Statistical methods for the investigation of highly-flexible proteins

Javier González-Delgado

Institut de Mathématiques de Toulouse, LAAS-CNRS

CBS seminar

May 11, 2023

Joint work with

Pierre Neuvial[1]

Juan Cortés[2]

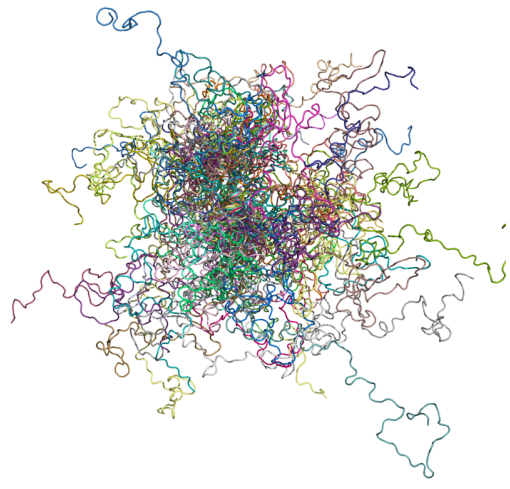
Pau Bernadó[3]

1. Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, Toulouse, France.
2. LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.
3. Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, France.

Understanding highly-flexible proteins

IDP : Amino acid sequence $\Longrightarrow$ Distribution of states.

$\overset{?}{\Longrightarrow}$



High variablity $\sim$ Probabilistic nature

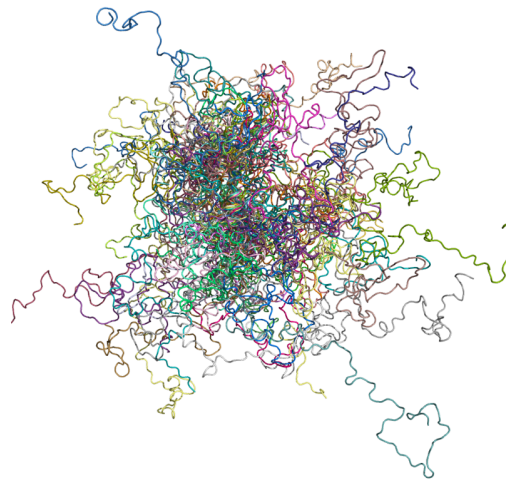
$\Downarrow$

Statistical interpretation

Understanding highly-flexible proteins

IDP : Amino acid sequence $\Longrightarrow$ Distribution of states.

$\overset{?}{\Longrightarrow}$



High variablity $\sim$ Probabilistic nature

$\Downarrow$

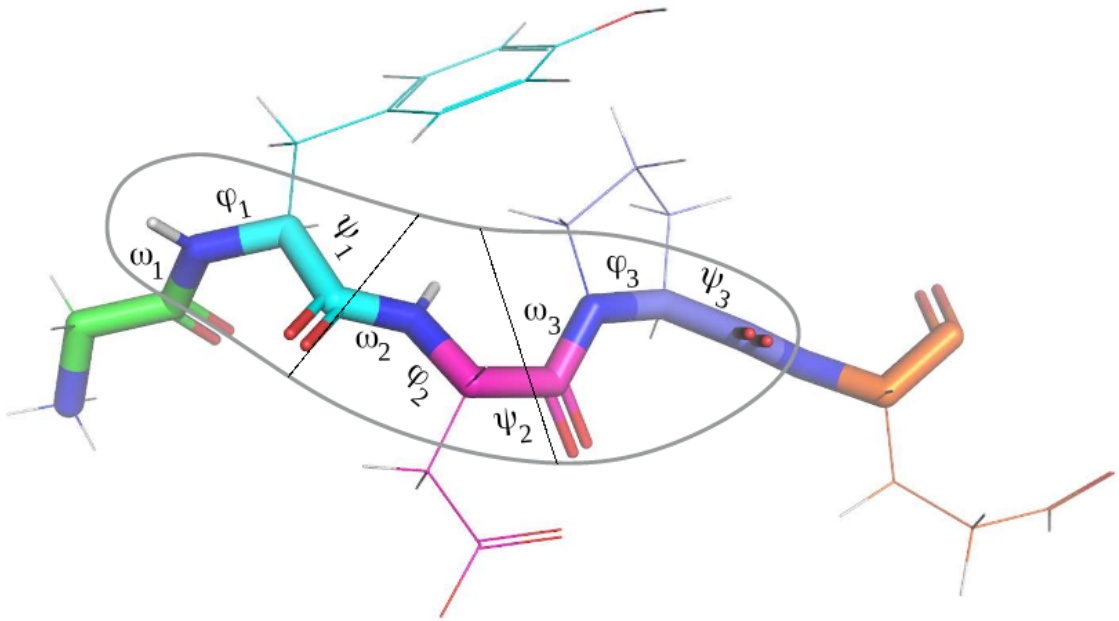
Statistical interpretation

Goal

Compare and **characterize** IDP ensembles at the **local** and **global** scale

Local comparison of IDP ensembles

Local protein conformation



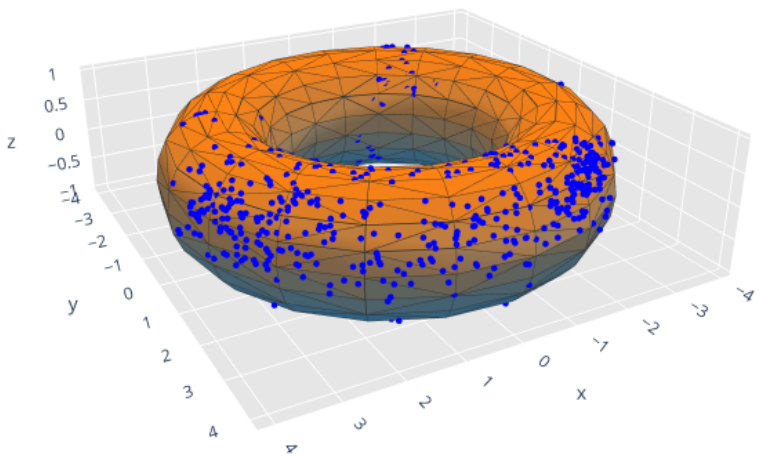
At each amino-acid, local conformation is given by the torsion dihedral angles[1]

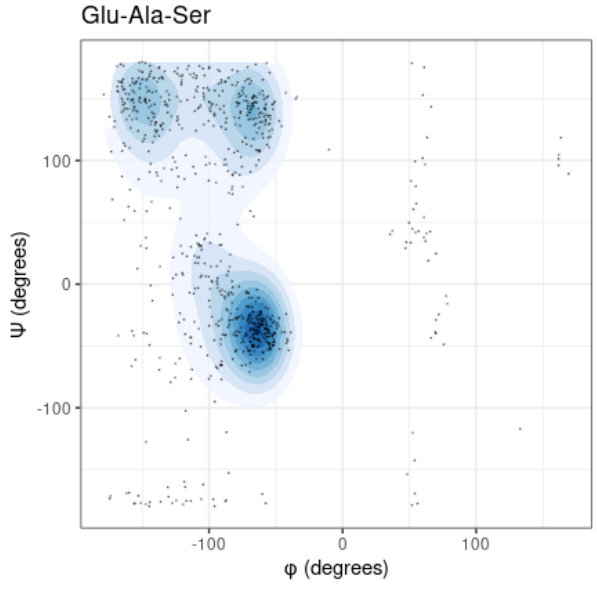
- $\omega_i \in \{0, \pi\}$ is fixed,
- $(\varphi_i, \psi_i) \in [-\pi, \pi] \times [-\pi, \pi]$ determine local conformation.

1. Ramachandran et al, 1963.

Local protein conformation

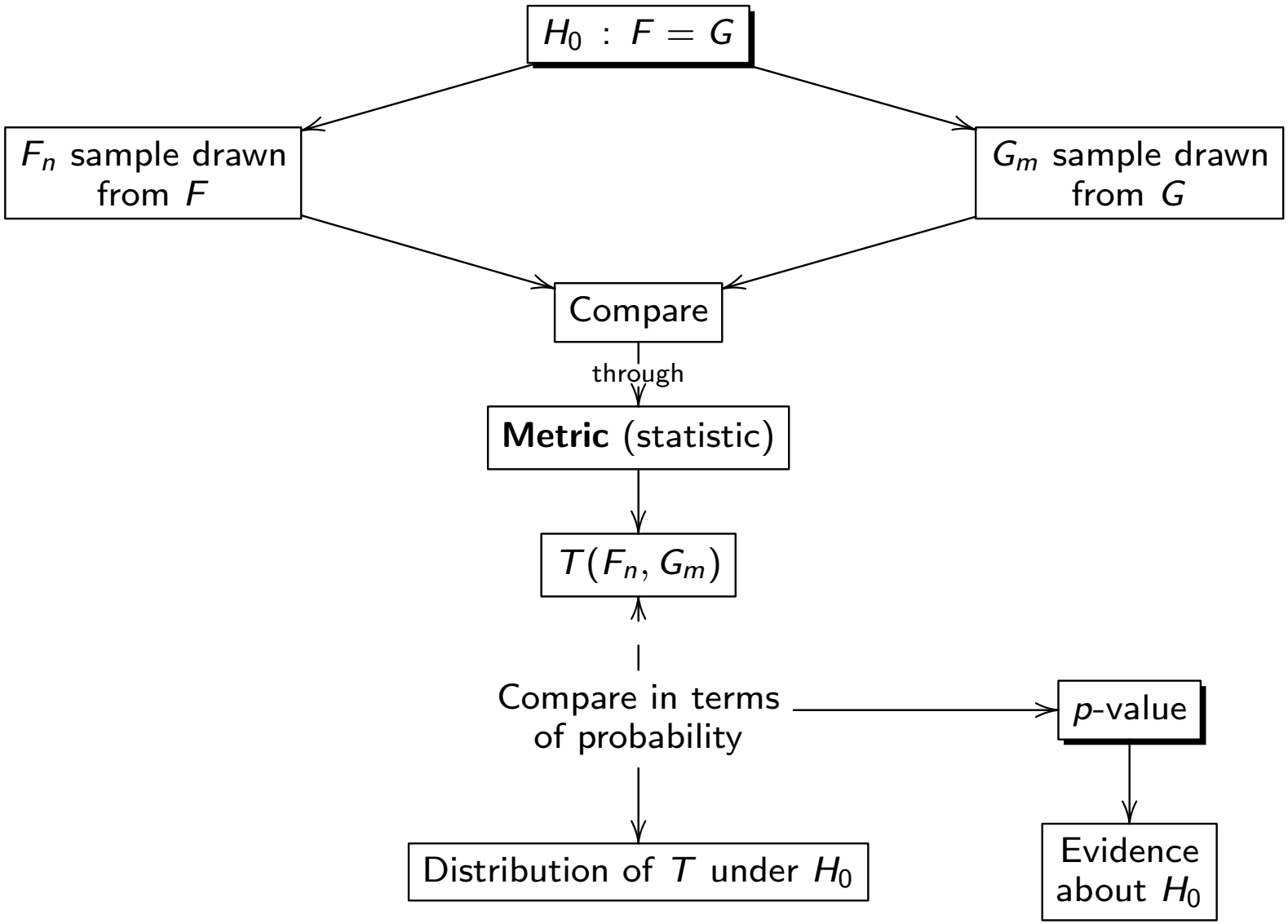
Local conformation is given by a **probability measure supported on** $\mathbb{T}^2$.





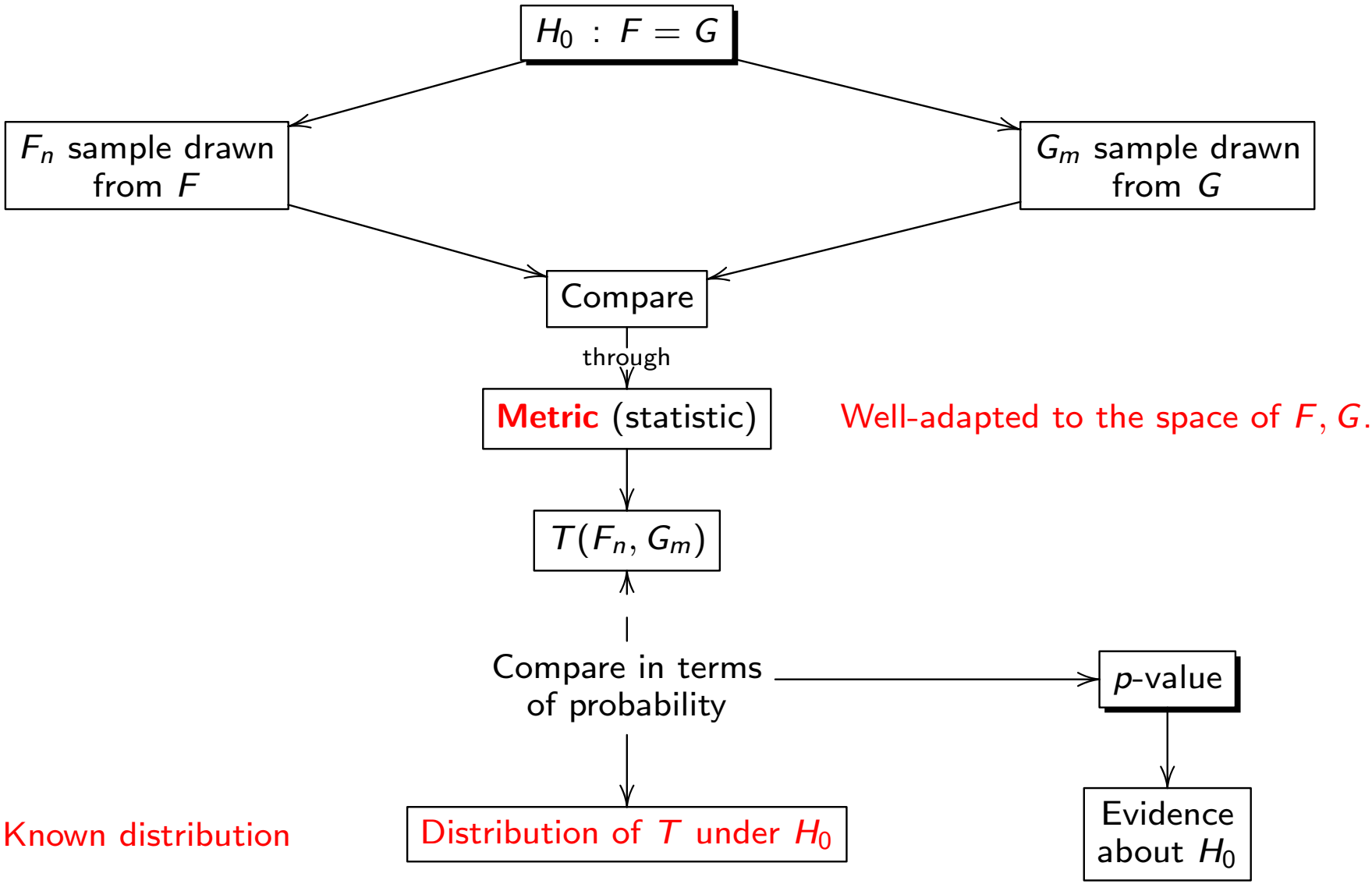
Two-sample goodness-of-fit test

Strategy



Two-sample goodness-of-fit test

Strategy

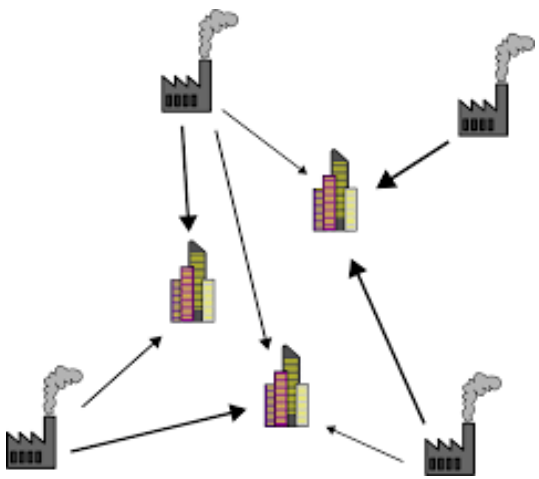


Optimal Transport Theory

From discrete to continuous

Optimal transport between discrete measures

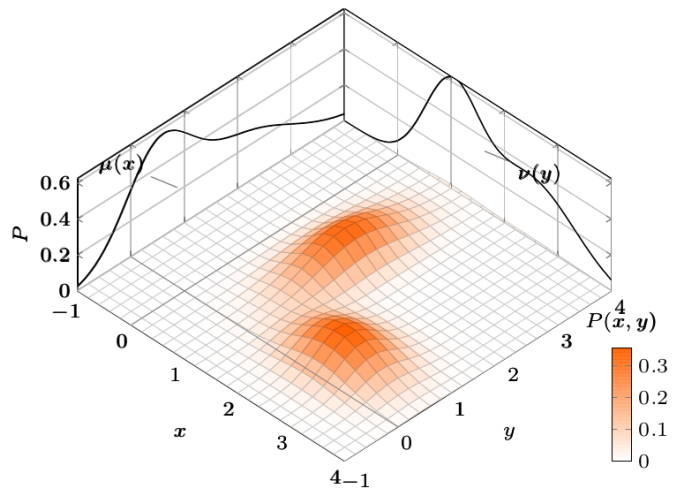
How to send all flavour produced in all factories to the all the bakeries in the city by **minimizing the total transport cost** ?



2. Image : Hugo Cisneros, 2021.

Optimal Transport Theory

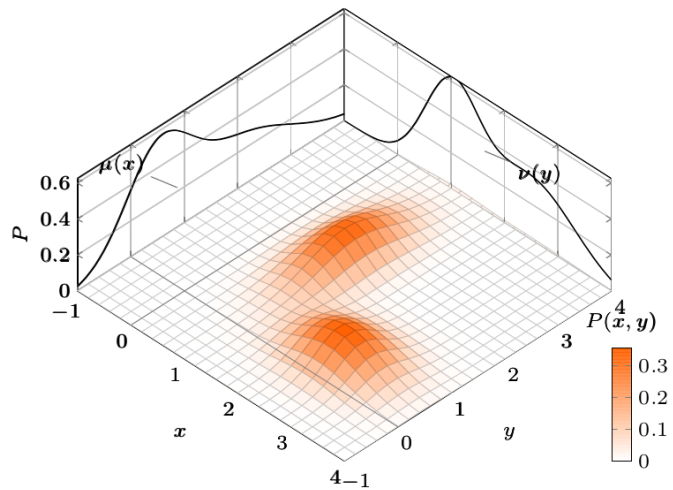
From discrete to continuous



3. Image : Marco Cuturi, 2019. On the several ways to regularize optimal transport. [Presentation]. Workshop 3 - CEB T1, 02/04/2019, Institut Henri Poincaré.

Optimal Transport Theory

From discrete to continuous



Kantorovich problem for arbitrary measures (1939)

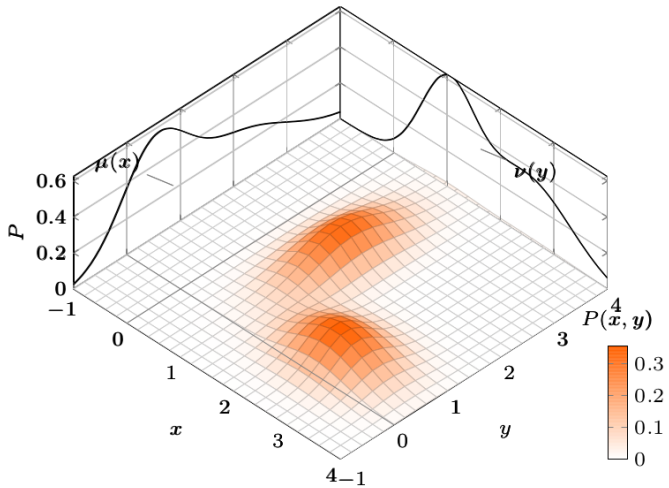
Let μ and ν be two arbitrary probability measures supported on two spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}^+$ a cost function.

$$\min_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y),$$

3. Image : Marco Cuturi, 2019. On the several ways to regularize optimal transport. [Presentation]. Workshop 3 - CEB T1, 02/04/2019, Institut Henri Poincaré.

Optimal Transport Theory

From discrete to continuous



Kantorovich problem for arbitrary measures (1939)

Let μ and ν be two arbitrary probability measures supported on two spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}^+$ a cost function.

$$\mathcal{W}_p(\mu, \nu) = \left(\min_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)^p \mathrm{d}\pi(x, y) \right)^{\frac{1}{p}}$$ that's a distance ![4]

3. Image : Marco Cuturi, 2019. On the several ways to regularize optimal transport. [Presentation]. Workshop 3 - CEB T1, 02/04/2019, Institut Henri Poincaré.

4. Villani, 2008.

Optimal Transport Theory

Wasserstein distance

p-Wasserstein distance between two arbitrary measures

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

Optimal Transport Theory

Wasserstein distance

p-Wasserstein distance between two arbitrary measures

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

Integrates the geometry of the underlying space !

We choose c as the **geodesic distance** on $\mathbb{T}^2$.

Optimal Transport Theory

Wasserstein distance

p-Wasserstein distance between two arbitrary measures

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

Integrates the geometry of the underlying space!

We choose c as the **geodesic distance** on $\mathbb{T}^2$.

Two-sample goodness-of-fit test for measures on the two-dimensional torus[5]

$H_0 : P = Q$ with P, Q distributions supported on $\mathbb{T}^2$, based on the Wasserstein distance between two "samples" drawn from P and Q.

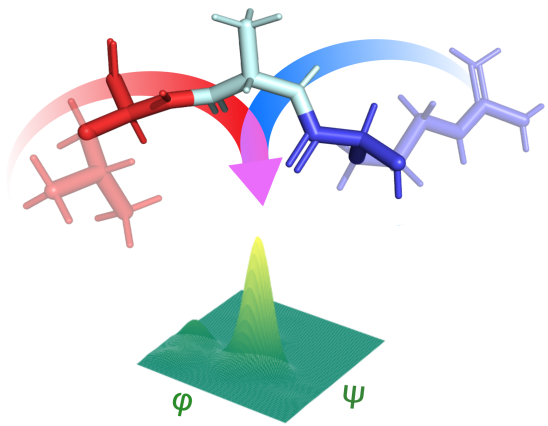
Implemented in the `R` package `torustest`.
https://github.com/gonzalez-delgado/torustest

5. GD, González-Sanz, Neuvial and Cortés, 2021.

Local comparison of IDP ensembles

Applications

Effect of neighboring amino-acids

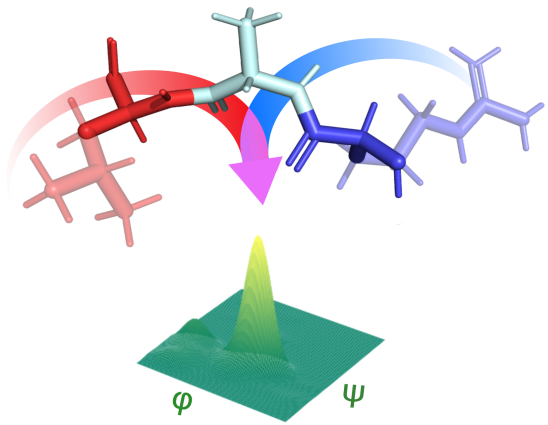


Reject of H_0 : Flory's Isolated Pair hypothesis (IPH)[6,7]

The identities of left and right amino-acids do not have an effect on (ϕ, ψ) distribution.

6. Flory et al., 1969.
7. GD, Bernadó, Neuvial and Cortés, 2022.

Effect of neighboring amino-acids



Reject of H_0 : Flory's Isolated Pair hypothesis (IPH)[6,7]

The identities of left and right amino-acids do not have an effect on (ϕ, ψ) distribution.

With a different testing procedure :

Reject of H_0 : The effects of neighboring amino-acids are independent[7]

Nearest neighbors have interdependent effects.

---

6. Flory et al., 1969.
7. GD, Bernadó, Neuvial and Cortés, 2022.

Multiple comparison of force-fields

with Matteo Paloni and Alessandro Barducci

(ϕ, ψ) simulations for GXG tripeptides

- X=G,A,F,K,E,S,L,
- amber19sb, amber99sb-disp, desamber, charmm36m force-fields,
- + the three "old" force-fields amber94, amber03, and charmm22,
- 3 time series per tripeptide and force-field.

All simulations also compared with

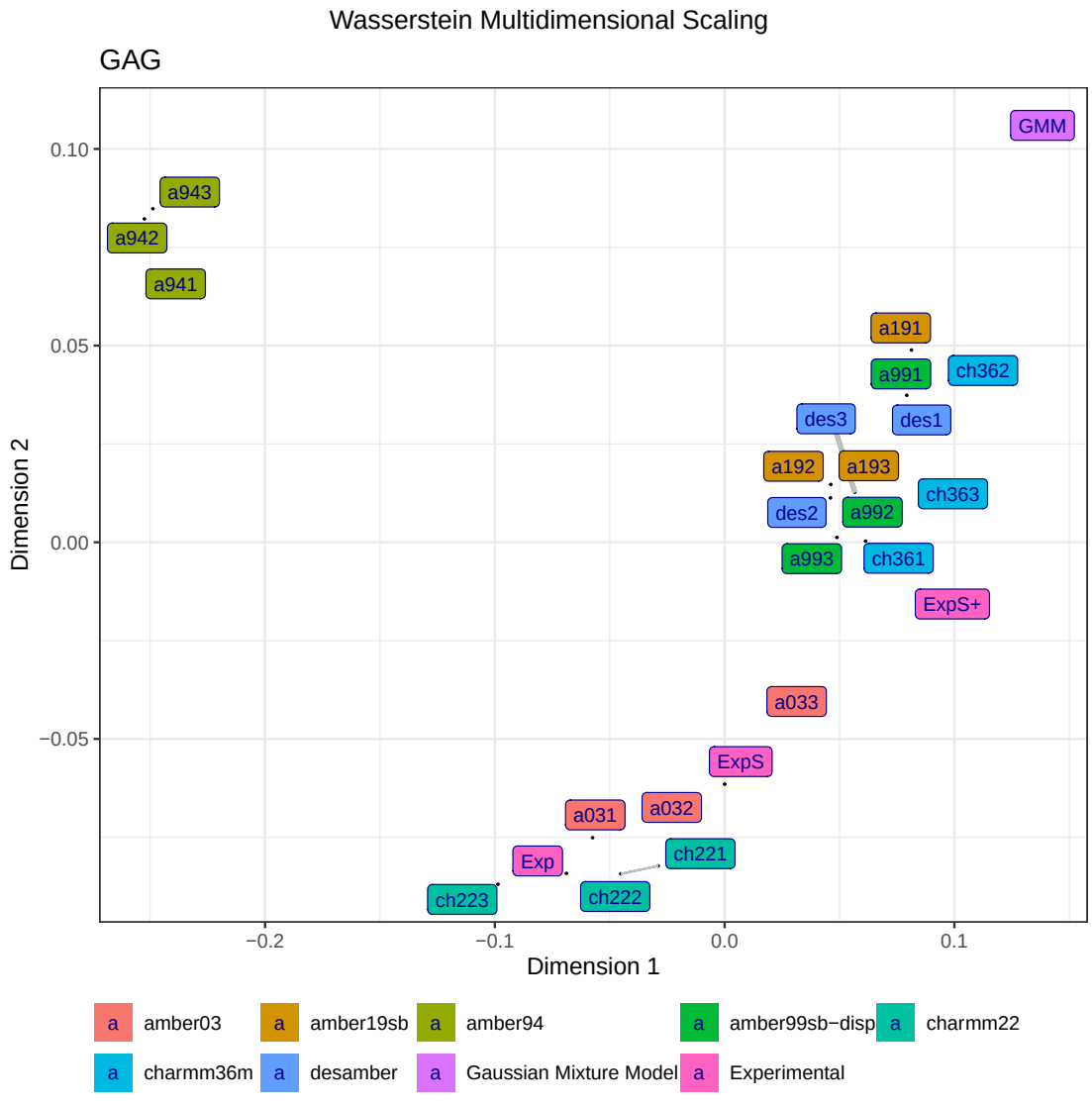
- Experimental data with different levels of solvent exposure,
- The Gaussian Mixture Model (GMM) introduced by Schweitzer-Stenner et al.

Methodology

1. Wasserstein distance + p-value between every pair of (ϕ, ψ) distributions,
2. Visual representation in 2-dimensions after multidimensional scaling.

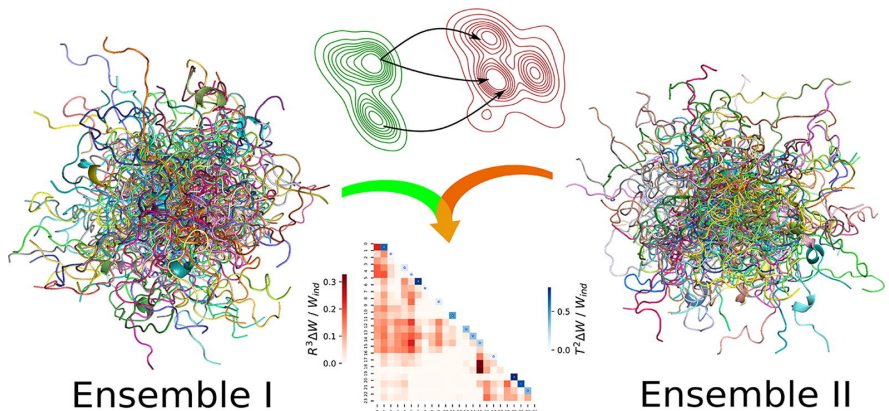
Multiple comparison of force-fields

Example



Global comparison of IDP ensembles

WASCO[8]



8. GD, Lindorff-Larsen, Bernadó, Neuvial, Cortés et al. 2023.

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

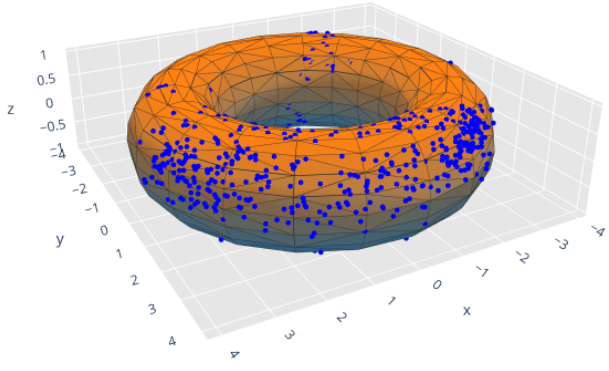
- We define the structure of an ensemble as a **set of probability distributions**, capturing its entire variability.
- The structures are compared using a **metric** that **integrates the geometry** of the conformational space.
- Allows residue-specific detection of global and local differences.
- An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

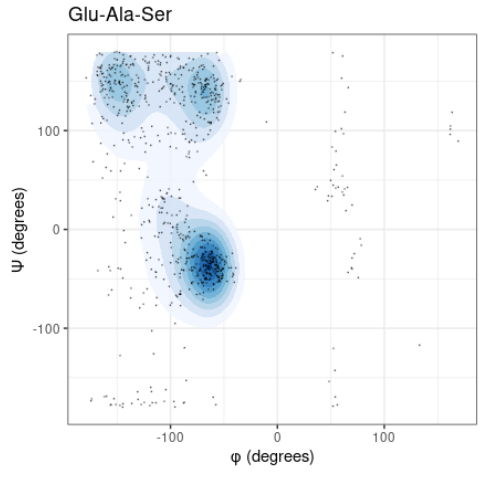
Conformational ensembles as a set of probability distributions

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

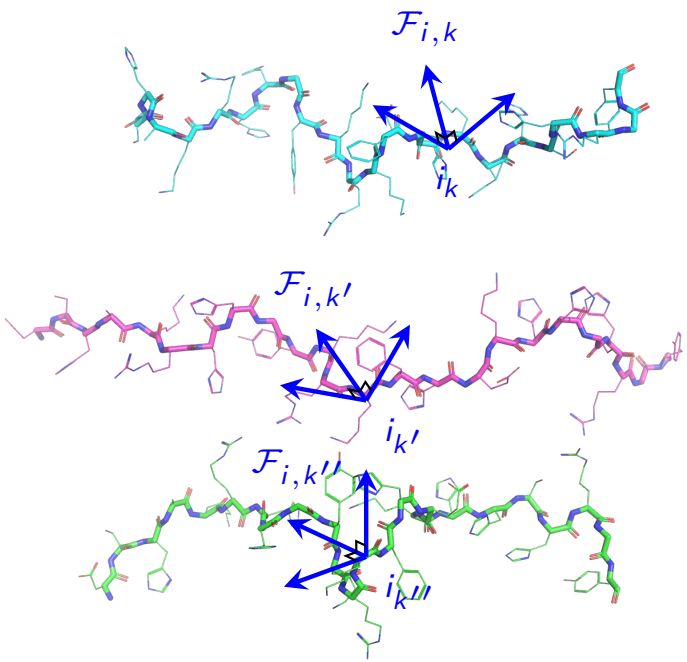
- We use the **relative positions** of residues (invariant under rigid-body motions).

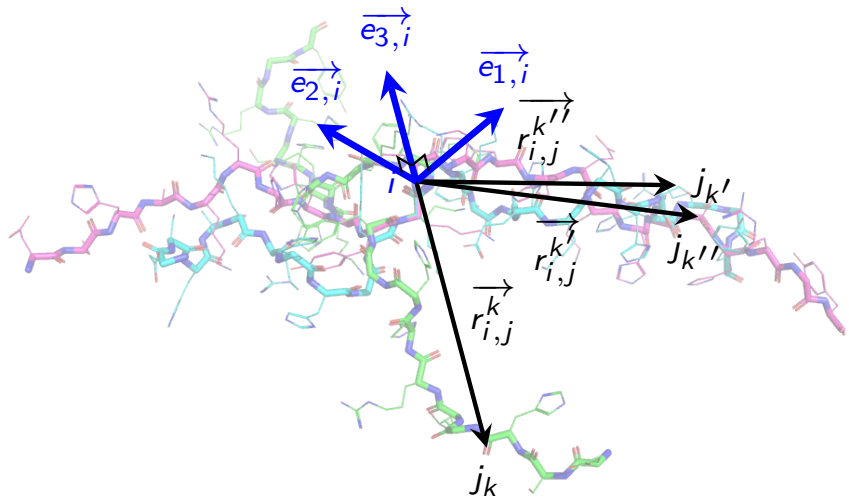
$$\left(\begin{array}{l} \text{We define the position of a given residue as the the position} \\ \text{of its } C_\beta \text{ atom when it exists and of its } C_\alpha \text{ atom otherwise.} \end{array} \right)$$

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 $\mathbb{R}^3$).



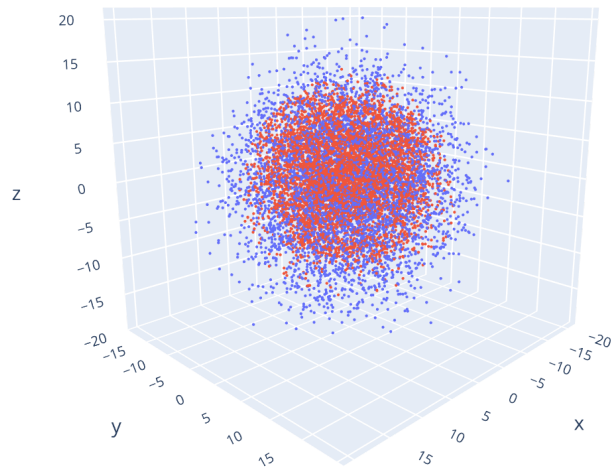


Conformational ensembles as a set of probability distributions

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

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

The comparison tool

Distances between local and global structures

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

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

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

Inter-ensemble ($\mathcal{W}_{\text{inter}}^{I,A,B}$) Intra-ensemble ($\mathcal{W}_{\text{intra}}^{I,A,B}$)

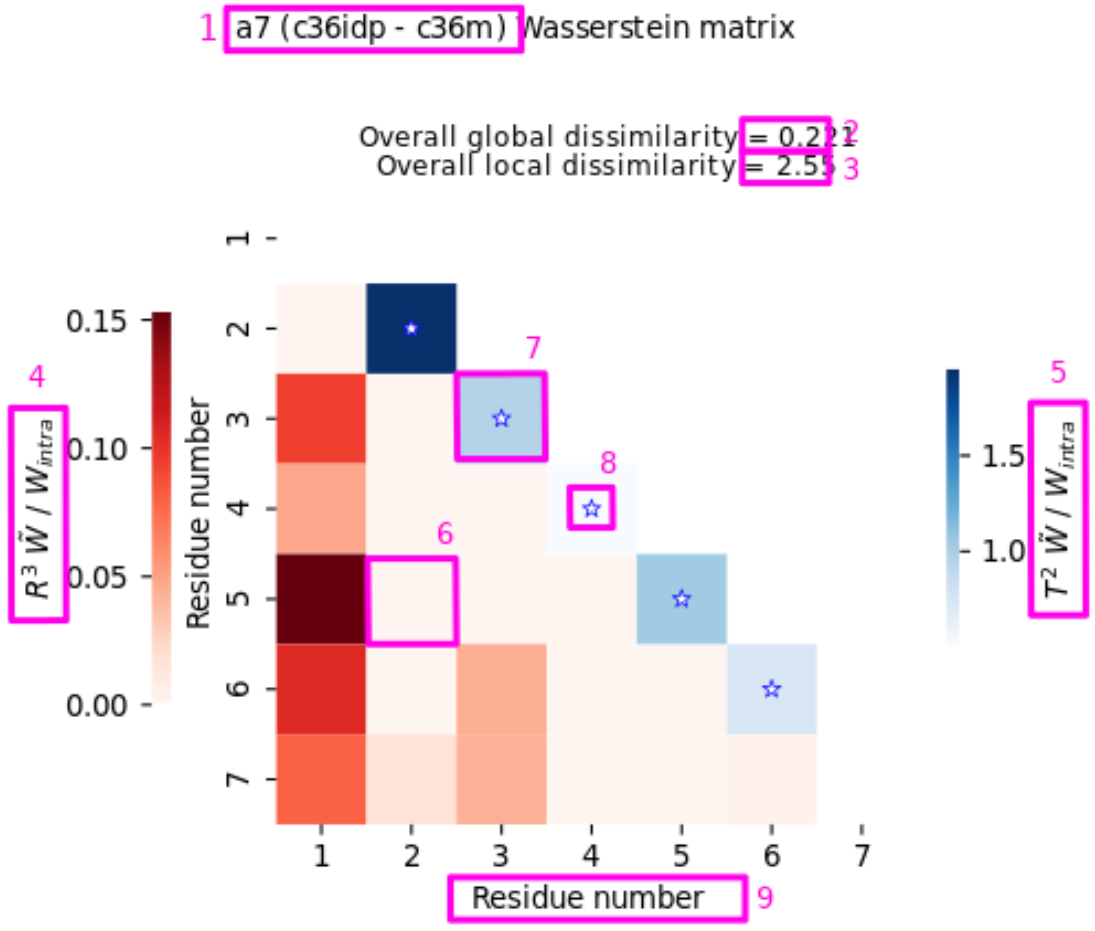
where, for any real number x, $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise.

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

The comparison tool

The Jupyter Notebook

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



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

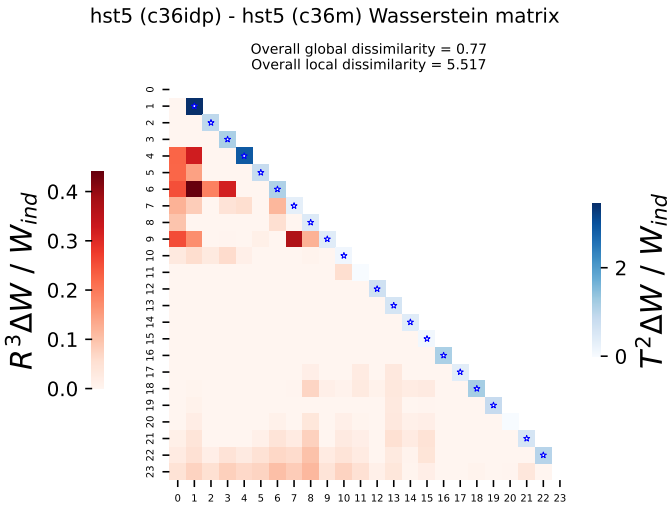
```
histatin_filtered_path = "/".join([path_to_notebook,'Examples','histatin_filtered'])
histatin_pool_path = "/".join([path_to_notebook,'Examples','histatin_pool'])

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)
```

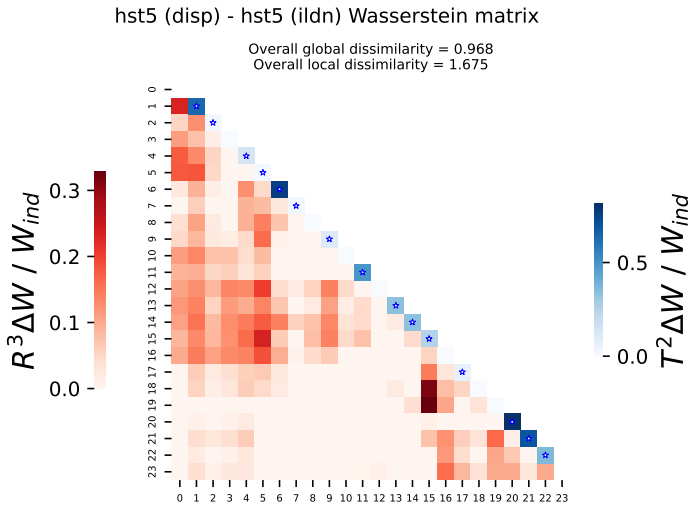
The ensemble is given as one .xtc file per replica with one .pdb file containing topology information

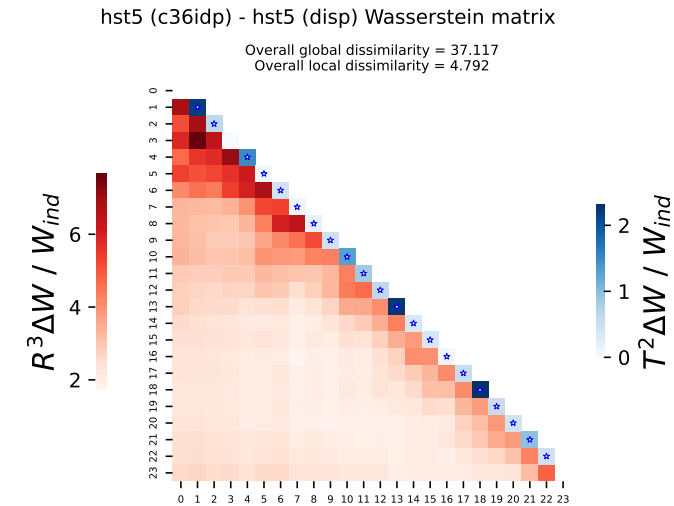
```
a7_c36idp_path = "/".join([path_to_notebook,'Examples','a7_c36idp'])
a7_c36m_path = "/".join([path_to_notebook,'Examples','a7_c36m'])

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)
```

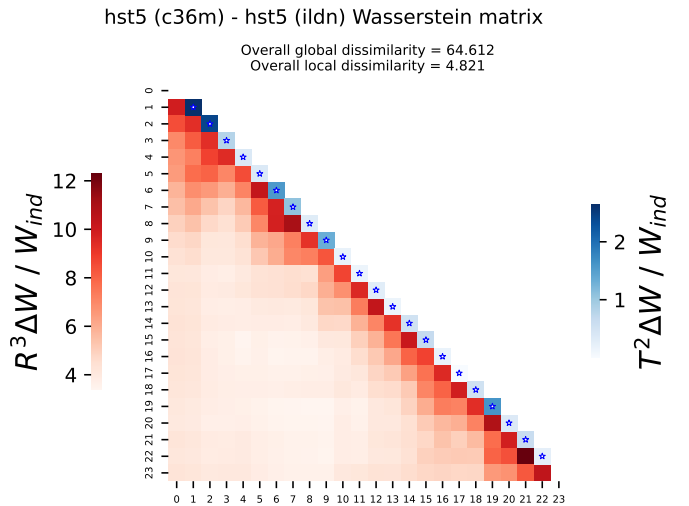
Comparison of force fields

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).

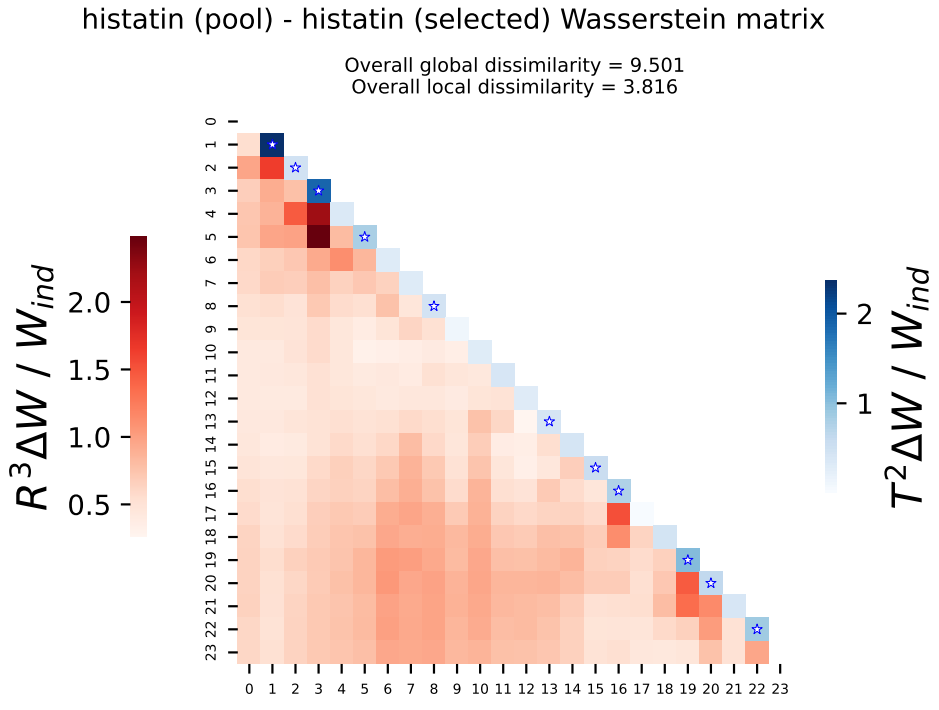








Histatin ensemble before and after filtering based on experimental SAXS data



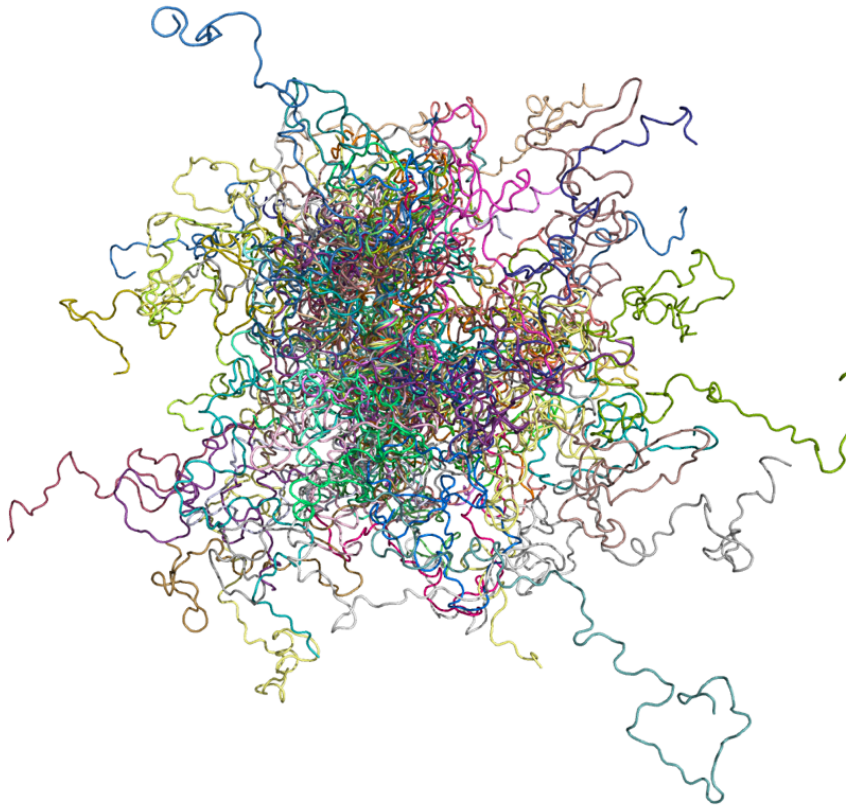
Global comparison of IDP ensembles

Clustering the conformational space

Global characterization of an IDP ensemble

Clustering the conformational space (ongoing work)

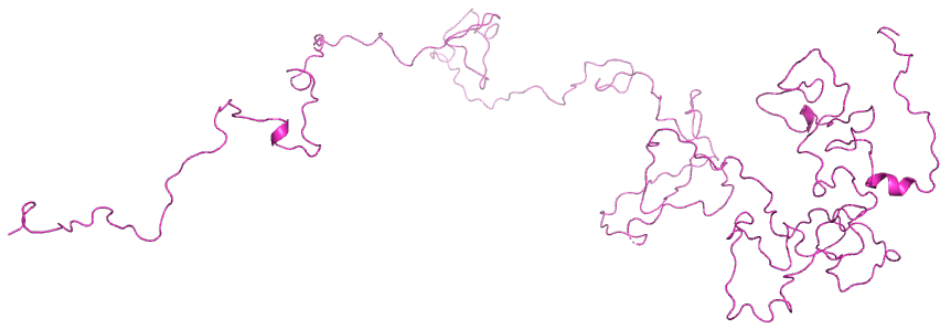
How to cluster the conformations of an IDP ensemble ?



Global characterization of an IDP ensemble

Clustering the conformational space (ongoing work)

How to transform a given conformation into a vector of *interpretable* descriptors ?



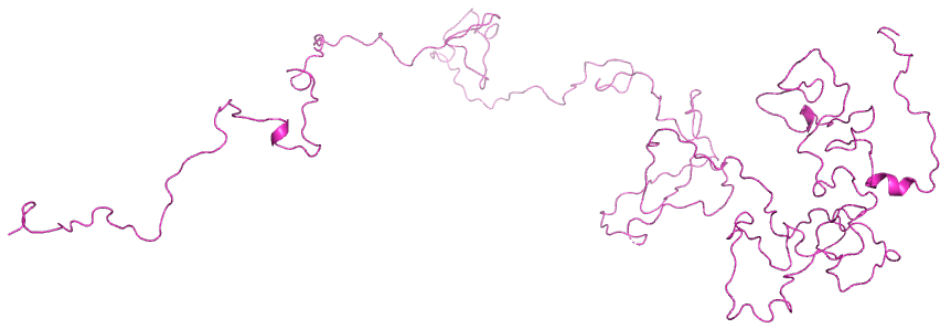
$$\Downarrow$$

$$(X_1, \ldots, X_p) \in \mathbb{R}^p$$

Global characterization of an IDP ensemble

Clustering the conformational space (ongoing work)

How to transform a given conformation into a vector of *interpretable* descriptors ?



⇓ Interaction between amino-acids

$$(X_1, \ldots, X_p) \in \mathbb{R}^p$$

Clustering the conformational space

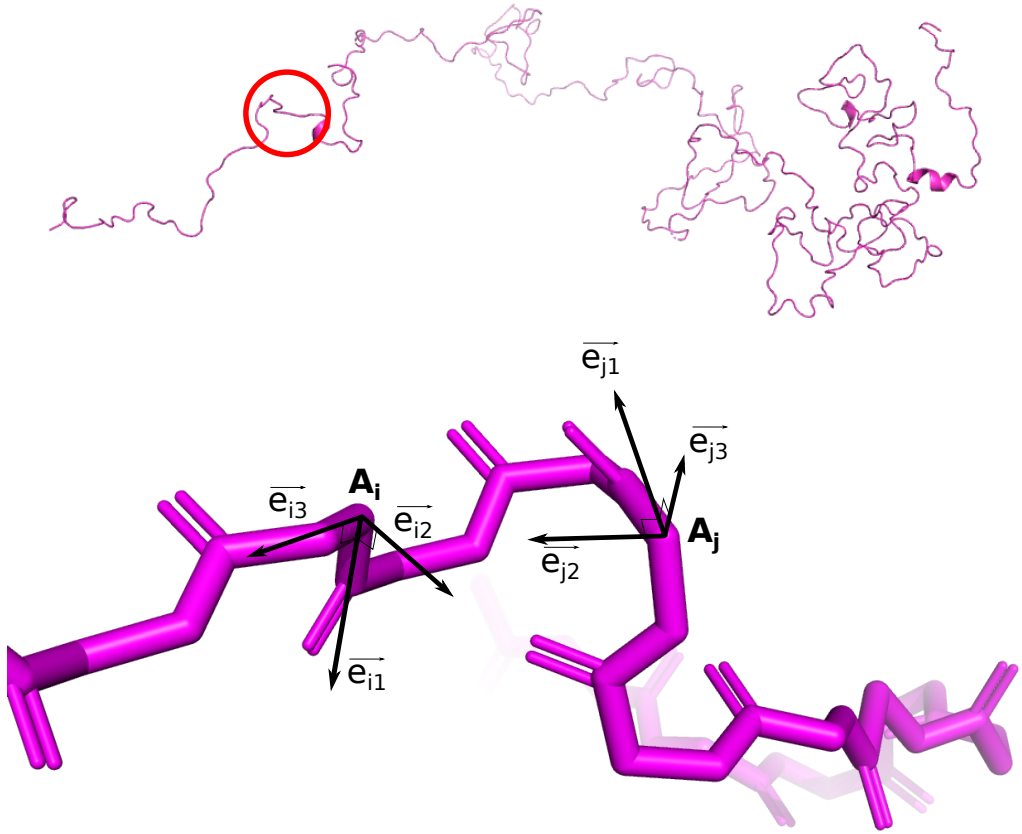
Definition of a pairwise distance between amino-acids

Clustering the conformational space

Definition of a pairwise distance between amino-acids

Clustering the conformational space

Definition of a pairwise distance between amino-acids



At each amino-acid A_i, we build a reference system $\mathcal{F}_i = \{\vec{e}_{i1}, \vec{e}_{i2}, \vec{e}_{i3}\}$ based on the relative position of its main atoms (N, C, C_α, C_β).

Clustering the conformational space

Definition of a pairwise distance between amino-acids

Interaction between amino-acids depends on

- Relative position,
- For short-range interactions : relative orientation.

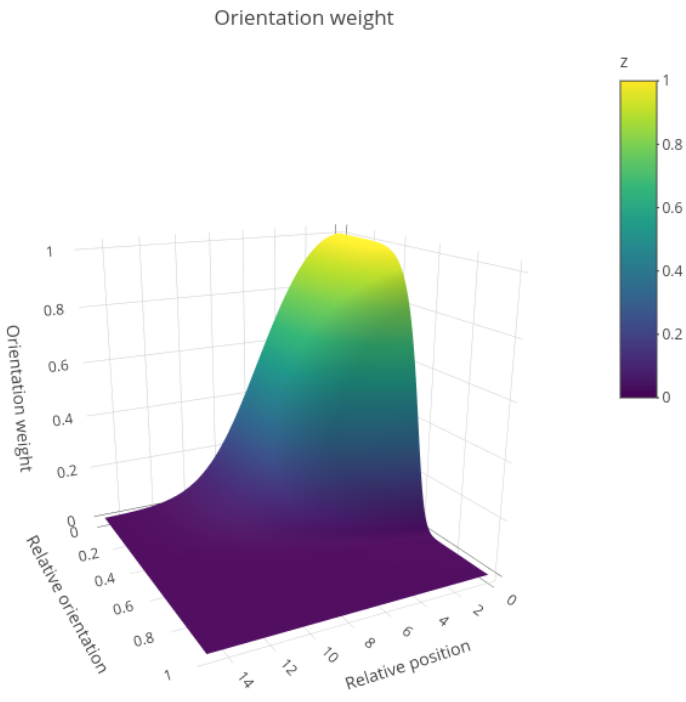
Pairwise distance between amino-acids

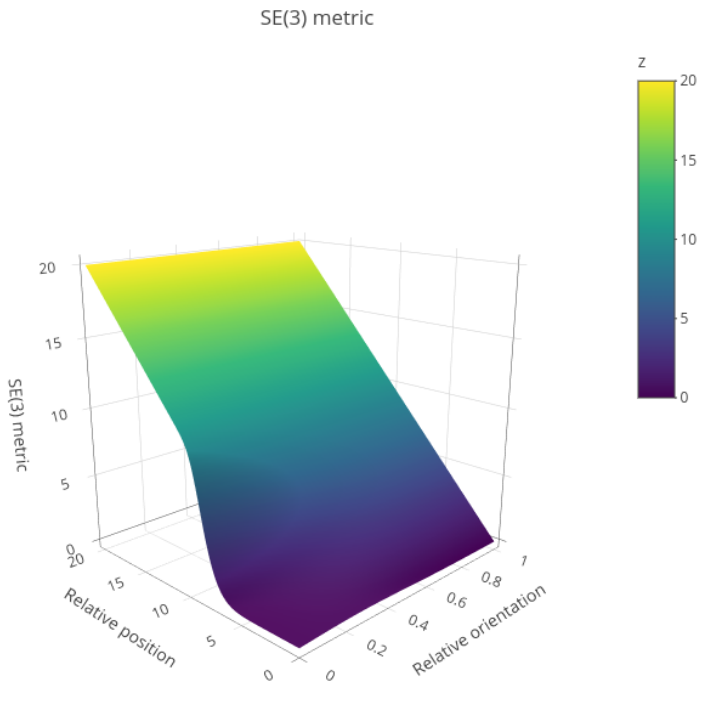
$$d^2(A_i, A_j) = (1-\omega)^2\, d^2_{\mathbb{R}^3}(A_i, A_j) + \omega^2\, d^2_{SO(3)}(\mathcal{F}_i, \mathcal{F}_j)$$

- $\omega = \omega\left(A_i, A_j, |i-j|, d_{\mathbb{R}^3}(A_i, A_j), d_{SO(3)}(\mathcal{F}_i, \mathcal{F}_j)\right)$.
- $d_{SO(3)}(\mathcal{F}_i, \mathcal{F}_j)$ measures the distance of $\left(\mathcal{F}_i, \mathcal{F}_j\right)$ to an *ideal* relative orientation in terms of interaction.

Pairwise distance between amino-acids

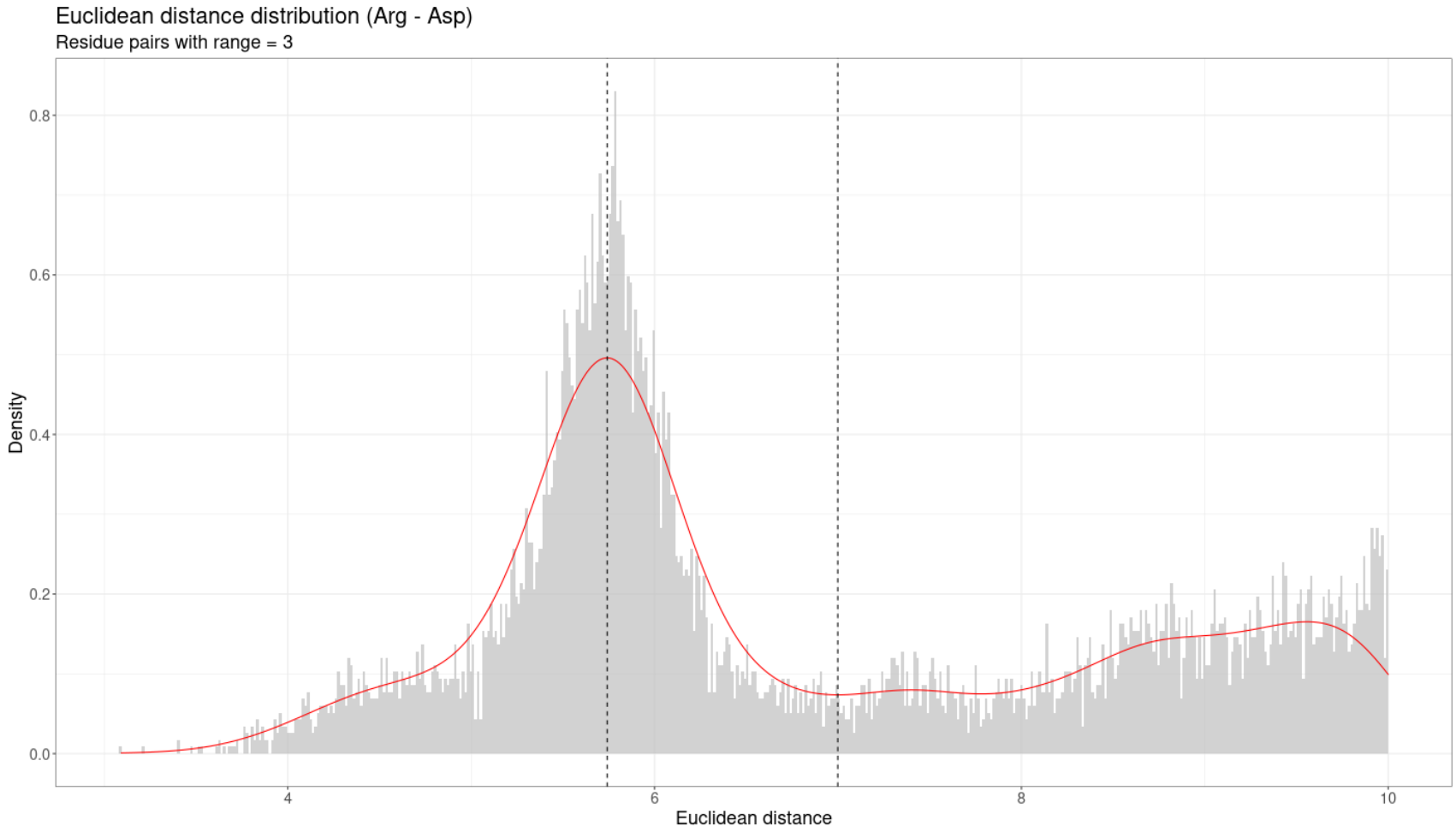
How to weight distance and relative orientation





Pairwise distance between amino-acids

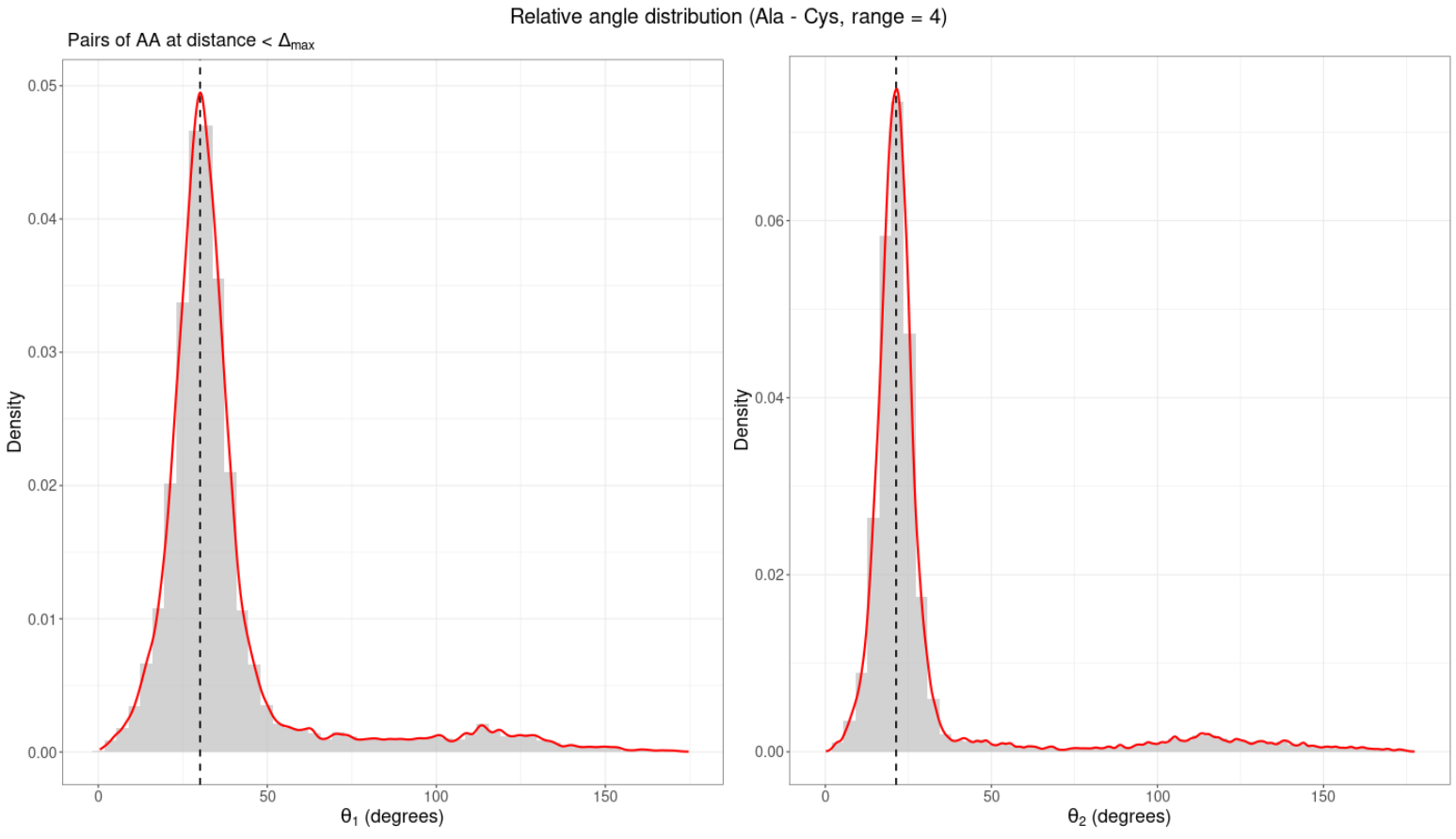
How to weight distance and relative orientation



Pairwise distance between amino-acids

How to define a distance between relative orientations

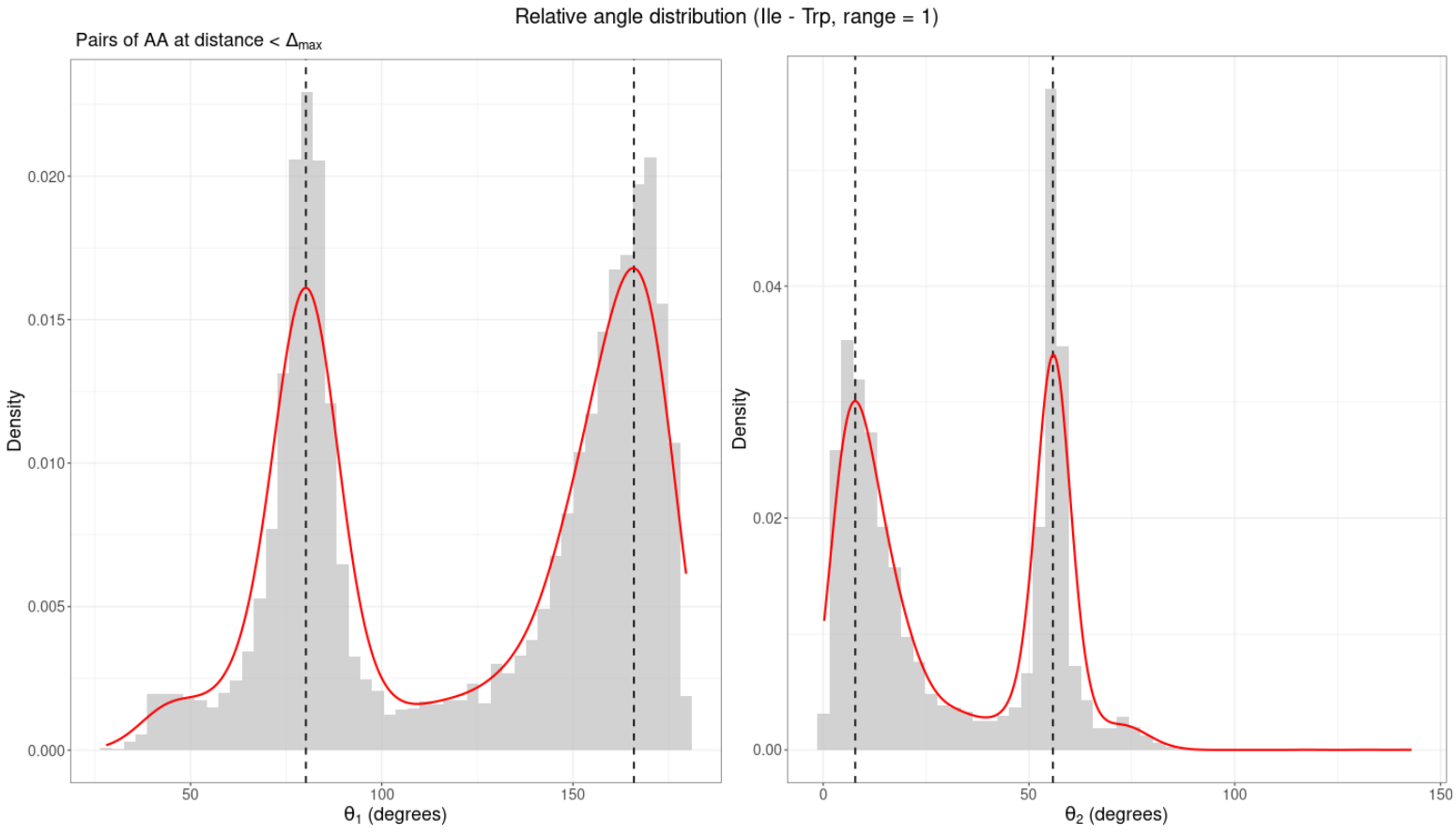
For every pair of amino-acids and every range ≤ 4, preferred relative orientations exist



Pairwise distance between amino-acids

How to define a distance between relative orientations

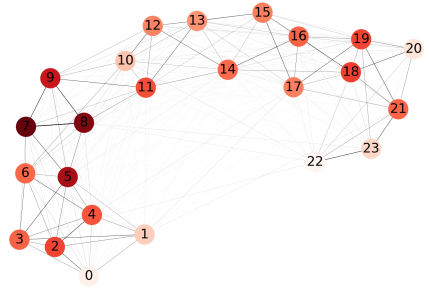
For every pair of amino-acids and every range ≤ 4, preferred relative orientations exist



Clustering the conformational space

Representing conformations as graphs

$$\Downarrow \quad d(A_i, A_j) \quad \forall\, i, j \in \{1, \ldots, L\}$$



Node $i = A_i$, Edge weight $\omega_{ij} = f(d(A_i, A_j))$ with f decreasing.

Clustering the conformational space

From a graph to a vector of features

Assortativity

Let Y be a node attribute. The Y-**assortativity** is the correlation of Y between pairs of connected nodes. It describes the preference for the graph's nodes to attach to others that have similar values of Y.

Vector of features

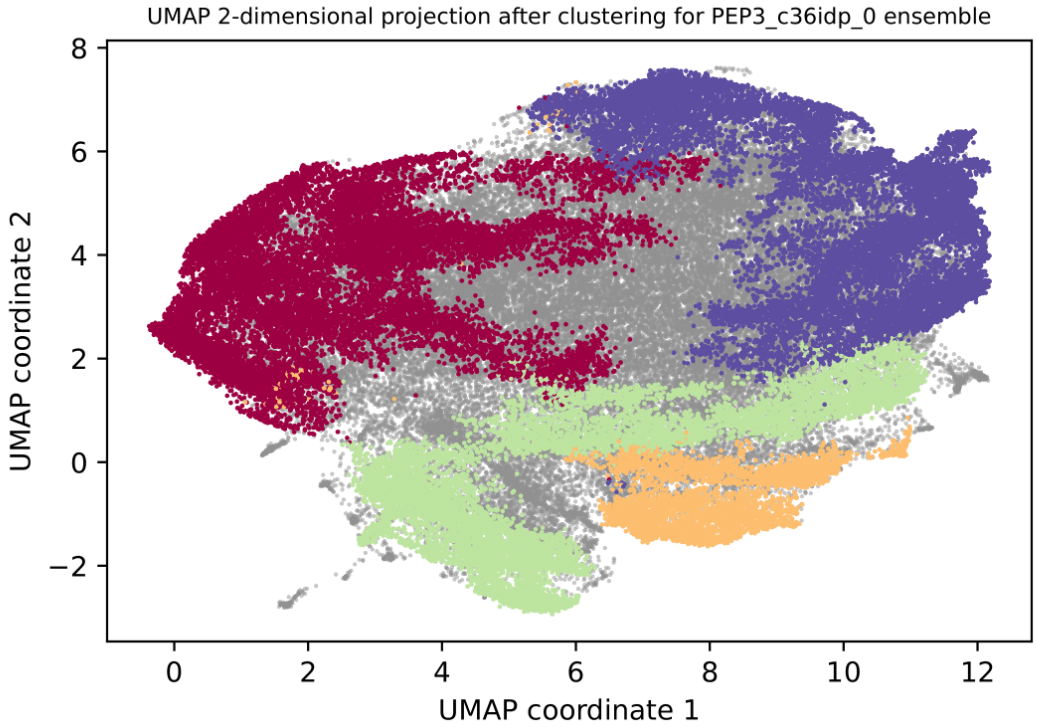
Let $Y_1, \ldots, Y_p$ be p amino-acid attributes having a significant effect on amino-acid interactions. We consider $X_i = \mathrm{Assortativity}(Y_i) \in (-1, 1)$ as the i-th conformational descriptor.

Perform clustering based on features $(X_1, \ldots, X_p) \in (-1, 1)^p$.

Preliminary results

MD simulation for PEP13 using CHARMM26IDPSFF force-field

Clustering 100.000 conformations with 47 features describing the physico-chemical properties mediating contacts.

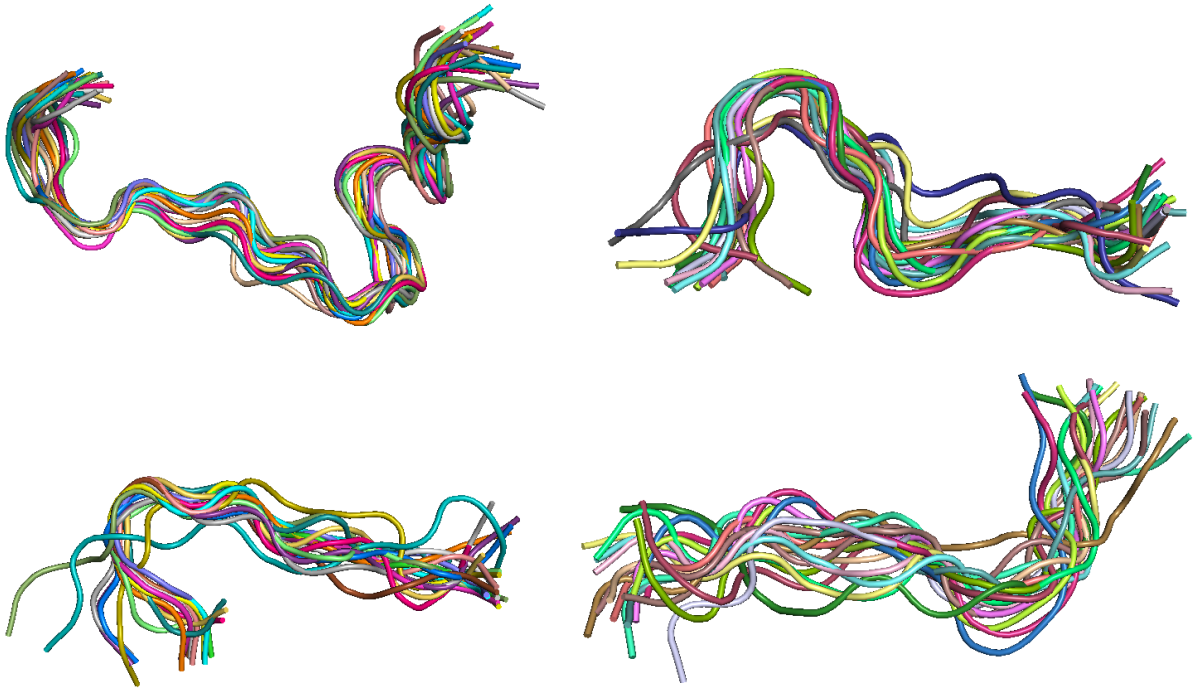


Projection to a 2-dimensional UMAP space (just for data visualization)

Preliminary results

MD simulation for PEP13 using CHARMM26IDPSFF force-field

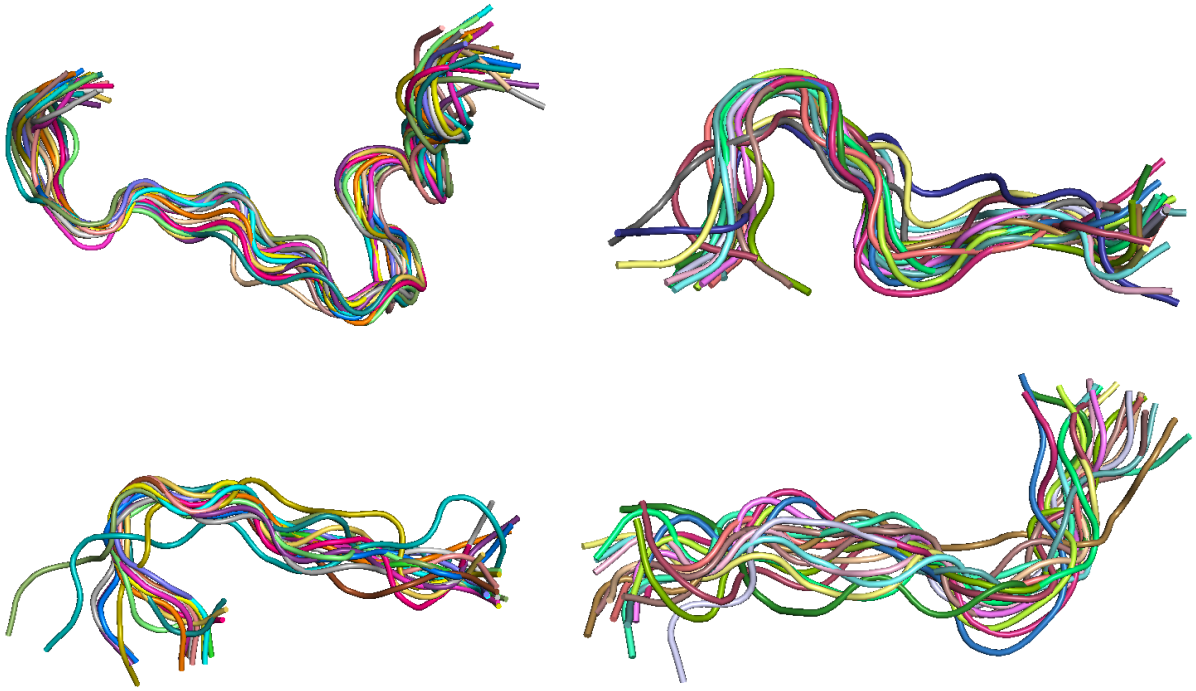
Looking at the 4 clusters...



Preliminary results

MD simulation for PEP13 using CHARMM26IDPSFF force-field

Looking at the 4 clusters...



Every clustering algorithm is sensitive to a family of parameters

Need of **post-clustering inference** analysis : assess whether two clusters are *really* different.

Thank you for your attention !

http://www.math.univ-toulouse.fr/∼jgonzale