Algorithms and Learning for Protein Science

Molecular kinematics, inverse problems, loop sampling

Frederic.Cazals@inria.fr

Overview

- Theory/algorithms
  - Exploring high dimensional spaces: RRT and HAR
  - Loop closure
- Protein science
  - Tripeptide/Triaxial loop closure
  - Loop sampling

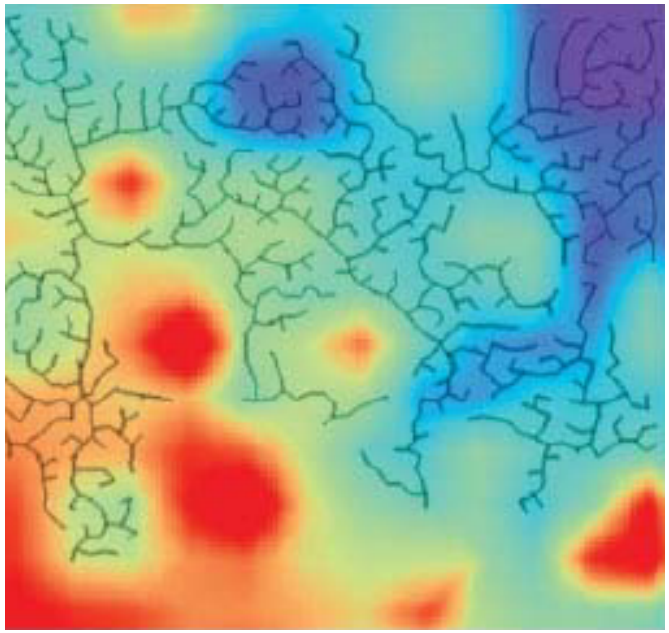
Loop sampling

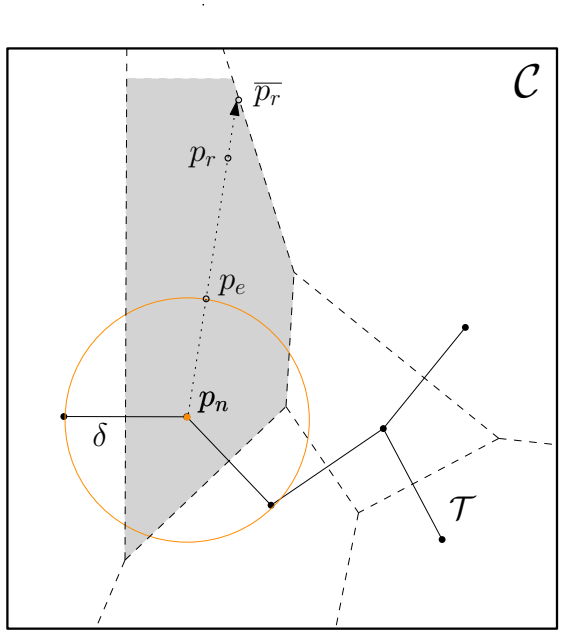
Exploring high dimensional spaces: two methods

Exploring Potential Energy Landscapes:

transition based rapidly exploring random trees (T-RRT)

- ▷ Goal: sample basins and transitions, avoiding trapping
- ▷ Algorithm growing a random tree favoring yet unexplored regions
  - node to be extended selection: *Voronoi* bias
  - node extension: interpolation + Metropolis criterion (+temperature tuning)
- ▷ Limitations: oblivious to local minima; not a Markov chain



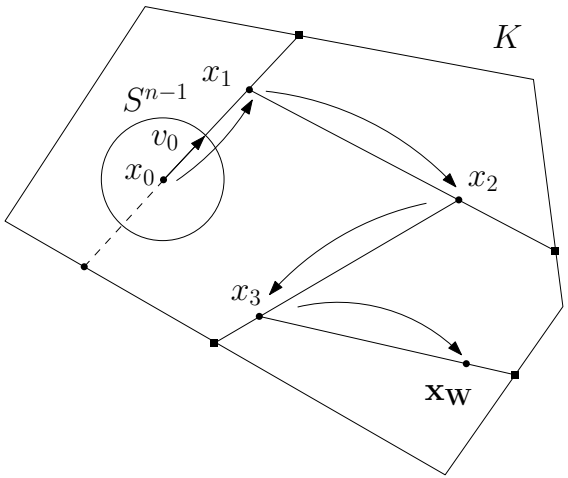


▷Ref: LaValle, Kuffner, IEEE ICRA 2000

▷Ref: Jaillet, Corcho, Pérez, Cortés, J. Comp. Chem, 2011

Random walk: hit-and-run

- ▷ Goal: sample point in K according to a prescribed density f in a polytope K
- ▷ (Random-direction) hit-and-run: random point x_W after W steps



- ▷ Iteratively:
  - pick a random vector
  - move to random point on the chord $l \cap K$, chosen from the distribution induced by f on l
- ▷ Comments:
  - risk of being trapped near a vertex
  - large W helps *forgetting* the origin x_0

▷ Thm (Berbee et al) The limit distribution induced by HR is uniform in K.

▷ Thm (Vempala et al) HR can be modified to sample an isotropic Gaussian (restricted to K).

▷ Thm (Lovász) Let r and R denote the radii of the largest inscribed and circumscribed balls for K. One sample generation: $O^\star(d^3)$.

▷Ref: Berbee et al, Math. Prog., 1987

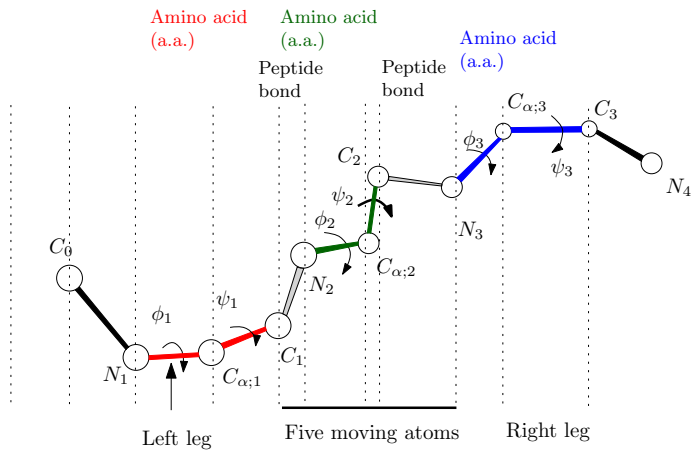
▷Ref: Lovász, Math. Prog. Ser. A, 1999

▷Ref: Lovász, Vempala, SIAM J Comp., 2006

Loop sampling

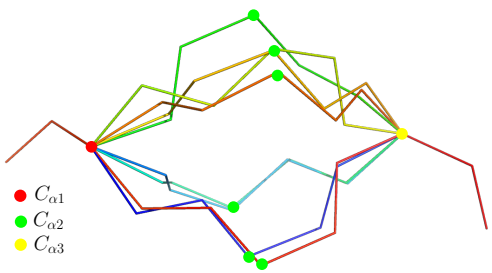
Tripeptide Loop Closure – TLC

- TLC: for 3 amino acids, fix all internal coordinates BUT the $(\phi_i, \psi_i)_{i=1,2,3}$ angles

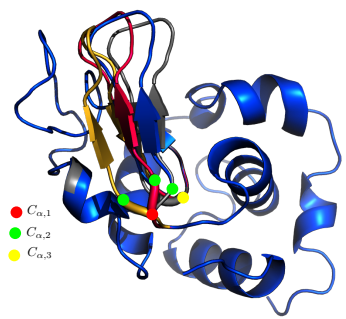


$\Rightarrow$ Find all possible values of the six angles $(\phi_i, \psi_i)_{i=1,2,3}$ compatible with the remaining fixed internal coordinates (bond lengths, valence angles, ω_i)

- Theorem: at most 16 solutions $\leftrightarrow$ real roots of a degree 16 polynomial
- The three amino acids may not be consecutive



3 consecutive a.a.

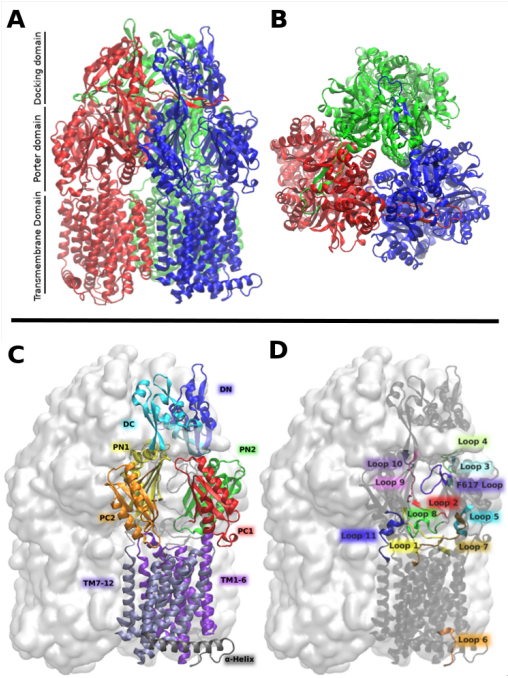


3 a.a. sandwiching SSE–CDRs

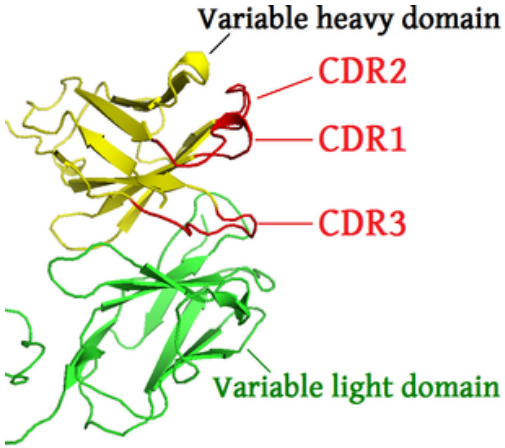
- Ref: Go and Scheraga, Macromolecules, 1970
- Ref: Coutsias et al, J. Comp. Chem., 2004

Loops: biological relevance and dynamics

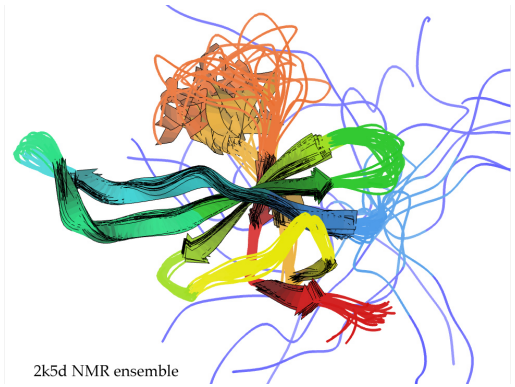
- Loops in biological processes



AcrB/RND



CDRs



IDP (2k5d/NMR)

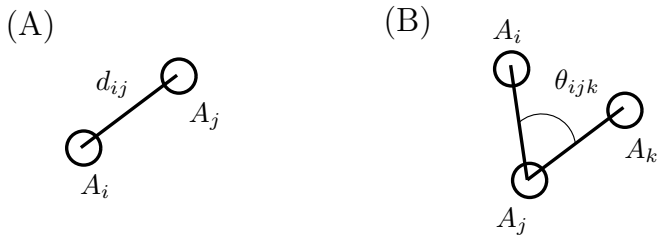
- Action modes
  - (Structure) Global dynamics: global motions of domains
  - (Thermodynamics) Localized dynamics of CDR in antibodies (binding affinity)
  - (Mix) IDP and more generally highly flexible regions

- Open problems: accurate predictions for structure / thermodynamics / kinetics

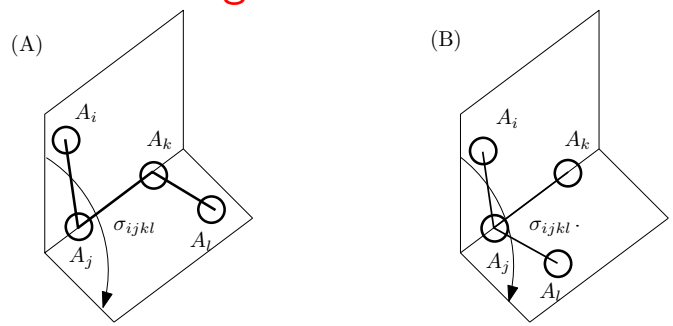
Geometric models: Cartesian and internal coordinates

▷ Cartesian versus internal coordinates: $\{x_i y_i z_i\}_i$ versus $\{d_{ij}, \theta_{ijk}, \sigma_{ijkl}\}$

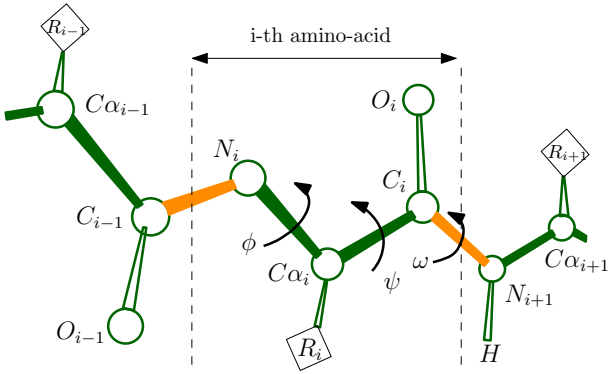
▷ Bond length and valence angle



▷ Dihedral angles



▷ Protein backbone

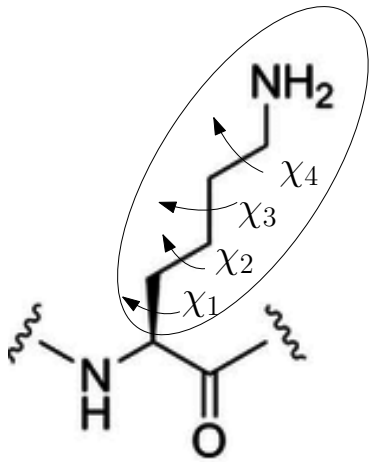


Ramachandran diagram, per a.a. type:

- bivariate distribution for (ϕ, ψ)

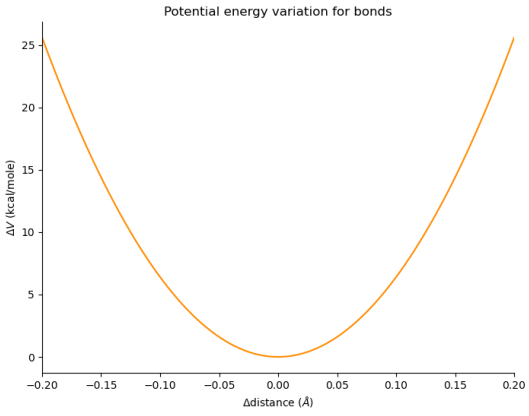
▷ Side chain: 20 natural amino acids

Exple: Lysine, 4 dihedral angles

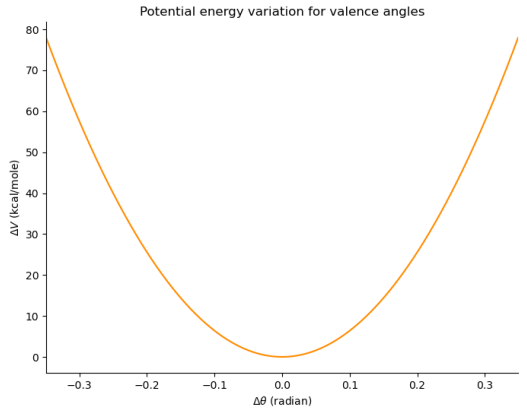


LYS

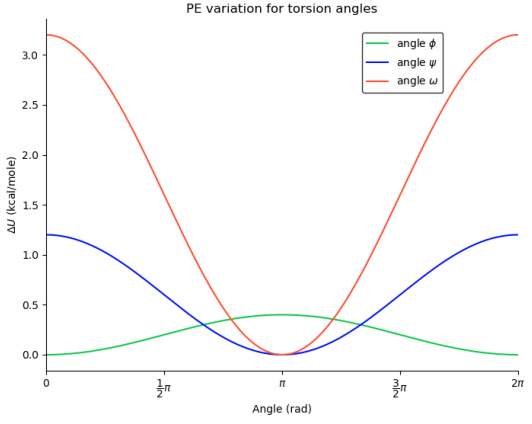
Softness of Internal coordinates –force constants from CHARMM 36



Bonds: $\delta d_{ij} \sim .2\text{Å} : \Delta V \sim 20 kcal/mol$



Valence angles: $\delta\theta_{ij} \sim 10^\circ : \Delta V \sim 20 kcal/mol$



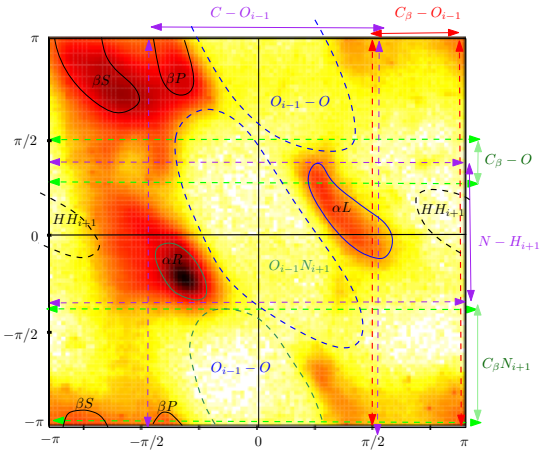
Torsion angles:$\Delta V \sim 3 - 4 kcal/mol$

Dihedral angles:

- are indeed *soft* coordinates, but. . .
- long range steric clashes,
- yield complicated inverse problems. for loop closure

The Ramachandran diagrams

▷ Ramachandran diagrams and populated regions

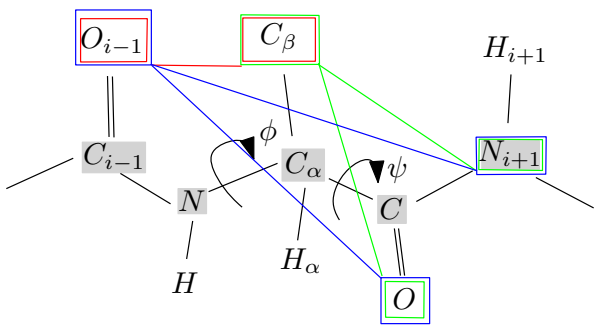


- Main regions: $\alpha L, \alpha R, \beta S, \beta P$
- Three prototypical diagrams
  - Glycine – no side chain/chiral C_α
  - Proline – side chain cycles on N
  - Others – with C_β and chiral C_α

▷ Distance constraints and the Ramachandran tetrahedron

$C1 : C_\beta - O_{i-1}$ $C2 : C_\beta - O + C_\beta N_{i+1}$

$C3 : O_{i-1} - O + O_{i-1} N_{i+1}$



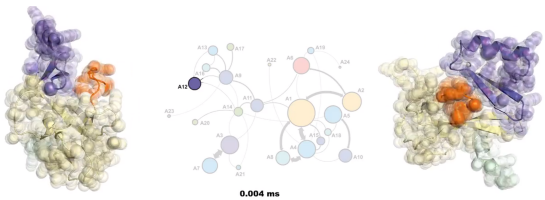
▷Ref: Stereochemistry of polypeptide chain configurations, JMB, 1963; Ramachandran et al

▷Ref: Revisiting the Ramachandran plot, Protein Science, 2003; Ho et al

Challenge *Dynamics of proteins*: specification

- Input: structure(s) of biomolecules + potential energy model
- Output
  - Thermodynamics: meta-stable states and observables
  - Kinetics: transition rates, Markov state models
- Time-scales
  - Biological time-scale $>$ millisecond
  - Integration time step in molecular dynamics: $\Delta t \sim 10^{-15} s$

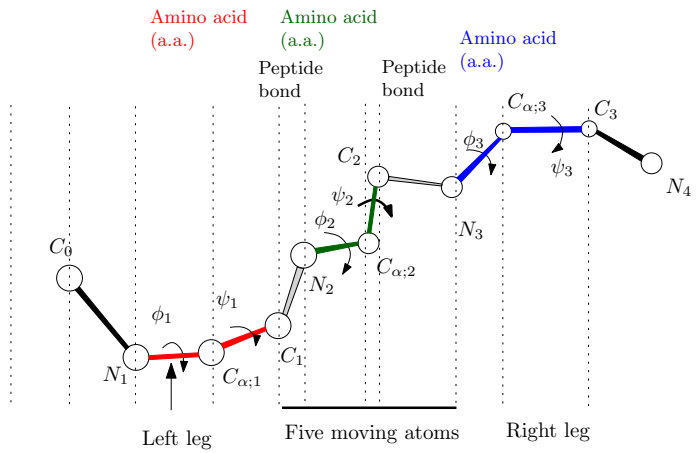
- 162 amino acids, $>$ 2000 atoms
- 5.058ms of simulation time
- $\sim$ 230 GPU years on NVIDIA GeForce GTX 980 processor



Ref: Chodera et al, eLife, 2019; Youtube link

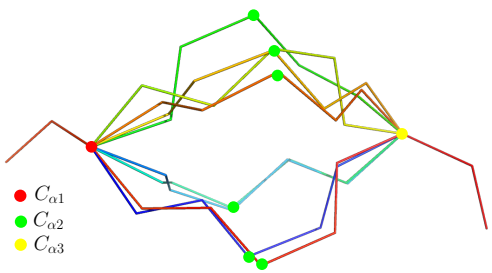
Tripeptide Loop Closure – TLC

- TLC: for 3 amino acids, fix all internal coordinates BUT the $(\phi_i, \psi_i)_{i=1,2,3}$ angles

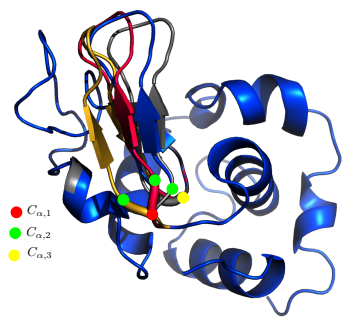


$\Rightarrow$ Find all possible values of the six angles $(\phi_i, \psi_i)_{i=1,2,3}$ compatible with the remaining fixed internal coordinates (bond lengths, valence angles, ω_i)

- Theorem: at most 16 solutions $\leftrightarrow$ real roots of a degree 16 polynomial
- The three amino acids may not be consecutive



3 consecutive a.a.

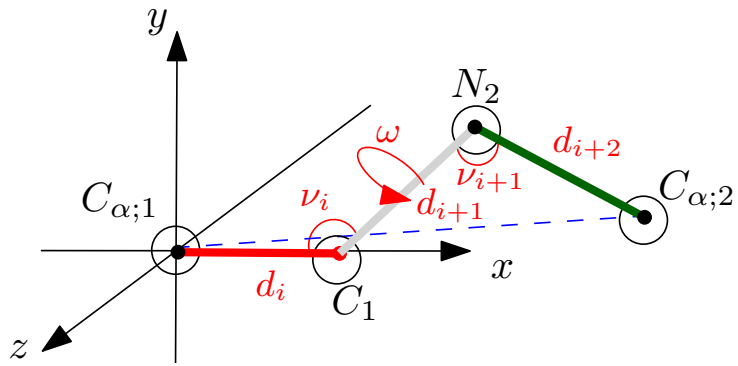


3 a.a. sandwiching SSE–CDRs

- Ref: Go and Scheraga, Macromolecules, 1970
- Ref: Coutsias et al, J. Comp. Chem., 2004

The peptide bond and peptide rigid bodies

▷ The peptide bond defines a rigid body:

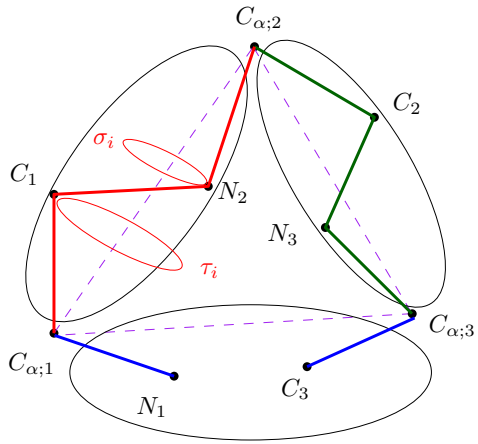


$$||C_{\alpha;2} - C_{\alpha;1}|| = Const$$

Internal coordinates fixed

- Bond lengths
- Valence angles
- ω angle

▷ The C_α triangle is rigid



- $C_{\alpha;2}$ belongs to the intersection of two spheres centered at $C_{\alpha;i}$ $C_{\alpha;i+2}$ $\Rightarrow$ C_α triangle has fixed geometry
- Legs fixed + C_α triangle rigid: rotate the three (colored) rigid bodies,

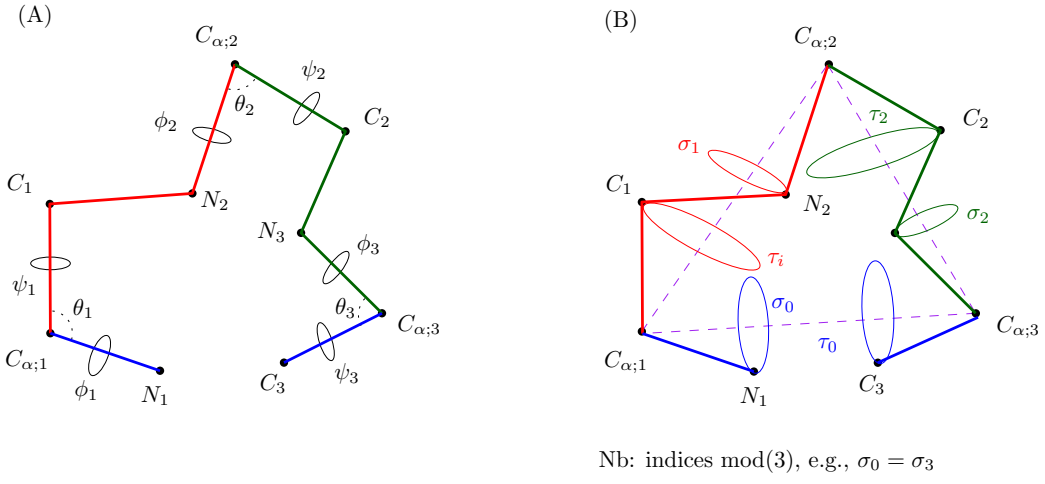
▷ Observations:
- one solves for six rotation angles $\{(\tau_i, \sigma_i)\}_{i=1,2,3}$
- **TLC parameterized in an angular space of dim. 12**

▷Ref: Coutsias et al, J. Comp. Chem., 2004

▷Ref: Cazals et al, Proteins, 2022

TLC model: from six to three angles

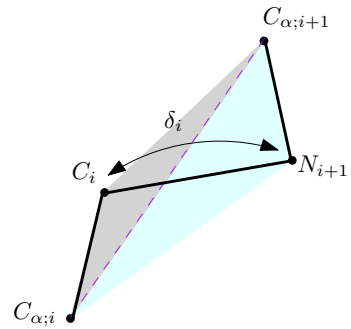
▷ Motions of the 3 rigid bodies: 6 angles



Nb: indices mod(3), e.g., $\sigma_0 = \sigma_3$

▷ . . . which are actually three

$$\sigma_i = \tau_i + \delta_i \qquad (1)$$



$\delta_i = \angle\text{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \text{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1})$

▷ Key ingredients of TLC:

- Initially: six dihedral angles $\{(\phi, \psi)\}_{\{i=1,2,3\}}$
- Then: three pairs $\{\delta_i, \tau_i\}$
- Finally: three angles τ_i

▷ The valence angle constraints: the θ_i angles at the $C_{\alpha;i}$s must remain constant.

$\Rightarrow$ It is the coupling introduced by the θ_i angles onto the rotation angles τ_i yields a degree 16 polynomial.

▷Ref: Coutsias et al, 2004

The three local frames

▷ Local frames and individual rotations:

- Defining invidividual rotations

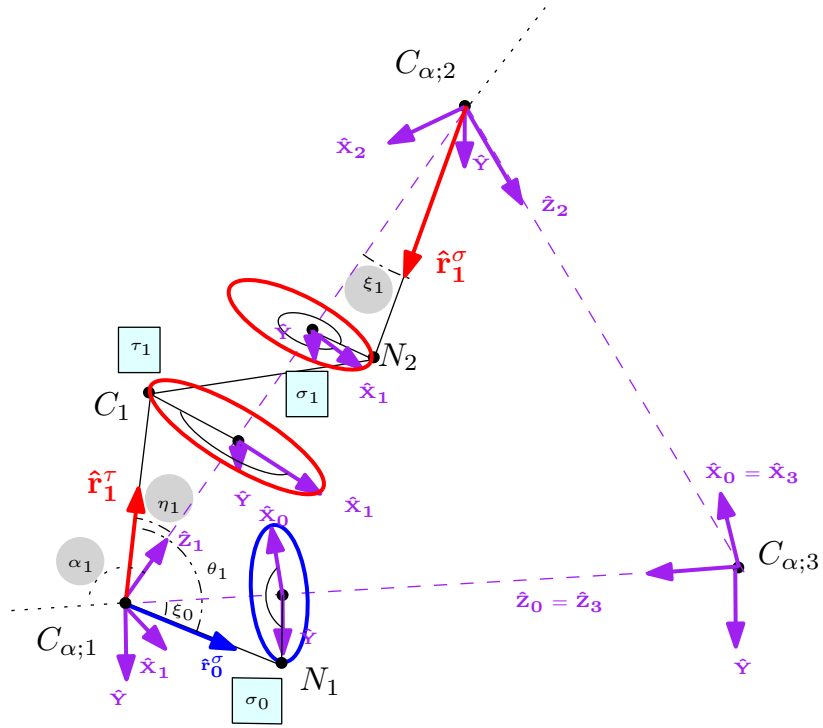
$\hat{\mathbf{r}}_\mathbf{i}^{\sigma}$; $\hat{\mathbf{r}}_\mathbf{i}^{\tau}$

- Orthonormal local frames:

Nb: $\hat{\mathbf{Z}}_\mathbf{i}$ = Unit vector along $\mathbf{C}_{\alpha;\mathbf{i}}\mathbf{C}_{\alpha;\mathbf{i+1}}$

$\hat{\mathbf{Y}}_\mathbf{i} \equiv \hat{\mathbf{Z}}_{\mathbf{i-1}} \times \hat{\mathbf{Z}}_\mathbf{i}$ Nb: $\hat{\mathbf{Y}}_\mathbf{i} = \hat{\mathbf{Y}}$

$\hat{\mathbf{X}}_\mathbf{i} = \hat{\mathbf{Y}}_\mathbf{i} \times \hat{\mathbf{Z}}_\mathbf{i} = (\hat{\mathbf{Z}}_\mathbf{i} \cdot \hat{\mathbf{Z}}_{\mathbf{i+2}})\hat{\mathbf{Z}}_\mathbf{i} - \hat{\mathbf{Z}}_{\mathbf{i+2}}$



▷ Angular description of the tripeptide: $4 \times 3 = 12$ angles

$$\begin{cases} \alpha_i & = \angle \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{Z}}_{\mathbf{i-1}} \\ \xi_i & = \angle - \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{r}}_\mathbf{i}^{\sigma} \\ \eta_i & = \angle \hat{\mathbf{Z}}_\mathbf{i}\hat{\mathbf{r}}_\mathbf{i}^{\tau} \\ \delta_i & = \angle \mathrm{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \mathrm{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1}) \end{cases} \tag{2}$$

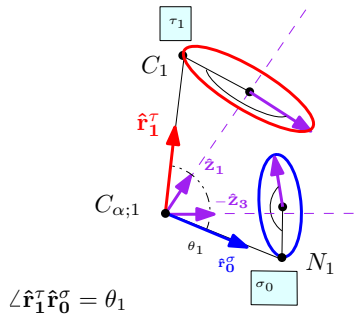
▷ Four tuple of angles for $C_{\alpha;i}$ of tripeptide T_k: $\mathsf{A}_{k,i} = \{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}$

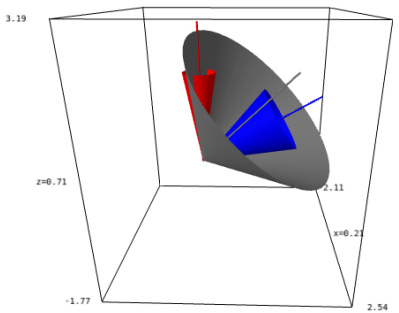
Rotations and dot product

Vectors in local frames; dot product in global frame

▷ Rotations of C_i and N_i: the two cones problem

- N_i, angle σ_i:
  vector $\hat{\mathrm{r}}^{\sigma}_{\mathrm{i}-1}$ about $\hat{\mathrm{Z}}_{\mathrm{i}-1}$
- C_i, angle τ_i:
  vector $\hat{\mathrm{r}}^{\tau}_{\mathrm{i}}$ about $\hat{\mathrm{Z}}_{\mathrm{i}}$





▷ Expressions of rotation vectors in local frames:

$$\text{In frame:}(\hat{\mathrm{X}}_{\mathrm{i}-1}, \hat{\mathrm{Y}}, \hat{\mathrm{Z}}_{\mathrm{i}-1}): \quad \hat{\mathrm{r}}^{\sigma}_{\mathrm{i}-1} = -\cos\xi_{i-1}\hat{\mathrm{Z}}_{\mathrm{i}-1} + \sin\xi_{\mathrm{i}-1}(\cos\sigma_{\mathrm{i}-1}\hat{\mathrm{X}}_{\mathrm{i}-1} + \sin\sigma_{\mathrm{i}-1}\hat{\mathrm{Y}}) \tag{3}$$

$$\text{In frame}(\hat{\mathrm{X}}_{\mathrm{i}}, \hat{\mathrm{Y}}, \hat{\mathrm{Z}}_{\mathrm{i}}): \quad \hat{\mathrm{r}}^{\tau}_{\mathrm{i}} = \cos\eta_i\hat{\mathrm{Z}}_{\mathrm{i}} + \sin\eta_{\mathrm{i}}(\cos\tau_{\mathrm{i}}\hat{\mathrm{X}}_{\mathrm{i}} + \sin\tau_{\mathrm{i}}\hat{\mathrm{Y}}) \tag{4}$$

▷ Valence angle constraint equation: θ_i kept constant

$$\begin{aligned}\langle \hat{\mathrm{r}}^{\sigma}_{\mathrm{i}-1}, \hat{\mathrm{r}}^{\tau}_{\mathrm{i}} \rangle = &-\cos\xi_{i-1}\cos\eta_i\cos\alpha_i \\ &-\cos\xi_{i-1}\sin\eta_i\cos\tau_i\sin\alpha_i \\ &-\cos\eta_i\sin\xi_{i-1}\cos\sigma_{i-1}\sin\alpha_i \\ &+\sin\xi_{i-1}\sin\eta_i(\cos\sigma_{i-1}\cos\tau_i\cos\alpha_i + \sin\sigma_{i-1}\sin\tau_i) \\ &= \cos\theta_i.\end{aligned} \tag{5, 6}$$

Algebra: the degree TLC solutions via the 16 polynomial

▷ Change of variables:

$$u_i = \tan(\tau_i/2), w_i = \tan(\sigma_i/2). \tag{7}$$

▷ Re-write the valence angle constraint – see also Eq. 21:

$$A_i w_{i-1}^2 u_i^2 + B_i w_{i-1}^2 + C_i w_{i-1} u_i + D_i u_i^2 + E_i = 0, \tag{8}$$

where the coefficients A_i, B_i, C_i, D_i, E_i depend on the angles $\theta_i, \alpha_i, \eta_i, \xi_{i-1}$.

▷ Perform another round of elimination for the w_{i-1} – coupling via δ_i: yields three three biquadratic polynomials in three variables, namely
$P_1(u_3, u_1), P_2(u_1, u_2), P_3(u_2, u_3)$

- By the Bernshtein-Kusnirenko-Khovanskii theorem, at most 16 solutions.
- The bound is tight.

▷ Using resultants: degree 16 polynomial in 1 variable

▷ Nb: the bound it tight.

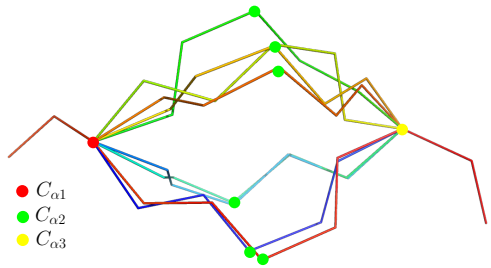
▷ Robust solutions: requires some care since π is involved

▷Ref: Cox,Little,O'Shea, Using algebraic geometry, 2005
https://en.wikipedia.org/wiki/Bernstein%E2%80%93Kushnirenko_theorem

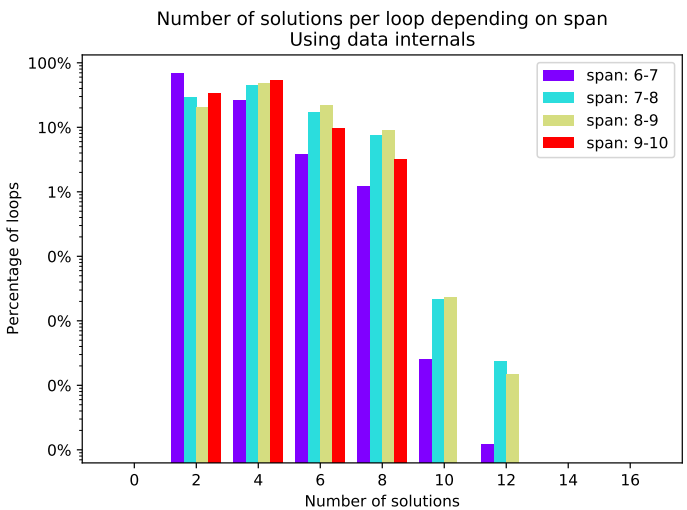
Loop sampling

TLC: number of solutions and atomic displacements

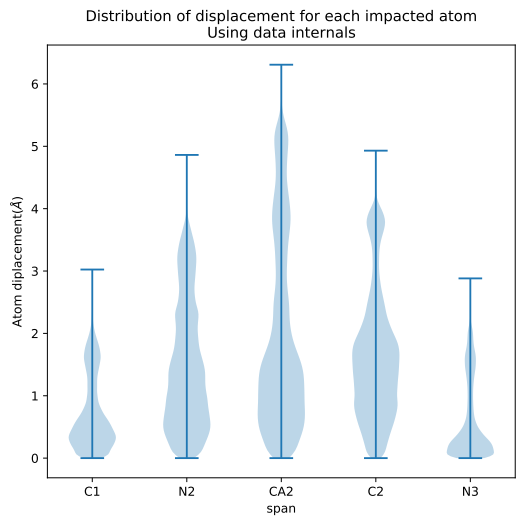
- Dataset: $\sim$ 2.6 million tripeptides in *loops* from high-resolution non redundant PDB structures



- # of solutions: function of span



- Atomic displacements

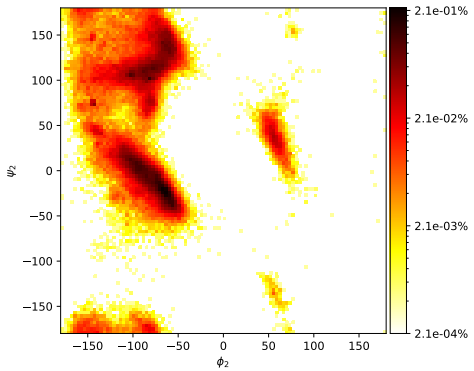


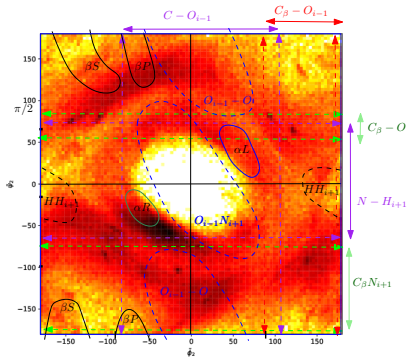
Interpolatory properties of TLC reconstructions

in the Ramachandran domains of the 3 amino acids

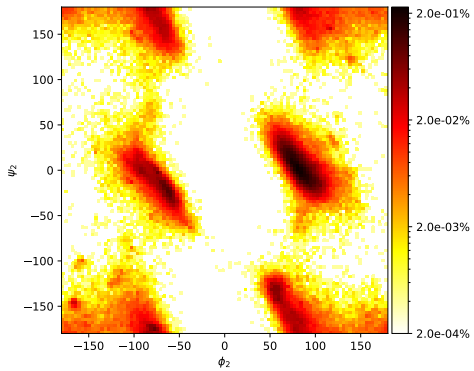
- Method: for the 3 Ramachandran domains (since 3 peptides):
  - compare the distribution of data versus reconstructions
  - distinguish on a per-class amino acid basis
- Ramachandran distributions

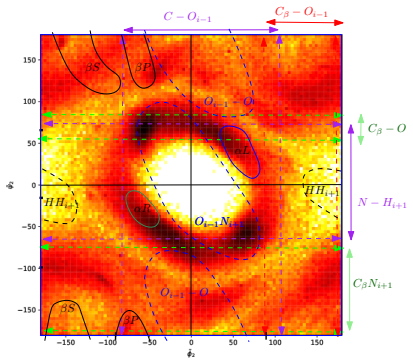
ASP





GLY





Data: $\mathcal{R}_{\mathcal{D},i}$ Reconstructions \ data: $\mathcal{R}_{\overline{\mathcal{D}},2}$

- NB: *transient* regions discovered – absent from crystals

Loop sampling

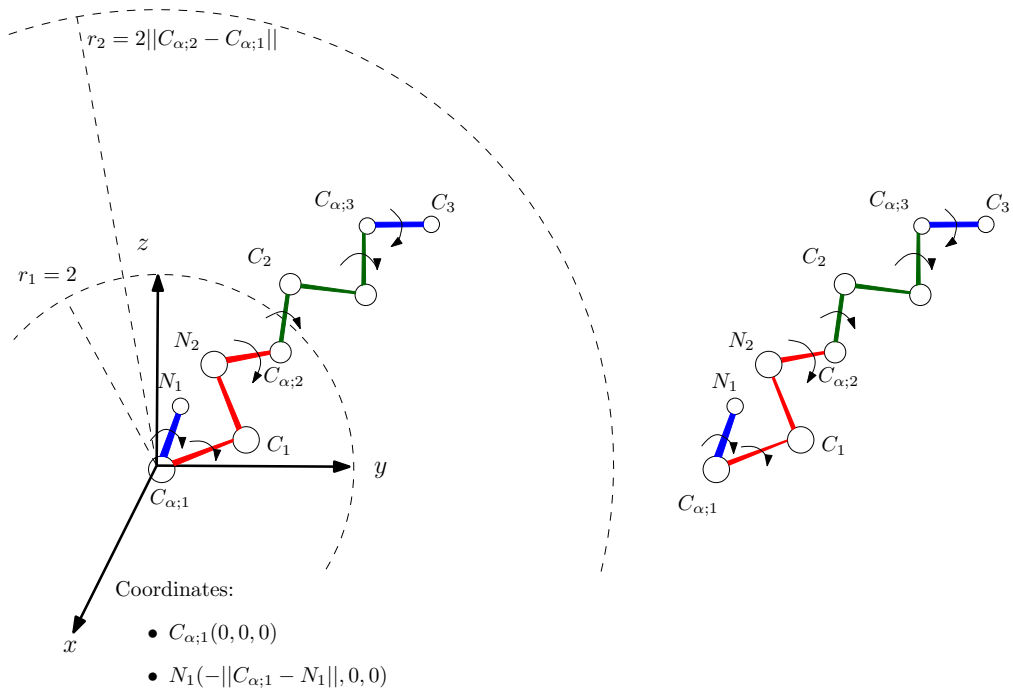
TLC with moving legs and embeddable tripeptides

▷ Geometric model:

- Tripeptide such that : left leg $N_i C_{\alpha;i}$ fixed, right leg $C_{\alpha;i+2} C_{i+2}$ free to move
- Six dihdedral angles $\{\phi_i, \psi_i\}$ free

▷ Question: provide necessary conditions on the position of the first and last segment–the **legs**, for the Tripeptide Loop Closure (TLC) algorithm to hold solutions.

▷ Nb: the relative position of legs suffices; in that case, position + orientation of $C_{\alpha;i+2} C_{i+2}$ yields a 5-dim search space.



Embedding tripeptides: recap

- Defining invidividual rotations

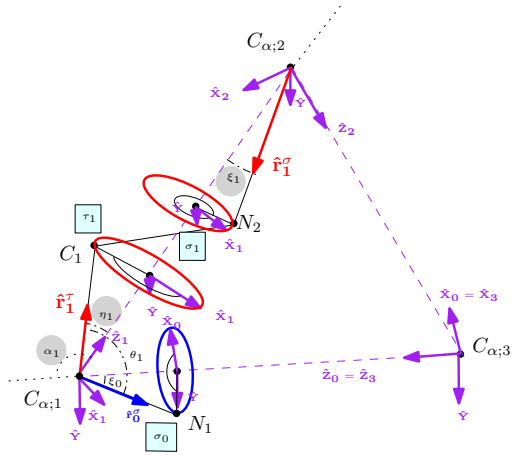
$\hat{\mathbf{r}}_{\mathbf{i}}^{\sigma}$; $\hat{\mathbf{r}}_{\mathbf{i}}^{\tau}$

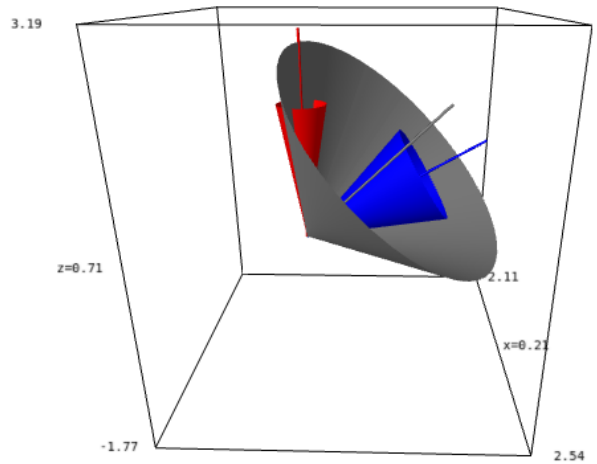
- Orthonormal local frames:

Nb: $\hat{\mathbf{Z}}_{\mathbf{i}}$ = Unit vector along $\mathbf{C}_{\alpha;\mathbf{i}}\mathbf{C}_{\alpha;\mathbf{i+1}}$

$\hat{\mathbf{Y}}_{\mathbf{i}} \equiv \hat{\mathbf{Z}}_{\mathbf{i-1}} \times \hat{\mathbf{Z}}_{\mathbf{i}}$ Nb: $\hat{\mathbf{Y}}_{\mathbf{i}} = \hat{\mathbf{Y}}$

$\hat{\mathbf{X}}_{\mathbf{i}} = \hat{\mathbf{Y}}_{\mathbf{i}} \times \hat{\mathbf{Z}}_{\mathbf{i}} = (\hat{\mathbf{Z}}_{\mathbf{i}} \cdot \hat{\mathbf{Z}}_{\mathbf{i+2}})\hat{\mathbf{Z}}_{\mathbf{i}} - \hat{\mathbf{Z}}_{\mathbf{i+2}}$





▷ 1. From the position of legs: compute $\{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}_{i\in\{1,2,3\}}$

▷ 2. TLC: find the (σ, τ) angles such that:

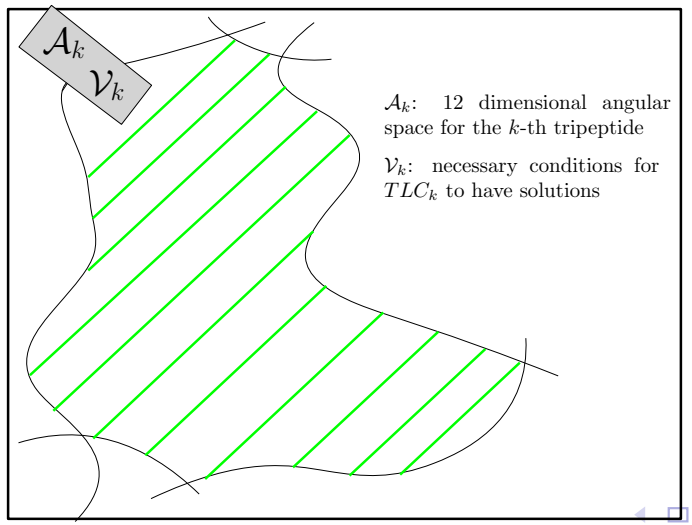
$$\langle \hat{\mathbf{r}}_{\mathbf{i-1}}^{\sigma}, \hat{\mathbf{r}}_{\mathbf{i}}^{\tau} \rangle = \cos\theta_i. \tag{9}$$

▷ Our goal:

- Conditioning of the solutions wrt the $\{\alpha, \xi, \eta, \delta\}$ via necessary conditions
- Ability to sample uniformly solutions given the necessary conditions

Sampling strategy based on validity intervals: overview

- Angular representations:
  - $C_{\alpha;i}$ from tripeptide T_k: four tuple of angles $\mathsf{A}_{k,i} = \{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}$ with $i \in \{1, 2, 3\}$
  - Tripeptide T_k, 12-dim angular space: $\mathsf{A}_k = \{\mathsf{A}_{k,1}, \mathsf{A}_{k,2}, \mathsf{A}_{k,3}\}$.
- Strategy:
  - Assume we have necessary conditions for the angles, as a finite set of **validity intervals** $I = \{[a_i, b_i]\}$
  - Assume each bound a_i or b_i is defined by an implicit equation in the 12 angular variables
  - By the implicit function theorem (assuming it applies): each equation corresponds to a hyper-surface
  - Domain enclosed by these hyper-surfaces: domain within which TLC solution lie



Validity Intervals and Depth One Validity Intervals (DOVI)

▷ Validity intervals: for each angle $\tau_{k,i}$, one can compute 2+2 intervals on S^1, representing (stringent) necessary conditions for TLC to admit solutions:

$$\text{(Initial)}\mathcal{I}_{\tau_{k,i}} = \{I_{\tau_{k,i}}\} \text{ with } I_{\tau_{k,i}} = [I_\tau^{\min}(\mathsf{A}_{k,i}), I_\tau^{\max}(\mathsf{A}_{k,i})] \tag{10}$$

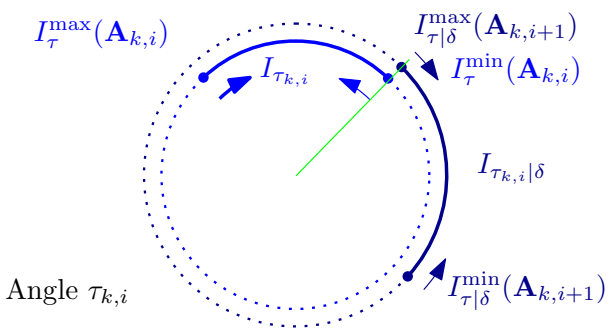
$$\text{(Rotated)}\mathcal{I}_{\tau_{k,i}|\delta} = \{I_{\tau_{k,i}|\delta}\} \text{ with } I_{\tau_{k,i}|\delta} = [I_{\tau|\delta}^{\min}(\mathsf{A}_{k,i+1}), I_{\tau|\delta}^{\max}(\mathsf{A}_{k,i+1})] \tag{11}$$

Indeed:

- $I_{\tau_{k,i}}$: obtained from the invariance of θ_i at $C_{\alpha;i}$
- $I_{\tau_{k,i}|\delta}$: obtained from $I_{\sigma_{k,i}}$ via the relation $\sigma_i = \tau_i + \delta_i$
- NB: 2 initial and 2 rotated: intervals bounds in $[0, \pi]$ + symmetry wrt C_α plane

▷ Intersection of validity intervals: necessary conditions expressed as intervals

$$I_{\tau_{k,i}} \cap I_{\tau_{k,i}|\delta}.$$



Limit case: implicit equation in the 12 dimensional space $\mathcal{A}_k$.

Limit case: $I_\tau^{\max}(\mathbf{A}_{k,i}) = I_{\tau|\delta}^{\min}(\mathbf{A}_{k,i+1})$

Validity domain for tripeptide T_k:

intersecting two initial and two rotated intervals

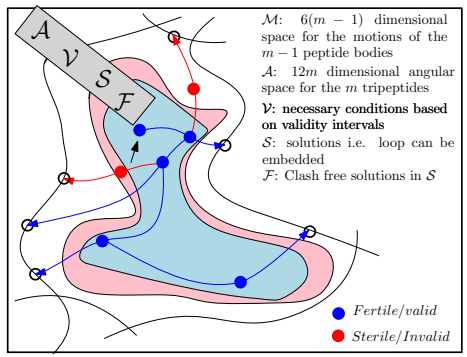
▷ Rigid body I of tripeptide T_k: angles tuples $\rightsquigarrow$ Depth One Validity Intervals

$$\mathrm{DOVI}_{\tau_{k,i}}(\cdot) : \mathcal{A}_k \mapsto \emptyset + (\mathcal{I}_{\tau_{k,i}} \cap \mathcal{I}_{\tau_{k,i}|\delta})^4. \tag{12}$$

▷ The *angular* validity domain $\mathcal{V}_k$ for T_k:

For the angle $\tau_{k,i}$: the domain $\mathcal{V}_k \subset \mathcal{A}_k$ such that

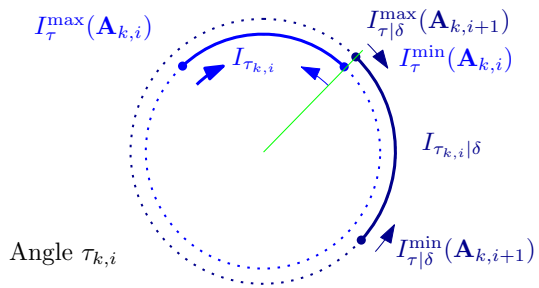
$$\forall k, \forall i, \forall a \in \mathcal{V}_k : \mathrm{DOVI}_{\tau_{k,i}}(a) \neq \emptyset.$$



▷ Non empty intersection for 2 intervals $I_{\tau_{k,i}} \in \mathcal{I}_{\tau_{k,i}}$ and $I_{\tau_{k,i}|\delta} \in \mathcal{I}_{\tau_{k,i}|\delta}$: conditions

$$\begin{cases} I_{\tau}^{\max}(\mathrm{A}_{k,i}) = I_{\tau|\delta}^{\min}(\mathrm{A}_{k,i+1}) \\ \text{or } I_{\tau}^{\min}(\mathrm{A}_{k,i}) = I_{\tau|\delta}^{\max}(\mathrm{A}_{k,i+1}) \end{cases}$$

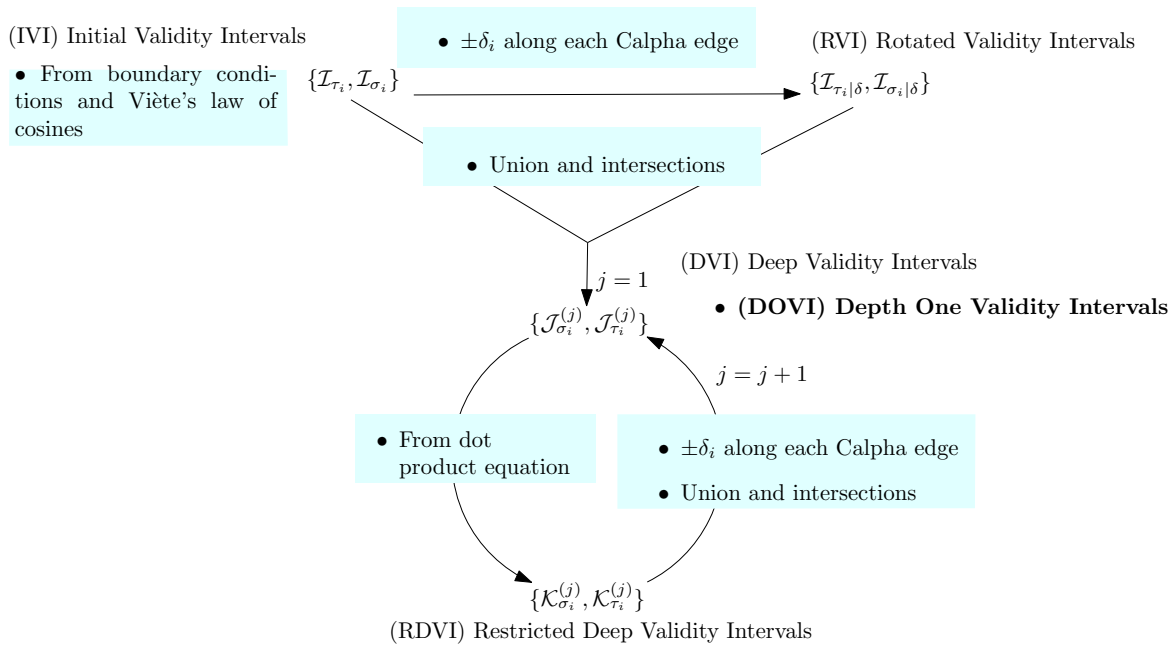
$\Rightarrow$ **two implicit equations in** $\mathcal{A}_k$: two sub-manifolds $\mathcal{V}_k$



Limit case: $I_{\tau}^{\max}(\mathbf{A}_{k,i}) = I_{\tau|\delta}^{\min}(\mathbf{A}_{k,i+1})$

Validity intervals: deep *i.e.* iterated VI

- Two types of constraints:
  - Coherence along each edge of the C_α triangle – via ω angle
  - Constraint on θ_i at each C_α
- A sequential and iterative construction: interval types used
  - Initial VI
  - Rotated VI
  - Deep VI and Restricted Deep VI



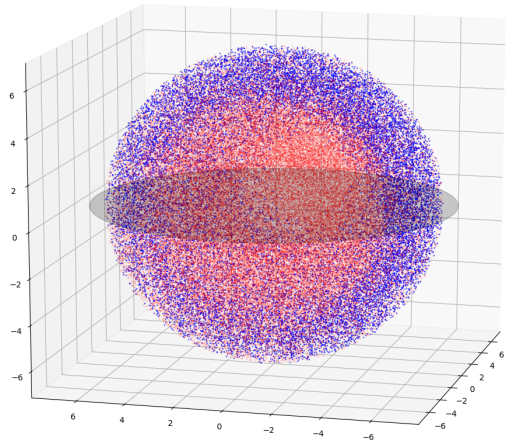
Stringency of necessary conditions: assessment

- Reminder: the search space is 5D
- Evaluation of the stringency of validity intervals:
  - Take random instances of peptides – in the 5D space
  - Identify positives (P) and negatives (N)
  - Given that N = True Negative + False Positives

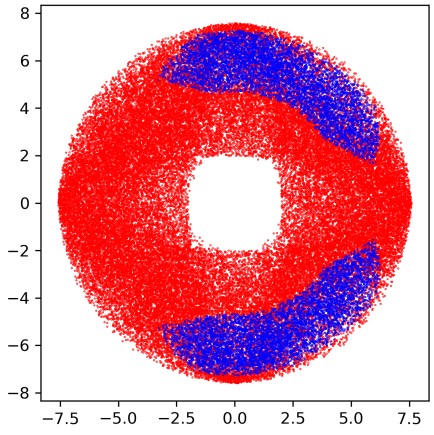
$$\text{Strigency of necessary condition } c : \frac{FP(c)}{N} \tag{13}$$

- Nb: projecting the 5D points into 3D: coordinates of $C_{\alpha;i+2}$

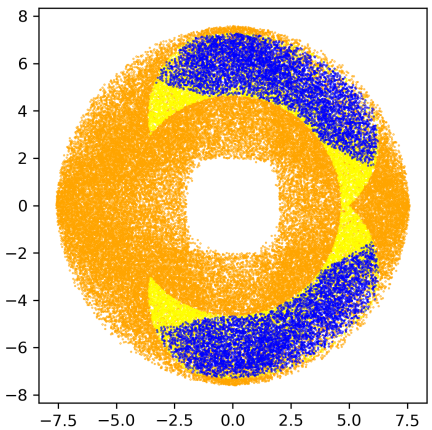
Stringency of necessary conditions: results



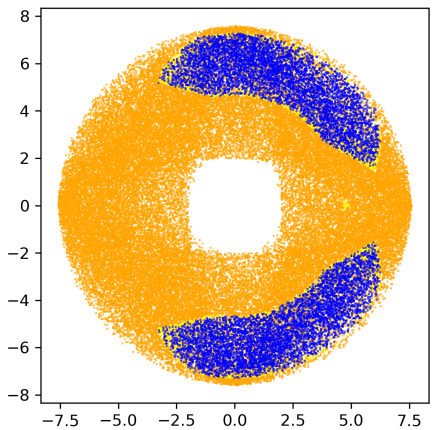
(A)



(B)



(C)



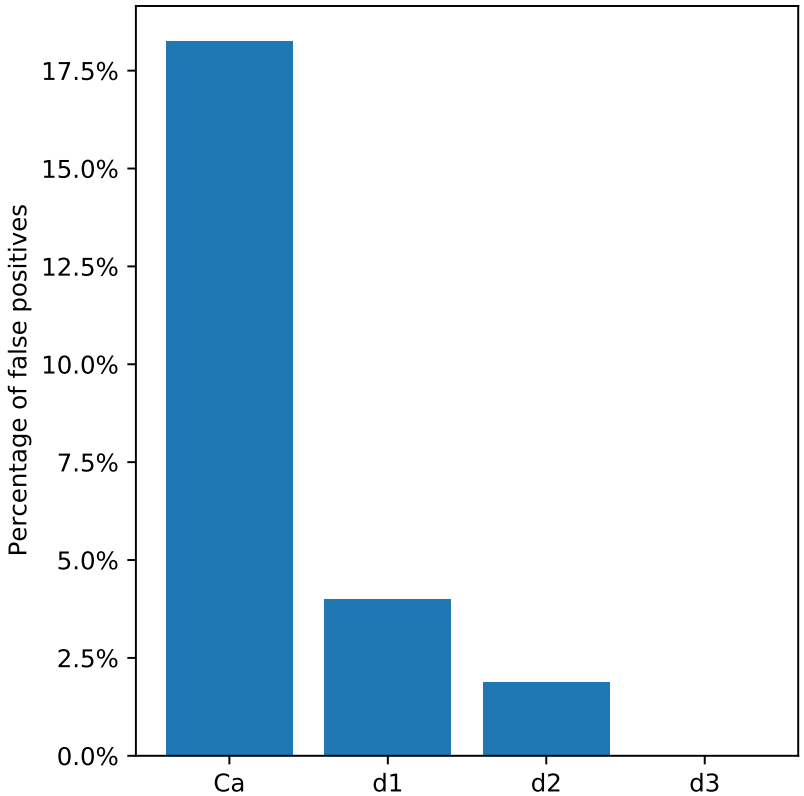
(D)

- (A,B) Random TLC instances: position of $C_{\alpha;i+2}$. Blue/red: fertile/sterile point.
- (C) C_α valence constraints: False Positives in yellow
- (D) Depth 1 validity intervals: False Positives in yellow

▷ Nb: FP reduced significantly...but beware of the bias due to the 3D projection!

Stringency of deep validity intervals

- Stringency: initial validity intervals + deep validity intervals



- Observation: the % of FP decreases
- Conjecture: the intervals converge towards the solutions of TLC

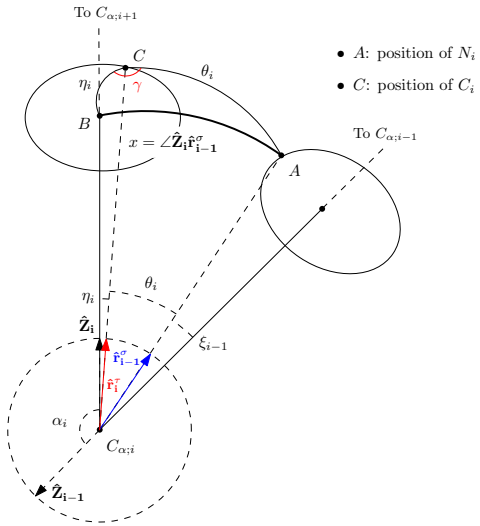
Initial Validity Intervals: bounds

▷ Obs: limit cases for the dot product $\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-1}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = \cos(\theta_{\mathbf{i}} \pm \eta_{\mathbf{i}})$.

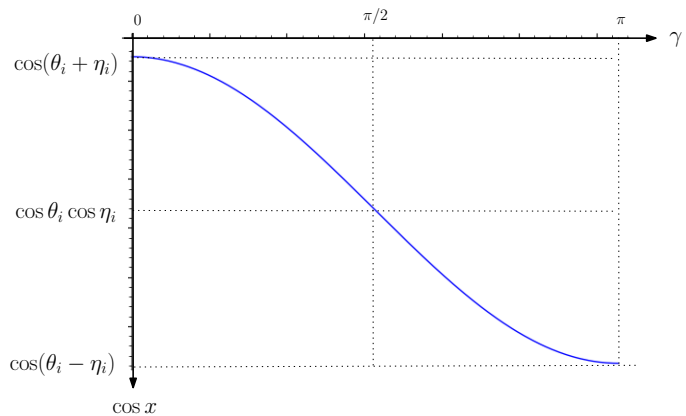
Proof: Viète's law of cosines for the spherical triangle ABC:

$$\cos x = \cos\theta_i \cos\eta_i + \sin\theta_i \sin\eta_i \cos\gamma. \tag{14}$$

Extreme values for $\gamma = 0, \pi$: $\cos(\theta_i \pm \eta_i)$



Viète formula of cosines



$\cos x$

▷ Final step:

- plug the extreme values into the dot product $\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-1}, \hat{\mathbf{r}}^{\tau}_{\mathbf{i}} \rangle$
- $\Rightarrow$ polynomial in cos, sin of the 12 angles + the 3 σs

Valence angle constraint: the case of σ_{i-1} (II)

▷ $\sigma_{i-1;-}$: first limit case
start with the dot product

$$\langle \hat{\mathbf{r}}^{\sigma}_{i-1}, \hat{\mathbf{Z}}_i \rangle = -\cos\sigma_{i-1}\sin\xi_{i-1}\sin\alpha_i - \cos\xi_{i-1}\cos\alpha_i. \tag{15}$$

$$\langle \hat{\mathbf{r}}^{\sigma}_{i-1;-}, \hat{\mathbf{Z}}_i \rangle = \cos(\theta_i + \eta_i) \tag{16}$$

from which we obtain

$$\begin{cases} S^- = \frac{+\cos(\theta_i - \eta_i) + \cos\xi_{i-1}\cos\alpha_i}{\sin\xi_{i-1}\sin\alpha_i} \\ \sigma_{i-1;-} = \arccos S^- \end{cases} \tag{17}$$

When $S^- \to 1^-, \sigma_{i-1;-} \to 0^+$. Therefore,

$$S^- > 1, \tag{18}$$

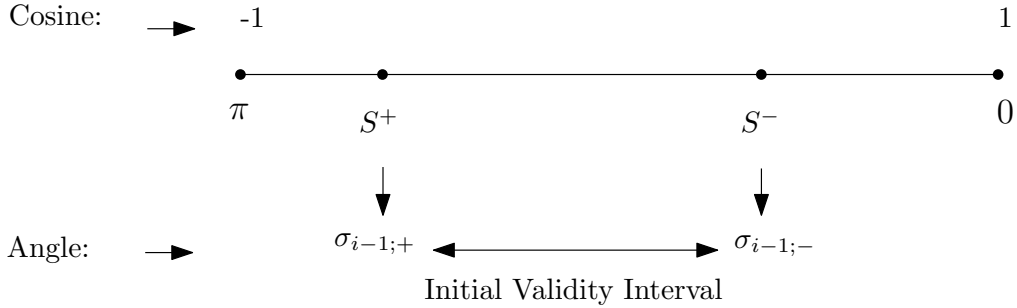
we set $\sigma_{i-1;-} = 0$, so that any value $\sigma_{i-1} \leq \sigma_{i-1;+}$ is valid.

▷ $\sigma_{i-1;+}$: mutatis mutandis

▷ Result: validity interval $I_{\sigma_{i-1}} = [\sigma_{i-1;-}, \sigma_{i-1;+}] \subset [0, \pi]$

C_α valence constraints

▷ Conditions to define the four extreme angles: the case of σ_{i-1}



Definition 1. (C_α valence constraints) The C_α *valence constraints* are the *necessary validity conditions* defined by :

- Angle $\sigma_{i-1;-}$: the condition $\sigma_{i;-} < \sigma_{i;+}$ requires $S^- \geq -1$.
- Angle $\sigma_{i-1;+}$: the condition $\sigma_{i;-} < \sigma_{i;+}$ requires $S^+ \leq 1$.
- Angle $\tau_{i;-}$: the condition $\tau_{i;-} < \tau_{i;+}$ requires $T^- \geq -1$.
- Angle $\tau_{i;+}$: the condition $\tau_{i;-} < \tau_{i;+}$ requires $T^+ \leq 1$.

For the constraint to be verified all these conditions must be valid for all three $\{(\sigma_{i-1}, \tau_i)\}$ pairs.

▷ Application: pick a tripeptide geometry $\{\alpha_i, \xi_i, \eta_i, \delta_i\}$, and check whether the four previous conditions are fulfilled.

Validity Intervals: Initial and Symmetric

Pairs of Validity Intervals

- Angle σ_{i-1}:
  - Validity interval $I_{\sigma_{i-1}} = [\sigma_{i-1;-}, \sigma_{i-1;+}] \subset [0, \pi]$
  - Symmetric interval with respect to the plane $C_{\alpha;i} C_{\alpha;i+1} C_{\alpha;i+2}$:

$$I^{'}_{\sigma_{i-1}} = [\sigma^{'}_{i-1;-}, \sigma^{'}_{i-1;+}] \stackrel{Def}{=} [2\pi - \sigma_{i-1;-}, 2\pi - \sigma_{i-1;+}].$$

    Nb: values in $(\pi, 2\pi]$.

- Angle τ_i: mutatis mutandis

Definition 2. (Initial validity intervals) The *initial validity interval* for σ_{i-1} are defined by:

$$\mathcal{I}_{\sigma_{i-1}} = I_{\sigma_{i-1}} \cup I^{'}_{\sigma_{i-1}} \tag{19}$$

Likewise, the *initial validity interval* for τ_i are defined by:

$$\mathcal{I}_{\tau_i} = I_{\tau_{i-1}} \cup I^{'}_{\tau_i}. \tag{20}$$

Extreme angles: visualization

- Dot product surface:

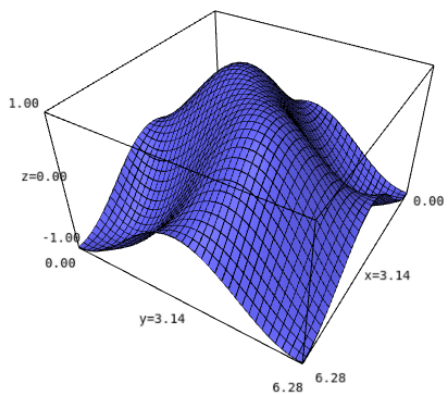
$$f(\sigma_{i-1}, \tau_i) = \langle \hat{\mathbf{r}}_{i-1}^{\sigma}, \hat{\mathbf{r}}_i^{\tau} \rangle \tag{21}$$

$$= -\cos\xi_{i-1}\cos\eta_i\cos\alpha_i \tag{22}$$
$$-\cos\xi_{i-1}\sin\eta_i\cos\tau_i\sin\alpha_i$$
$$-\cos\eta_i\sin\xi_{i-1}\cos\sigma_{i-1}\sin\alpha_i$$
$$+\sin\xi_{i-1}\sin\eta_i(\cos\sigma_{i-1}\cos\tau_i\cos\alpha_i + \sin\sigma_{i-1}\sin\tau_i)$$

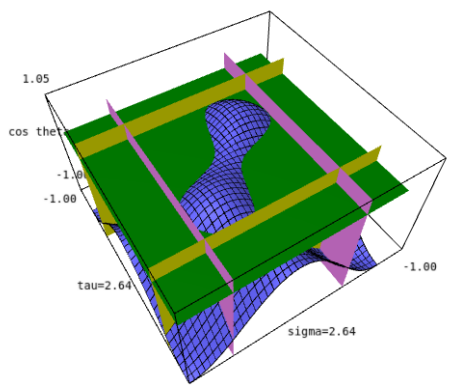
$$= \cos\theta_i \tag{23}$$

  - angles $\sigma_{i-1;-}$ and $\sigma_{i-1;+}$ correspond to planes orthogonal to the σ_{i-1}; dito for $\tau_{i;-}$ and $\tau_{i;+}$

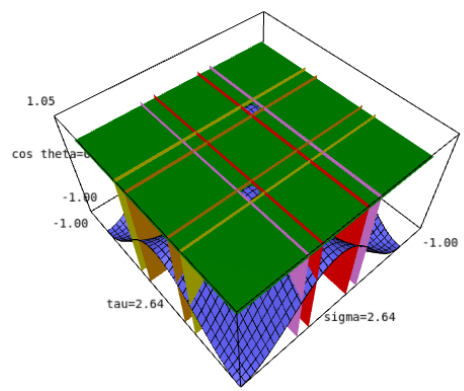
- Dot product surface and extreme angles $\sigma_{i-1;-}, \sigma_{i-1;+}, \tau_{i-1;-}, \tau_{i-1;+}$



(A)



(B)



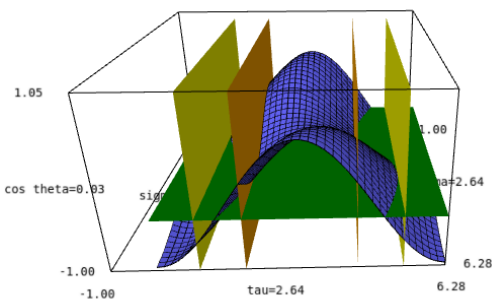
(C)

Nb: $\alpha_i = 100, \chi_{i-1} = 50, \eta_i = 50$ **(A)** Whole surface **(B)** With horizontal plane $\cos\theta_i = \cos 9°$. Intersection curve: 1 c.c. **(C)** With horizontal plane $\cos\theta_i = \cos 35°$. Intersection curve: 2 c.c.

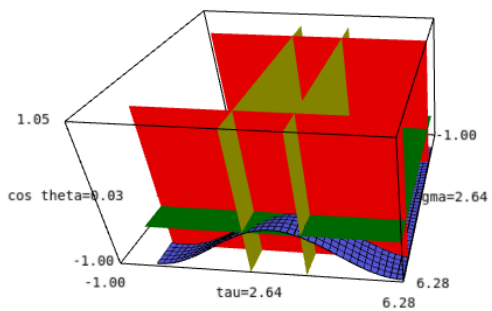
Dot surfaces and their classification

Definition 3. (Signature at C_α) Consider the endpoints of the validity intervals, in this order $\sigma_{i-1;-}, \sigma_{i-1;+}, \tau_{i;-}, \tau_{i;+}$. The *signature* of a TLC problem is a string in $\{N, P, Z\}^4$ –one letter for each each extreme angle, with the following convention:

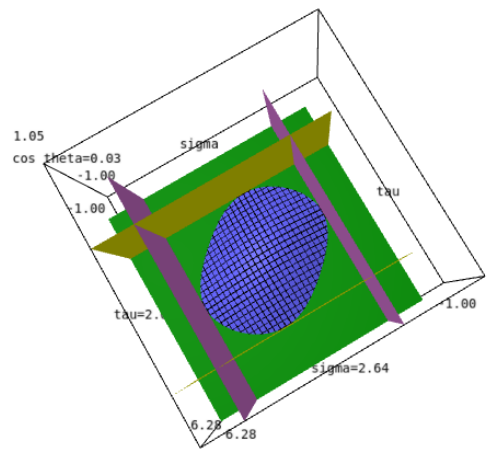
- letter N for $cos(endpoint) < -1$,
- letter P for $cos(endpoint) > 1$,
- letter Z for $-1 < cos(endpoint) < 1$.



PNZZ



PZZN



ZNZN

Dot surfaces and validity intervals for the dataset of random TLC instances. (A) The 7 signatures (Def. 3) in terms of extreme angles for the data set of random TLC instances. In all cases, the green plane corresponds to $\cos\theta_i = \cos 111.6°$. A signature reads as follows: N:negative ie dot product < -1; Z: zero ie dot product $\in [-1, 1]$; P: positive ie dot product > 1. **(B)** Validity intervals.

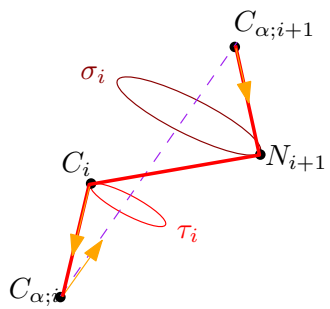
Rotated validity intervals (I)

- Along C_α edge:

$$\sigma_i = \tau_i + \delta_i. \tag{24}$$

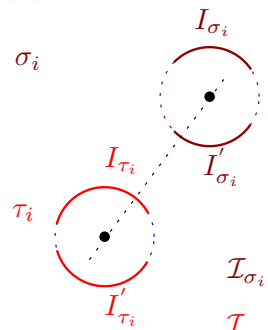
- Rotated interval for an angle: obtained from the value of its twin angle (from τ_i for σ_i, and vice-versa)

(A)



$\delta_i = \angle\text{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \text{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1})$

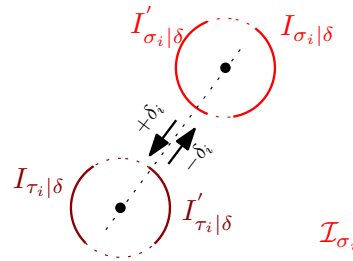
(B)



$\mathcal{I}_{\sigma_i} = \{I_{\sigma_i}, I'_{\sigma_i}\}$

$\mathcal{I}_{\tau_i} = \{I_{\tau_i}, I'_{\tau_i}\}$

(C)



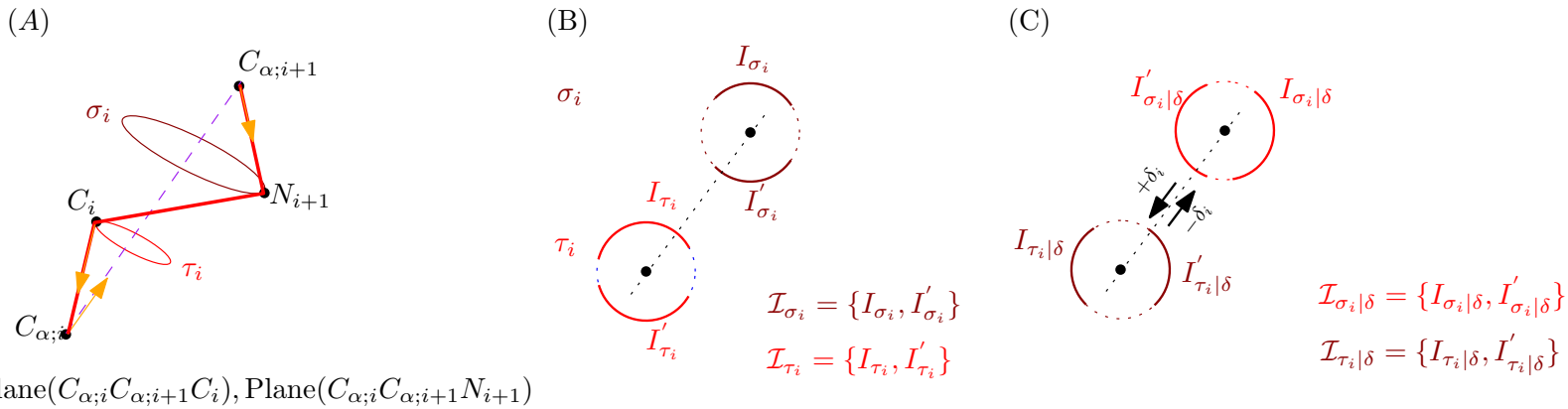
$\mathcal{I}_{\sigma_i|\delta} = \{I_{\sigma_i|\delta}, I'_{\sigma_i|\delta}\}$

$\mathcal{I}_{\tau_i|\delta} = \{I_{\tau_i|\delta}, I'_{\tau_i|\delta}\}$

Rotated validity intervals (II)

Definition 4. (Rotated validity intervals) The rotated validity intervals for the angles and τ_i are defined by:

- for σ_{i-1}: $\mathcal{I}_{\sigma_{i-1}|\delta} = I_{\sigma_{i-1}|\delta} \cup I'_{\sigma_{i-1}|\delta}$ with:
  - $I_{\sigma_{i-1}|\delta}$: interval for σ_{i-1} obtained by applying Eq. (24) to $I_{\tau_{i-1}}$. (Nb: uses the edge $C_{\alpha;i}C_{\alpha;i-1}$ of the C_α triangle.)
  - $I'_{\sigma_{i-1}|\delta}$: interval for σ_{i-1} obtained by applying Eq. (24) to $I'_{\tau_{i-1}}$. (Nb: uses the edge $C_{\alpha;i}C_{\alpha;i-1}$ of the C_α triangle.)
- for τ_i: dito



$\delta_i = \angle\text{Plane}(C_{\alpha;i}C_{\alpha;i+1}C_i), \text{Plane}(C_{\alpha;i}C_{\alpha;i+1}N_{i+1})$

Deep Validity Intervals: depth 1

- Intervals obtained so far:
  - The conditions on σ_{i-1} and τ_i inherent to the conservation of the valence angles (Eq. (26)).
  - The conditions exploiting rotated validity intervals, stemming from Eq. (24)
- Combination: intervals combined as follows $(I_{\cdot_{i-1}}, I'_{\cdot_{i-1}}) \times (I_{\cdot_{i-1}|\delta}, I'_{\cdot_{i-1}|\delta})$, which yields *depth one validity intervals*:

Definition 5. (Depth one validity intervals) The *depth 1 inter-angular interval* set $\mathcal{J}^{(1)}_{\sigma_{i-1}}$ for σ_{i-1}:

$$\mathcal{J}^{(1)}_{\sigma_{i-1}} = (I_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I_{\sigma_{i-1}|\delta}) \cup (I'_{\sigma_{i-1}} \cap I'_{\sigma_{i-1}|\delta}) \quad (25)$$

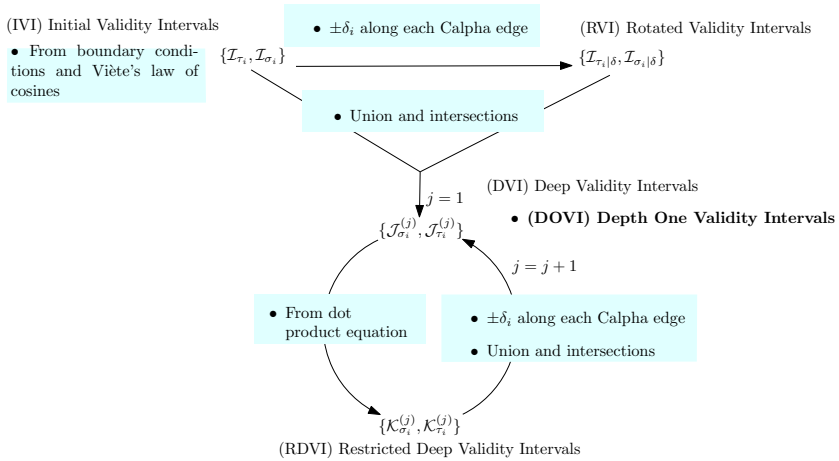
depth 1 inter-angular interval set $\mathcal{J}^{(1)}_{\tau_i}$ for τ_i: dito.

Definition 6. (Depth 1 inter-angular constraint) The *depth 1 inter-angular constraint* for σ_{i-1} is $\mathcal{J}^{(1)}_{\sigma_{i-1}} \neq \emptyset$.

The *depth 1 inter-angular constraint* for τ_i is: $\mathcal{J}^{(1)}_{\tau_i} \neq \emptyset$.

For the constraint to be verified all these conditions must be valid for all three $\{(\tau_i, \sigma_{i-1})\}$ pairs.

Depth-n validity constraints: outline



▷ Depth 1 validity intervals:

- Initialization via the limit conditions – from Viète law of cosines:

$$\begin{cases} \langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-\mathbf{1};-}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = \cos(\theta_{\mathbf{i}} + \eta_{\mathbf{i}}), \\ \langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-\mathbf{1};+}, \hat{\mathbf{Z}}_{\mathbf{i}} \rangle = \cos(\theta_{\mathbf{i}} - \eta_{\mathbf{i}}) \end{cases}$$

- Then refinement thanks to intersections with Rotated validity intervals

▷ Depth-n validity intervals:

- Given a DVI of depth j (initially, $j = 1$), apply the valence angle constraint to obtain the twin interval on τ_i from σ_{i-1} and vice-versa, using

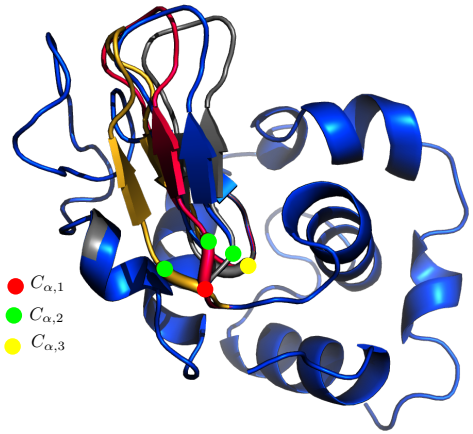
$$\langle \hat{\mathbf{r}}^{\sigma}_{\mathbf{i}-\mathbf{1}}, \hat{\mathbf{r}}^{\tau}_{\mathbf{i}} \rangle = \cos\theta_i. \tag{26}$$

- Iterate

Loop sampling

Metaphor: two problems with global+local components

Paris / San Francisco / Stanford:
30' + 30' minutes



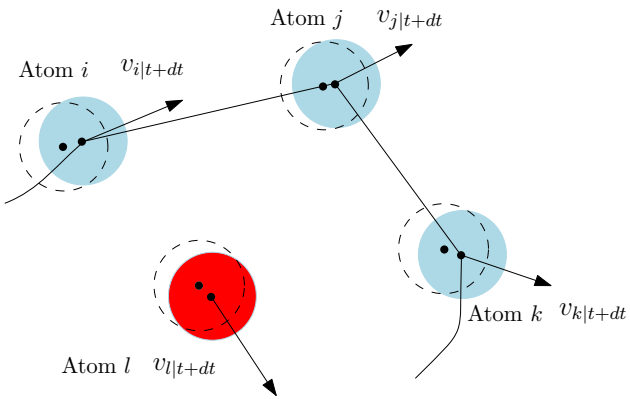
Biomolecules: identifying stable states and their probabilities

- First leg:
  - Paris to San Francisco airport (SFO): ???
  - Biomolecules, finding large amplitude conformational changes between states: my methods based on inverse problems
- Second leg:
  - SFO to Stanford: shuttle, cab
  - Biomolecules: studying equilibria with molecular dynamics

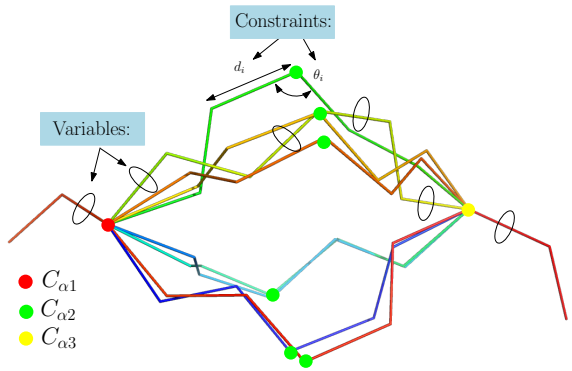
$\Rightarrow$ our methods vs classical methods: complementary

Next gen sampling: scientific punchline, originality and risks

- Three classes of techniques to study the dynamics of biomolecules:
  - Direct problems: molecular dynamics
  - Inverse problems of the loop closure type using internal coordinates
  - (Deep learning based: no massive data at hand–at this stage)



Molecular dynamics, time-steps of 10^{-15}s:
$\|\Delta x_i\| \sim 1/100$Å

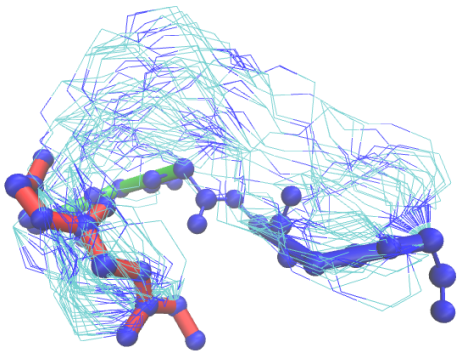


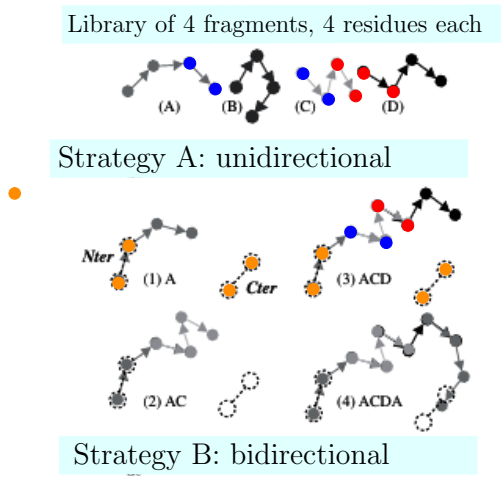
Inverse problems, typical changes:
$\|\Delta x_i\| \sim 1 - 10$Å

- Using internal coordinates: originality
  - Fast methods to predict large amplitude conformational changes
    - NB: geometric proxy/priors for classical methods such as MD
- Using internal coordinates: caveats
  - Risks: model accuracy (solvent, side chains), statistical biases
  - Gains: unmatched diversity and speed

Protein Loop Sampling: main approaches

- ▷ Classical approaches:
  - ~~Molecular dynamics~~: cost + handling loop closure
  - Non rigid geometry–but solution space is continuous (manifold)
  - Data driven/combinatorial greedy methods + inverse kinematics
  - Dihedral angles only/rigid geometry + inverse kinematics (TLC)





- ▷ Open questions:
  - Global loop parameterization amenable to sampling: all a.a. on equal footing
  - Uniform sampling in $\{(\phi, \psi)\}$ angle space,
  - Connection with thermodynamics,
  - Complexity: how hard are these problems?

▷Ref: Dod et al 1983; Cortés and Siméon, 2004; Levitt, Guibas et al, 2005; Snoeyink et al, 2005; Latombe et al, 2005; Cortés et al 2019, etc

▷Ref: Cazals et al; 2022

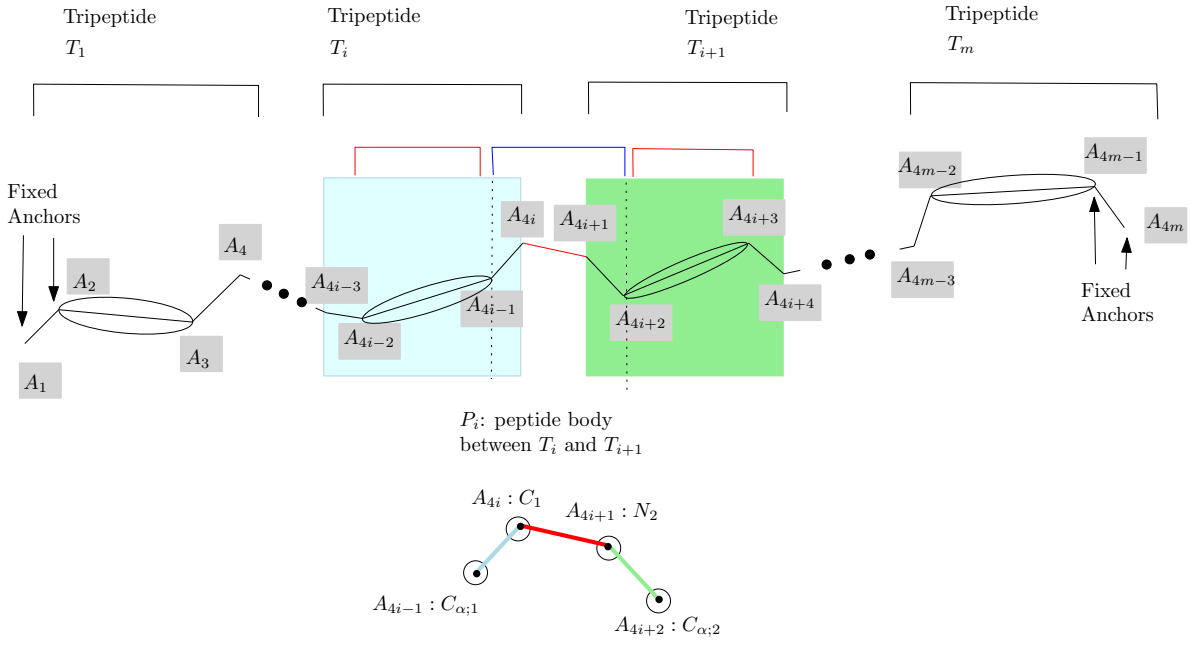
Loop sampling: difficulties and main approaches

- Main difficulties
  - Space of solutions: a continuous space – if #dihedral angles > 6
  - *Walking* on this constrained *manifold*: geometrically/numerically difficult
  - Incremental construction based on tripeptides: combinatorial explosion
- A mixed discrete - continuous approach
  - `Rosetta` KIC for a chain with n amino acids: perturb the dihedral angles of $n-3$ a.a.; then close the chain on the last 3 with TLC
  - Concatenation of solutions yielded by tripeptides: grow chains from left and right; close with TLC
- The problem remains difficult:
  - Practice: orphan loops in databases / IDPs
  - Theory: no global parametric solution

▷Ref: Kolodny, Guibas, Levitt, Koehl, 2005
▷Ref: Kortemme et al, Nat. Methods, 2009
▷Ref: Cortes et al, Bioinformatics, 2018
▷Ref: Deane et al, Bioinormatics, 2018
▷Ref: Cazals, O'Donnell; Submitted

TLC teleportation, rigid motions, and frames

▷ Loop decomposition into tripeptides and connecting peptide bodies



- Tripeptide: 9 atoms, 5 moving via *teleportation*
- Peptide body connecting two tripeptides: rigid ... whence rigid motions
- Consequence: two classes of citizens
  - peptide bodies within tripeptides
  - peptide bodies connecting tripeptides

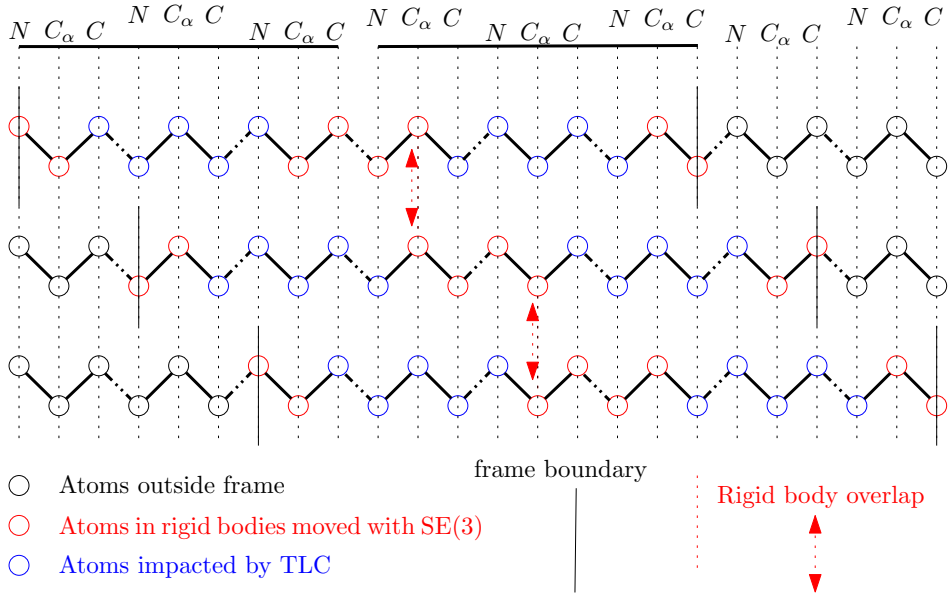
$\Rightarrow$ corrected via frame shifting

Frames involved in whole loop sampling

Definition 7. Subset of the loop to which individual TLC are applied.

- Frame shifting:
  - frame 1 starting at the first a.a. always contains n tripeptides regardless of N;
  - frame 2 at the second peptide contains $n-1$ tripeptides if $N \mod 3 = 0$ and n;
  - frame 3 at the third peptide contains n if $N \mod 3 = 2$ and $n-1$ otherwise.

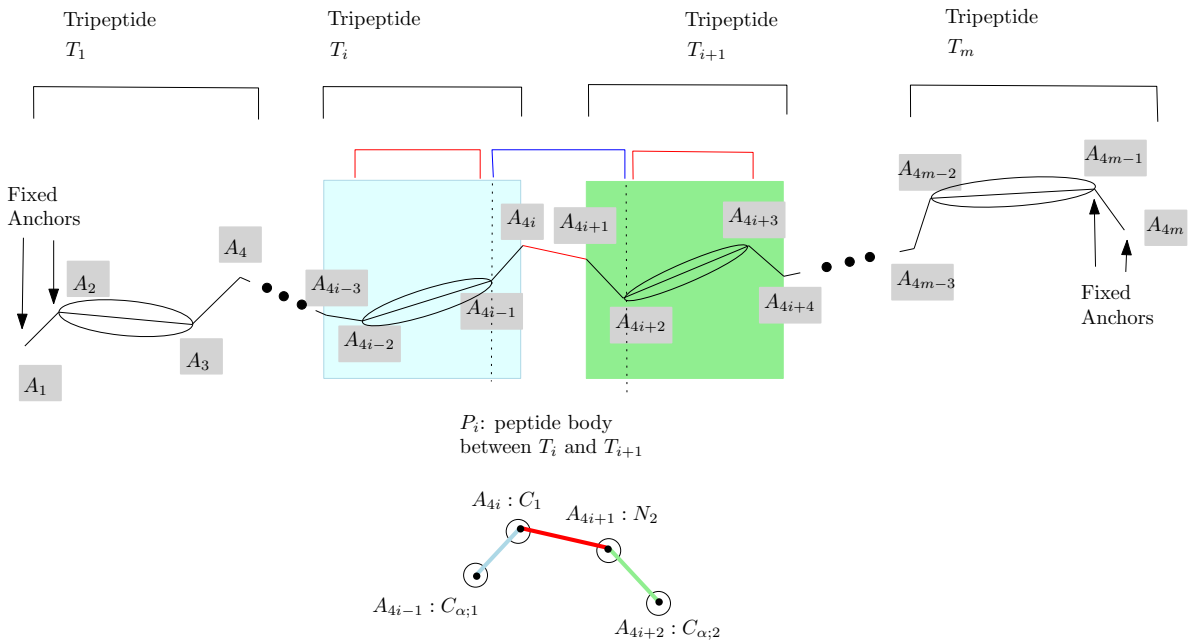
- The three frames obtained for a chain of $N = 3n+2$ a.a., with $n = 2$



Global geometric model

- Loop studied L: $M = 3 \times m$ amino, m tripeptides: $L = T_1, \ldots, T_m$
- Loop decomposition: rigid peptide bodies and their complements

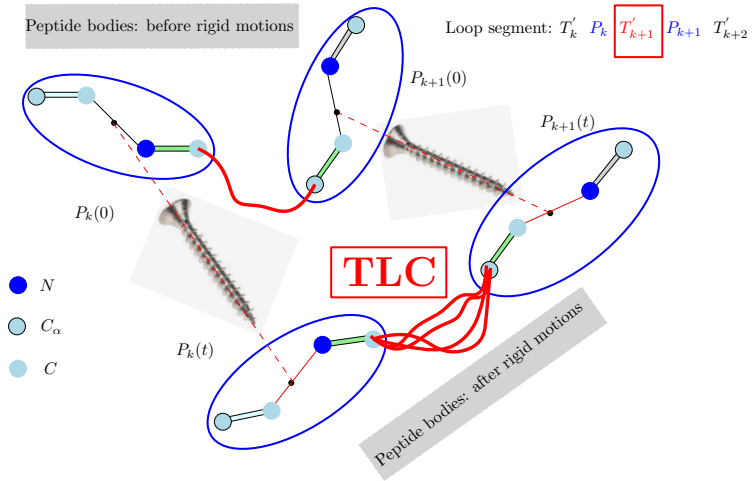
$$L = P_0 \; T_1^{'} \; P_1 \; \ldots \; P_{k-1} \; T_k^{'} \; P_k \; \ldots \; P_{m-1} T_m^{'} P_m. \tag{27}$$



- Parametric space:
  - For one peptide body: $SE(3) = SO(3) \times \mathbb{R}^3$
  - For one tripeptide: solution space of TLC. . . except that
    - The angular parameterization of TLC $\{\alpha, \xi, \eta, \delta\}$: depends on $SE(3) \times SE(3)$ since the left and right legs come from P_{i-1} and P_{i-1}

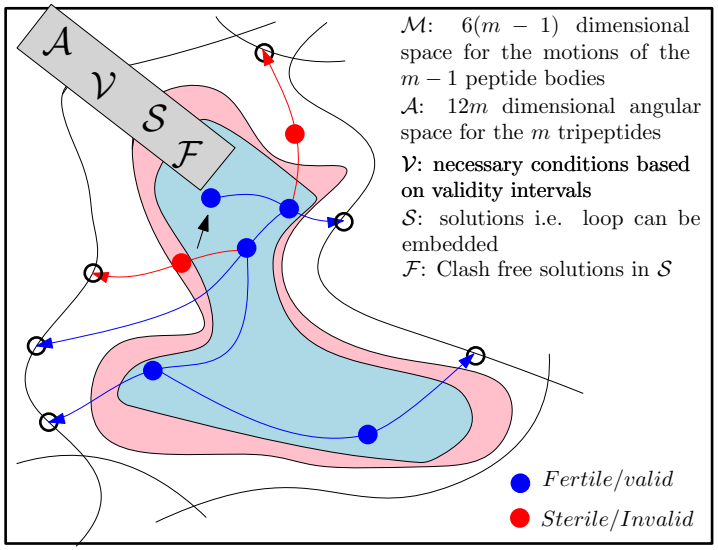
Sampling one frame: spaces involved and main idea

- Loop decomposition into: rigid peptide bodies and tripeptides cores



$$L = P_0\ T'_1\ P_1\ \ldots$$
$$P_k\ T'_{k+1}\ P_{k+1}\ \ldots$$
$$P_{m-1}\ T'_m P_m.$$

- Random sampling of loop conformations using Hit-and-Run:



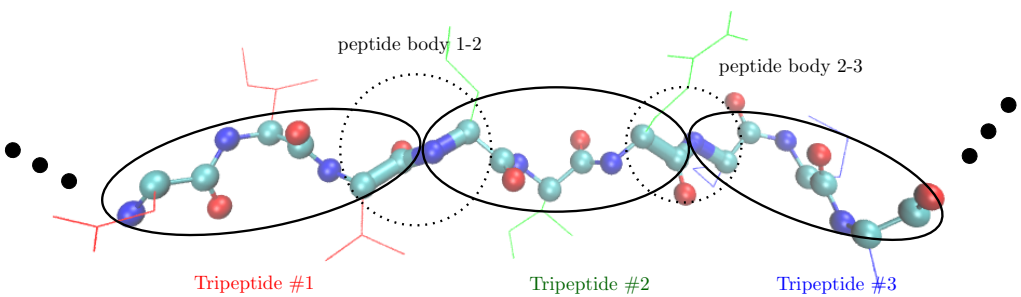
- Aim: perform rejection sampling in a region $\mathcal{V}$ containing all valid loop geometries.
- How: with Hit-and-Run in a domain characterizing necessary conditions – cf validity intervals

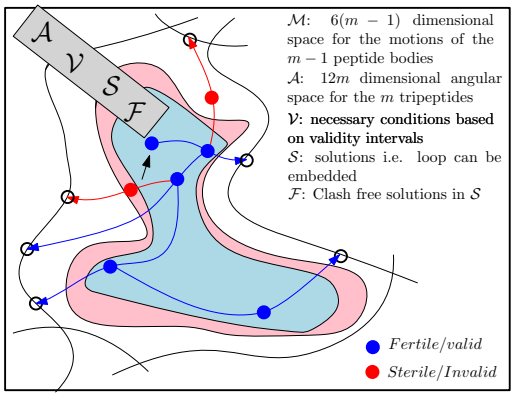
Sampling one frame: spaces involved and solution sketch

▷ Global parameterization of the conformational space of the loop: based on rigid bodies associated with peptide bonds

- $\mathcal{M}$: motion space for the $m-1$ peptide bodies, essentially $(SE(3))^{m-1}$
- $\mathcal{A}$: $12m$-dimensional angular space coding the geometry of tripeptides
- $\mathcal{V}$: domain bounded by 24 hyper-surfaces in $\mathcal{A}$, corresponding to Validity Constraints Necessary Constraints for TLC to admit solutions
- $\mathcal{S}$: the fertile space, where TLC admits one solution for each tripeptide
- $\mathcal{F}$: clash free solutions in $\mathcal{S}$ for $\{N, C_\alpha, C, O, C_\beta\}$ pairs

▷ Number of solutions: $\prod_i$(num solutions tripeptide i)





Angular representations: tripeptide and loop

- Angular representation of a tripeptide: the 2×4 angles

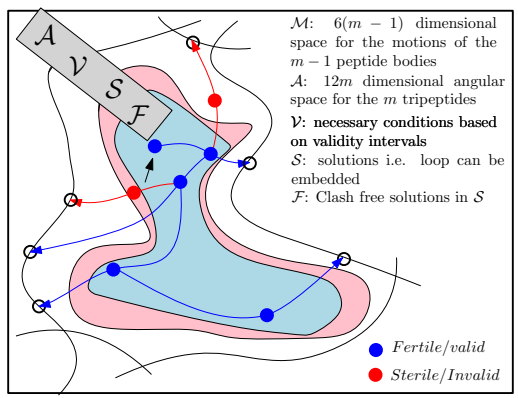
Definition 8. Let $\mathrm{A}_{k,i} = \{\alpha_{k,i}, \eta_{k,i}, \xi_{k,i-1}, \delta_{k,i-1}\}$ be the set of angles associated with $C_{\alpha;i}$ in the k-th tripeptide T_k.
The *angular representation* of a tripeptide T_k is the 12-tuple $\mathrm{A}_k = \{\mathrm{A}_{k,1}, \mathrm{A}_{k,2}, \mathrm{A}_{k,3}\}$.
The corresponding 12-dimensional space is denoted $\mathcal{A}_k$.

Definition 9. (Angular conformational space $\mathcal{A}$) The *angular conformational space* of the loop L is the $12m$ dimensional space defined by the product of the m angular space of the individual tripeptides:

$$\mathcal{A} \stackrel{Def}{=} \prod_{k=1}^{m} \mathcal{A}_k. \tag{28}$$

Validity domain for the whole chain L with m tripeptides

- Angles τ: $3m$ angles τ (3 for each tripeptide)
- Recap per angle τ:
  - For one angle: at most 4 Depth One Validity Intervals (DOVI)
  - For each DOVI: 2 sub-manifolds of $\mathcal{A}_k$ defined by limit cases; yields (at most) 8 sub-manifolds in $\mathcal{A}_k$.
- For one tripeptide: 3 τ angles $\Rightarrow$ 24 constraint surfaces in the 12 dimensional angular space $\mathcal{A}_k$.
- Enumerating constraints:
  - One tripeptide: 24
  - Whole loop: $24m$



Motion space for peptide bodies

Moving peptide bodies with rigid motions

- ▷ Configuration spaces for motions:
  - One peptide body: $\mathcal{R} : (S^2 \times [0, A)) \times (S^2 \times [0, 2\pi)) \subset SE(3)$
  - The $m-1$ peptide bodies in the loop L: $\mathcal{M} = \mathcal{R}^{m-1}$
- ▷ Peptide body motions: sample $m-1$ independent screw motions (translation+rotation)
- ▷ Overall linear interpolation $r \in \mathcal{M}$: between the identity and the rigid motion corresponding to r:

$$\text{Ray}(V) = \{\gamma(t) = Id + tV, \text{ with } \gamma(0) = Id\}. \tag{29}$$

- ▷ Restriction to each peptide body: defines a rigid transformation

$$\gamma_k : [0, 1] \mapsto SE(3), \gamma_k(0) = Id, \tag{30}$$

- ▷ Position of the k-th peptide body $P_k(t)$ at time t:

$$P_k(t) = \gamma_k(t) P_k(0). \tag{31}$$

Algorithm overview

▷ For a given angle τ:

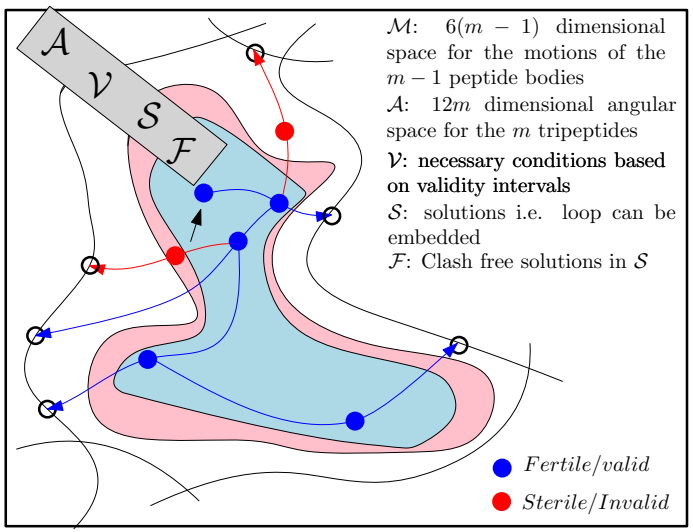
$$t \text{ determines the positions of peptide bodies whence tripeptide legs} \tag{32}$$

$$\rightsquigarrow \text{ kinetic angular representation } \mathrm{A}_i(t) \text{ of } T_k \tag{33}$$

$$\rightsquigarrow \text{ kinetic validity intervals } I_{\tau_{k,i}}(t), I_{\tau_{k,i}|\delta}(t) \tag{34}$$

▷ Example condition for kinetic depth 1 validity interval to be $\neq \emptyset$:

$$I_{\tau}^{\max}(\mathrm{A}_{k,i}(t)) = I_{\tau|\delta}^{\min}(\mathrm{A}_{k,i+1}(t)) \tag{35}$$



▷ Algorithm overview:

- For each angle $\tau_{k,i}$: find the closest intersection with the 24 hyper-surfaces, along the 1D curve defined by the rigid motion interpolation.
- Let t_{max} be the corresponding value of t: draw $t_s \leftarrow \mathrm{Uniform}(0, t_{max})$
- Apply the rigid transforms defined by t_s to the $m-1$ peptide bodies
- Solve the m individual TLC problems

Sampling algorithm for one frame: pseudo-code

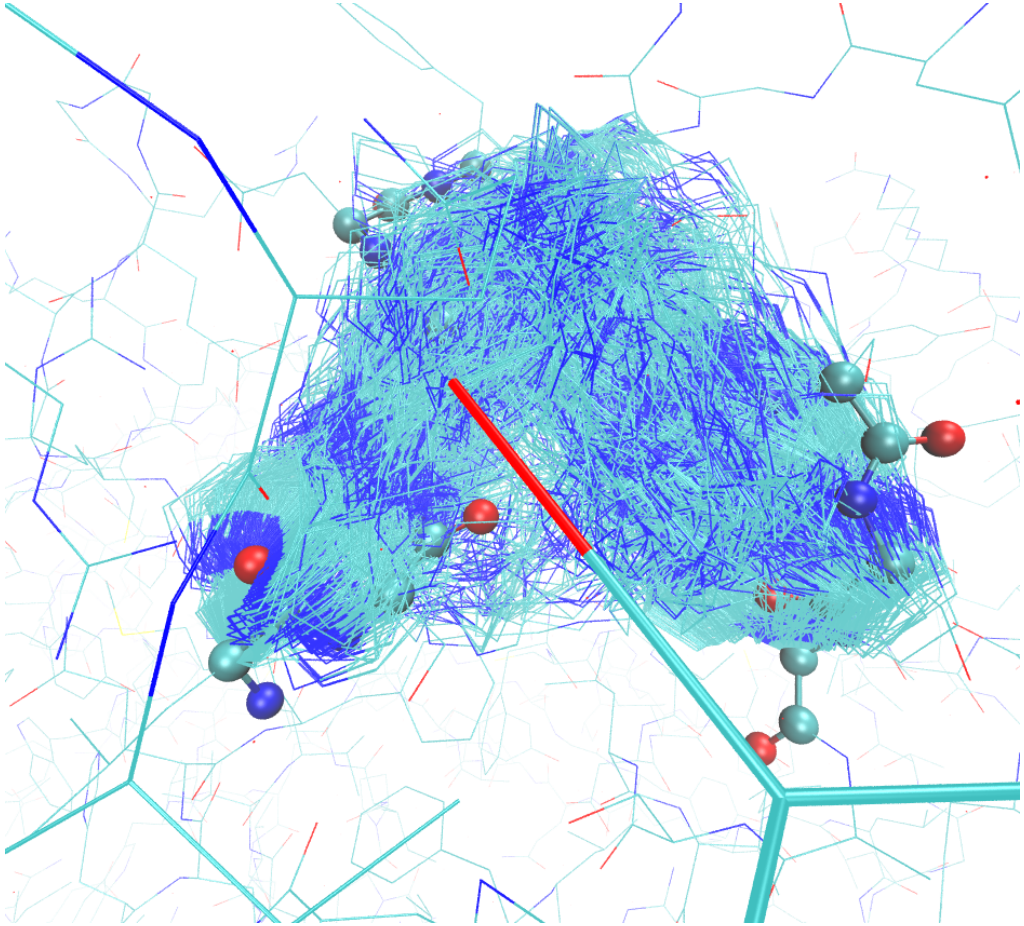
Input: p_j: point from which the move is made; corresponds to $t = 0$
Output: a point $\in \mathcal{S}$
Var t_{max}: initialized using the smallest value of $t > 0$ breaking triangular inequality in a given tripeptide
V: Random direction (Eq. 29)
for i $\in \{1, ..., m\}$ **do**
    for l $\in \{1, 2, 3\}$ **do**
        // Angle $\tau_{k,i}$: process the (at most) 24 equations
        $S = \{t_{max}\}$
        // Process all interval pairs
        for $I_{\tau_{k,i}}(t) \in \mathcal{I}_{\tau_{k,i}}(t)$ **do**
            for $I_{\tau_{k,i}|\delta}(t) \in \mathcal{I}_{\tau_{k,i}|\delta}(t)$ **do**
                $S_{tmp} \leftarrow$ numerical solutions for Eq. **??** and **??** $t \in [t_{min}, t_{max}]$
                $S = S \cup S_{tmp}$
            end for
        end for
        Sort S by ascending order
        Let t_k be the k-th element of S
        $u_k := \frac{t_k + t_{k+1}}{2}$
        $k = 1$
        // Stop when no validity interval can be defined for $\tau_{k,i}$
        while $\text{DOVI}_{\tau_{k,i}}(u_k) \neq \emptyset$ **do**
            $t_{max} = t_k$
            $k = k + 1$
        end while
    end for

Algorithms and parameters

- Unmixed loop sampler $\mathrm{ULS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$:
  - $One|All$ a flag indicating how many solutions are retained at each embedding step,
  - N_{ES} the number of embedding steps,
  - N_V the number of random trajectories followed in motion space,
  - N_{OR} the output rate (the number of steps in-between the ones where conformations get harvested),
  - p_0: the starting configuration.

- Mixed loop sampler $\mathbb{MLS}^{N_V;N_{OR}}_{One|All;N_{ES}}[p_0]$: every other step, the loop is shifted by 1 or 2 units to also sample the peptide bodies.

VMD demo



Loops sampling: ϕ, ψ and ω

- Typical values of the torsion angle ω:
  - SSE?
  - loops?

Loops sampling: ϕ, ψ and ω

- Typical values of the torsion angle ω:
  - SSE? $\pi \pm 2 - 3^\circ$
  - loops? $\pi \pm 15^\circ$

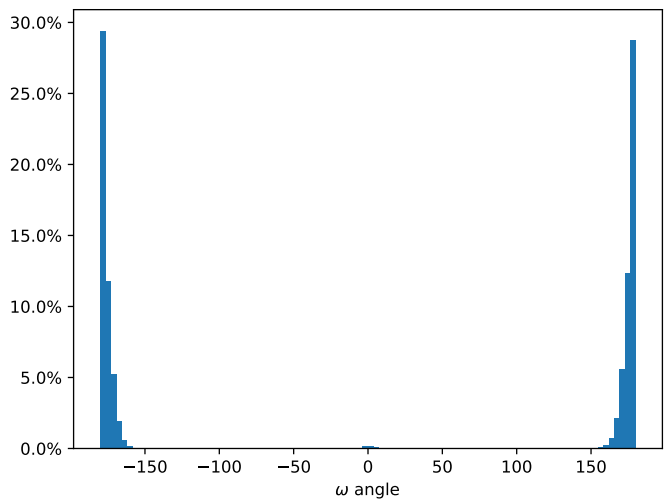
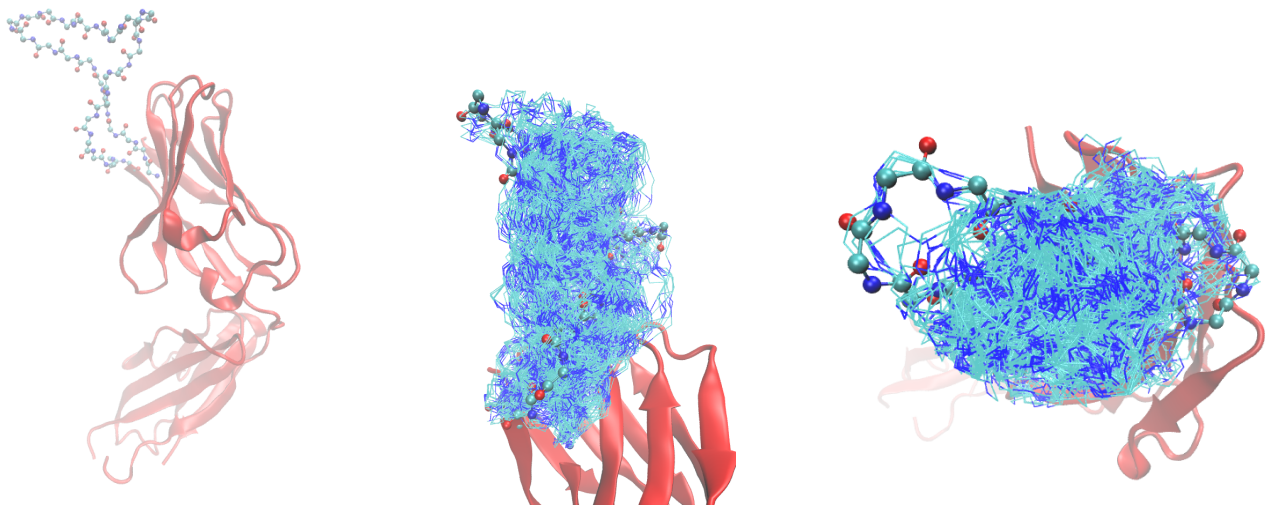


Illustration: CDR-H3-HIV, 30 amino acids

▷ System:

- The loop is a complementarity-determining region (CDR-H3) from PG16, an antibody with neutralization effect on HIV-1.
- pdbid: 3mme, chain A; residues: 93-100, 100A-100T, 101, 102.



Conformations generated by algorithm $\mathbb{MLS}^{1;1}_{One;250}$. **(A)** Variable domain (red) and the 30 a.a. long CDR3. **(B,C)** Side/top view of 250 conformations.

▷ Generation speed: $\sim$ 10 conformations per second

Results: sampling and study of fluctuations

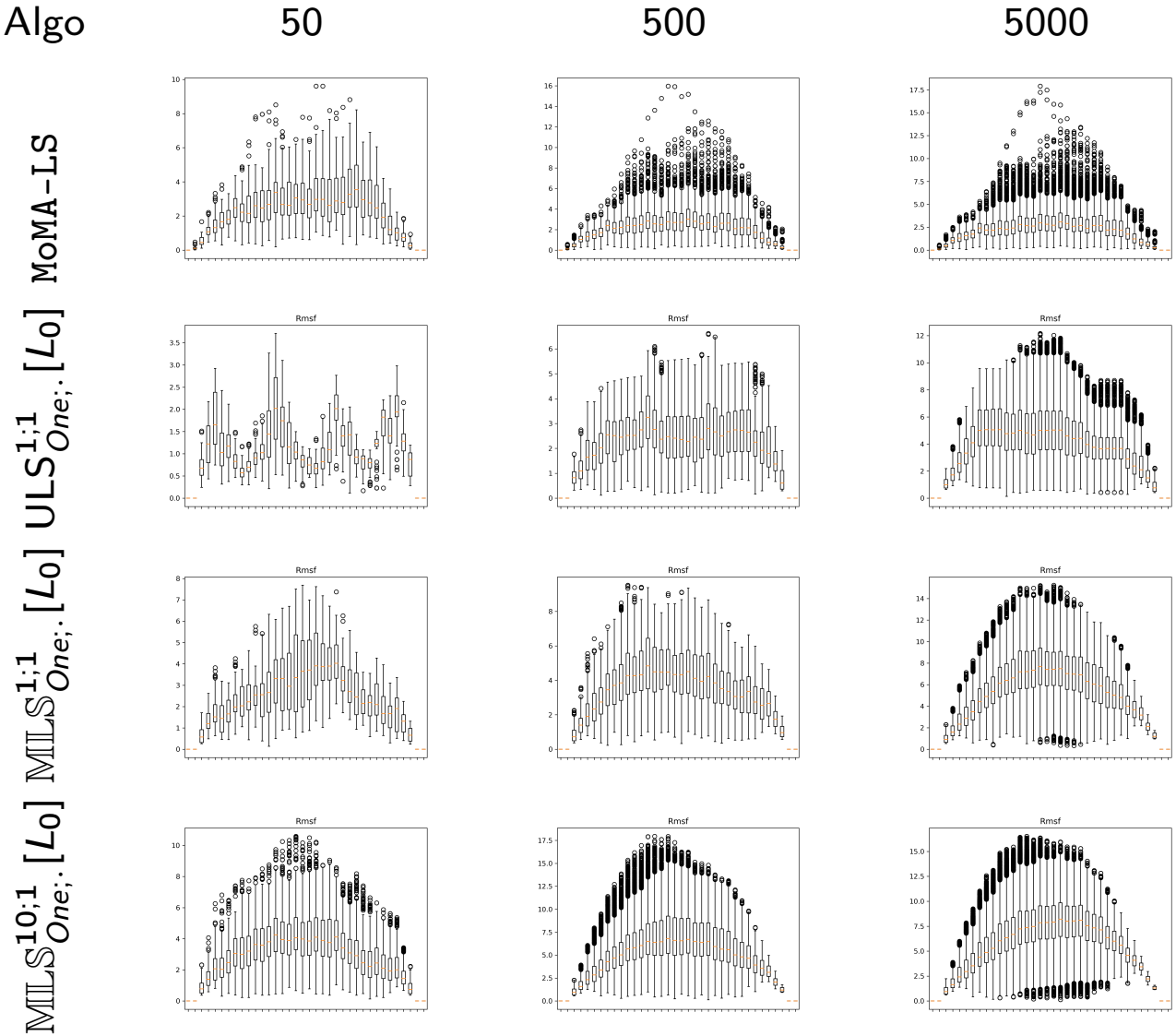
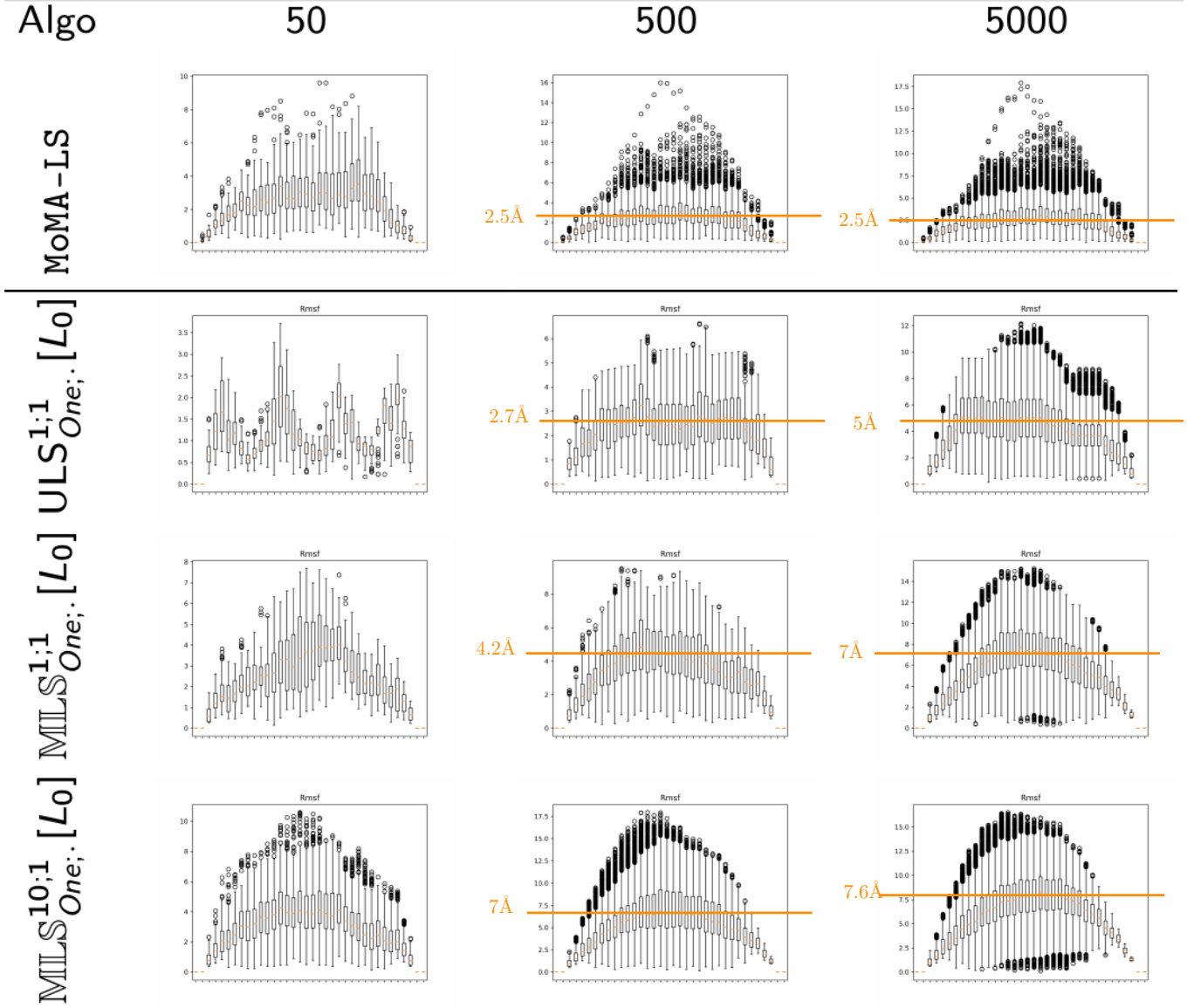


Figure: **Backbone RMSF (36 atoms) for the 12 amino acid long loop PTPN9-MEG2.**

Results: sampling and study of fluctuations



Backbone RMSF (36 atoms) for the 12 amino acid long loop PTPN9-MEG2.

Outlook

- Key features:
  - First global parametric model of protein loops amenable to effective sampling strategies a-la Hit-and-Run
  - Results: on par or better with state-of-the-art methods
    - Atomic fluctuations along the loop
    - Mutual reachability for existing conformations
  - Insights on the intrinsic difficulty of the problem–via random walks and curved polytopes
- Open problems:
  - Uniformity of sampling (Theorem)
  - Connexion to micro-canonical ensembles and densities of states
  - Sampling with side chains

Loop sampling

Open problems

- Tightness of the Depth-N Validity Constraints
- Uniformity of the sampling in solution space
- Mixing dihedral angles and the remaining internal coordinates

Bibliography

A. Chevallier, F. Cazals, and P. Fearnhead.

Efficient computation of the the volume of a polytope in high-dimensions using piecewise deterministic markov processes.

In *AISTATS*, 2022.

A. Chevallier, S. Pion, and F. Cazals.

Improved polytope volume calculations based on Hamiltonian Monte Carlo with boundary reflections and sweet arithmetics.

J. of Computational Geometry, NA(NA), 2022.

T. O'Donnell, C.H. Robert, and F. Cazals.

Tripeptide loop closure: a detailed study of reconstructions based on Ramachandran distributions.

Proteins: structure, function, and bioinformatics, 90(3):858–868, 2022.

T. O'Donnell and F. Cazals.

Geometric constraints within tripeptides and the existence of tripeptide reconstructions.

Technical report, 2022.

T. O'Donnell and F. Cazals.

Protein loops sampling based on a global parameterization of the backbone conformational space.

Technical report, 2022.