcal{W}_{ii}$ are marked with a star if their associated p-value (6) is less than the significance level $\alpha = 0.05$.

9. The axes labels correspond to the residue position, counting from the N-terminal, relative to the sequence segment that is being compared (and not to the absolute position in the entire sequence).

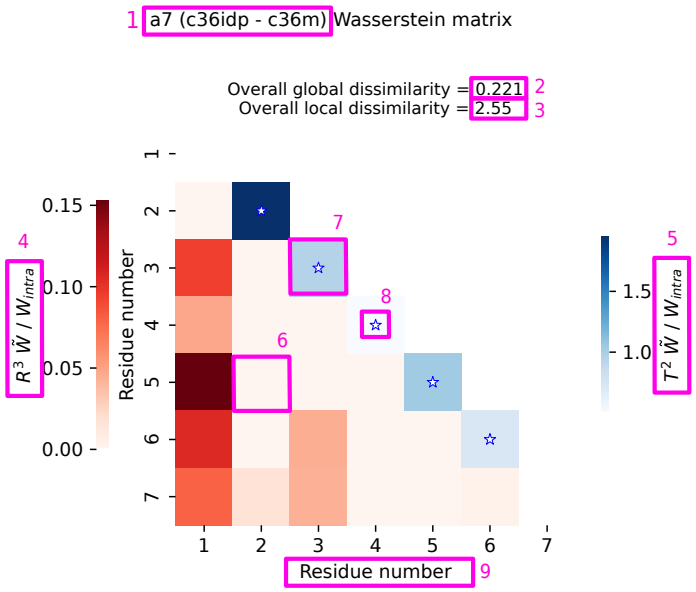


Figure S4: Schematic representation of the output of WASCO. All the elements marked with numbers are described in Section S1.3.

S2 Additional results

S2.1 Comparison of PEP3 ensembles produced by MD simulations using different force-fields

We replicated the analysis described in Section 3.1 for MD simulations of PEP3 using the same force-fields. Results are presented in Figure S5. Here, the discrimination between the two force-field families is not observed. Nonetheless, we still observe that structures simulated with disp and ildn are very close in Wasserstein distance (Figure S5b). Indeed, the overall global dissimilarity is substantially smaller than these of the remaining comparisons. Only inter-ensemble corrected differences representing about 20% of the intra-ensemble ones appear for residues at the C-terminus. The distances between c36idp and c36m are now higher than for Hst5, and corrected differences of the same magnitude than the intra-ensemble ones appear in the interior of the matrix. The same behavior is observed when comparing force-fields of different groups for PEP3. See, for instance, that substantial differences arise between relative positions of residues at opposite terminus in panel (d), which are highly weighted when computing the overall global discrepancy. One intriguing observation is that while there are substantial differences between disp and ildn (and between c36idp and c36m), simulations with c36idp and c36m used the same water model (the CHARMM-modified TIP3P water model) and the disp and ildn simulations also used very similar water models (TIP4P-D and a slightly variant of this) [5]. Overall, these results are complementary to those presented in [5], which mainly focused on secondary structure differences among ensembles, and they show the ability of WASCO to identify differences at both local and global scales.

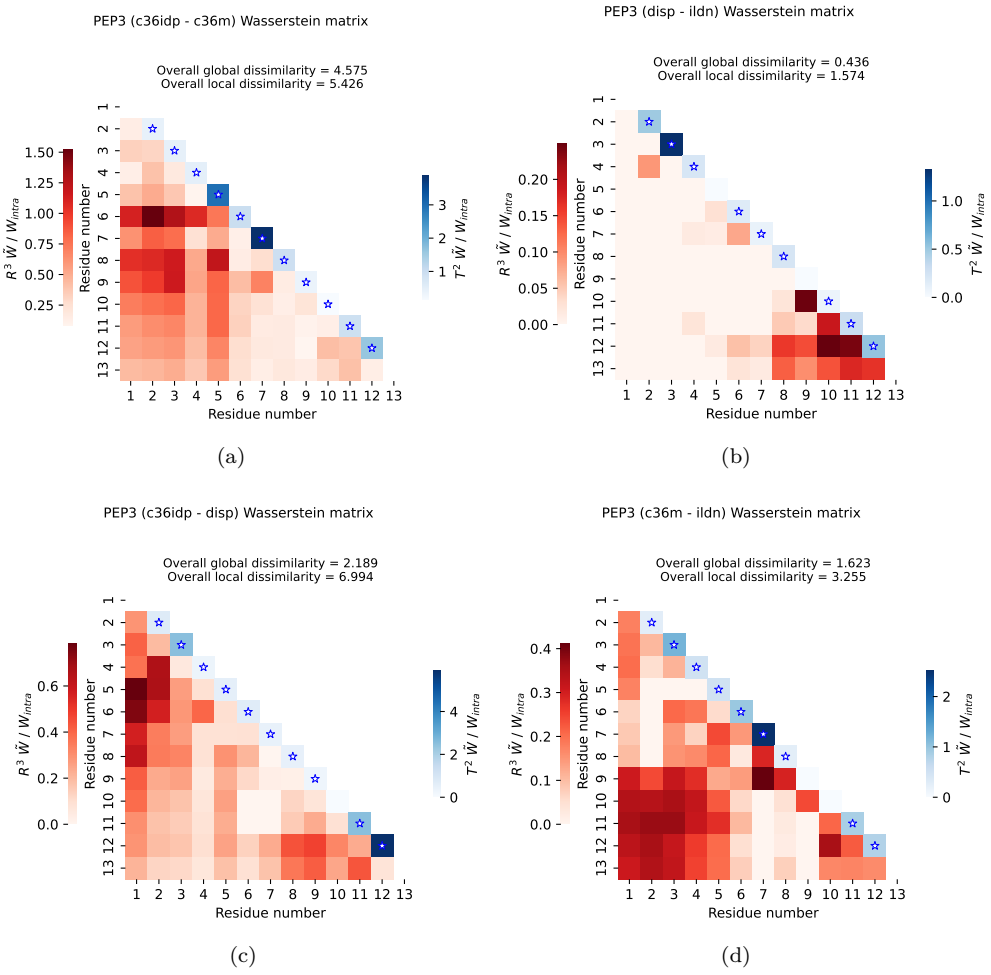


Figure S5: Comparison of Molecular Dynamics simulations of PEP3 ensemble using different force fields. The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower-triangle (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 $\alpha = 0.05$).

S2.2 Assessment of the convergence of MD simulations

Ensemble comparisons have previously been used to assess convergence in MD simulations of folded proteins [6–8]. We here propose to use the overall ensemble distances (defined in Section 2.4.3) to examine the convergence of an MD simulation of a disordered protein. Moreover, this can be done on-the-fly to assess whether the simulation can be stopped. Let T denote the current simulation time and let $0 < t_1 < t_2 < \cdots < t_k = T$ be k time points. If we denote by A_t the conformational ensemble simulated up to time t, we can compute the *online* overall distances

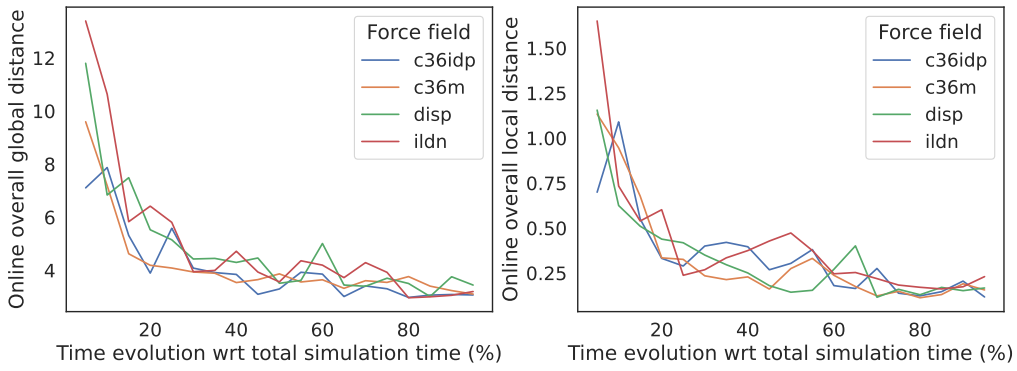
$$\mathcal{OW}_i^l = \mathcal{OW}^{l,A_{t_{i-1}},A_{t_i}}, \tag{4}$$

defined in (11) of the main text, for all $i = 2, \ldots, k$. Analogously, we compute the *online* overall global distances

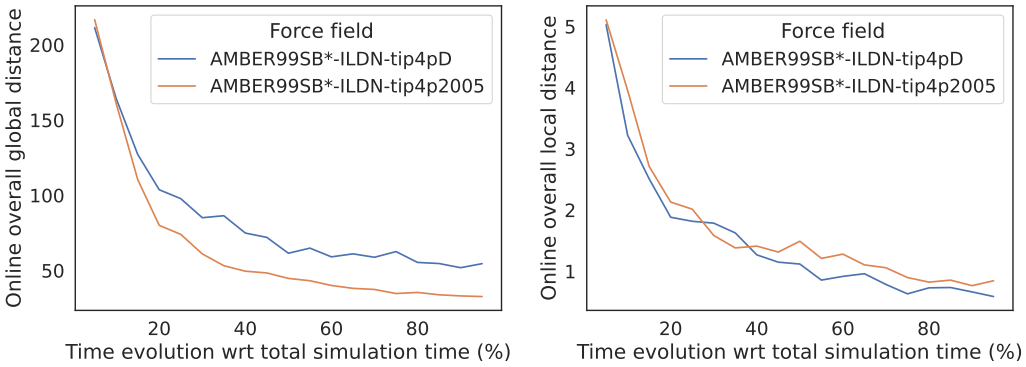
$$\mathcal{OW}_i^g = \mathcal{OW}^{g, A_{t_{i-1}}, A_{t_i}}, \tag{5}$$

as defined in (12) of the main text.

For each i, $\mathcal{OW}_i^l$ (resp. $\mathcal{OW}_i^g$) corresponds to the overall local (resp. global) distance between the ensemble from $t = 0$ to $t = t_i$ and the ensemble from $t = 0$ to $t = t_{i-1}$. In other words, (4) (resp. (5)) is the distance between the ensembles simulated up to time t_{i-i} and up to time t_i. Consequently, it quantifies whether the new simulated trajectories between t_{i-1} and t_i yielded a non-negligible contribution to the ensemble structure (if (4) is not small) or, otherwise, whether proceeding the simulation up to t_i does not yield any substantial contribution (if (4) is close to zero). Then, the representation of $\mathcal{OW}_i^l$, $\mathcal{OW}_i^g$ with respect to the t_i indicates whether the simulation has converged or not. Note that the distances $\mathcal{OW}_i^l$, $\mathcal{OW}_i^g$ can never be equal to zero, as they are empirical distances which *converge* to zero when the sample size tends to infinity. Therefore, the profiles will approach a non-zero plateau under convergence, whose ordinate will decrease when sample size increases. The criteria to assume convergence will be therefore



(a) PEP3



(b) K-18 domain of Tau

Figure S6: (a) Online convergence analysis for PEP3 ensemble simulated with force-fields c36idp, c36m, disp and ildn. (b) Online convergence analysis for K-18 domain of Tau ensemble simulated with AMBER ff99SB*-ILDN and TIP4P-D water model. In abscissas, the percentage of simulation time, divided in 20 equally spaced time intervals. In ordinates, the overall distances between the ensembles simulated at the extremes of the time intervals. The left (resp. right) column presents the evolution of $\mathcal{OW}_i^g$ (resp. $\mathcal{OW}_i^l$) with respect to time.

the reach of such a plateau at a *reasonable* ordinate, meaning that it must be small enough if sample sizes are considerably large. Nevertheless, this criteria provides a stronger evidence of non-convergence, as the achievement of an asymptote for (5), even if necessary, may not be sufficient to guarantee convergence. If we resolve that the simulation must keep going until time $T' > T$, it suffices to add $\mathcal{OW}^{l,A_T,A_{T'}}$ and $\mathcal{OW}^{g,A_T,A_{T'}}$ to each curve and recheck.

Figure S6a presents the evolution of the online overall distances for PEP3 simulated with the four force-fields introduced in Section 3.1. We observe that all the curves exhibit an asymptote at a value close to zero after 80% of simulation time, which is compatible with convergence in all cases. This is not the case for the simulation in Figure S6b, corresponding to a 1,000 ns simulation of the K-18 domain of Tau using the AMBER ff99SB*-ILDN force-field and the TIP4P-D water model (Sthitadhi Maiti and Matthias Heyden, unpublished). Here, we clearly observe that curves do not reach an asymptote and present a decreasing behavior during all the time evolution. This result was expected due to the length of the protein (129 amino acids) and the reduced simulated time.

S2.3 Comparison of ensembles using distance matrices

As it is discussed in Section 1, the use of average descriptors to compare IDP ensembles may yield a substantial loss of information when the underlying distributions describing their structure exhibit a high and complex variability. The work presented in [9] computes the median C_α-C_α distance for every pair of residues $i < j$, denoted $\bar{d}_{ij}$, as well as its corresponding standard deviation, denoted σ_{ij}. If $\bar{d}^A_{ij}$, σ^A_{ij} (resp. $\bar{d}^B_{ij}$, σ^B_{ij}) denote the previously defined descriptors for ensemble A (resp. B), the difference between both ensembles is given by a matrix with entries M_{ij}, where

$$M_{ij} = \begin{cases} \Delta\bar{d}_{ij} = |\bar{d}^A_{ij} - \bar{d}^B_{ij}| & \text{if} \quad i < j, \\ \Delta\sigma_{ij} = |\sigma^A_{ij} - \sigma^B_{ij}| & \text{if} \quad j > i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In [9], the entries M_{ij} are neglected if they are not significantly different from zero (according to a Mann-Whitney-Wilcoxon test for the distance distributions). Here, we skipped this step for simplicity. We computed the matrix with entries M_{ij} for the comparison analysis presented in Section 3.1, using one replica per ensemble. The counterpart of Figure 2 is depicted in Figure S7. As could be anticipated, the conclusions stated in Section 3.1 are difficult to extract from the matrices in Figure S7. First, the overall behaviour between force-fields suggested by Figure 2 is not observed in the distance matrices, as the corresponding color scales do not present significant discrepancies in the distance magnitudes between comparisons (see, on the contrary, the differences between rows in Figure 2). When looking at the differences located in the interior of the matrices, some similarities might arise between Figures 2 and S7 for the top left comparison (c36idp vs. c36m), where the more important discrepancies appear between residues close to the N-terminus. However, the remaining comparisons exhibit contradictory behaviors between both methods, as the regions where the more relevant discrepancies appear differ. See, notably, comparisons on the bottom row. In Figure 2, only residues close to each other present big changes on their relative position, and no discrepancies are found in the interior region of the matrix. The opposite behavior is found in Figure S7. The fact that the distance matrix (6) ignores the uncertainty (intra-ensemble distances) might partially explain the encountered discrepancies between methods.

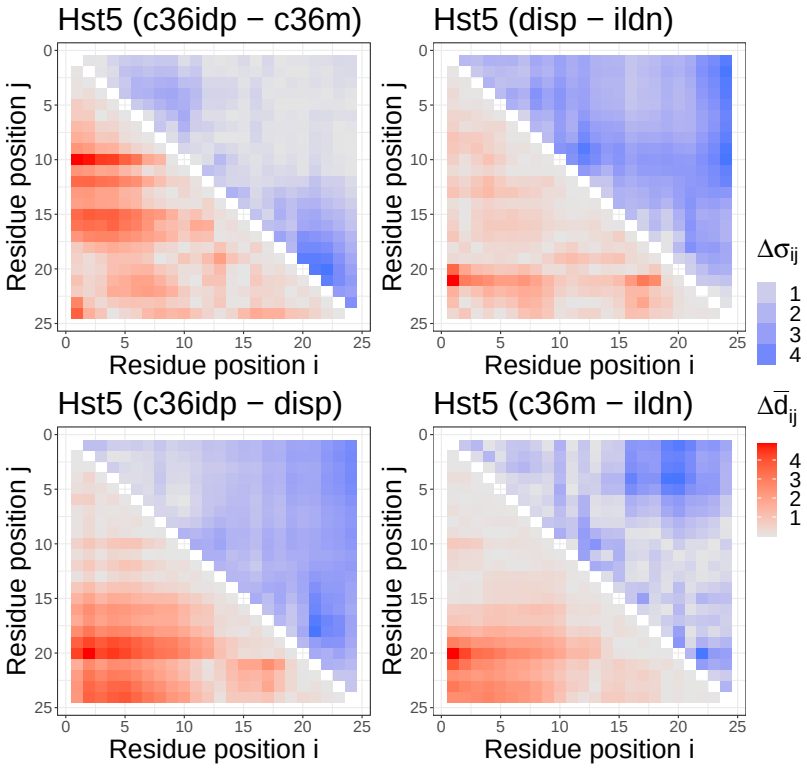


Figure S7: Comparison of Molecular Dynamics simulations of Hst5 ensemble using different force-fields, using the methodology described in [9]. The matrix entries correspond to the absolute differences defined in (6).

S3 Supplementary figures

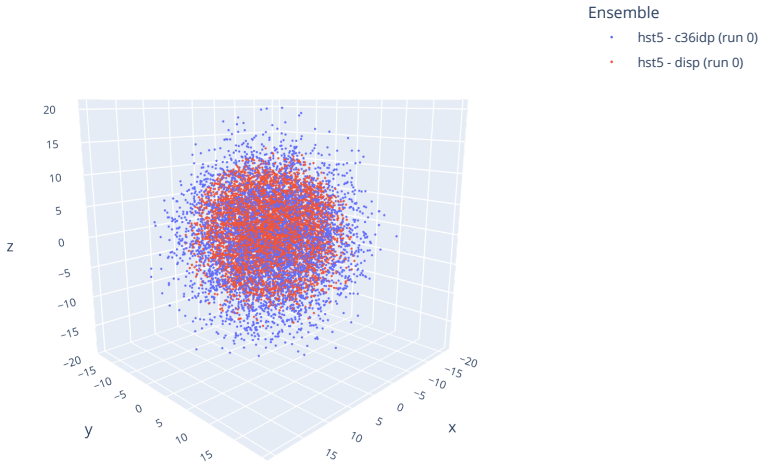
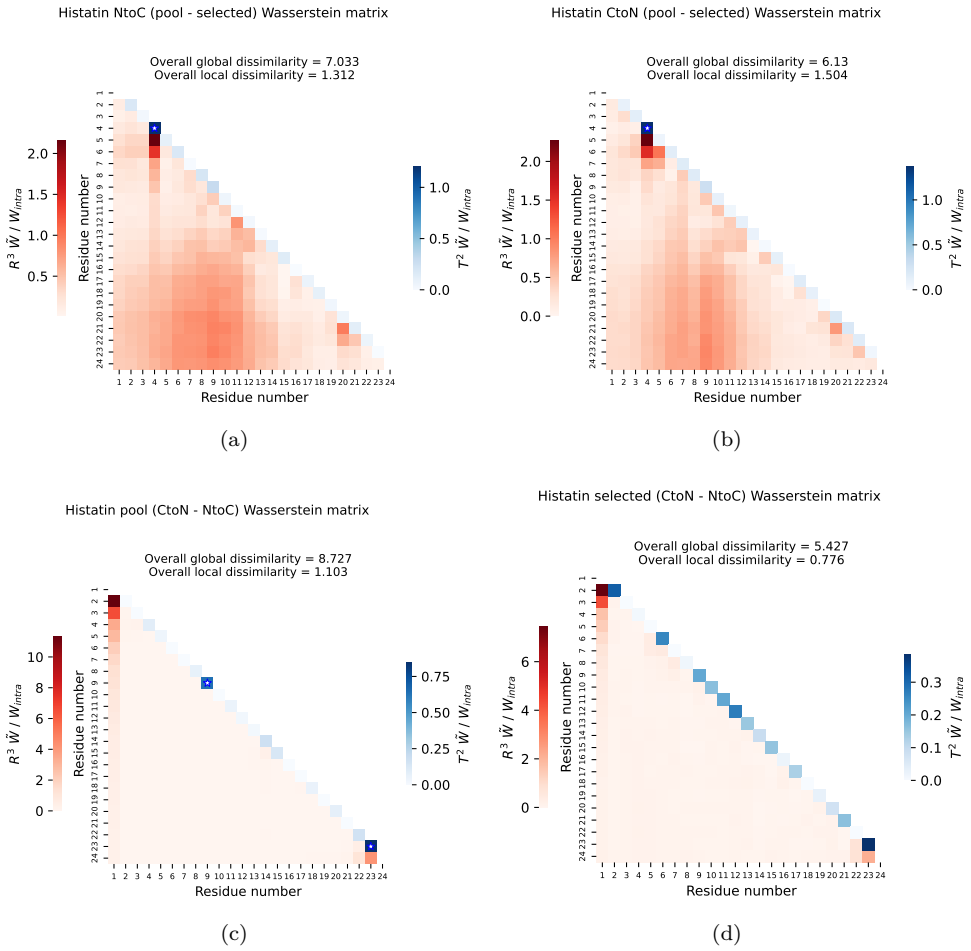


Figure S8: Two samples of $\overrightarrow{R}_{3,10}$ corresponding to a pair of ensembles of Hst5 simulated with force-fields CHARMM36IDPSFF (c36idp) and AMBER ff99SB-disp (disp). Each sample is represented by a point cloud in the three-dimensional euclidean space.



(a) (b) (c) (d)

Figure S9: comparison of Hst5 ensembles before and after filtering with experimental SAXS data. The ensemble was simulated from (a) N-to-C or from (b) C-to-N. (c) Comparison of Hst5 ensembles generated from N-to-C and C-to-N. (d) comparison of the N-to-C and C-to-N SAXS refined. In all matrices, The color scale $\tilde{W}/W_{\text{intra}}$ corresponds to the score (10), representing the relative difference between the inter-ensemble distances and the uncertainty. The coefficients in the lower triangle (in red) corresponds to the global differences. Coefficients along the diagonal (in blue) correspond to the local differences. Blue stars indicate that the corresponding local corrected distance is significantly different from zero (the associated p-value (6) is smaller than $\alpha = 0.05$).

References

[1] Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2018). SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res* **47**, D1 (11), D475–D481.

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

[3] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *J Mach Learn Res* **22**, 78, 1–8.

[4] Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., and Schmitzer, B. (2020). *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. R package version 0.12-2, https://cran.r-project.org/package=transport.

[5] 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.

[6] Hess, B. (2002). Convergence of sampling in protein simulations. *Phys Rev E* **65**, 3, 031910.

[7] 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.

[8] 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.

[9] 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.