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

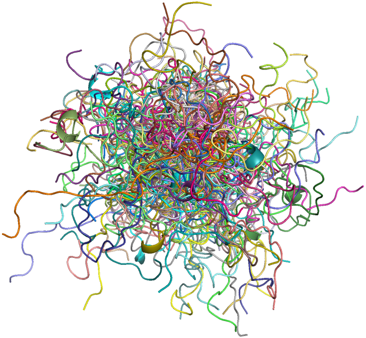
Javier González-Delgado[1,2], Amin Sagar[3], Christophe Zanon[2], Kresten Lindorff-Larsen[4], Pau Bernadó[3], Pierre Neuvial[1] and Juan Cortés[2]

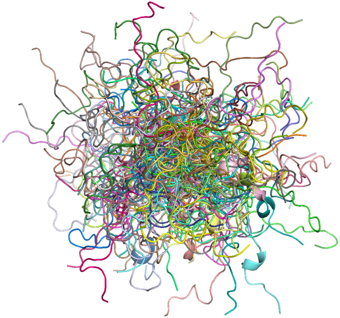
1. Institut de Mathématiques de Toulouse, 2. LAAS-CNRS,
3. Centre de Biologie Structurale, 4. The Linderstrøm-Lang Centre for Protein Science.

AlgoSB 2022: Intrinsic Disorder in Proteins

November 25, 2022

Goal: comparing a pair of IDP ensembles





State of the art

Comparison of proteins

For rigid proteins

- **Optimal rigid body superposition** (Rao and Rossmann, 1973). Minimization of Root-Mean-Square-Deviation (RMSD). Questioning the interpretation of RMSD as an absolute metric (Maiorov and Crippen, 1994).
- Extension to ensemble version (Brüschweiler, 2003).

State of the art

Comparison of proteins

For rigid proteins

- **Optimal rigid body superposition** (Rao and Rossmann, 1973). Minimization of Root-Mean-Square-Deviation (RMSD). Questioning the interpretation of RMSD as an absolute metric (Maiorov and Crippen, 1994).
- Extension to ensemble version (Brüschweiler, 2003).

For energy landscapes

- RSMD-based metric between ensembles of ordered systems (Lindorff-Larsen and Ferkinghoff-Borg, 2009).
- Graph-based representation of the conformational space based on a set of low-energy conformations. Comparison using Wasserstein distance (Cazals et al., 2015).

For disordered structures

- **Averaged conformational properties** over ensembles as informative descriptors of their functionality (e.g. pairwise distances (Lazar et al., 2020)).

In this work

- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.

In this work

- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.

In this work

- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.

In this work

- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.
- An **overall distance** between the pair of ensembles can be computed.

In this work

- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.
- An **overall distance** between the pair of ensembles can be computed.
- Non-parametric framework (no model assumptions).

In this work

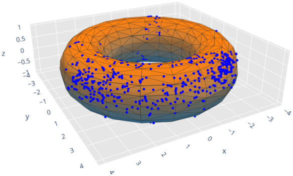
- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.
- An **overall distance** between the pair of ensembles can be computed.
- Non-parametric framework (no model assumptions).
- No intermediate/approximation steps (e.g. clustering, dimensionality reduction...).

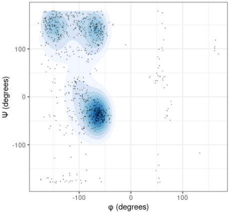
Conformational ensembles as a set of probability distributions

Local structure

Dihedral angles distributions

For the residue at the i-th position, with $i = 1, \ldots, L$, its dihedral angles (ϕ_i, ψ_i) follow a probability distribution $P_i^l \in \mathcal{P}(\mathbb{T}^2)$.



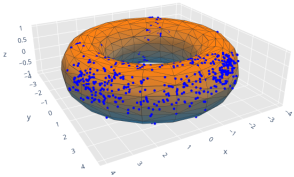


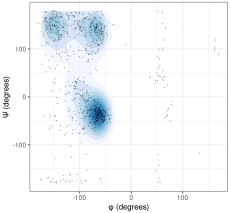
Conformational ensembles as a set of probability distributions

Local structure

Dihedral angles distributions

For the residue at the i-th position, with $i = 1, \ldots, L$, its dihedral angles (ϕ_i, ψ_i) follow a probability distribution $P_i^l \in \mathcal{P}(\mathbb{T}^2)$.





Local structure

We define the **local structure** of an ensemble as the L-tuple

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

Conformational ensembles as a set of probability distributions

Global structure

Conformational ensembles as a set of probability distributions

Global structure

Defining a global structure

- We use the **relative positions** of residues (invariant under rigid-body motions).

$$\left(\begin{array}{l}\text{We define the position of a given residue as the the position}\\ \text{of its } C_\beta \text{ atom when it exists and of its } C_\alpha \text{ atom otherwise.}\end{array}\right)$$

Conformational ensembles as a set of probability distributions

Global structure

Defining a global structure

- We use the **relative positions** of residues (invariant under rigid-body motions).

  (We define the position of a given residue as the the position of its C_β atom when it exists and of its C_α atom otherwise.)

  Idea: for every residue *i* along the sequence:

  1 Define a residue-specific reference frame at *i* for every conformation,

Conformational ensembles as a set of probability distributions

Global structure

Defining a global structure

- We use the **relative positions** of residues (invariant under rigid-body motions).

  (We define the position of a given residue as the the position of its C_β atom when it exists and of its C_α atom otherwise.)

  Idea: for every residue i along the sequence:

  1. Define a residue-specific reference frame at i for every conformation,
  2. Superimpose all reference frames ⇔ superimpose all the conformations,

Conformational ensembles as a set of probability distributions

Global structure

Defining a global structure

- We use the **relative positions** of residues (invariant under rigid-body motions).

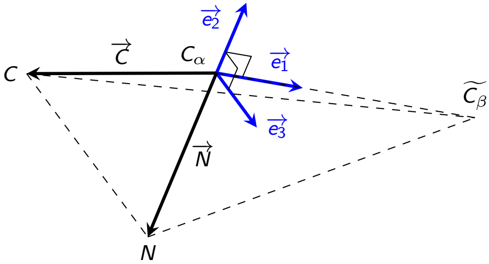
$$\left(\begin{array}{l}\text{We define the position of a given residue as the the position} \\ \text{of its } C_\beta \text{ atom when it exists and of its } C_\alpha \text{ atom otherwise.}\end{array}\right)$$

Idea: for every residue i along the sequence:

1. Define a residue-specific reference frame at i for every conformation,
2. Superimpose all reference frames $\Leftrightarrow$ superimpose all the conformations,
3. Access to the distribution of the relative position of any other residue $j \neq i$ with respect to i (point cloud in $\mathbb{R}^3$).

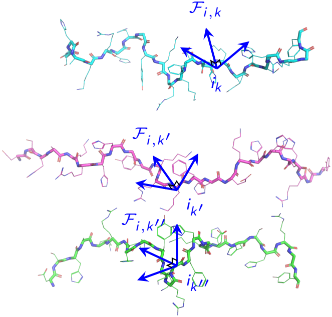
Global structure

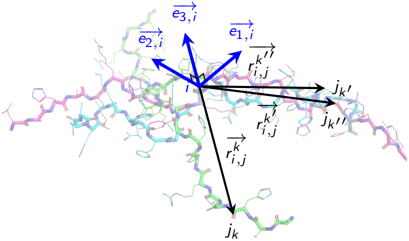
Reference frame overview



Global structure

Superposition of all the conformations



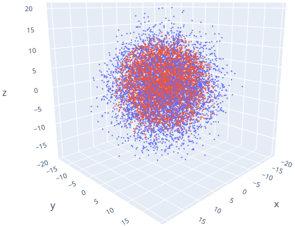

$\overrightarrow{e_{3,i}}$
$\overrightarrow{e_{2,i}}$
$\overrightarrow{e_{1,i}}$
$\overrightarrow{r^{k''}_{i,j}}$
i
$j_{k'}$
$j_{k''}$
$\overrightarrow{r^{k'}_{i,j}}$
$\overrightarrow{r^{k}_{i,j}}$
j_k

Conformational ensembles as a set of probability distributions

Global structure

Relative position distributions are point clouds in $\mathbb{R}^3$

For each pair of residues $i \neq j$, we denote as $P^g_{i,j}$ the probability distribution of their relative positions, which is supported on $\mathbb{R}^3$.

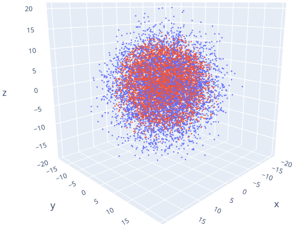


Conformational ensembles as a set of probability distributions

Global structure

Relative position distributions are point clouds in $\mathbb{R}^3$

For each pair of residues $i \neq j$, we denote as $P^g_{i,j}$ the probability distribution of their relative positions, which is supported on $\mathbb{R}^3$.



Global structure

We define the **global structure** of an ensemble as the $L(L-1)/2$-tuple

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

Distance between local/global structures

Desired properties in a metric

1. Satisfying the **mathematical properties that define a distance** (being 0 if an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Respecting (or, even better, integrating) the **geometry of the underlying space**.

Distance between local/global structures

Desired properties in a metric

1. Satisfying the **mathematical properties that define a distance** (being 0 if an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Respecting (or, even better, integrating) the **geometry of the underlying space**.

In the litterature...

- Hellinger distance to compare (ϕ, ψ) distributions (Ting et al., 2019). Ignores the geometry of the ground space (its periodicity).

Distance between local/global structures

Desired properties in a metric

1. Satisfying the **mathematical properties that define a distance** (being 0 if an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Respecting (or, even better, integrating) the **geometry of the underlying space**.

In the litterature...

- Hellinger distance to compare (ϕ, ψ) distributions (Ting et al., 2019). Ignores the geometry of the ground space (its periodicity).
- Symmetrized Kullback-Leibler (KL) divergence to compare ensembles of ordered systems (Lindorff-Larsen and Ferkinghoff-Borg, 2009). Misses the geometrical reliability, does not satisfy triangle inequality.

Distance between local/global structures

Desired properties in a metric

1. Satisfying the **mathematical properties that define a distance** (being 0 if an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Respecting (or, even better, integrating) the **geometry of the underlying space**.

In the litterature...

- Hellinger distance to compare (ϕ, ψ) distributions (Ting et al., 2019). Ignores the geometry of the ground space (its periodicity).
- Symmetrized Kullback-Leibler (KL) divergence to compare ensembles of ordered systems (Lindorff-Larsen and Ferkinghoff-Borg, 2009). Misses the geometrical reliability, does not satisfy triangle inequality.

Here: Wasserstein distance

- Satisfies 1 and 2,

Distance between local/global structures

Desired properties in a metric

1. Satisfying the **mathematical properties that define a distance** (being 0 if an only if the two compared distributions are identical, symmetry and triangle inequality),
2. Respecting (or, even better, integrating) the **geometry of the underlying space**.

In the litterature...

- Hellinger distance to compare (ϕ, ψ) distributions (Ting et al., 2019). Ignores the geometry of the ground space (its periodicity).
- Symmetrized Kullback-Leibler (KL) divergence to compare ensembles of ordered systems (Lindorff-Larsen and Ferkinghoff-Borg, 2009). Misses the geometrical reliability, does not satisfy triangle inequality.

Here: Wasserstein distance

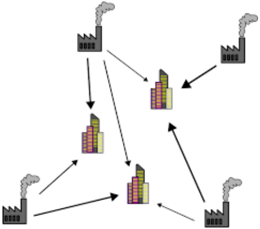
- Satisfies 1 and 2,
- Physical interpretation: minimum transportation cost needed to reconfigure the mass of one probability distribution to recover the other.

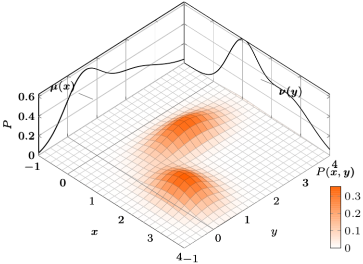
Distance between local/global structures

Wasserstein distance

Optimal Transport between two probability measures (Monge 1781, Kantorovich 1939)

Optimal way (in terms of transportation cost) to redistribute the mass of one probability distribution to recover the other.



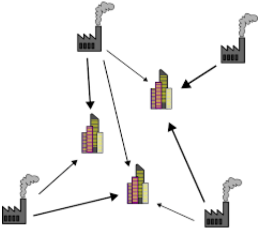


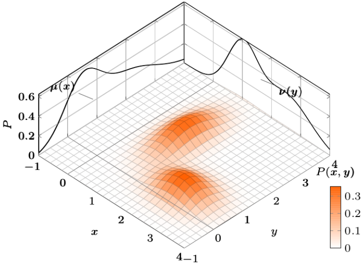
Distance between local/global structures

Wasserstein distance

Optimal Transport between two probability measures (Monge 1781, Kantorovich 1939)

Optimal way (in terms of transportation cost) to redistribute the mass of one probability distribution to recover the other.





p-Wasserstein distance between two arbitrary measures

$$\mathcal{W}_p^p(\mu,\nu)=\min_{\pi\in\mathcal{U}(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)^p \mathrm{d}\pi(x,y)=\min_{(X,Y)}\left\{\mathbb{E}_{(X,Y)}(c(X,Y)^p)\,:\,X\sim\mu\ \ Y\sim\nu\right\}.$$

The comparison tool

Definition and representation

Consider two ensembles A, B, associated to two sequences of equal length L.

Difference between local structures

We define the **difference between local structures** of A and B as the L-tuple of Wasserstein distances

$$(\mathcal{W}_1^{l,A,B}, \ldots, \mathcal{W}_L^{l,A,B}) = \left(\mathcal{W}(P_1^{l,A}, P_1^{l,B}), \ldots, \mathcal{W}(P_L^{l,A}, P_L^{l,B}) \right),$$

where $P_i^{l,A}$ (resp. $P_i^{l,B}$) denotes the i-th distribution of the local structure of ensemble A (resp. B).

The comparison tool

Definition and representation

Consider two ensembles A, B, associated to two sequences of equal length L.

Difference between local structures

We define the **difference between local structures** of A and B as the L-tuple of Wasserstein distances

$$(\mathcal{W}_1^{l,A,B}, \ldots, \mathcal{W}_L^{l,A,B}) = \left(\mathcal{W}(P_1^{l,A}, P_1^{l,B}), \ldots, \mathcal{W}(P_L^{l,A}, P_L^{l,B})\right),$$

where $P_i^{l,A}$ (resp. $P_i^{l,B}$) denotes the i-th distribution of the local structure of ensemble A (resp. B).

Difference between local structures: significance

To each $W_i^{l,A,B}$ we can associate a p-value, accounting for the statistical significance of the distance ($\sim$ the plausibility of the *true* distance to be equal to zero).

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

The comparison tool

Definition and representation

Consider two ensembles A, B, associated to two sequences of equal length L.

Difference between local structures

We define the **difference between local structures** of A and B as the L-tuple of Wasserstein distances

$$(\mathcal{W}_1^{l,A,B}, \ldots, \mathcal{W}_L^{l,A,B}) = \left(\mathcal{W}(P_1^{l,A}, P_1^{l,B}), \ldots, \mathcal{W}(P_L^{l,A}, P_L^{l,B})\right),$$

where $P_i^{l,A}$ (resp. $P_i^{l,B}$) denotes the i-th distribution of the local structure of ensemble A (resp. B).

Difference between global structures

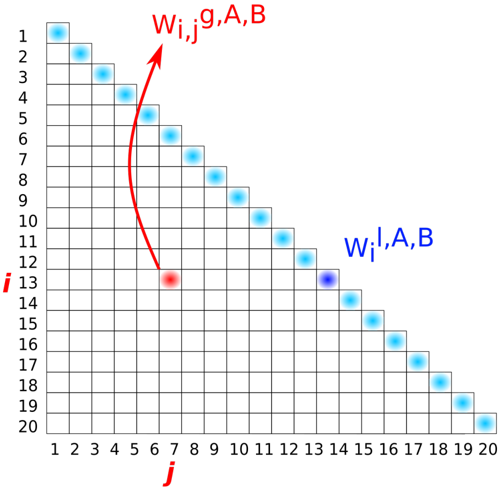
We define the **difference between global structures** of A and B as the $L(L-1)/2$-tuple

$$(\mathcal{W}_{1,2}^{g,A,B}, \ldots, \mathcal{W}_{L-1,L}^{g,A,B}) = \left(\mathcal{W}(P_{1,2}^{g,A}, P_{1,2}^{g,B}), \ldots, \mathcal{W}(P_{L-1,L}^{g,A}, P_{L-1,L}^{g,B})\right),$$

where $P_{i,j}^{g,A}$ (resp. $P_{i,j}^{g,B}$) denotes the i,j distribution of the global structure of ensemble A (resp. B).

The comparison tool

Matrix representation



The comparison tool

Account for uncertainty

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 comparison tool

Account for uncertainty

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

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

where each corrected distance, for each $i = 1, \ldots, L$, is defined as

$$\widetilde{\mathcal{W}}_i^{l,A,B} = \left(\underbrace{\boxed{\frac{1}{n_l} \sum_{s=1}^{n_l} \mathcal{W}_i^{l,A_s,B_s}}}_{\text{Inter-ensemble } (\mathcal{W}_{\text{inter}}^{l,A,B})} - \underbrace{\boxed{\frac{1}{2(n_l - 1)} \sum_{s=2}^{n_l} \left(\mathcal{W}_i^{l,A_1,A_s} + \mathcal{W}_i^{l,B_1,B_s} \right)}}_{\text{Intra-ensemble } (\mathcal{W}_{\text{intra}}^{l,A,B})} \right)_+$$

where, for any real number x, $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise.

The comparison tool

Account for uncertainty

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

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

where each corrected distance, for each $i = 1, \ldots, L$, is defined as

$$\widetilde{\mathcal{W}}_i^{l,A,B} = \left(\underset{\text{Inter-ensemble } (\mathcal{W}^{l,A,B}_{\text{inter}})}{\boxed{\frac{1}{n_l} \sum_{s=1}^{n_l} \mathcal{W}_i^{l,A_s,B_s}}} - \underset{\text{Intra-ensemble } (\mathcal{W}^{l,A,B}_{\text{intra}})}{\boxed{\frac{1}{2(n_l-1)} \sum_{s=2}^{n_l} \left(\mathcal{W}_i^{l,A_1,A_s} + \mathcal{W}_i^{l,B_1,B_s} \right)}} \right)_+$$

where, for any real number x, $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise.

- Noise reduction coming from uncertainty,
- Stand out residue-specific differences in the matrix representation.

The comparison tool

An interpretable scale

Definition of a continuous informative scale

Use the noise or **uncertainty as a reference** to which compare the inter-ensemble distances, reflecting in which proportion they exceed the “default” intra-ensemble ones.

The comparison tool

An interpretable scale

Definition of a continuous informative scale

Use the noise or **uncertainty as a reference** to which compare the inter-ensemble distances, reflecting in which proportion they exceed the “default” intra-ensemble ones.

The score

$$\frac{\widetilde{\mathcal{W}}_i^{l,A,B}}{\mathcal{W}_{\mathrm{intra}}^{l,A,B}} = \frac{\mathcal{W}_{\mathrm{inter}}^{l,A,B} - \mathcal{W}_{\mathrm{intra}}^{l,A,B}}{\mathcal{W}_{\mathrm{intra}}^{l,A,B}}$$

is the **proportion of the intra-ensemble difference** that represents the **corrected distance** between both structures (how big are inter-ensemble distances when compared to intra-ensemble ones).

The comparison tool

Overall distance between a pair of ensembles

Remark

If $d_1, \ldots, d_L$ are L distances defined on L metric spaces $\mathcal{X}_1, \ldots, \mathcal{X}_L$, the function $\sqrt{d_1^2 + \cdots + d_L^2}$ is a distance on the product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

The comparison tool

Overall distance between a pair of ensembles

Remark

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 $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

Overall local discrepancy

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

The comparison tool

Overall distance between a pair of ensembles

Remark

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 $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

Overall local discrepancy

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

Overall global discrepancy

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

where $w_{ij} = w(|i-j|)$ is an increasing function of $|i-j|$.

The comparison tool

Overall distance between a pair of ensembles

Remark

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 $\mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

Overall local discrepancy

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

Overall global discrepancy

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

where $w_{ij} = w(|i-j|)$ is an increasing function of $|i-j|$.

Remark

If the corrected distances are used to define the overall discrepancies, triangle inequality is no longer satisfied.

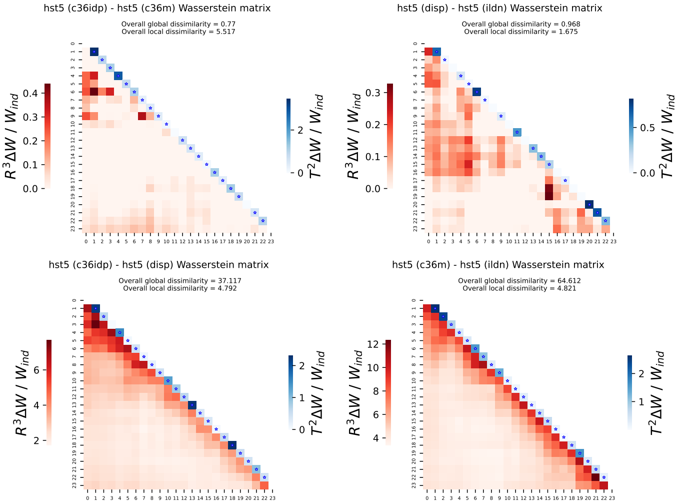
Results

Some applications of WASCO

- Comparisons of MD simulations using different force fields
- Effect of filtering based on SAXS experimental data
- Assessing the convergence of a MD simulation

Comparison of force fields

Results of MD simulations (Jephthah *et al.* 2021) for Hst5 using four different force-fields: AMBER ff99SB-disp (disp), AMBER ff99SB-ILDN (ildn), CHARMM36IDPSFF (c36idp), and CHARMM36m (c36m).



hst5 (c36idp) - hst5 (c36m) Wasserstein matrix

Overall global dissimilarity = 0.77
Overall local dissimilarity = 5.517

hst5 (disp) - hst5 (ildn) Wasserstein matrix

Overall global dissimilarity = 0.968
Overall local dissimilarity = 1.675

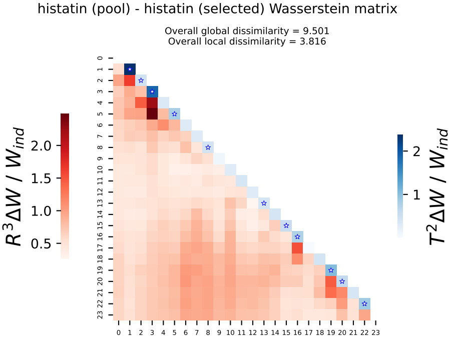
hst5 (c36idp) - hst5 (disp) Wasserstein matrix

Overall global dissimilarity = 37.117
Overall local dissimilarity = 4.792

hst5 (c36m) - hst5 (ildn) Wasserstein matrix

Overall global dissimilarity = 64.612
Overall local dissimilarity = 4.821

Histatin ensemble before and after filtering based on experimental SAXS data



histatin (pool) - histatin (selected) Wasserstein matrix

Overall global dissimilarity = 9.501
Overall local dissimilarity = 3.816

Using the overall distance to assess the convergence of a MD simulation

- Let T denote the current simulation time,
- Let $0 < t_1 < t_2 < \cdots < t_k = T$ be k time points.

If we denote A_t the conformational ensemble simulated at time t, we can compute the overall distances

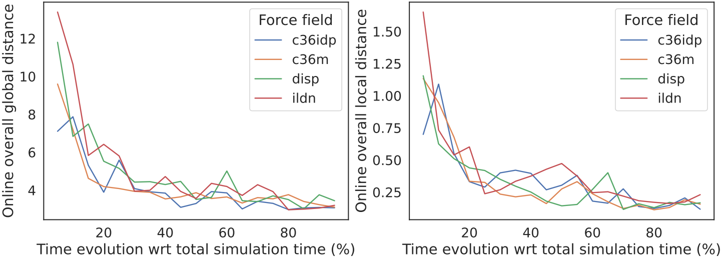
$$\mathcal{OW}_i^l = \mathcal{OW}^{l, A_{t_{i-1}}, A_{t_i}} \quad \text{for all } i = 2, \dots, k.$$

Analogously, we compute the overall global distances

$$\mathcal{OW}_i^g = \mathcal{OW}^{g, A_{t_{i-1}}, A_{t_i}} \quad \text{for all } i = 2, \dots, k.$$

Then, representing the $\mathcal{OW}_i^l$, $\mathcal{OW}_i^g$ with respect to the t_i will indicate whether the simulation has converged if the curve has “stabilized” (i.e. attained an asymptote at zero).

Convergence of a MD simulation (I)



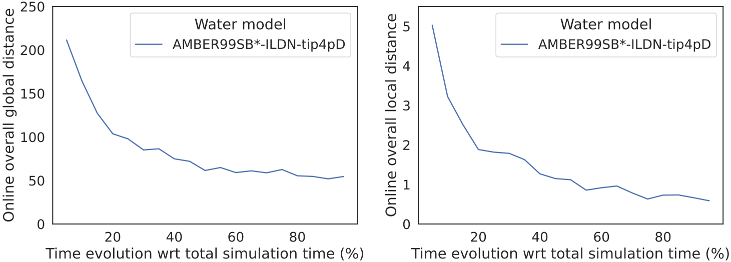
Online convergence analysis for PEP3 ensemble simulated with force-fields c36idp, c36m, disp and ildn.

Convergence ⇔ Asymptote at zero

Convergence of a MD simulation (II)

Another example: K-18 domain of Tau

Converging ensembles of IDPs of this length is very very hard...



Online convergence analysis for K-18 domain of Tau ensemble simulated with AMBER99SB*-ILDN-tip4pD water models.

No convergence ⇔ No asymptote at zero

Conclusions

- Novel approach to compare ensembles
- Specifically conceived for disordered systems (without a well-characterized energy landscape)
- Implemented in python, open source
- Drawback: computationally expensive for large systems (unfeasible if $L \gg 200$, $n_A, n_B \geq 10^5$)
- Future work: adapt WASCO to coarse-grain models and large ensembles