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

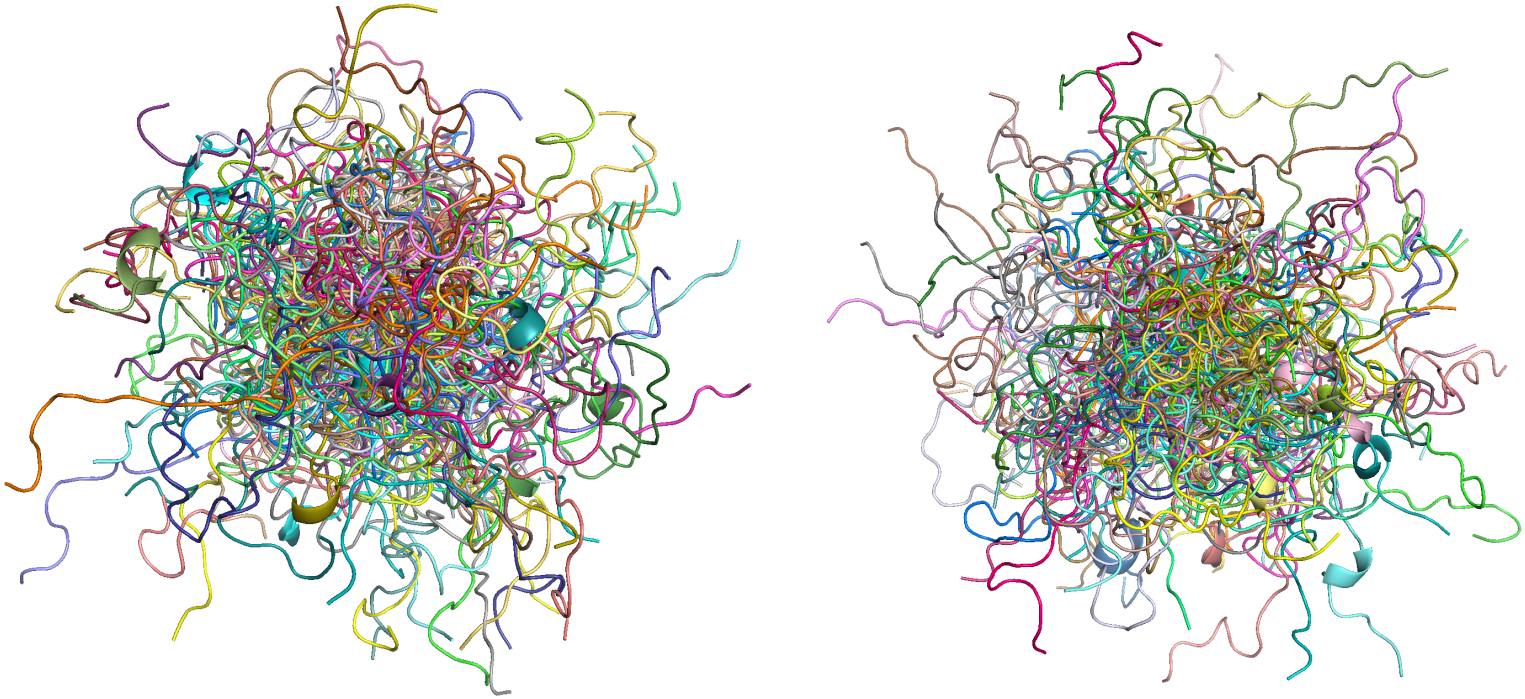
Javier González-Delgado[1,2], Amin Sagar[3], Christophe Zanon[2], Kresten Lindorff-Larsen[4], Pau Bernadó[3], Pierre Neuvial[1] and Juan Cortés[2]

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

GGMM 2023 - Young Modellers Conference

May 15-17, 2023

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).

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.
- 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.
- An **overall distance** between the pair of ensembles can be computed.

In this work

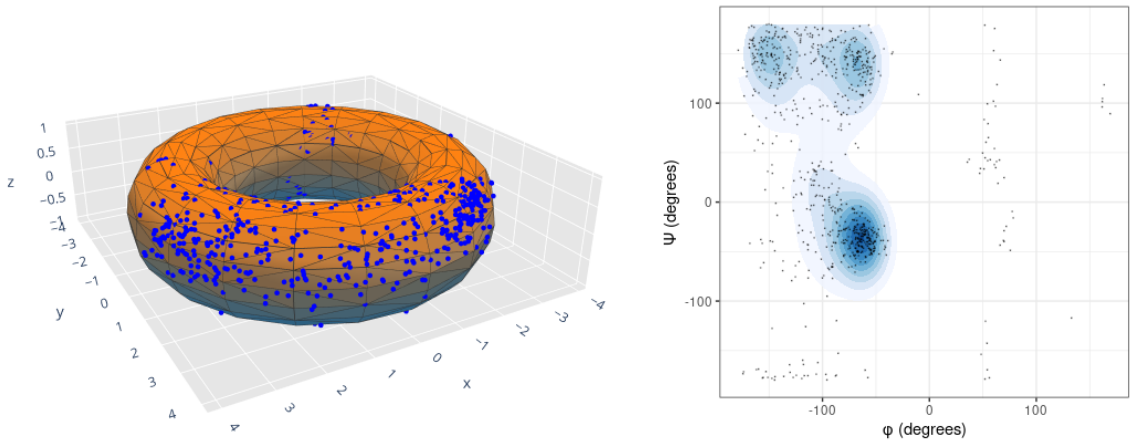
- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.
- An **overall distance** between the pair of ensembles can be computed.
- Non-parametric framework (no model assumptions).
- 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 $P_i^l \in \mathcal{P}(\mathbb{T}^2)$.



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

Defining a global structure

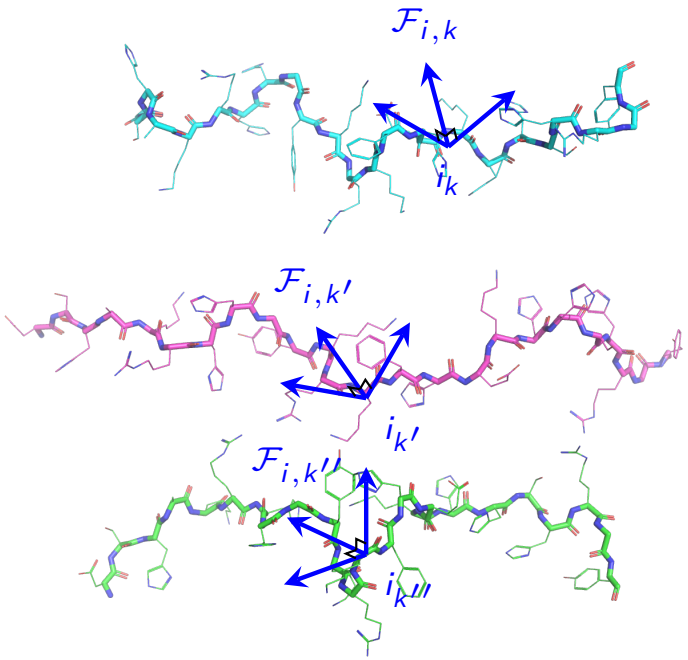
- We use the **relative positions** of residues (invariant under rigid-body motions).

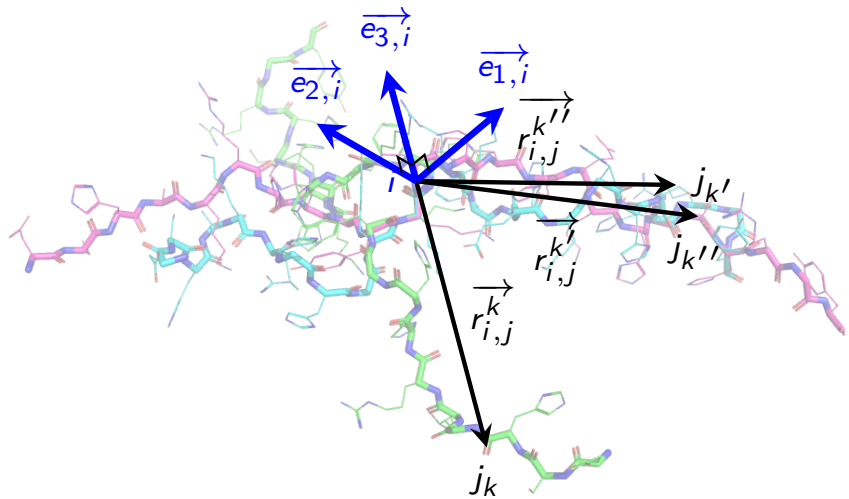
$$\left(\begin{array}{l} \text{We define the position of a given residue as the the position} \\ \text{of its } C_\beta \text{ atom when it exists and of its } C_\alpha \text{ atom otherwise.} \end{array} \right)$$

Global structure

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 $\mathbb{R}^3$).



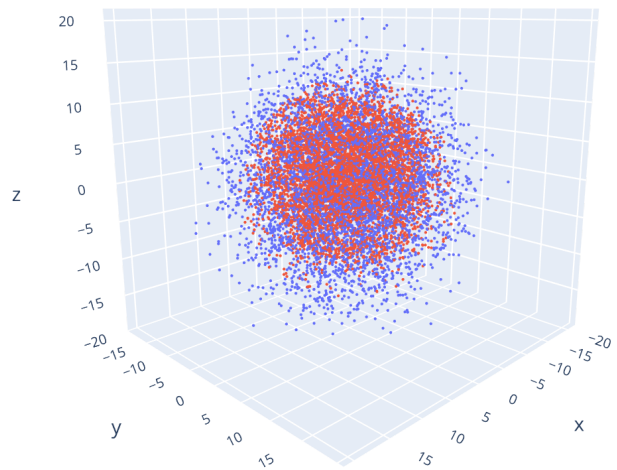


Conformational ensembles as a set of probability distributions

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

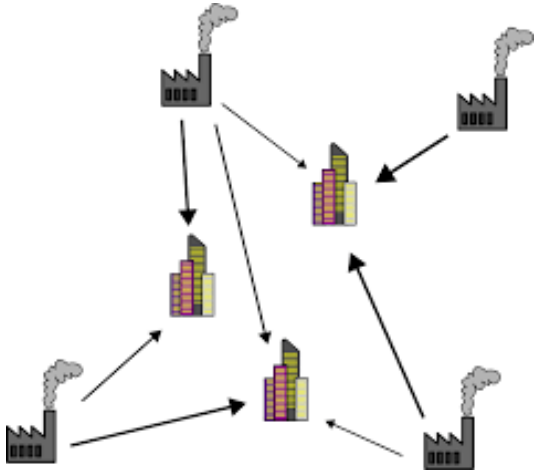
$$(P^g_{1,2}, P^g_{1,3}, \ldots, P^g_{L-1,L}), \quad P^g_{i,j} \in \mathcal{P}(\mathbb{R}^3) \quad \text{for all } i = 1, \ldots, L-1,\ j = i+1, \ldots, L.$$

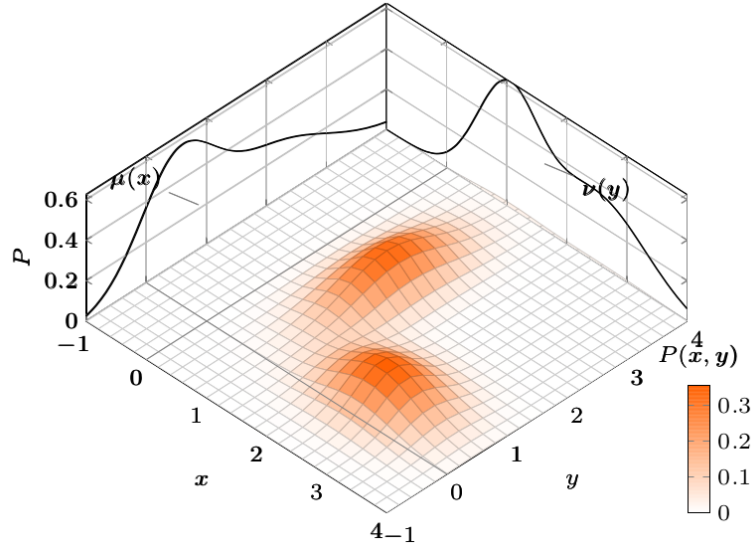
Distance between local/global structures

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

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

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

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

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

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 **corrected difference between local structures** of A and B is defined as the L-tuple

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

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

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

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.

The comparison tool

The Jupyter Notebook

https://gitlab.laas.fr/moma/WASCO

The ensemble is given as a folder per replica containing one .pdb file per conformation

```
histatin_filtered_path = "/".join([path_to_notebook,'Examples','histatin_filtered'])
histatin_pool_path = "/".join([path_to_notebook,'Examples','histatin_pool'])

comparison_tool(ensemble_1_path = histatin_filtered_path,
                ensemble_1_name = 'histatin_filtered',
                ensemble_2_path = histatin_pool_path,
                ensemble_2_name = 'histatin_pool',
                results_path = None,
                start_1 = None, end_1 = None,
                start_2 = None, end_2 = None)
```

The ensemble is given as one .xtc file per replica with one .pdb file containing topology information

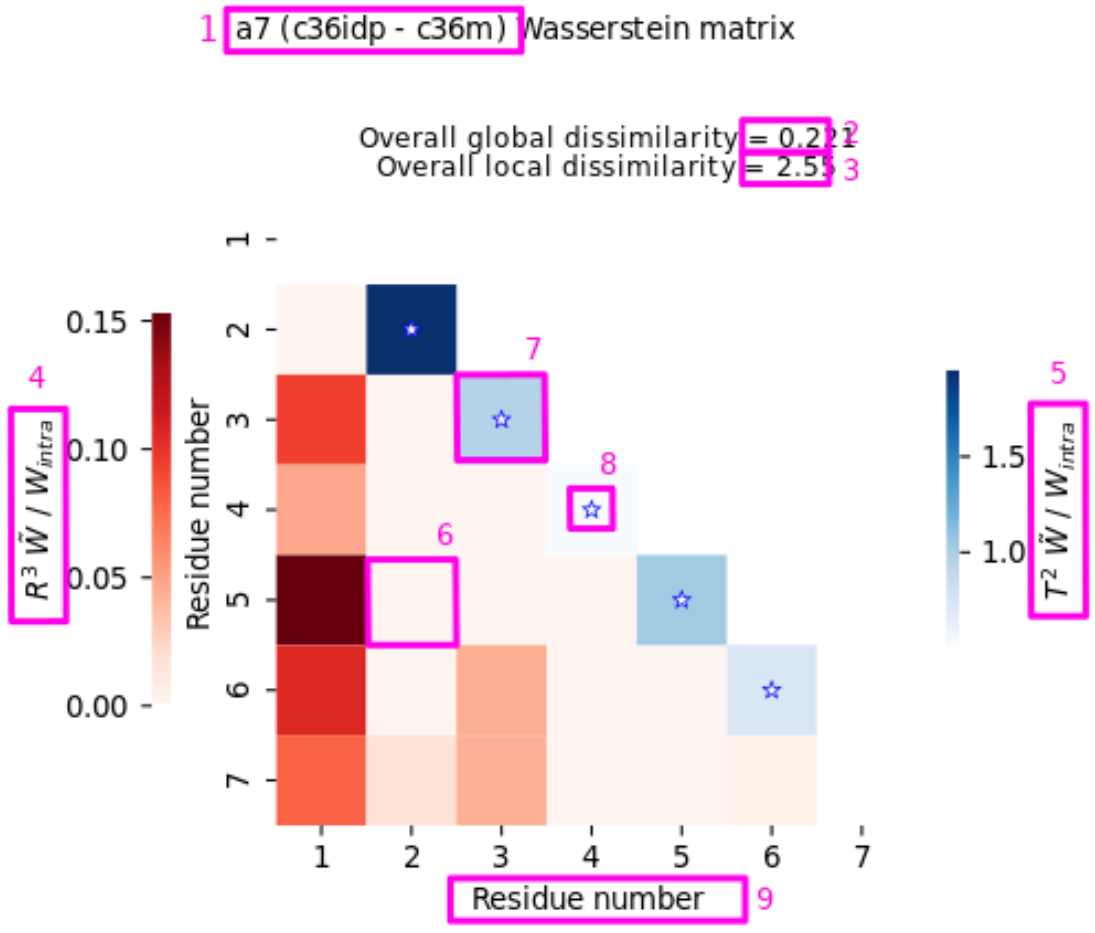
```
a7_c36idp_path = "/".join([path_to_notebook,'Examples','a7_c36idp'])
a7_c36m_path = "/".join([path_to_notebook,'Examples','a7_c36m'])

comparison_tool(ensemble_1_path = a7_c36idp_path,
                ensemble_1_name = 'a7_c36idp',
                ensemble_2_path = a7_c36m_path,
                ensemble_2_name = 'a7_c36m',
                results_path = None)
```

The comparison tool

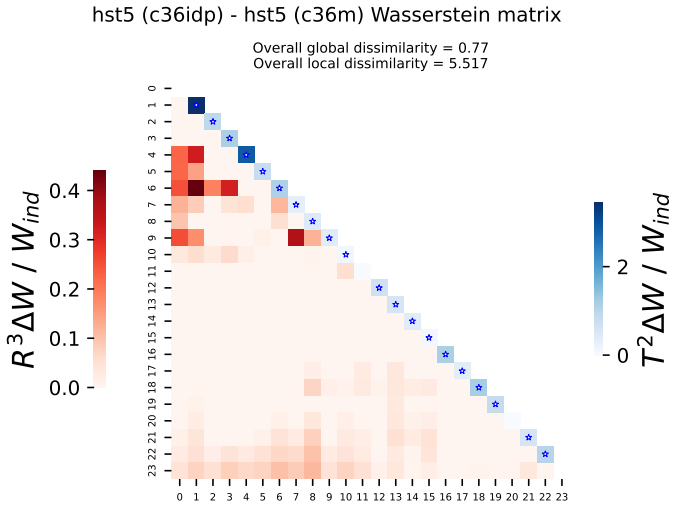
The Jupyter Notebook

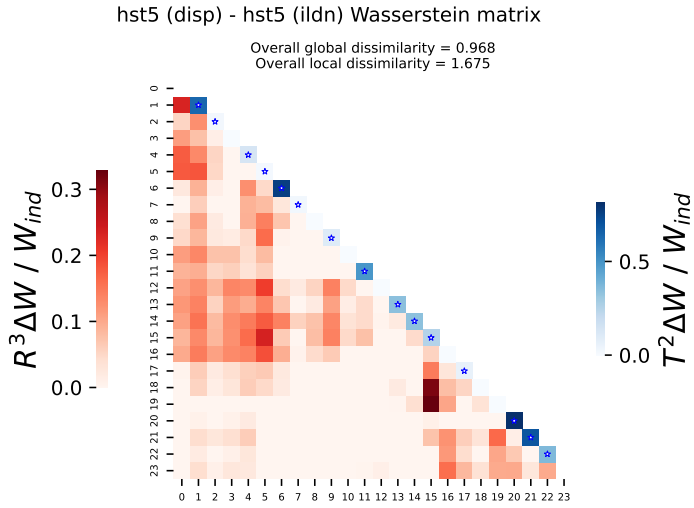
https://gitlab.laas.fr/moma/WASCO

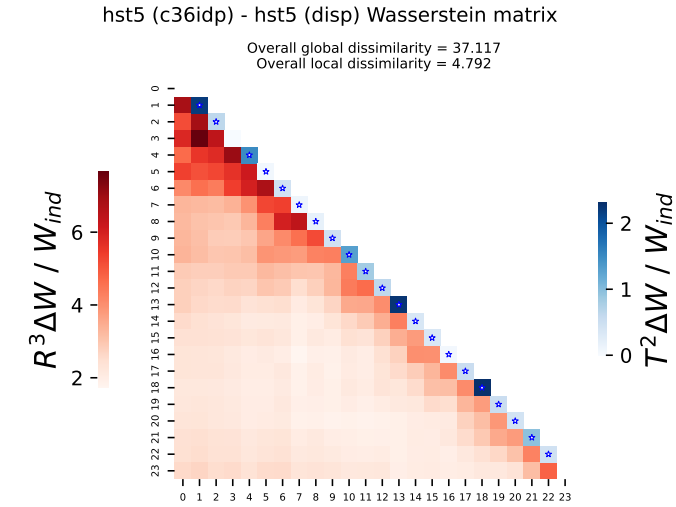


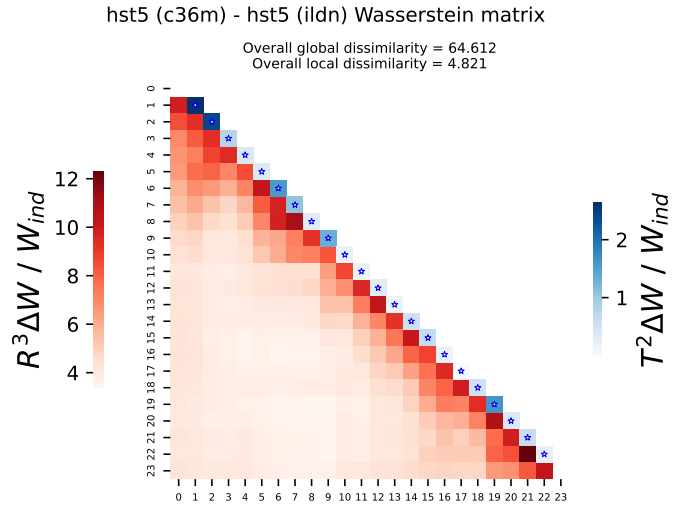
Comparison of force fields

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).

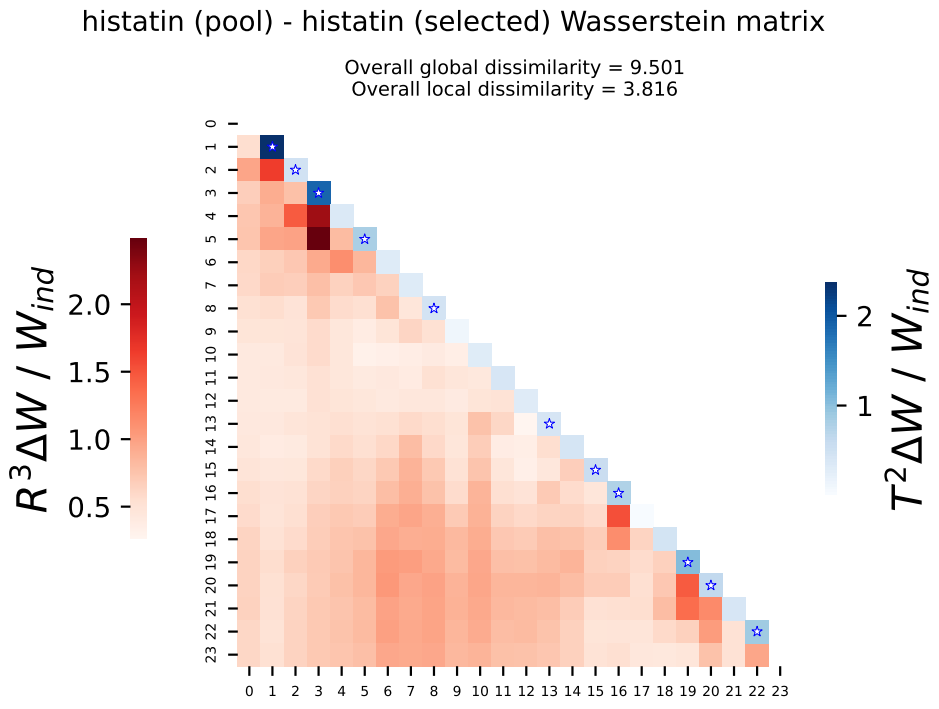








Histatin ensemble before and after filtering based on experimental SAXS data



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-grained models and large ensembles.

Paper: Javier González-Delgado et al. *WASCO: A Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins*, Journal of Molecular Biology, 2023.