Several stories about tripeptides:
basic components for modeling flexible protein structures

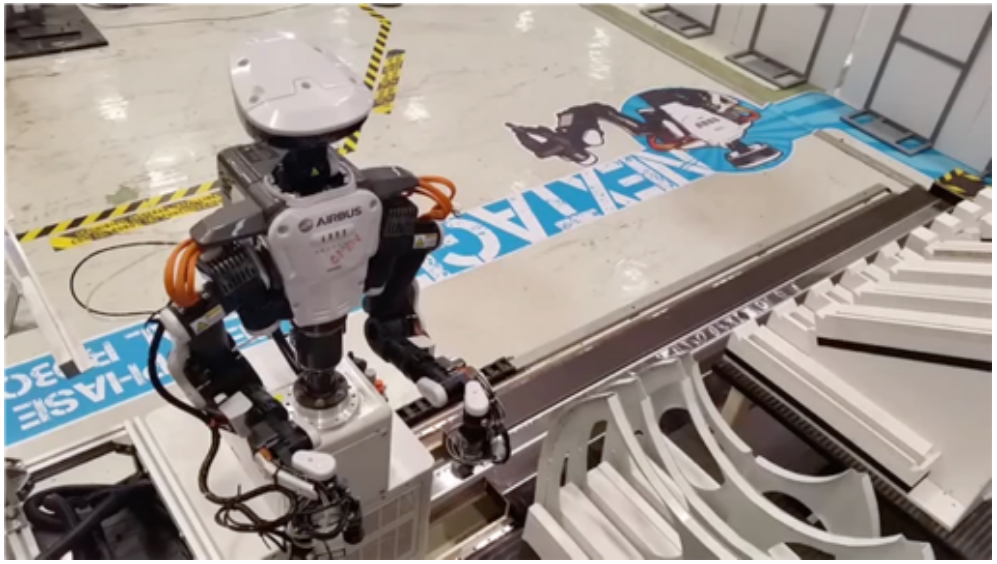
Juan Cortés

BSC - Barcelona, January 2020

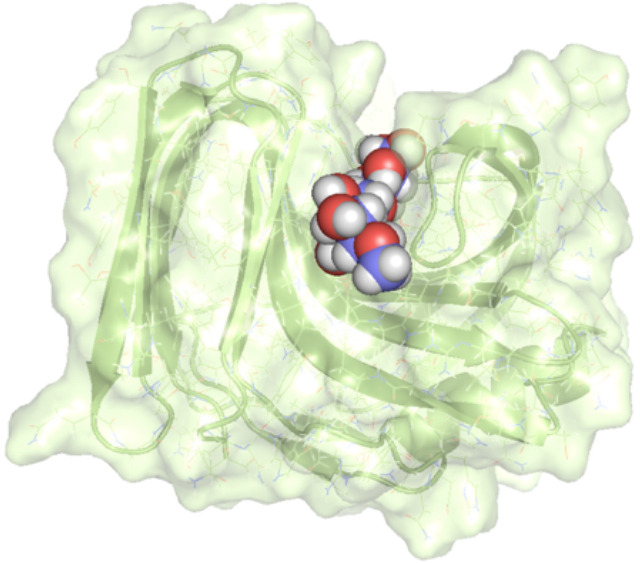
Once upon a time …

[Cortés *et al.*, *J Comput Chem*, 2004]

From **robot** motion planning…



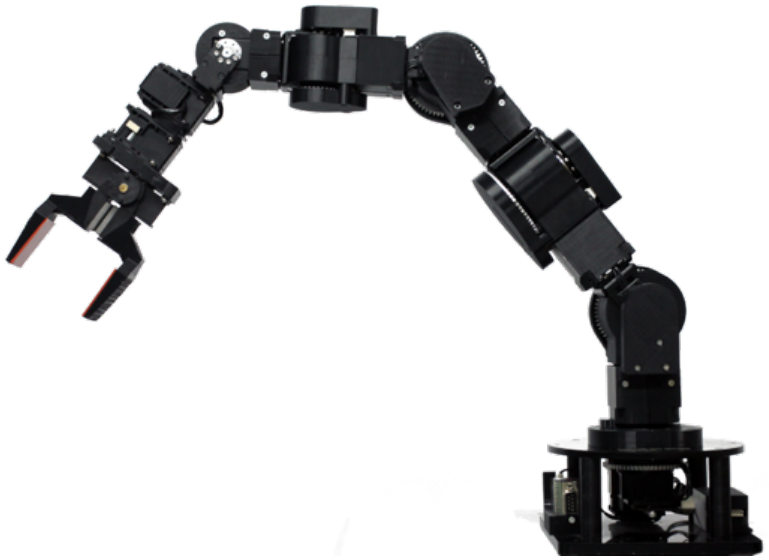
… to **protein** loop modeling



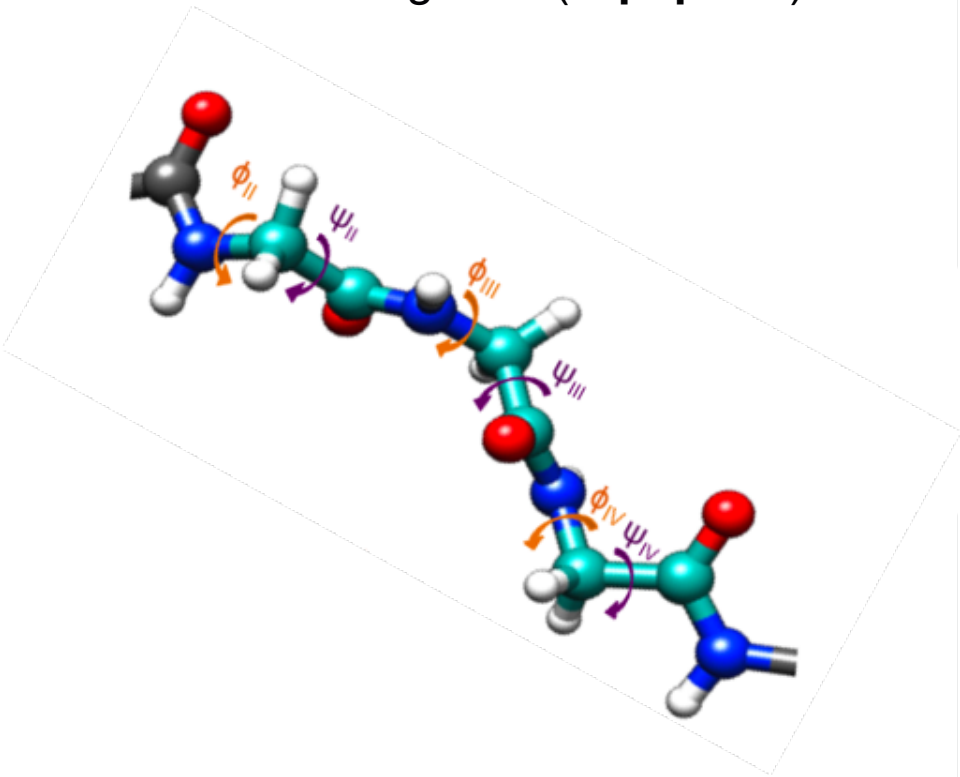
tripeptides are a clear link between the two domains

Tripeptides as basic mechanistic components

Robotic arm (**6R mechanism**)

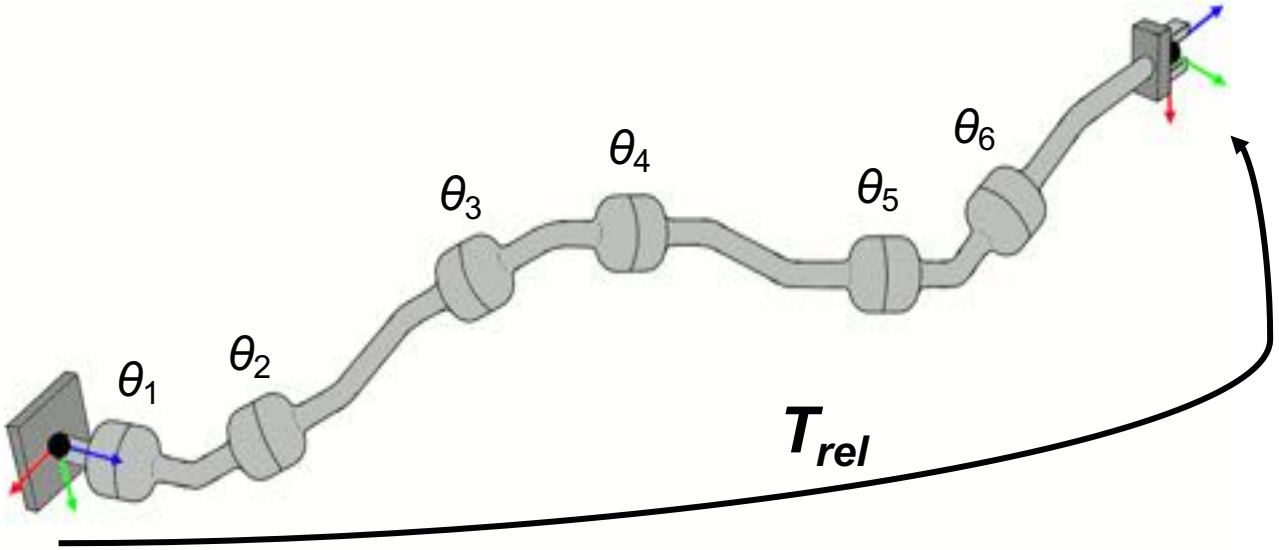


3 amino-acid fragment (**tripeptide**)



6 degrees of freedom : minimum number for full mobility

The Inverse Kinematics (IK) Problem



Given the relative pose T_{rel} of the *end-effector* with respect to the *base-frame* find the values of the joint variables θ_i

$$T_{rel} = {}^0T_1(\theta_1)\ {}^1T_2(\theta_2)\ {}^2T_3(\theta_3)\ {}^3T_4(\theta_4)\ {}^4T_5(\theta_5)\ {}^5T_6(\theta_6)$$

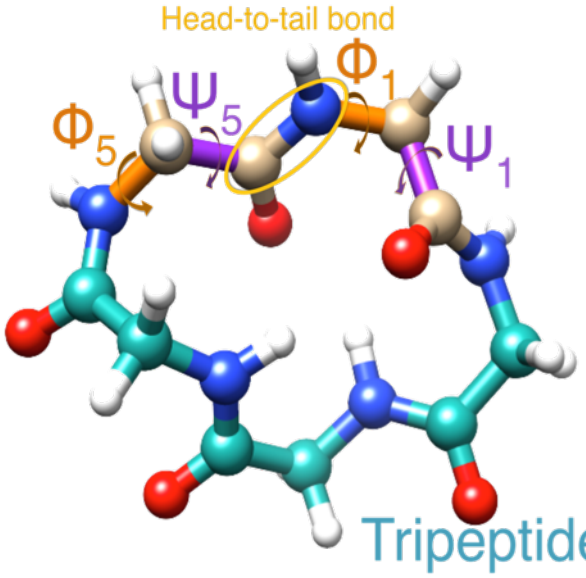
→ **Efficient closed-form solvers** [Renaud, 2000]

“Straightforward” application: Sampling of small cyclic peptides

[Jusot *et al.*, JCIM 2018]

Joint work with J. Chomilier and D. Stratmann (Sorbonne U., Paris)

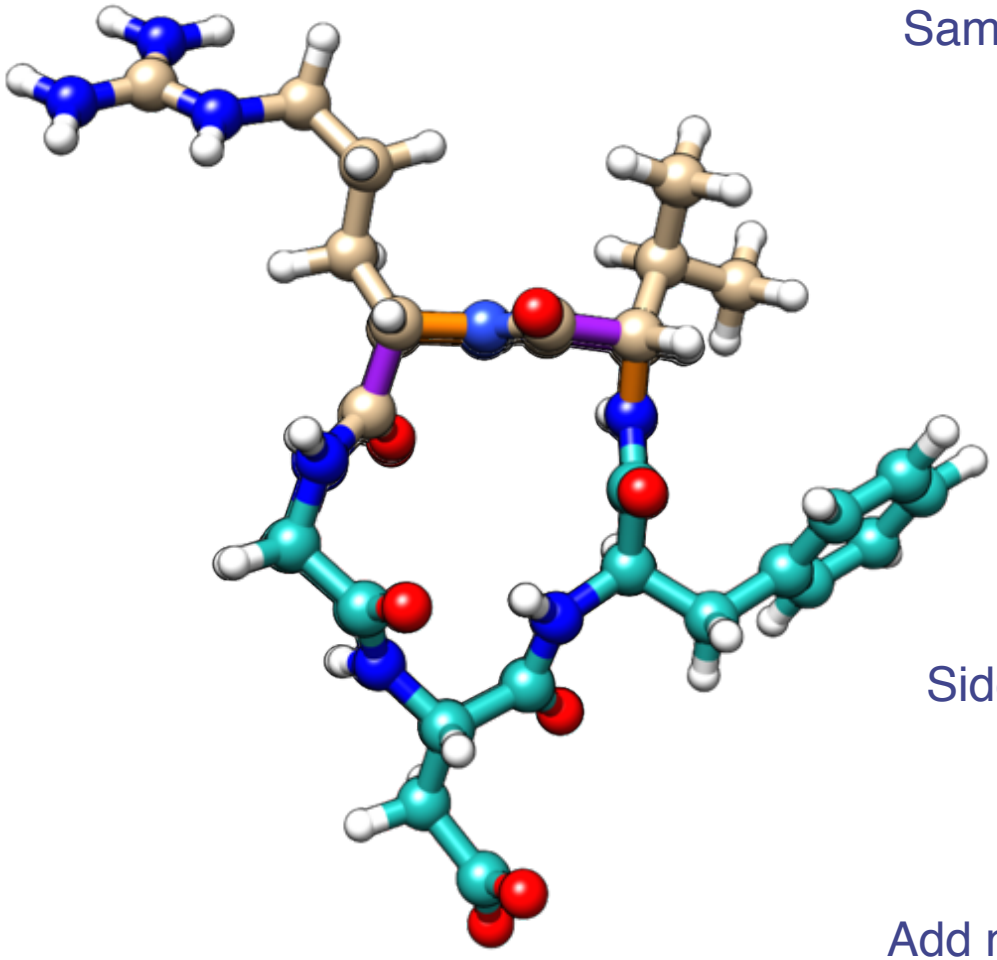
Cyclic-pentapeptide



Conformational space (considering only Φ, Ψ)
→ 4-dimension manifold in a 10-dimension space
Can be sampled exhaustively !

“Straightforward” application: Sampling of small cyclic peptides

[Jusot *et al.*, JCIM 2018]



Sampling of Φ_1, Ψ_1, Φ_5, Ψ_5 **(+ all ω)**

↓

Inverse kinematics:
- 0 solution
- 1 to 16 solutions

↓

Check collisions

↓

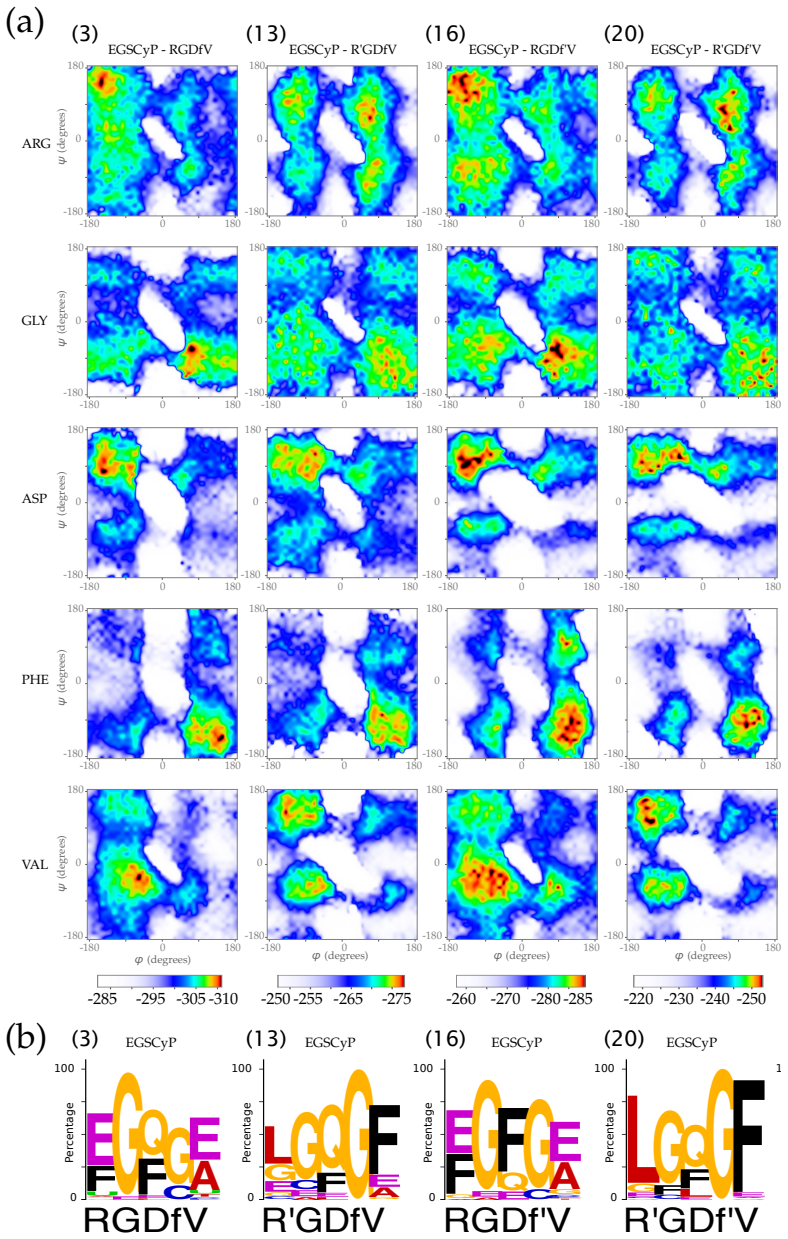
Side chains addition(SCWRL)

↓

Relaxation (Amber)

Add node to adjacency graph

“Straightforward” application: Sampling of small cyclic peptides



[Jusot *et al.*, JCIM 2018]

- Analysis of the effect of chemically-modifications or D amino acids
- Potential applications for drug design

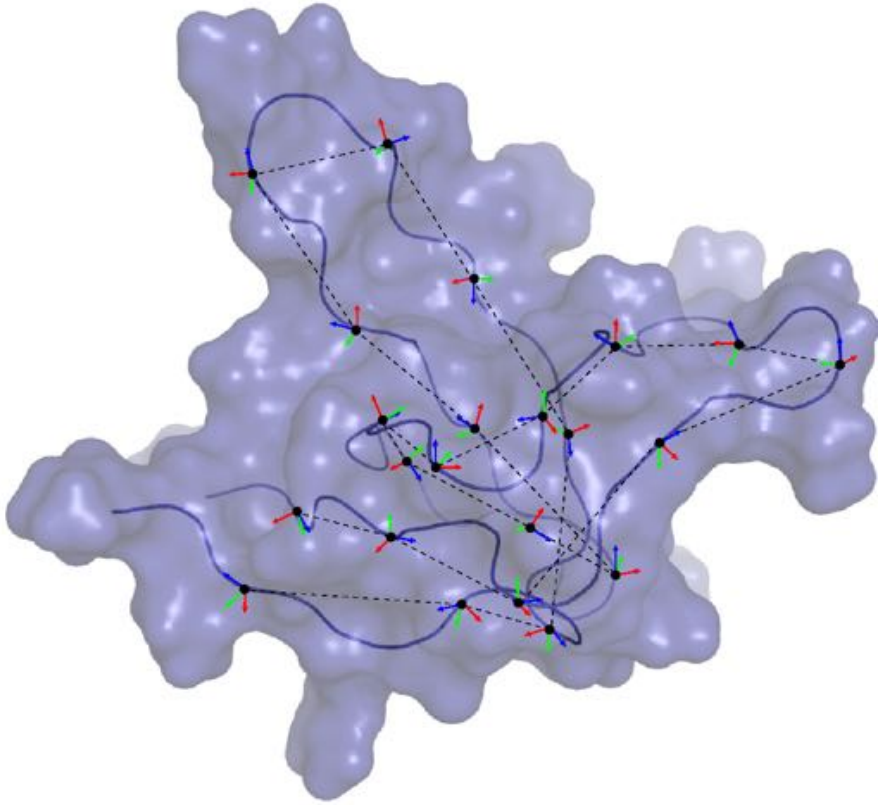
Conformational sampling using the tripeptide-based model

1. Sample/perturb the pose of the base-frame of tripeptides (*oriented **particles***)

   ➔ Coarse-grained model

1. Solve **IK** for tripeptides
   Semi-analytical solver
   (adapted from [Renaud, 2000])

   ➔ All-atom model

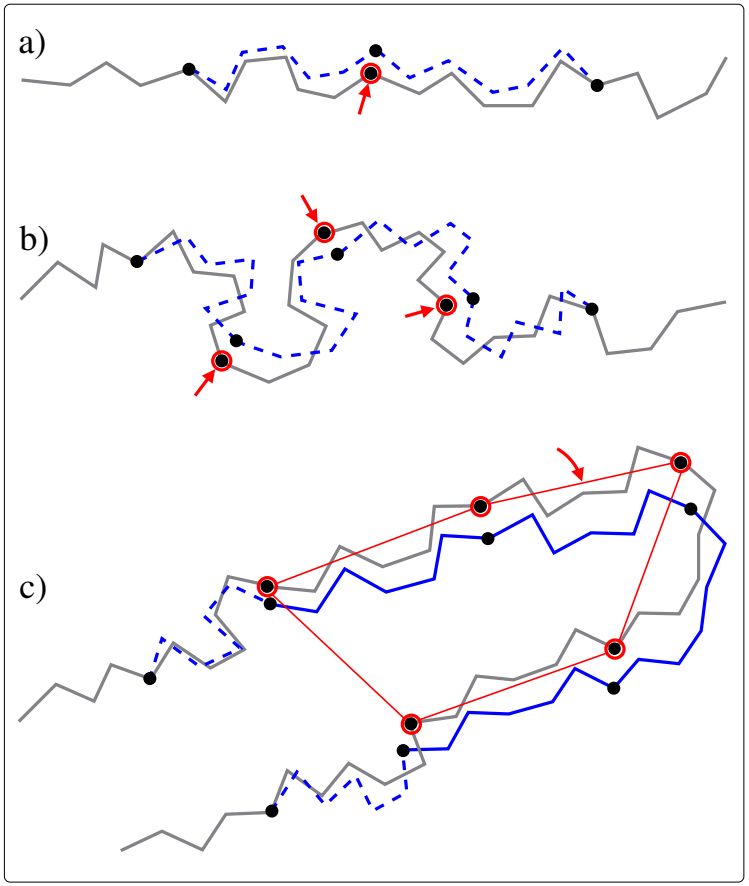


Tripeptide-based Monte Carlo move classes

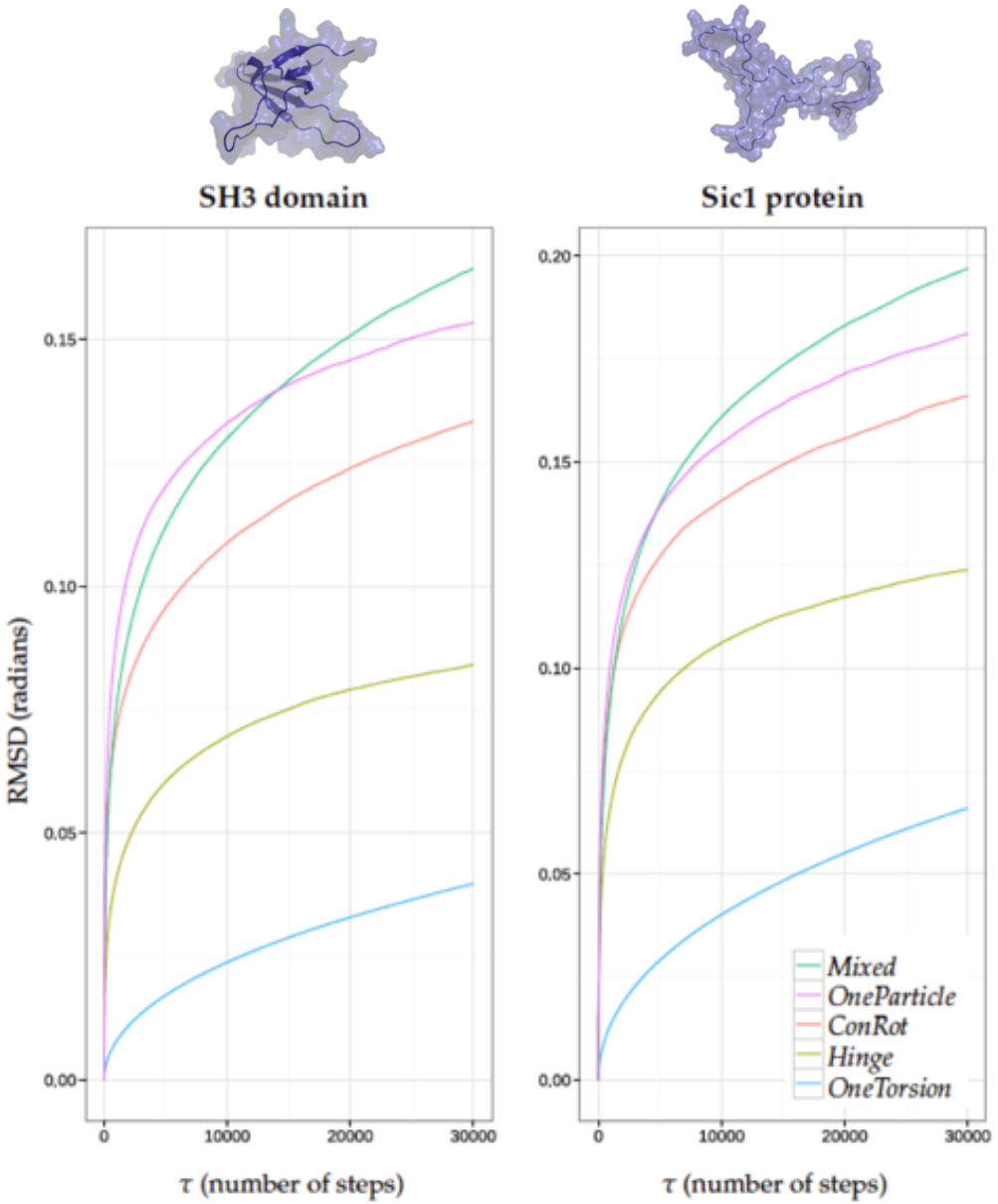
[Denarie *et al.*, Molecules 2018]

Local (fixed-end) backbone perturbation methods

- **One particle moves**
  - Perturb one particle
  - IK for two tripeptides
  - Similar to ConRot [Dodd *et al.*, 1993]
- **Flexible fragment moves**
  - Perturb *n* consecutive particles
  - IK for *n*+1 tripeptides
  - Similar to CCL [Canutescu *et al.*, 2003]
- **Rigid-body block moves ("hinge")**
  - Perturb *n* particles as a rigid-body
  - IK for two tripeptides
  - Similar to CRRUBAR [Betancourt, 2005]



Tripeptide-based Monte Carlo move classes

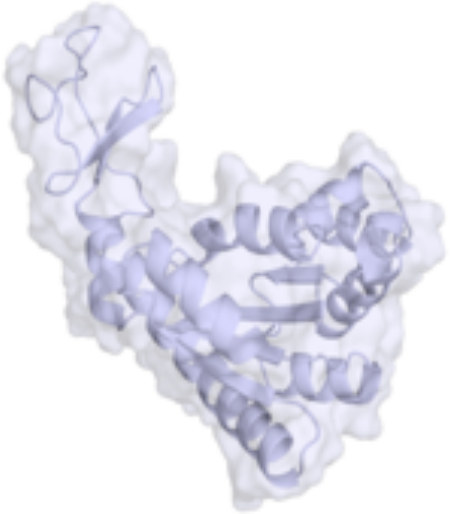


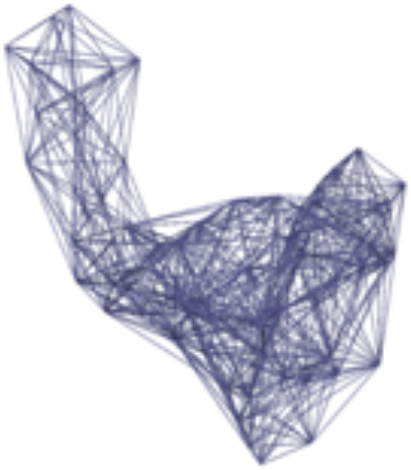
[Denarie *et al.*, Molecules 2018]

- Fixed-ends moves perform better than classical *pivot* moves
- *Mixing* moves (easy thanks to the tripeptide-based decomposition) outperforms individual moves

[Al-Bluwi *et al.*, BMC Struct Biol, 2013]

Tripeptide-based Elastic Network Model

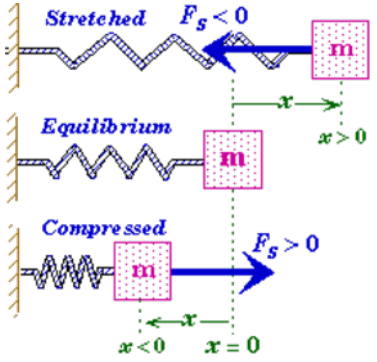




- ***Topological*** force field

  Springs between particles

$$V(x)=\frac{1}{2}k\left(x-x_0\right)^2$$



Guided Sampling using Normal Model Analysis

[Al-Bluwi *et al.*, BMC Struct Biol, 2013]

Tripeptide-based Elastic Network Model

- Diagonalization of the Hessian matrix :
  - Eigenvectors → Vibration directions
  - Eigenvalues → Vibration frequencies
- Low-frequency modes ➡ collective motions

$$\mathbf{H} = \begin{pmatrix} \left(\frac{\partial^2 V}{\partial r_1 \partial r_1}\right)_0 & \left(\frac{\partial^2 V}{\partial r_1 \partial r_2}\right)_0 & \cdots & \left(\frac{\partial^2 V}{\partial r_1 \partial r_{3N}}\right)_0 \\ \left(\frac{\partial^2 V}{\partial r_1 \partial r_2}\right)_0 & \left(\frac{\partial^2 V}{\partial r_2 \partial r_2}\right)_0 & \cdots & \left(\frac{\partial^2 V}{\partial r_2 \partial r_{3N}}\right)_0 \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{\partial^2 V}{\partial r_1 \partial r_{3N}}\right)_0 & \cdots & \cdots & \left(\frac{\partial^2 V}{\partial r_{3N} \partial r_{3N}}\right)_0 \end{pmatrix}$$

Guided Sampling using Normal Model Analysis

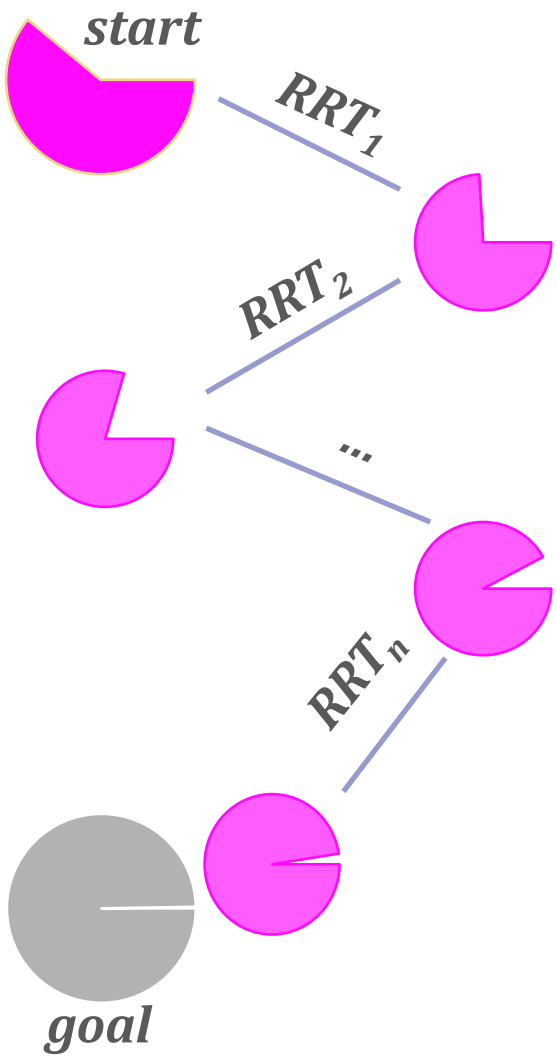
[Al-Bluwi *et al.*, BMC Struct Biol, 2013]

Tripeptide-based Elastic Network Model

- Advantages :
  - Reduces the size of the hessian matrix by a factor of 3 (compared to Cα-ENM)
  - Preserves the full flexibility of the protein (compared to rigid-block approaches)

Application : Sample Conformational Transition Pathways

[Al-Bluwi *et al.*, BMC Struct Biol, 2013]

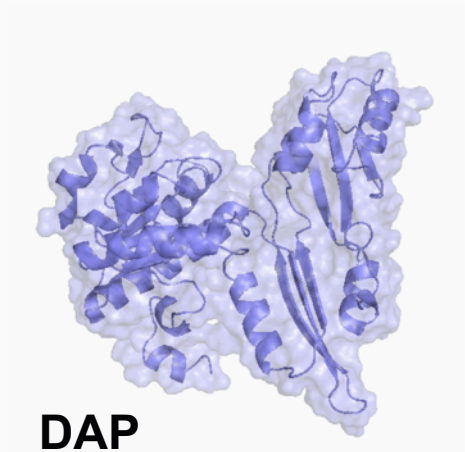


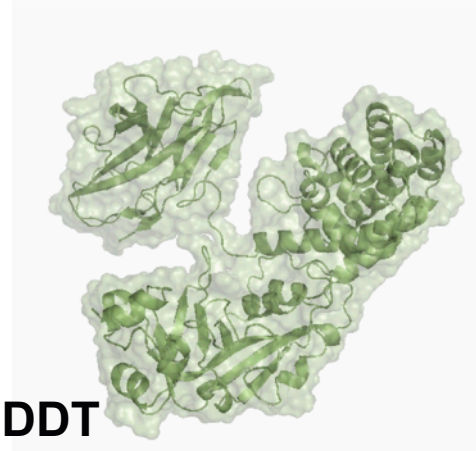
Algorithm 1: COMPUTE_PATHWAY

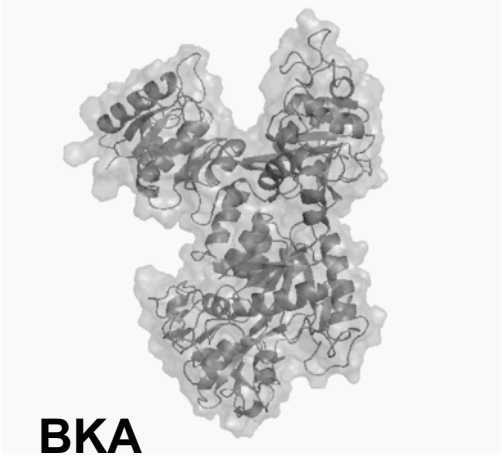
input : Initial conformation q_{init}, final conformation q_{goal} and minimum distance to target d_{target}
output : The transition pathway p
begin
  $q_{root} \leftarrow q_{init}$;
  while RMSD(q_{root}, q_{goal}) $> d_{target}$ **do**
    $m \leftarrow$ COMPUTE_NORMALMODES(q_{root});
    $t \leftarrow$ BUILD_RRT(m, q_{root}, q_{goal});
    $q_{close} \leftarrow$ CLOSESTTOTARGET(t, q_{goal});
    $q_{root} \leftarrow$ MINIMIZE(q_{close});
    $p \leftarrow$ CONCATENATE(p, q_{root});
end

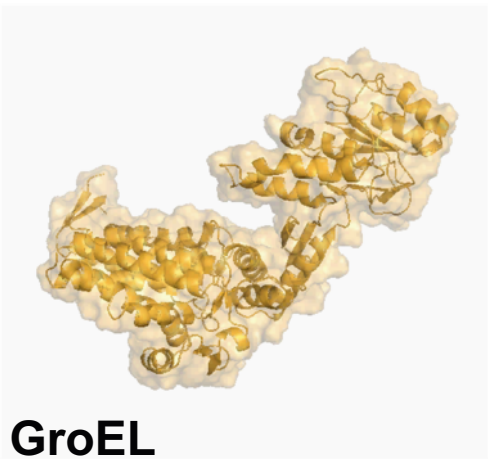
Application : Sample Conformational Transition Pathways

[Al-Bluwi *et al.*, BMC Struct Biol, 2013]







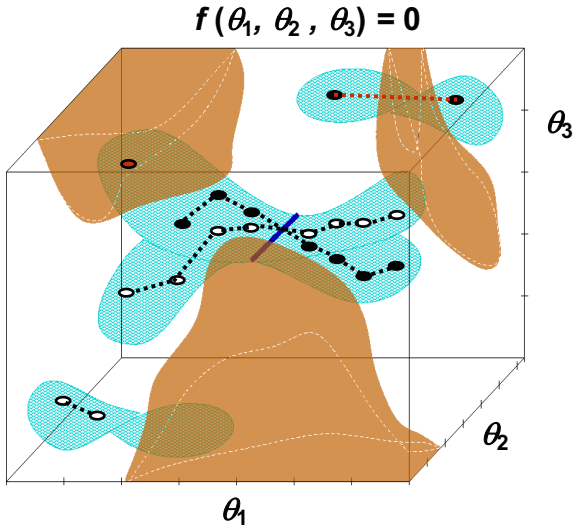


- Sigle-core CPU time : 2 hours – 2 days (*can be further improved*)

Sampling Long Protein Loops

Challenging problem

How to sample submanifolds?



Several approaches:

- ***Optimization-based*: Coordinate Cyclic Descent (CCD)** [Welman, 1993]
  - Application to proteins: [Canutescu et al, *Protein Sci.*, 2003]
- ***Semi-analytical*: Random Loop Generator (RLG)** [Cortés et al, 2002]
  - Application to proteins: [Cortés et al, *J. Comput. Chem.*, 2004]

Sampling Long Protein Loops: *Our first approach*

[Cortés *et al.*, *J Comput Chem*, 2004]

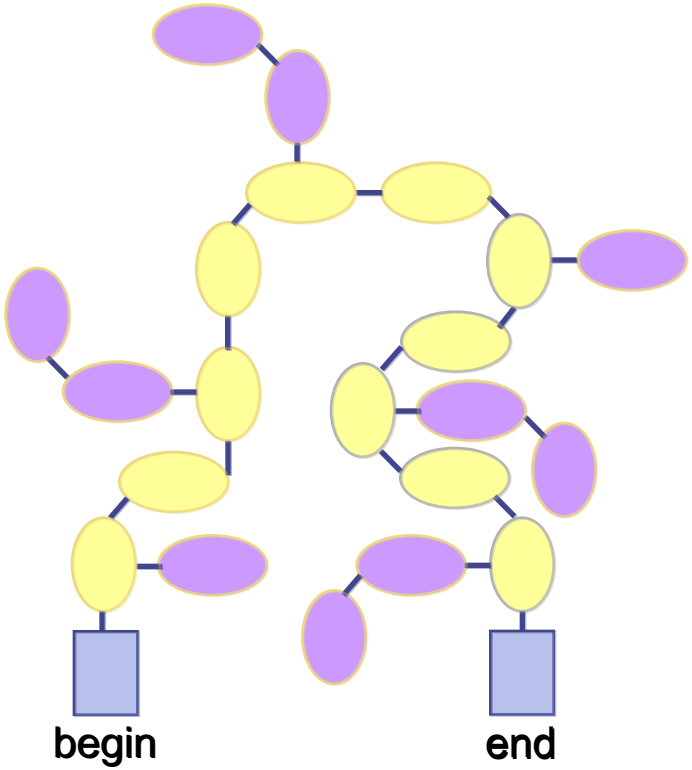
Random Loop Generator (RLG)

Main Chain (Backbone)

→ Independent sub-chain
- RLG algorithm + CollCheck

→ Dependent sub-chain
- General 6R IK [Renaud, 2000] + CollCheck

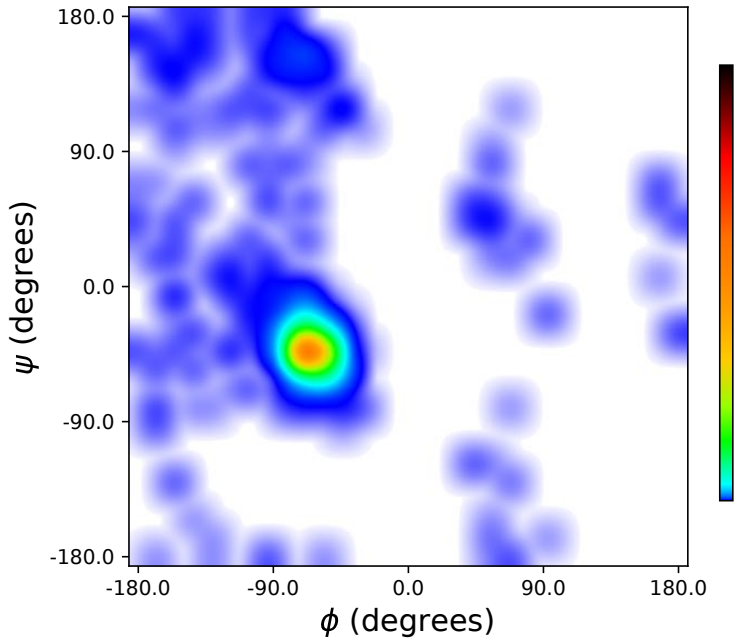
Side Chains
- Random Sampling + CollCheck



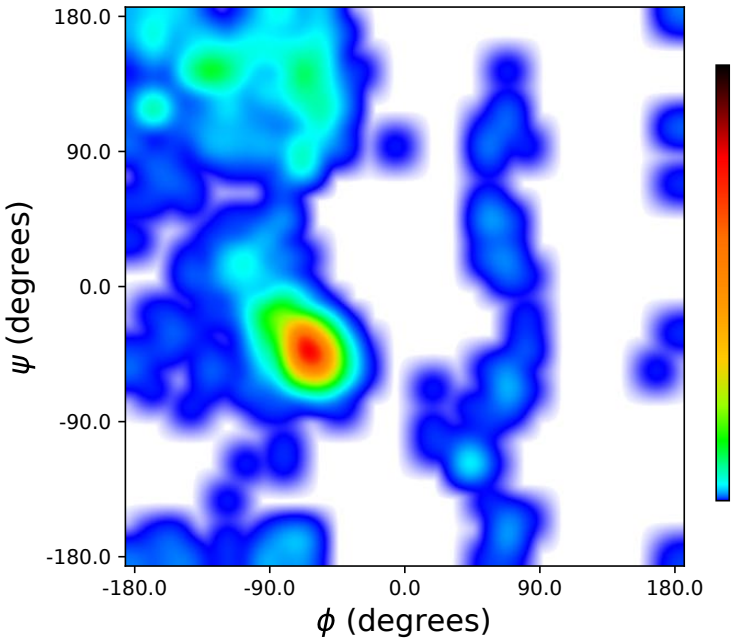
Residues sampled individually (*excepting for the last tripeptide subject to IK*)

Local sequence matters

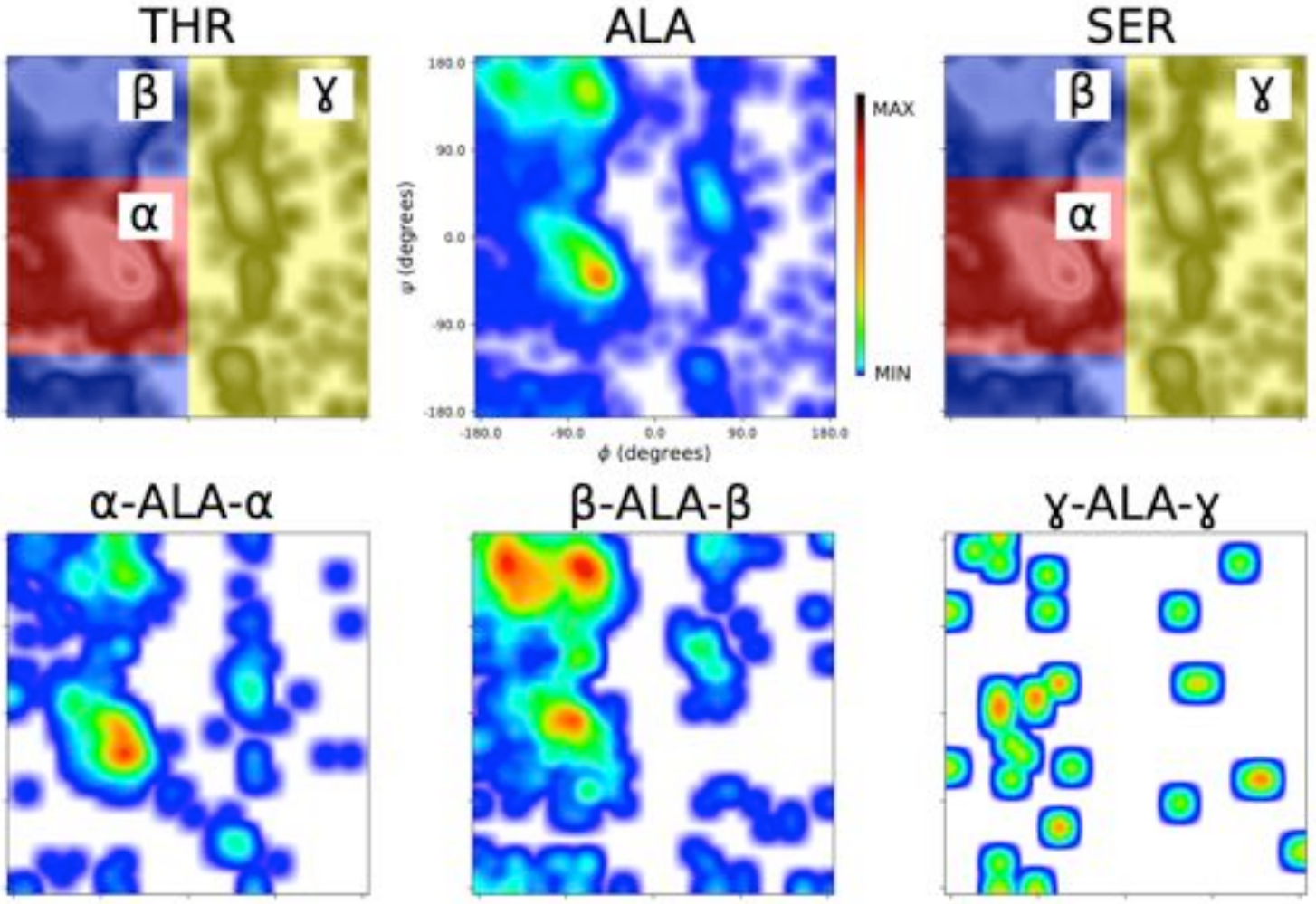
ALA-ALA-ALA



ALA-ALA-GLU

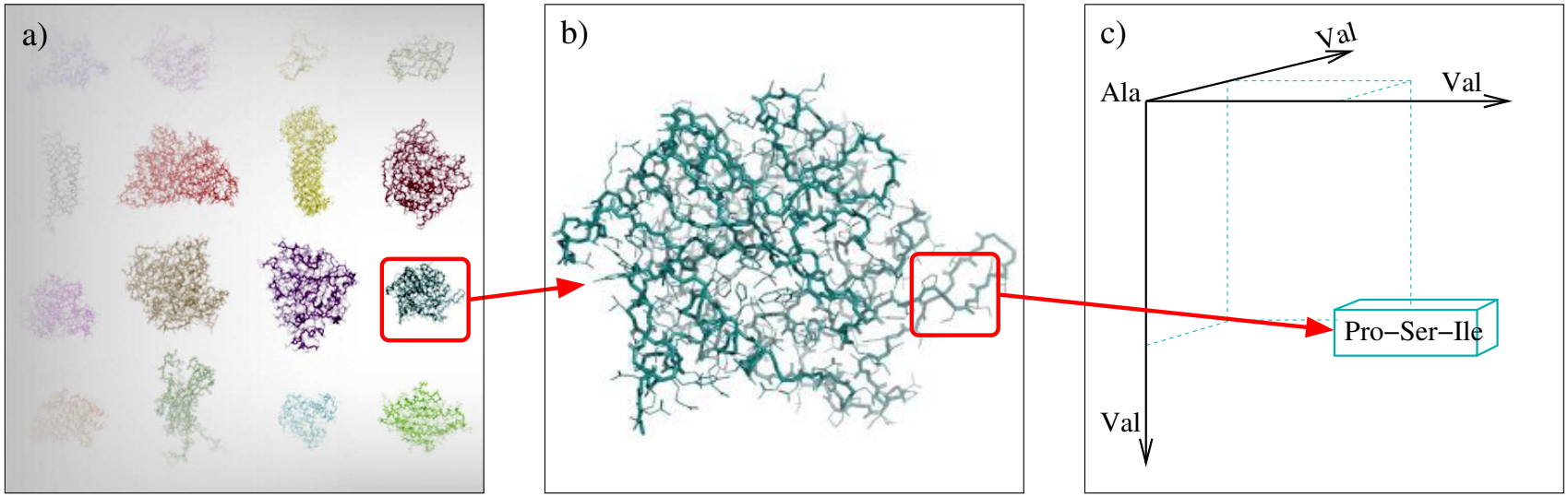


Local structure matters



Tripeptide conformations from experimentally-determined high-resolution protein structures

➡ Extracted from SCOPe 95% identity



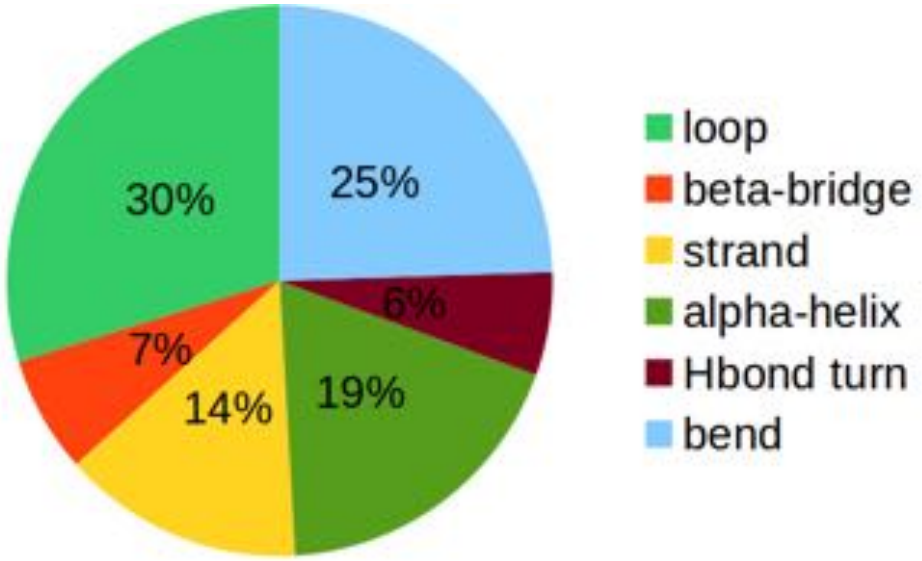
Tripeptide conformations from experimentally-determined high-resolution protein structures

➡ Extracted from SCOPe - 95% identity

81.749 scanned protein domains

Full database
tripeptides = 8.914.832

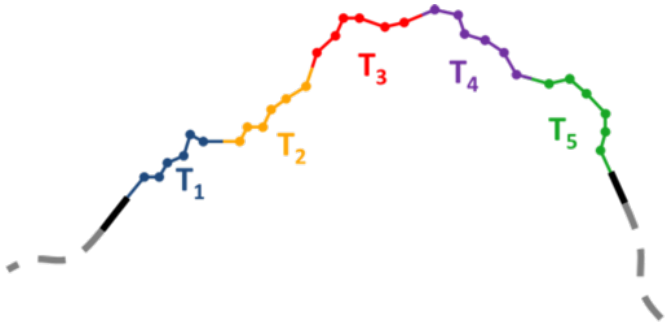
Coil database (LTS)
tripeptides = 3.645.381



[Barozet *et al.*, Bioinformatics 2019]

Main principle:

- Loop divided in 3-residue-long fragments (tripeptides)

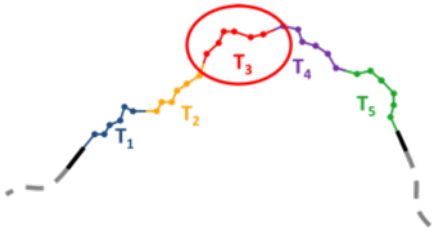


- Database of tripeptide fragments from known structures

[Barozet *et al.*, Bioinformatics 2019]

One iteration:

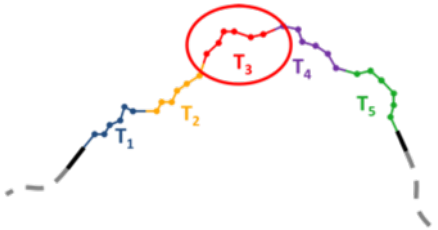
- Selection of a tripeptide for closure
- Iterative construction of the two loop ends, with backtracking



[Barozet *et al.*, Bioinformatics 2019]

One iteration:

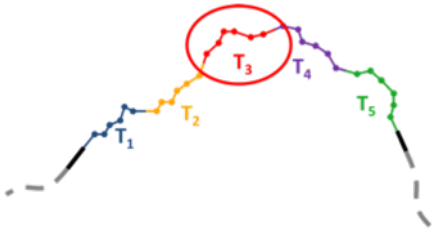
- Selection of a tripeptide for closure
- Iterative construction of the two loop ends, with backtracking

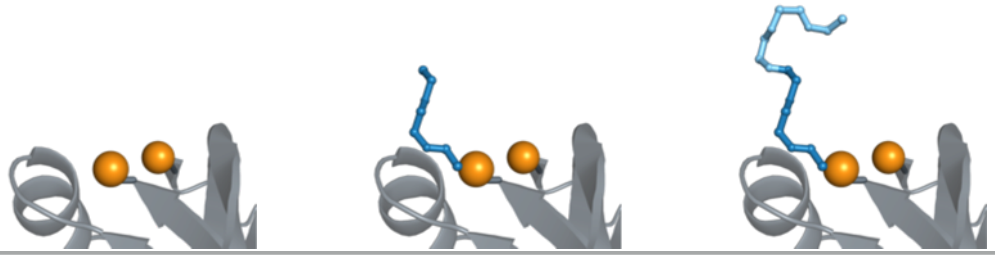


[Barozet *et al.*, Bioinformatics 2019]

One iteration:

- Selection of a tripeptide for closure
- Iterative construction of the two loop ends, with backtracking

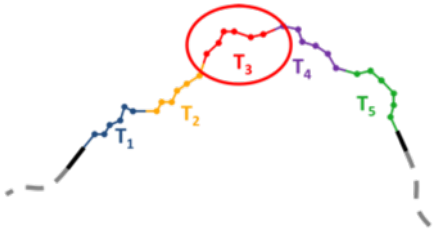


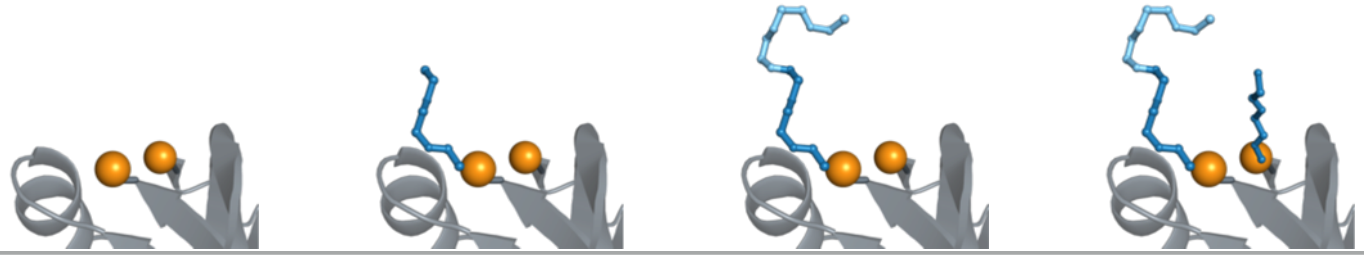


[Barozet *et al.*, Bioinformatics 2019]

One iteration:

- Selection of a tripeptide for closure
- Iterative construction of the two loop ends, with backtracking

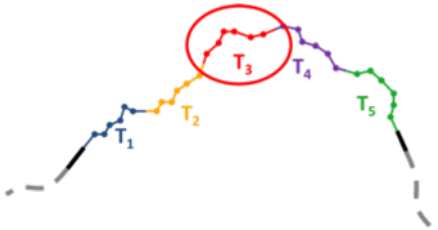


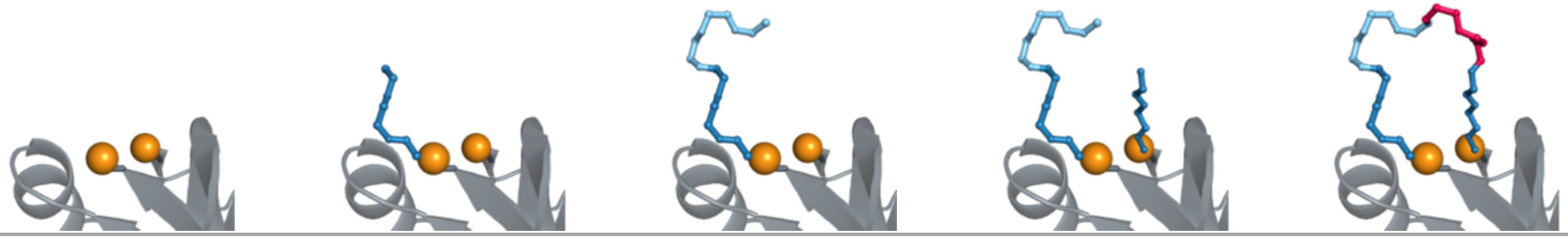


[Barozet *et al.*, Bioinformatics 2019]

One iteration:

- Selection of a tripeptide for closure
- Iterative construction of the two loop ends, with backtracking
- Loop closure using inverse kinematics (IK)



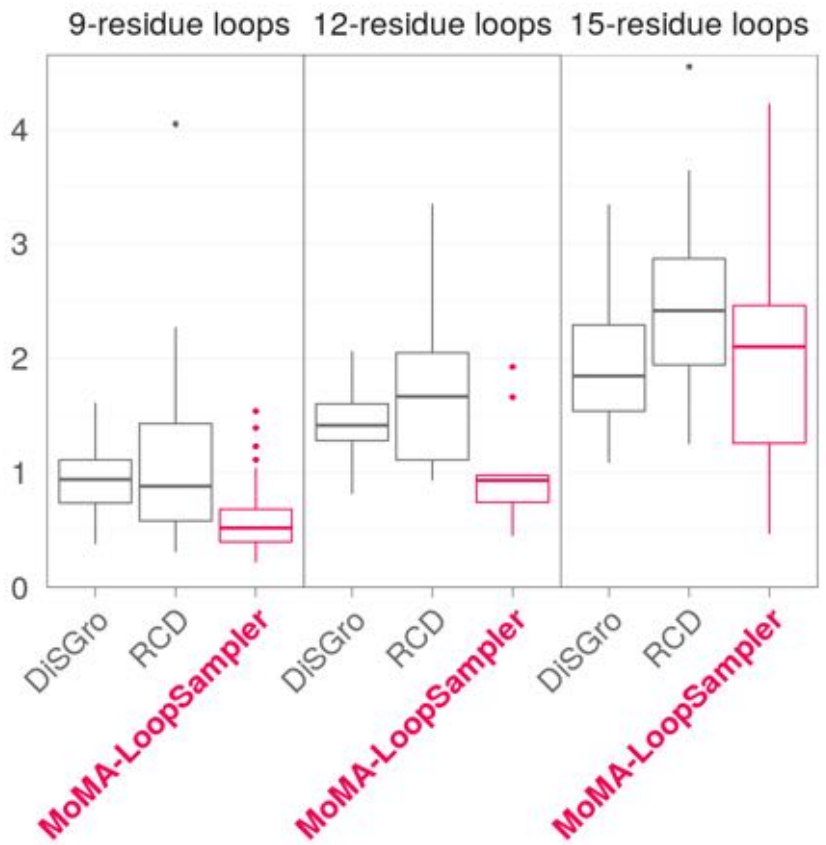


A machine-learning approach to protein loop sampling

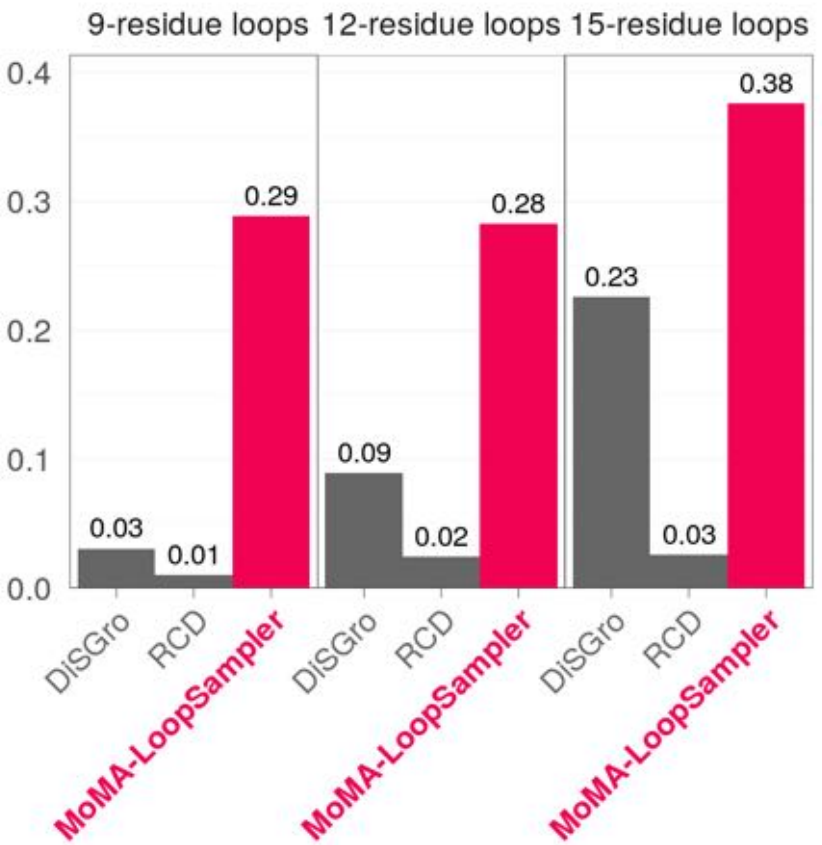
[Barozet *et al.*, Bioinformatics 2019]

MoMA-LoopSampler yields better results than state-of-the-art methods

Min RMSD to native (Å)



Median running time (s)

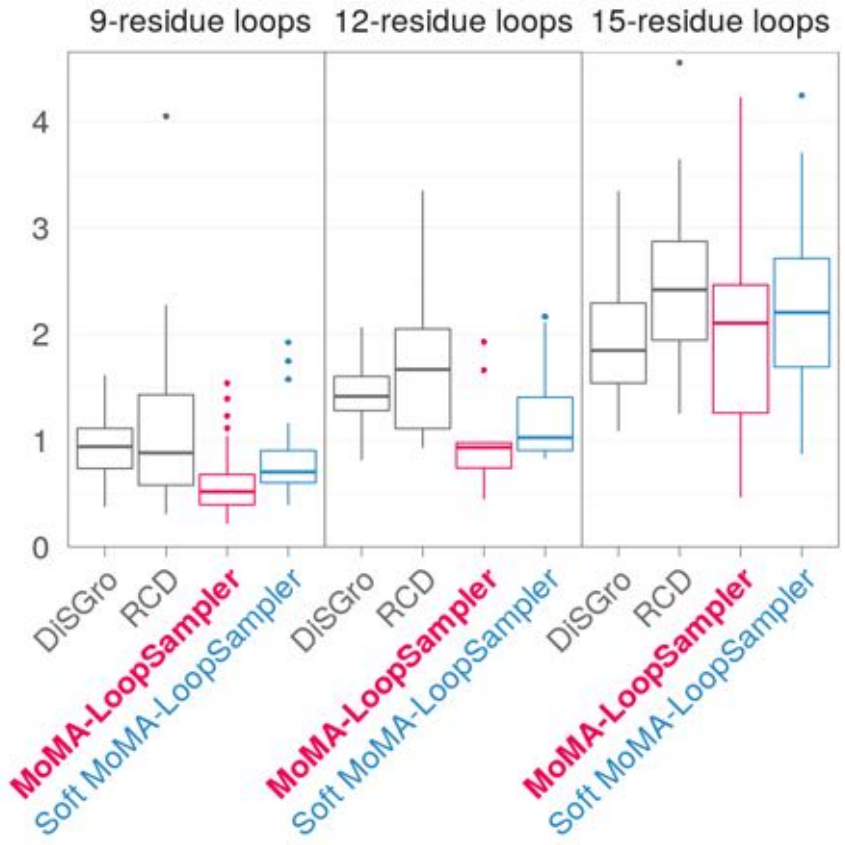


A machine-learning approach to protein loop sampling

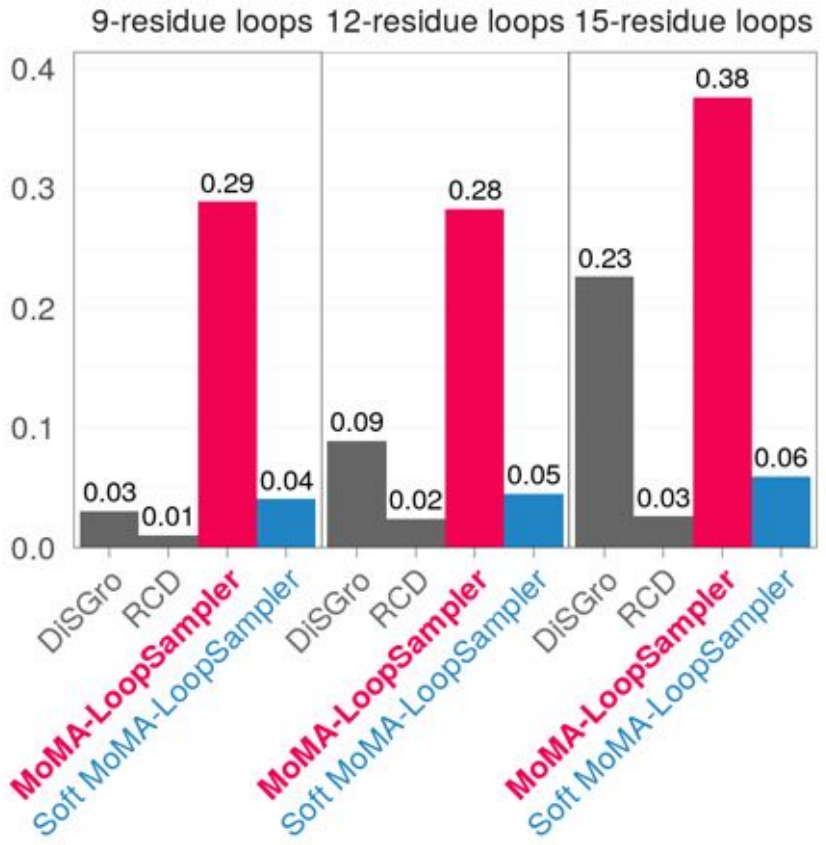
[Barozet *et al.*, Bioinformatics 2019]

Longer running times are due to the enforcement of stricter constraints

Min RMSD to native (Å)



Median running time (s)

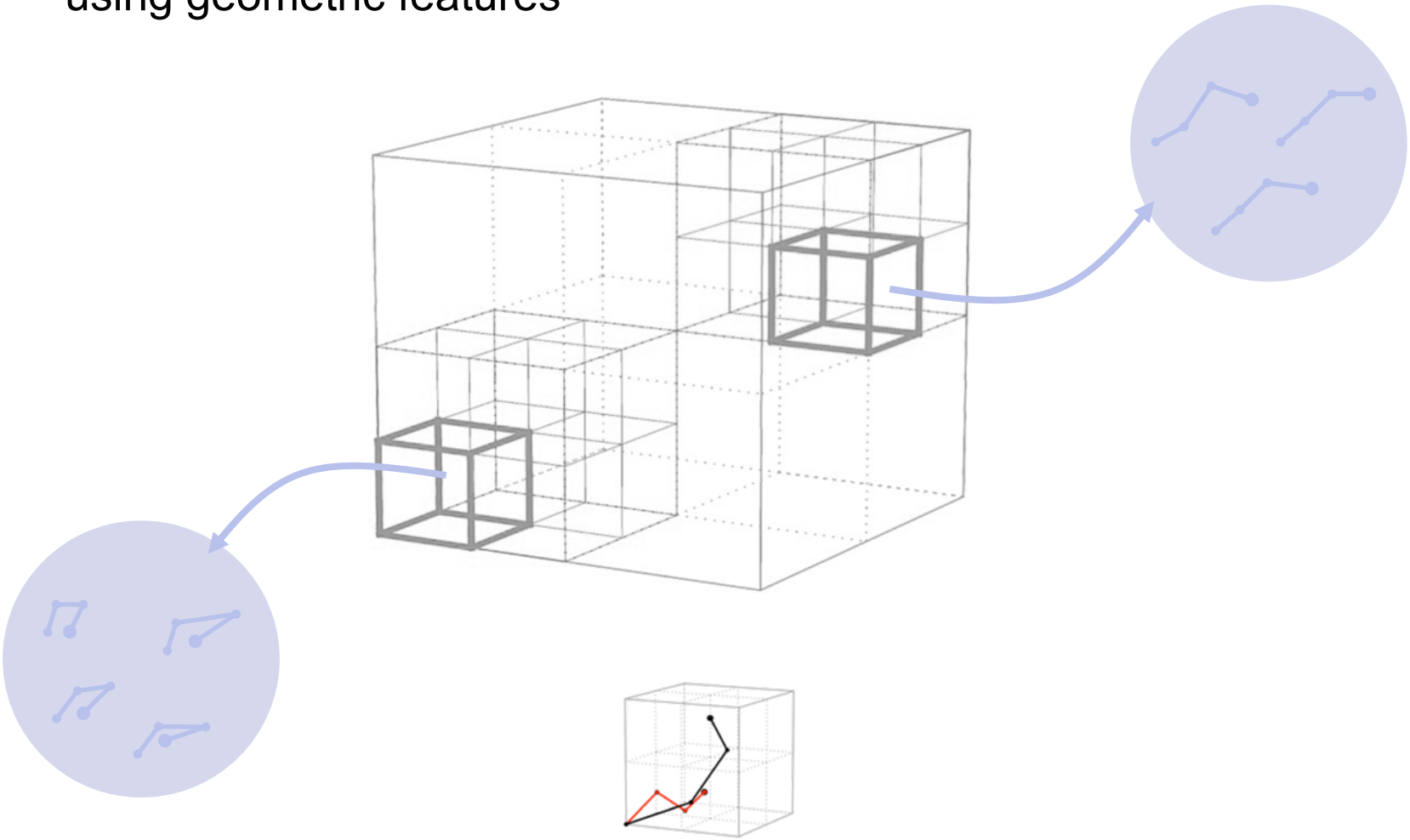


A machine-learning approach to protein loop sampling

[Barozet *et al.*, Bioinformatics 2019]

A **reinforcement-learning-based approach** to enhance sampling

- Learning data-structure based on hierarchical classification using geometric features

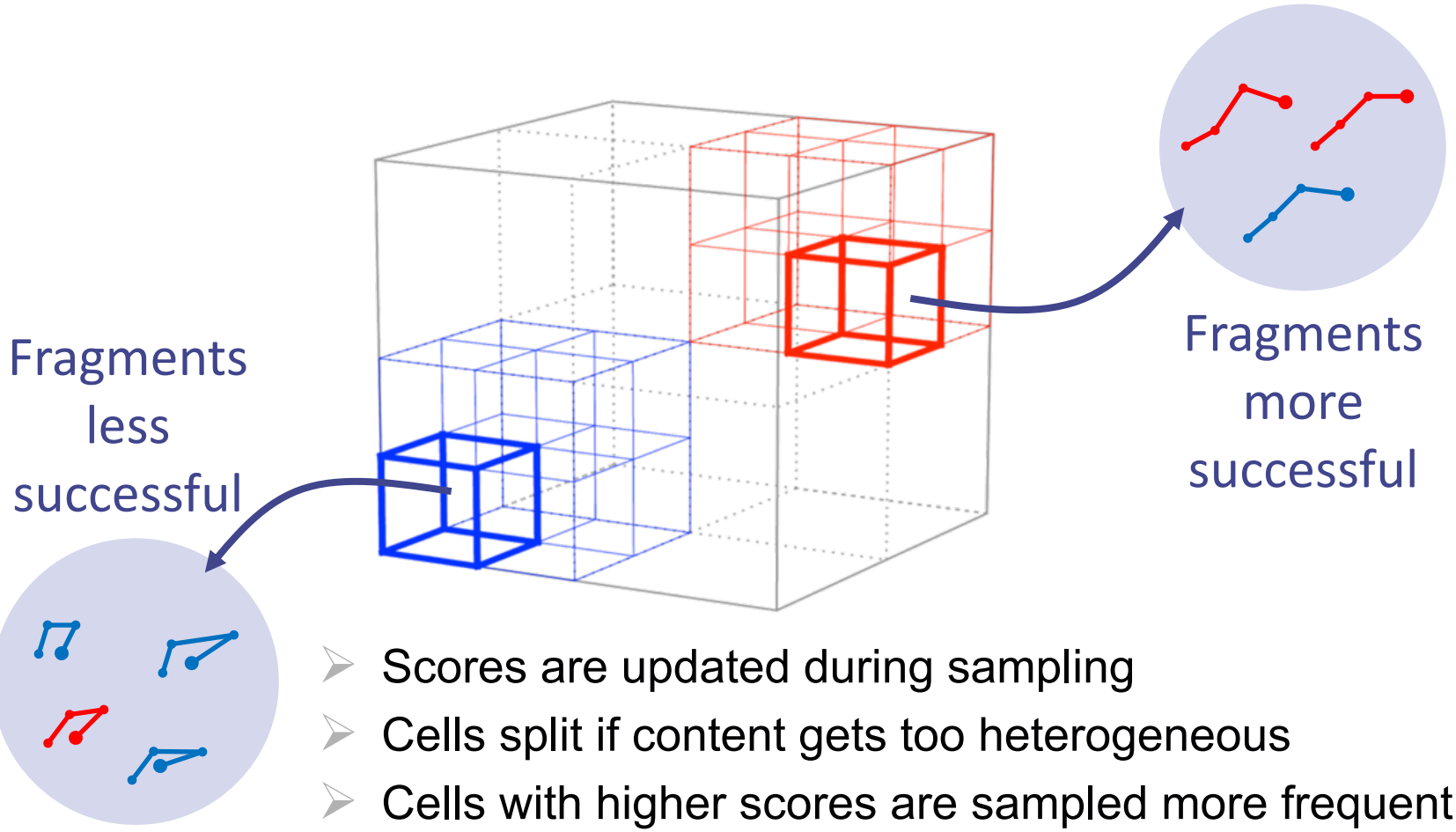


A machine-learning approach to protein loop sampling

[Barozet *et al.*, Bioinformatics 2019]

A **reinforcement-learning-based approach** to enhance sampling

- RL-based sampling heuristic



- Scores are updated during sampling
- Cells split if content gets too heterogeneous
- Cells with higher scores are sampled more frequently

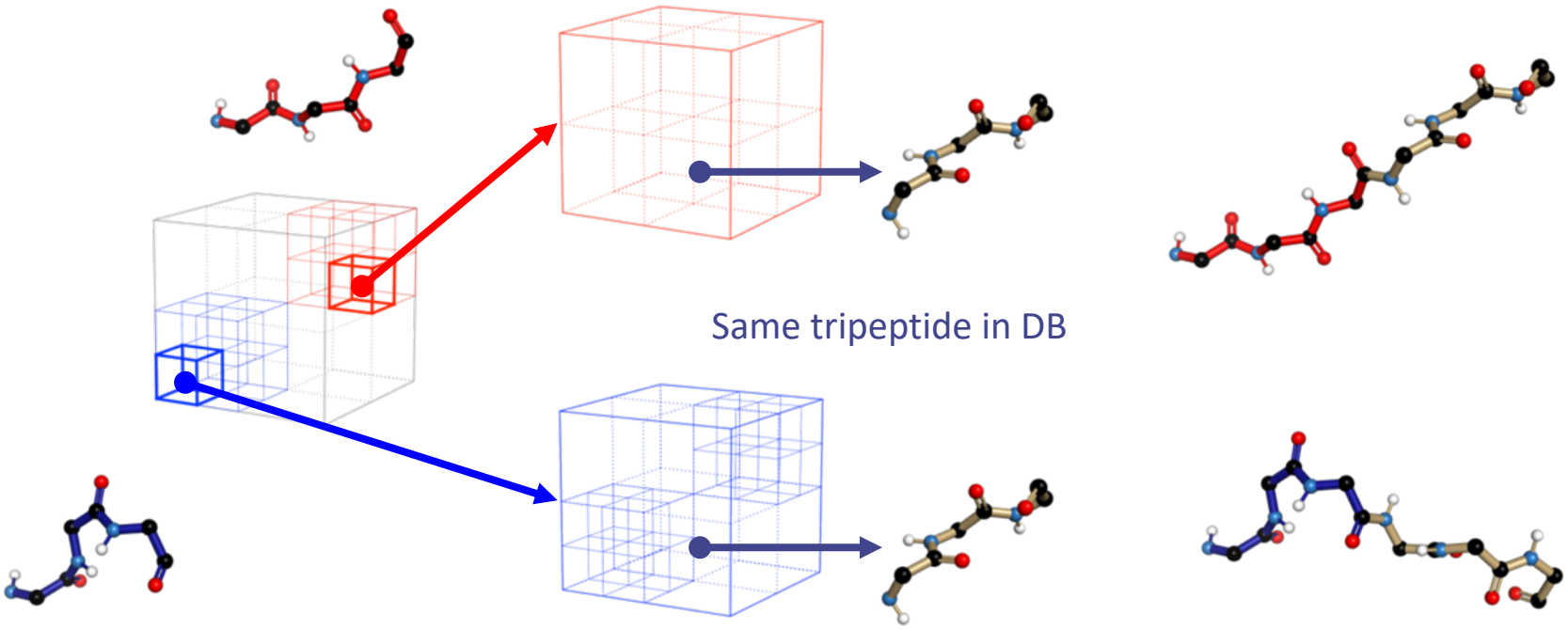
A machine-learning approach to protein loop sampling

[Barozet *et al.*, Bioinformatics 2019]

A **reinforcement-learning-based approach** to enhance sampling

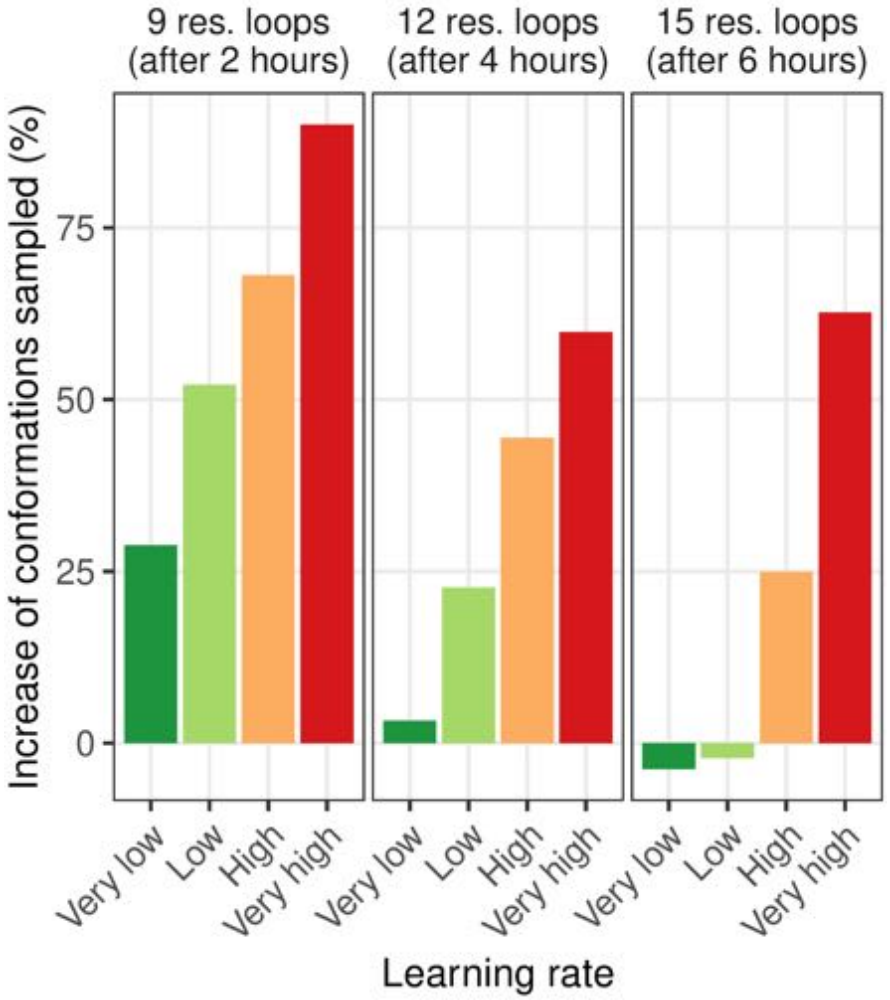
- The full learning structure involves chained trees

Tripeptide i	Tripeptide i+1	Concatenation



[Barozet *et al.*, Bioinformatics 2019]

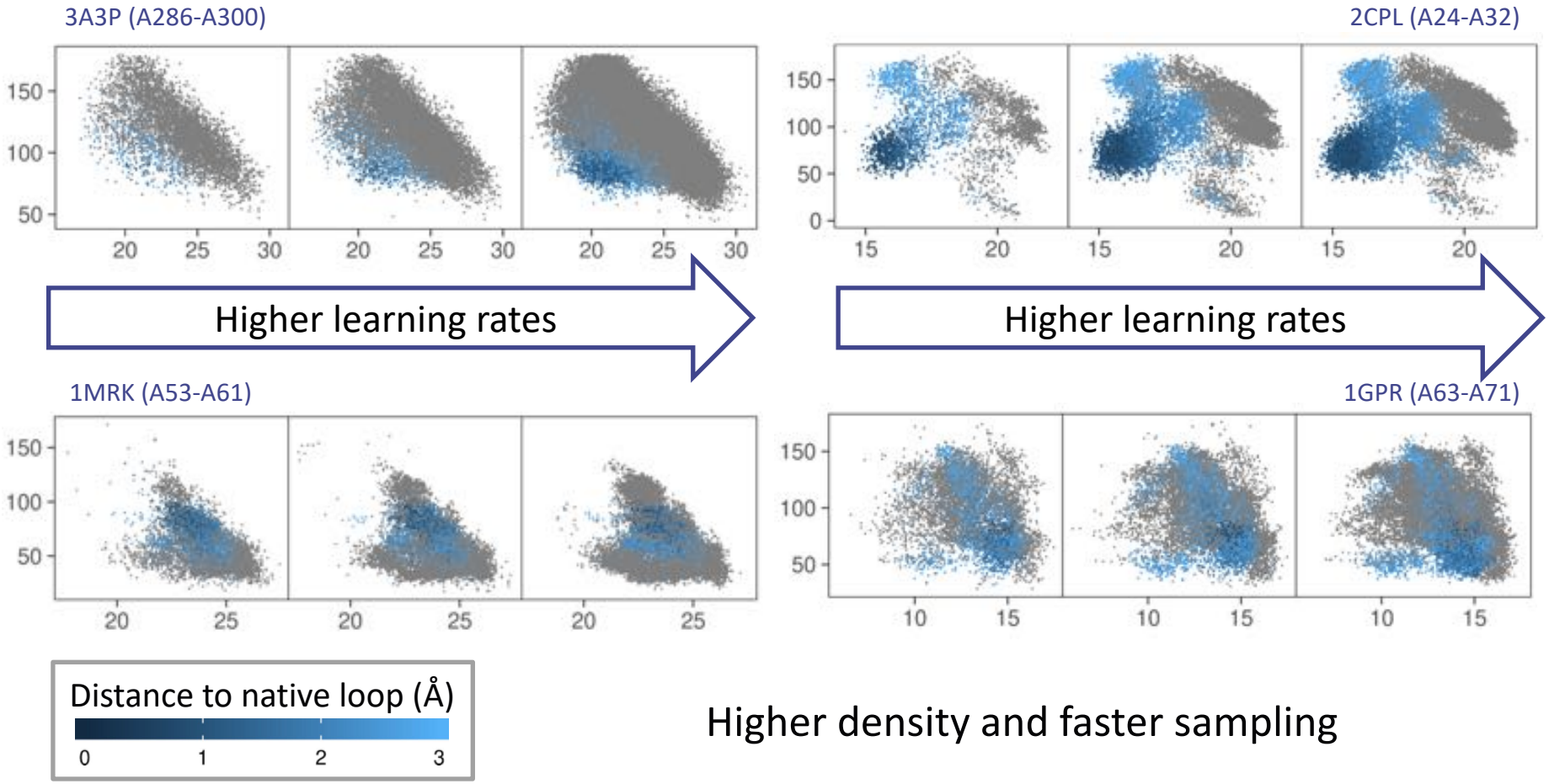
Reinforcement learning speeds up sampling



A machine-learning approach to protein loop sampling

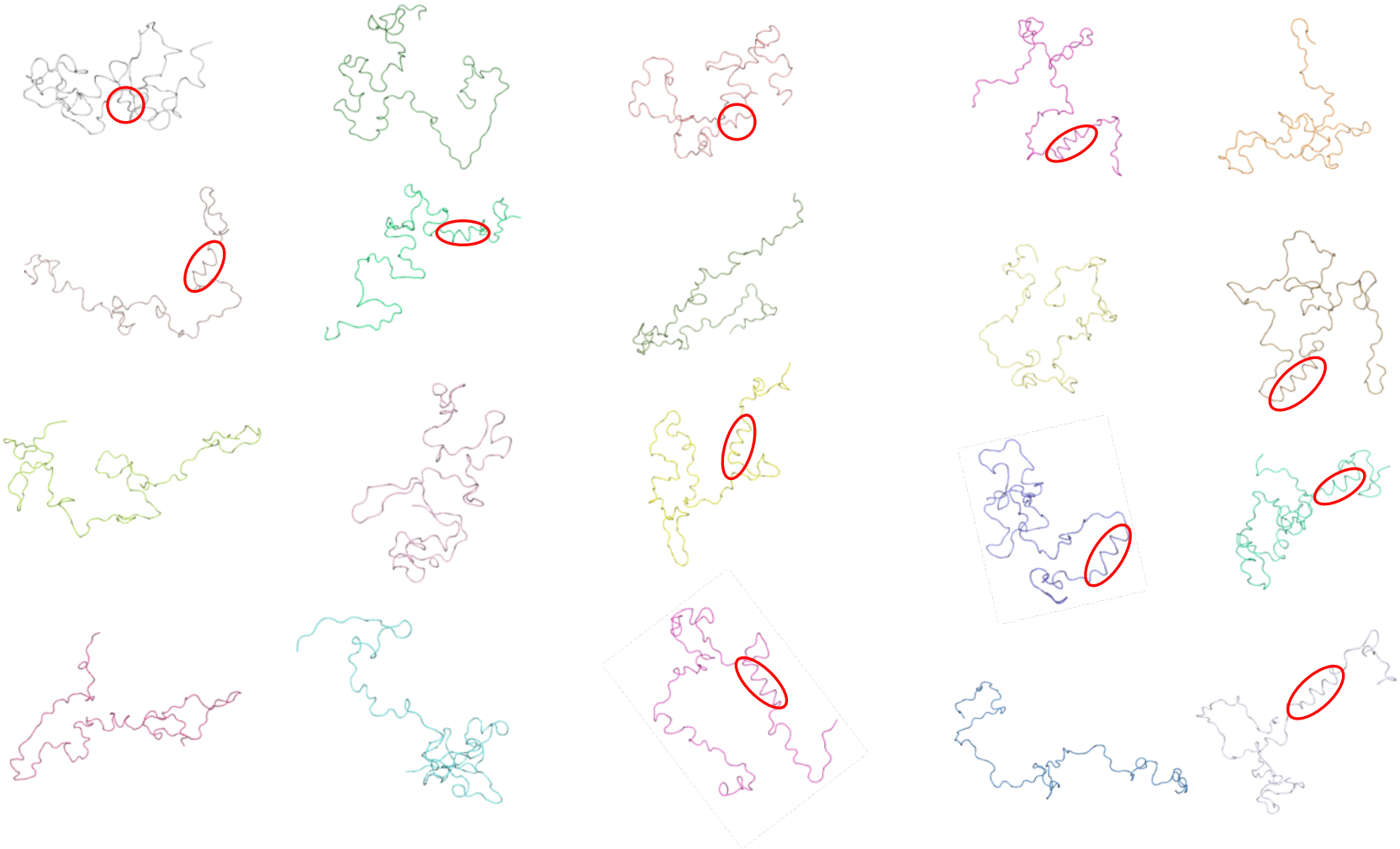
[Barozet *et al.*, Bioinformatics 2019]

Reinforcement learning preserves diversity

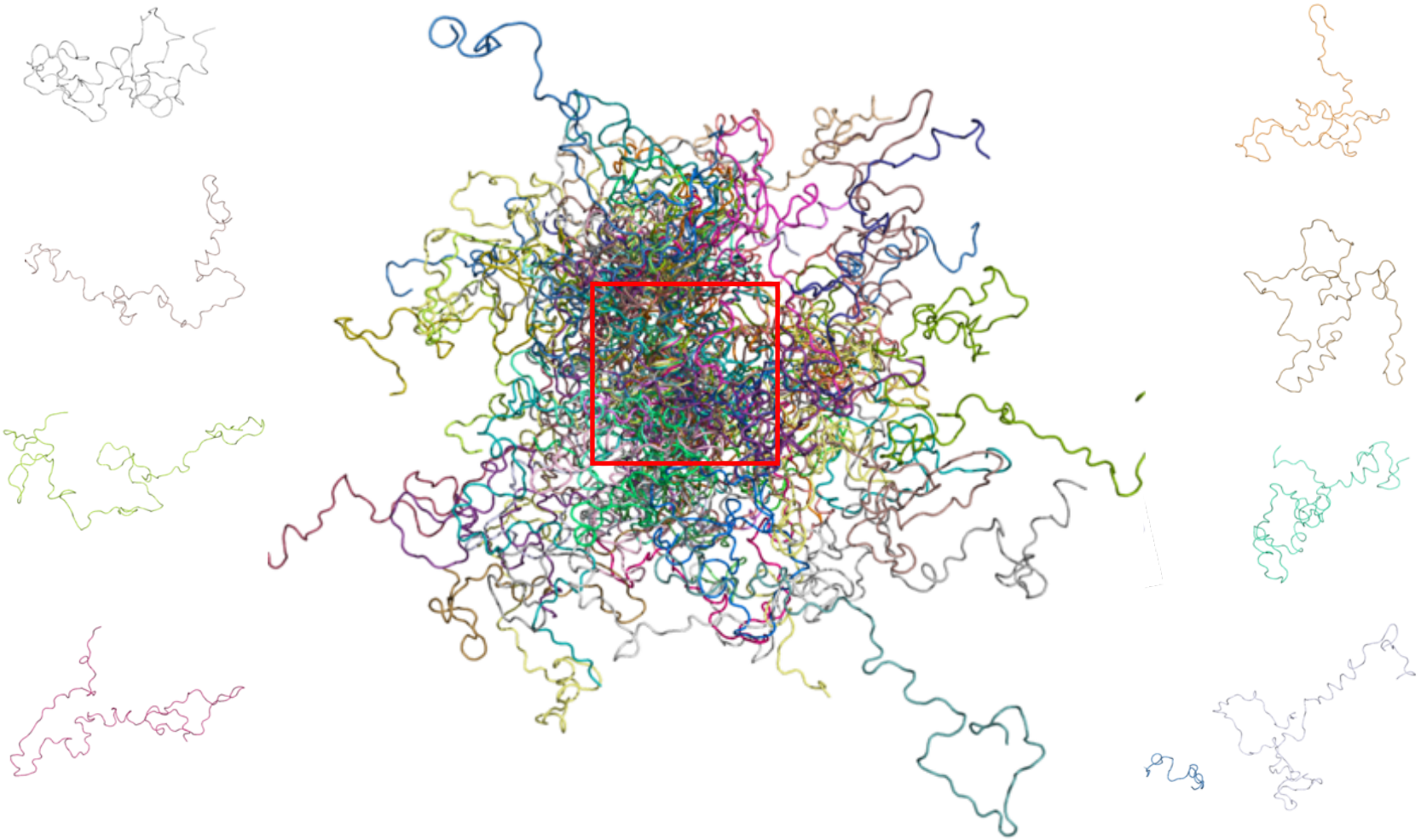


Higher density and faster sampling

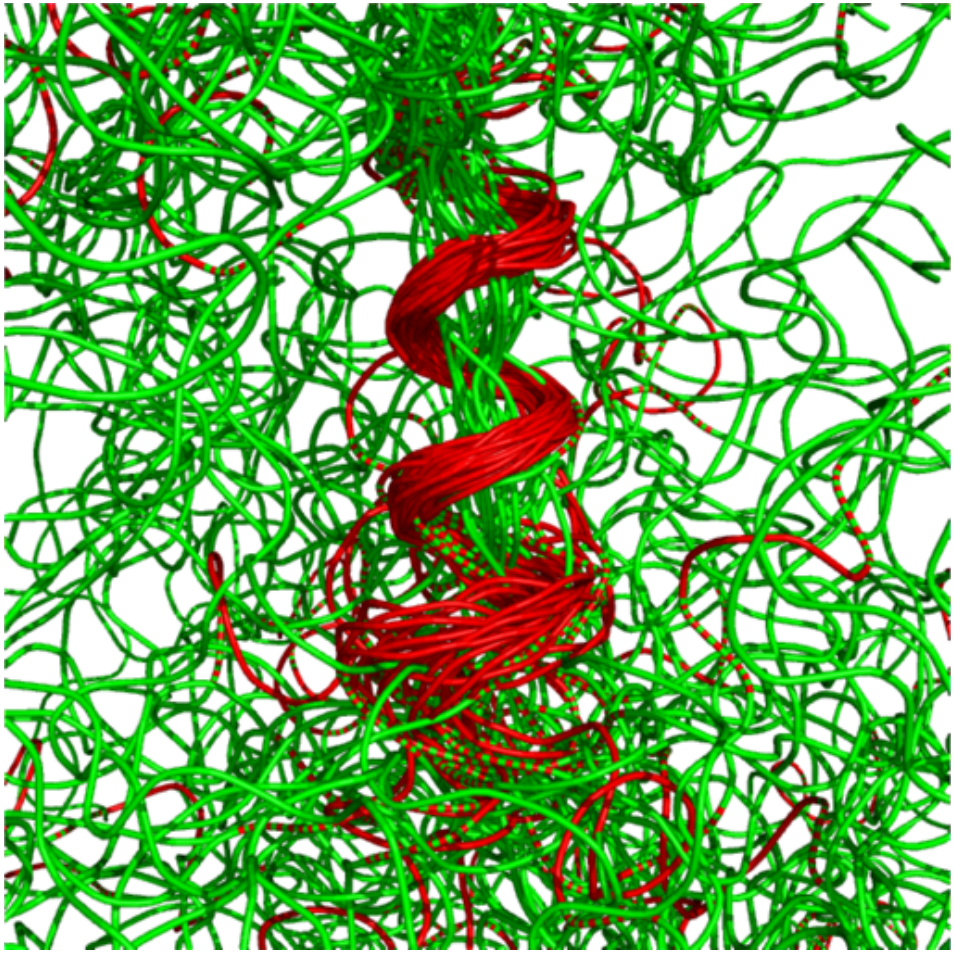
IDPs use to involve (partial) order inside structural disorder



IDPs use to involve (partial) order inside structural disorder



IDPs use to involve (partial) order inside structural disorder

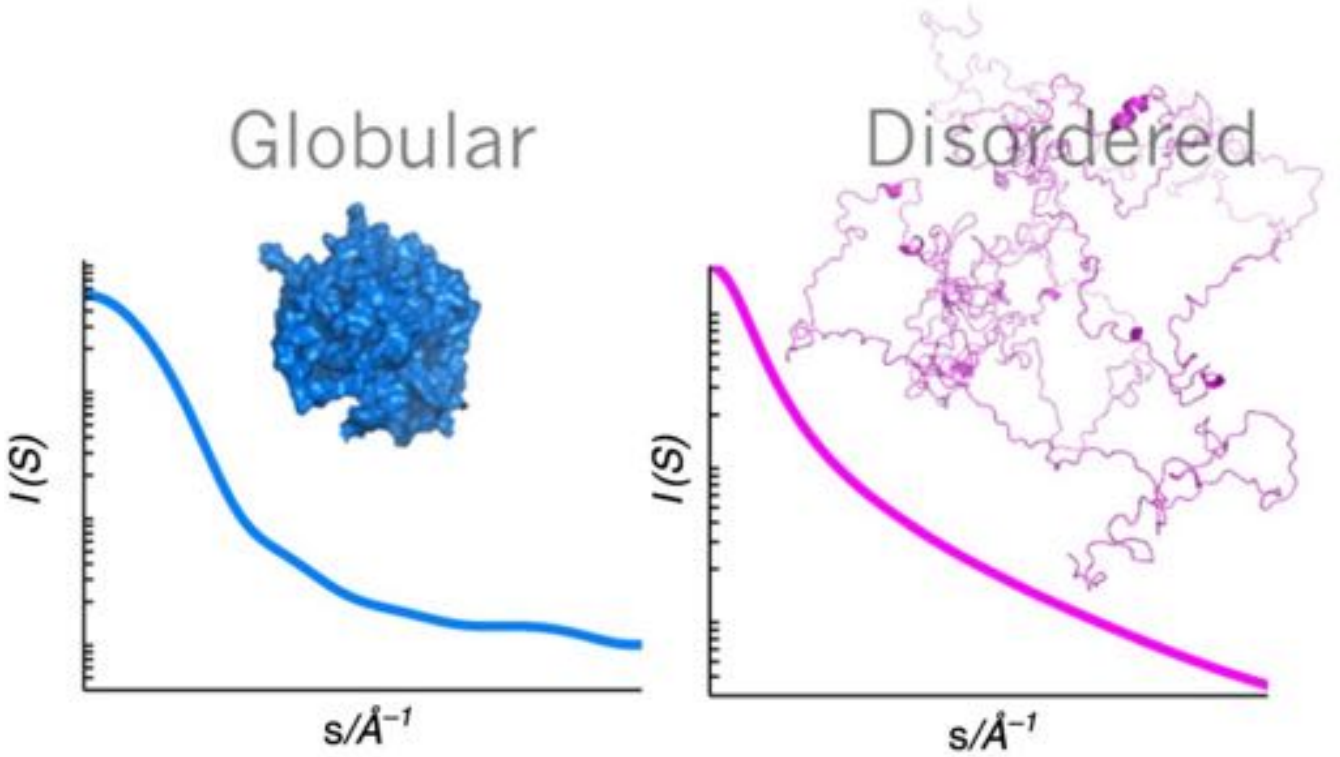


Collaboration with Pau Bernadó (CBS, Montpellier)

Using small angle X-ray scattering (**SAXS**)

[Cordeiro *et al*., *COSB*, 2017]

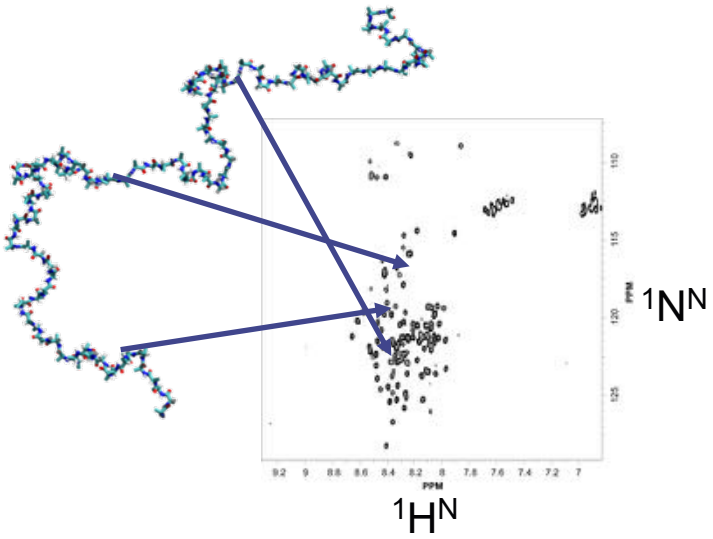
Information about overall shape



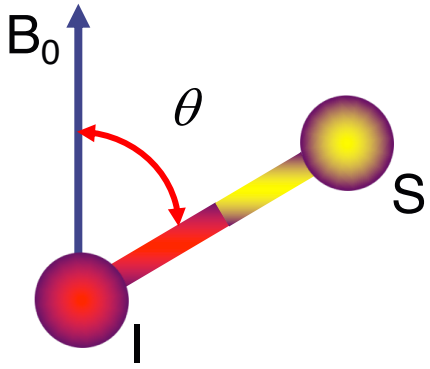
Collaboration with Pau Bernadó (CBS, Montpellier)

Using nuclear magnetic resonance (**NMR**)

Local information (at the single residue level)



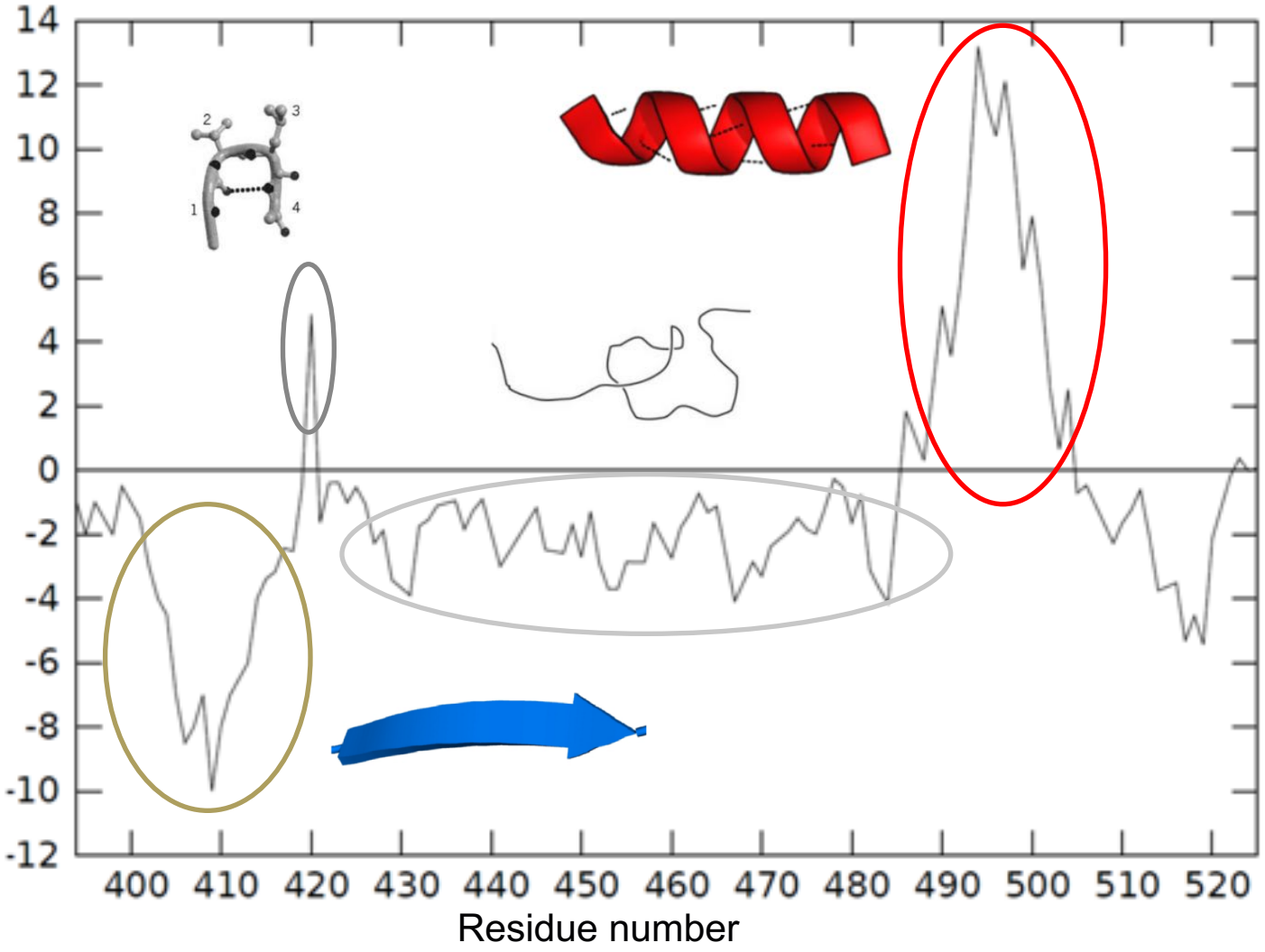
Each active nucleus can be identified



$$D_{IS}(\theta) = \frac{\hbar \gamma_I \gamma_S}{4\pi^2 r_{IS}^3}[1 - 3\cos^2\theta]$$

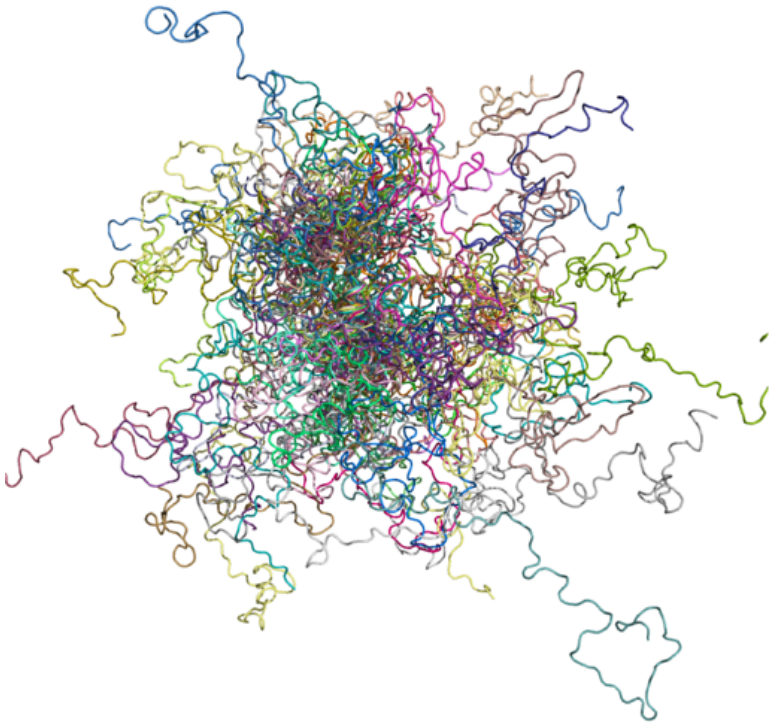
Residual dipolar couplings (**RDCs**) depend on the orientation of internuclear vectors relative to the alignment frame

RDC profiles provide relevant structural information

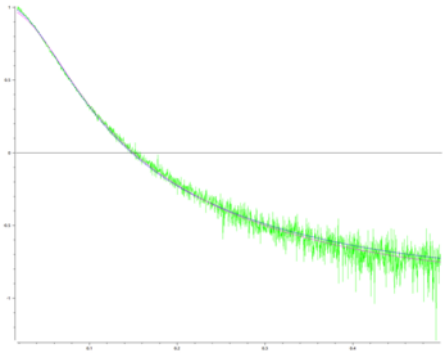


Computational methods are required to construct atomistic models fitting experimental data

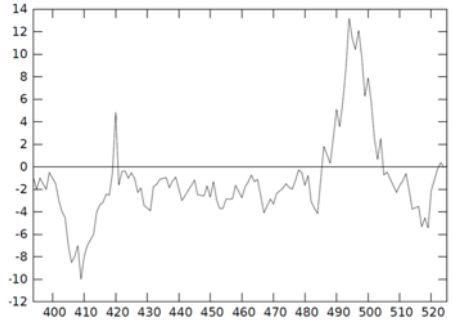
Conformational ensemble model



SAXS data

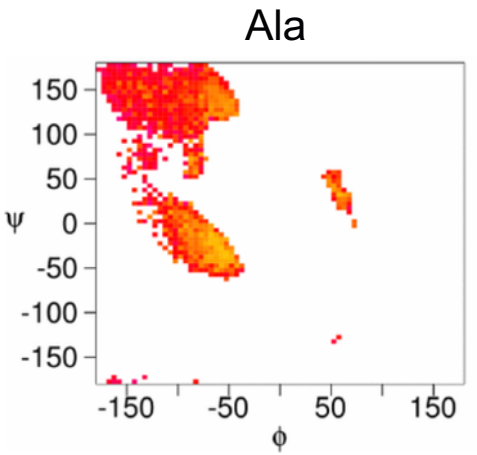


NMR data



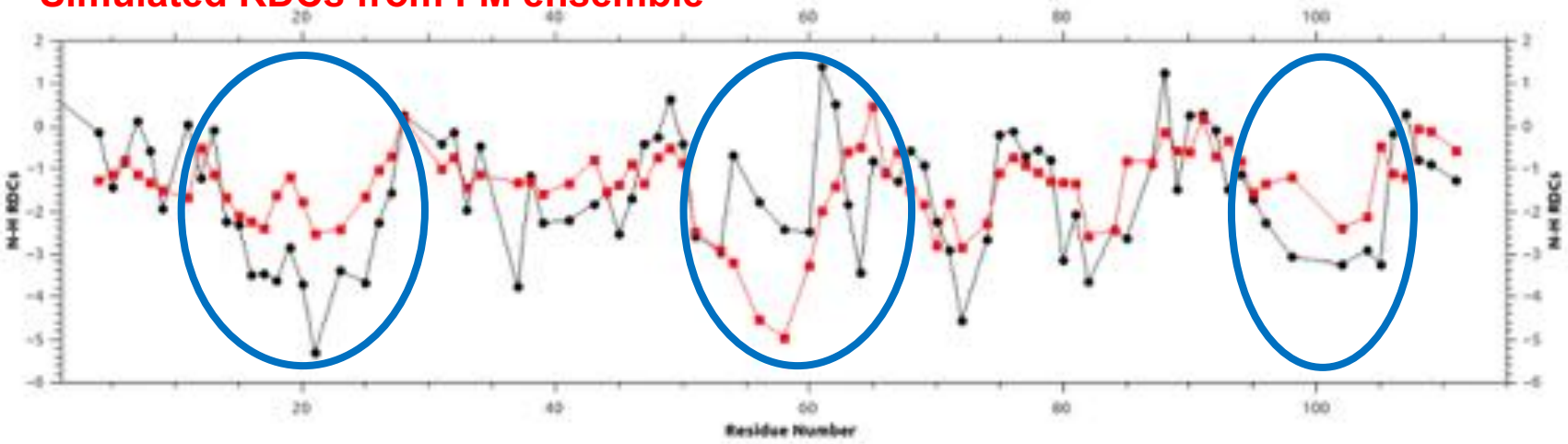
Flexible Meccano exploits statistical information at the **single residue level** to sample IDP's conformations

[Bernadó *et al.*, *PNAS*, 2005]
[Ozenne *et al.*, *Bioinformatics*, 2012]



- Accurate description of coil regions
- Fails to represent partially-structured regions

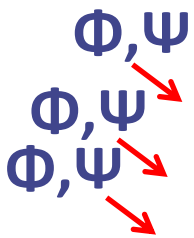
Experimental RDCs
Simulated RDCs from FM ensemble



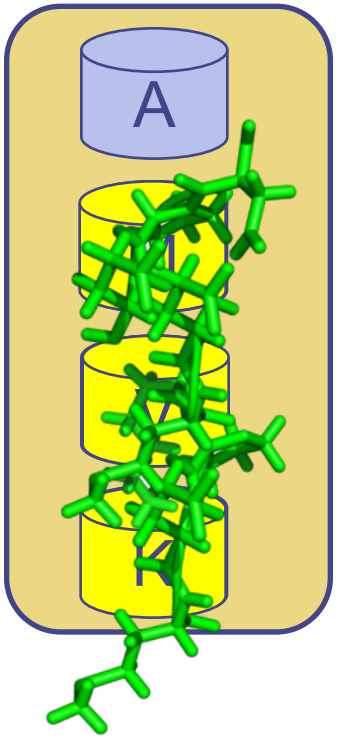
[Estaña *et al.*, *Structure*, 2019]

Single-residue-based sampling strategy (SRS)
(similar to Flexible-Meccano)

Protein sequence

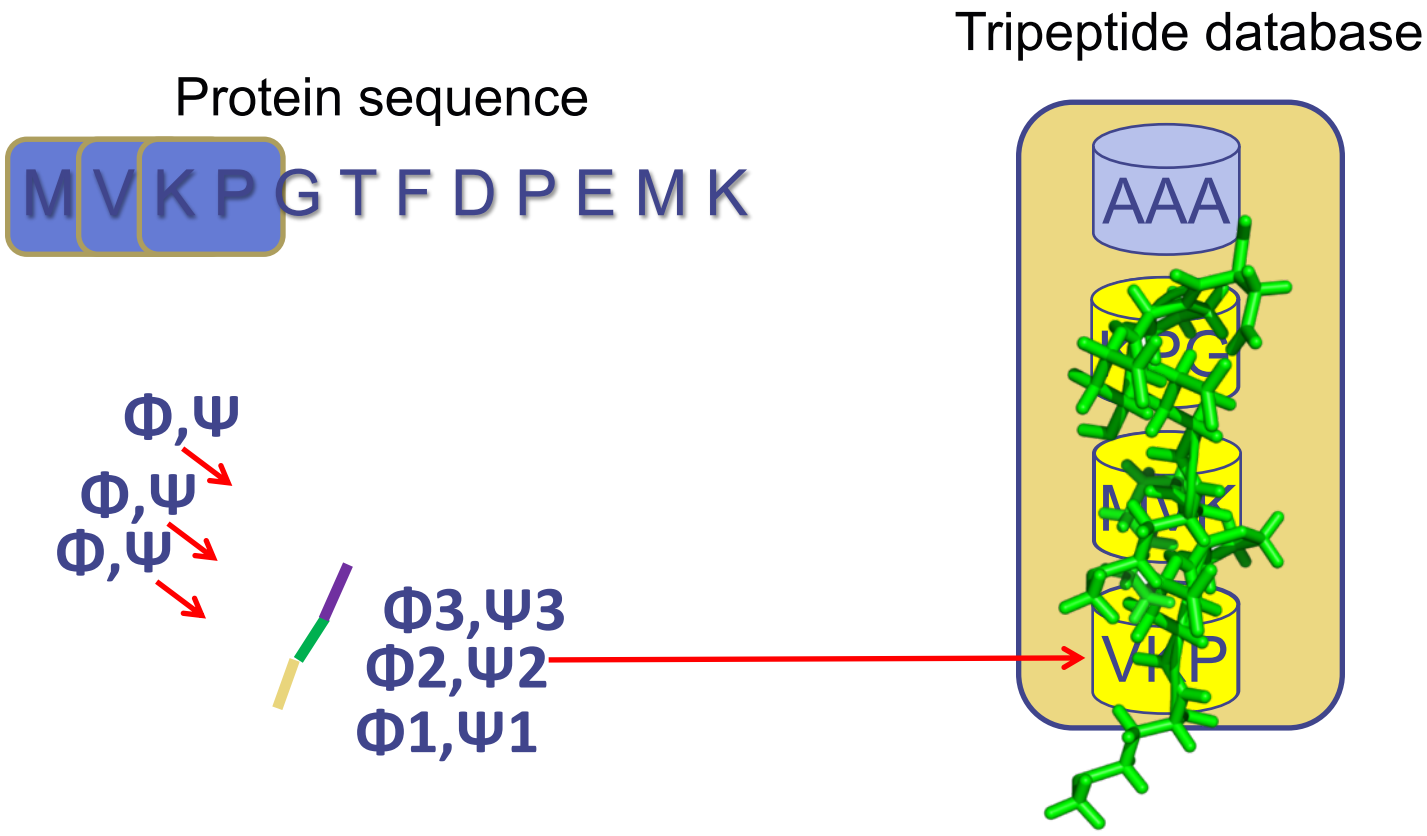


Tripeptide database

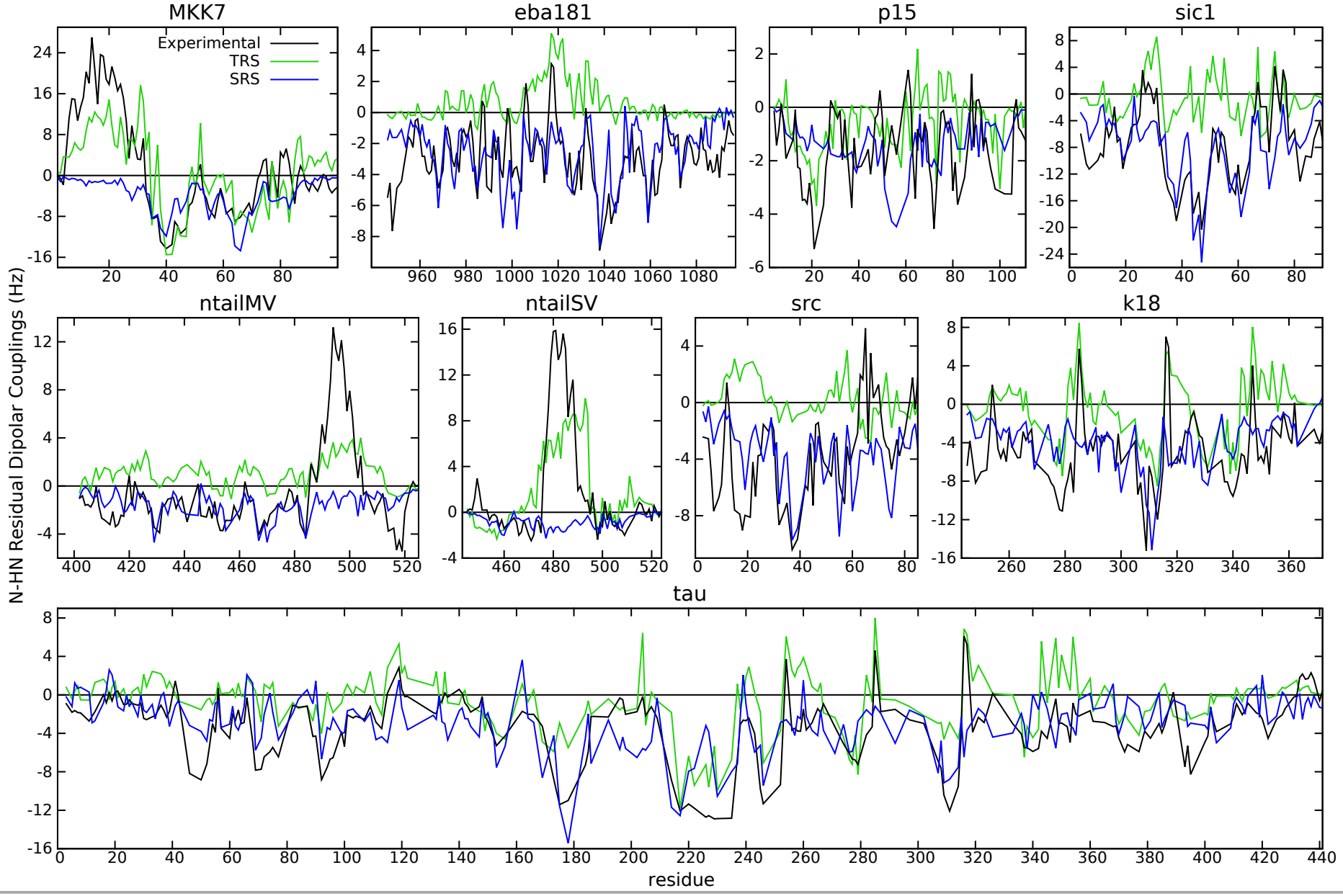


[Estaña *et al.*, *Structure*, 2019]

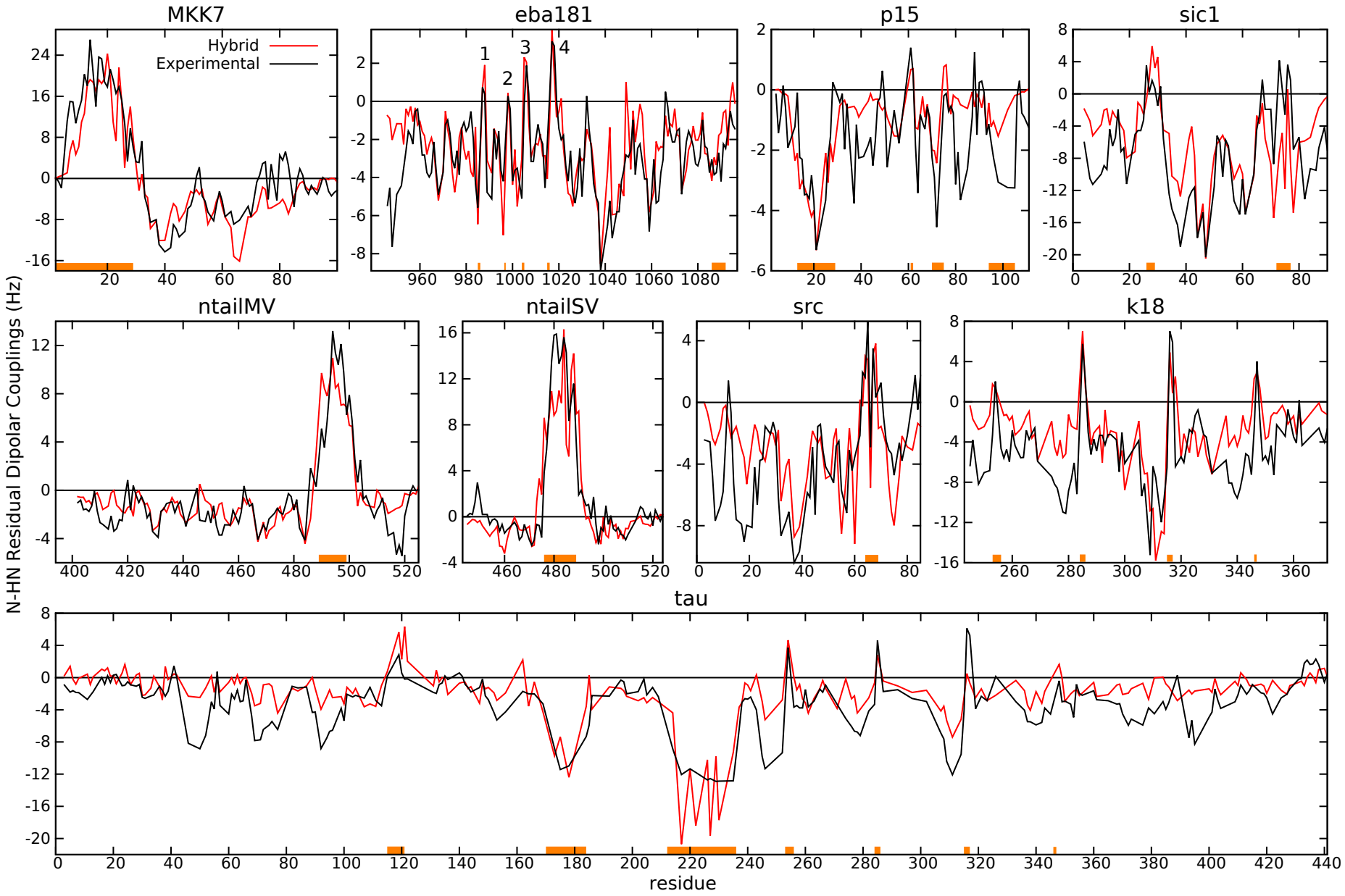
Three-residue-based sampling strategy (TRS)



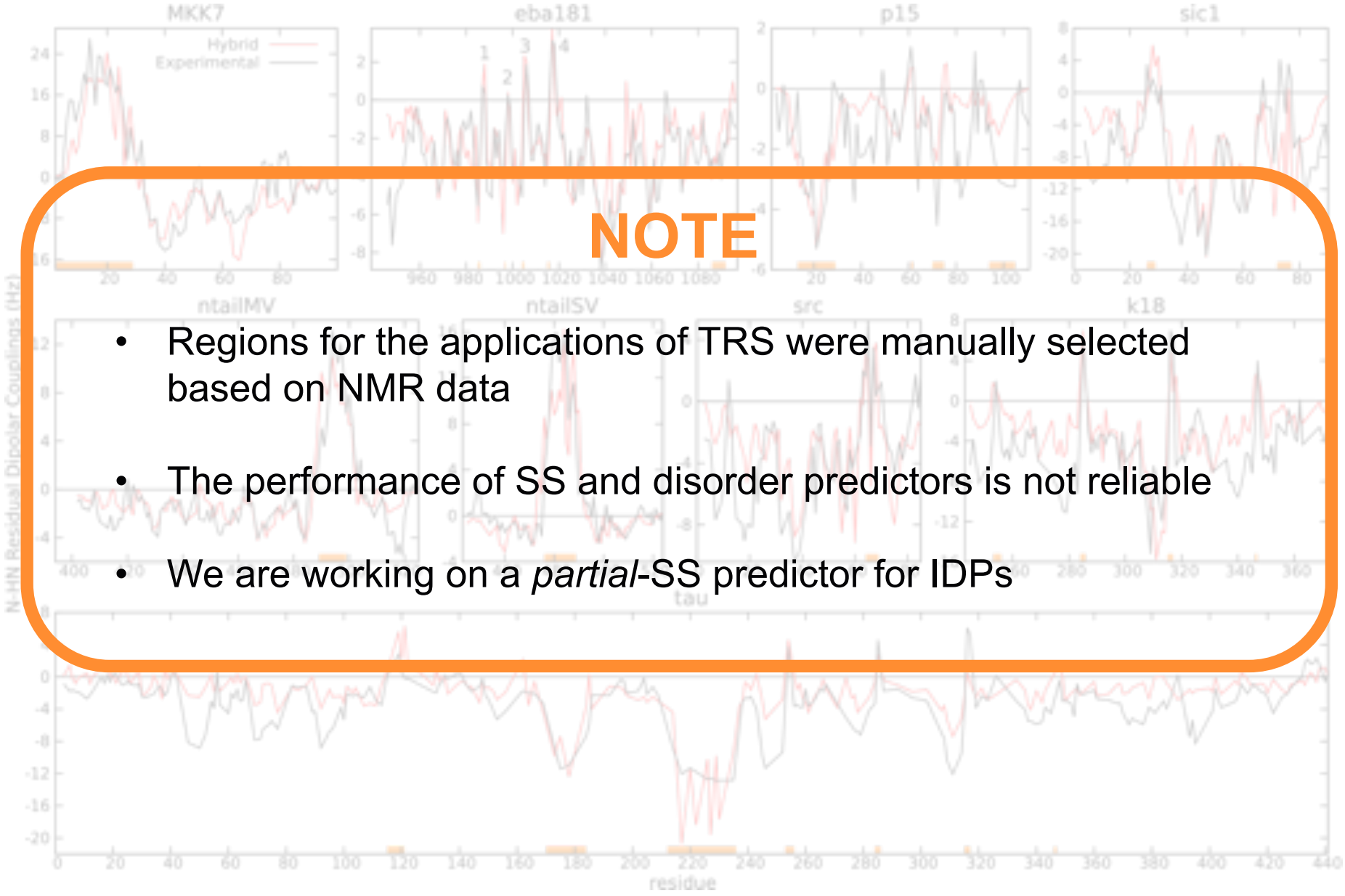
Results **SRS** vs **TRS** : RDC profiles



Results **Hybrid SRS-TRS** : RDC profiles



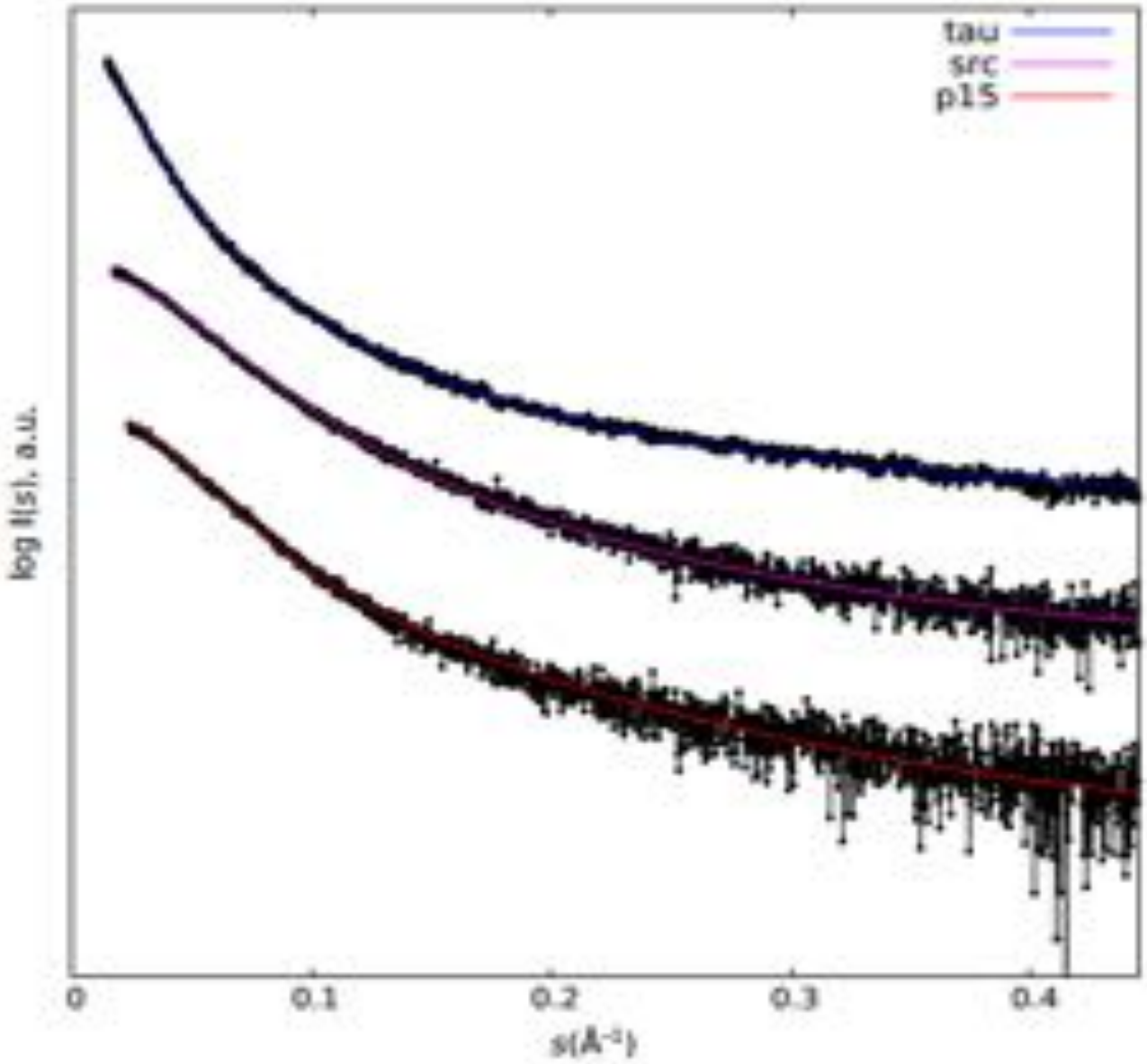
Results **Hybrid SRS-TRS** : RDC profiles



NOTE

- Regions for the applications of TRS were manually selected based on NMR data
- The performance of SS and disorder predictors is not reliable
- We are working on a *partial*-SS predictor for IDPs

Results **Hybrid SRS-TRS** : SAXS profiles

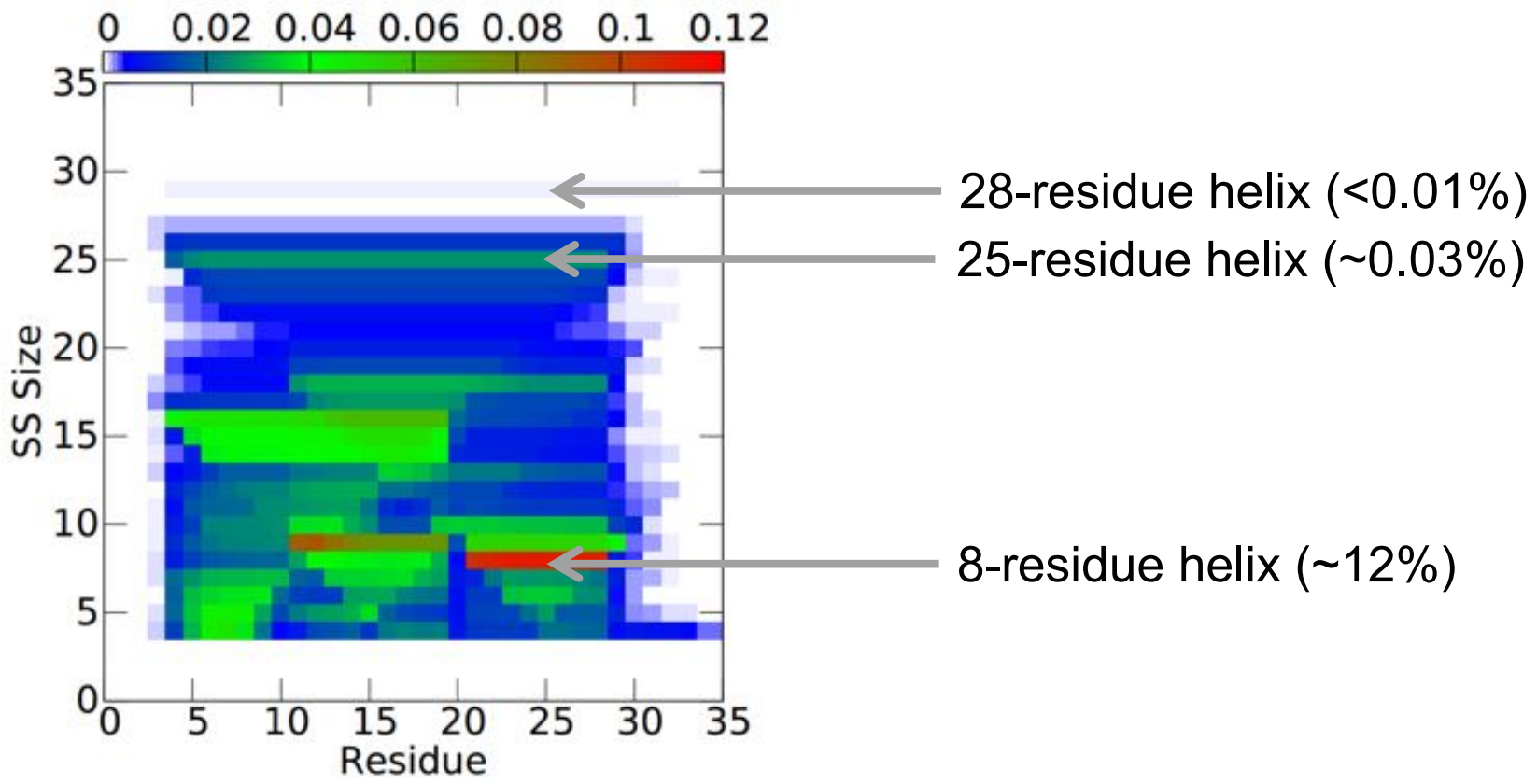


Characterization of local secondary structural elements

[Estaña *et al.*, *Structure*, 2019]

N-terminal region of MKK7:
Partially-helical region, involving from 4 up to 28 residues

Secondary-structure-map (SS-map) [Iglesias *et al.*, 2013]



To be continued …

- **Folding of structural elements**
  [Estaña *et al., Molecules* 2019]

- **Secondary structure propensity prediction**
  [Estaña *et al., in preparation*]

- **IDP design**
  ANR 2020 project proposal : LAAS, CBS, IMT

Questions ?

juan.cortes@laas.fr